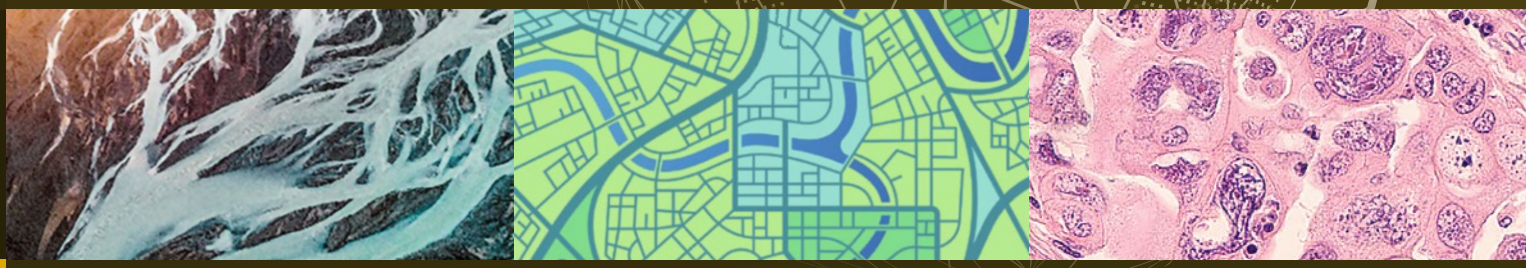
Fazlay S. Faruque
Editor



Geospatial Technology for Human Well-Being and Health

Springer

Geospatial Technology for Human Well-Being and Health

Fazlay S. Faruque
Editor

Geospatial Technology for Human Well-Being and Health

Springer

Editor
Fazlay S. Faruque
Department of Preventive Medicine
University of Mississippi Medical Center
Jackson, MS, USA

ISBN 978-3-030-71376-8 ISBN 978-3-030-71377-5 (eBook)
https://doi.org/10.1007/978-3-030-71377-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my mother Shahida Karim and father late Abdul Karim, who wanted me to grow up to help others, but I failed measurably.

Contents

About the Contributors

James G. Acker, PhD received his PhD in chemical oceanography from the University of South Florida in 1988. In 1995, he came to the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) to provide user assistance for Sea-viewing Wide Field-of-view Sensor (SeaWiFS) ocean color data. He has worked with and promoted the use of the Giovanni system since 2003 and has published several papers demonstrating the use of the system for various topics in Earth science. He also authored the book *The Color of the Atmosphere with the Ocean Below: A History of NASA's Ocean Color Missions* for the NASA Science and History Divisions.

Leslie Allsopp, PhD is an Assistant Professor at the University of North Texas Health Science Center, Texas College of Osteopathic Medicine, Department of Pediatrics and Women's Health. She is currently the Principal Investigator and Project Manager for Asthma 411, a school-based asthma initiative that has been adopted by eleven Independent School Districts, with over 300 participating campuses. Her research and public health interests include neighborhood-level contributors to health disparities and the dissemination and implementation of evidence-based community health initiatives. Her earlier professional experience includes providing primary health care services to low-resource, urban communities as a Family Nurse Practitioner.

Carlos Yáñez-Arenas, PhD is a full-time professor/researcher at the Universidad Nacional Autónoma de México. He was appointed as a postdoctoral researcher at the University of Kansas biodiversity institute, where he worked with the amazing biodiversity informatics and ecological niche modeling group led by Dr. Townsend Peterson and Dr. Jorge Soberón. Dr. Yañez obtained his PhD in 2013 from the Instituto de Ecología, A.C. under the guidance of Dr. Enrique Martinez-Meyer and Dr. Salvador Mandujano. He completed his master's (2009) at the Instituto de Ecología, A.C., and his bachelor thesis at the Universidad Autónoma de Yucatán (2007).

Seyed Mojtaba Hosseini Bamakan, PhD is an Assistant Professor in the Department of Industrial Management and Data Science Research Center at the Yazd University. He was a postdoctoral researcher at Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS). He received his PhD in Data Science from the University of Chinese Academy of Sciences (UCAS). His publications are mostly in the areas of knowledge-based systems, expert systems with applications, energy, neurocomputing, and computational design and engineering. His current research interests include blockchain technology, business intelligence, data mining, and intelligent optimization techniques.

Timothy R. Boe, MS is a Geographer with the US EPA's Homeland Security Research Program. His work primarily focuses on response and cleanup issues following chemical, biological, radiological, and nuclear (CBRN) incidents. He has also been developing computer-based decision support tools to aid decision makers in responding to wide-area contamination incidents. Before joining the EPA, he worked as an Oak Ridge Institute for Science and Education (ORISE) Fellow, where he conducted research on wide-area CBRN remediation.

M. Lucília Carvalho, PhD now retired, was an Associate Professor in the Department of Statistics and Operational Research at the Lisbon University, where she taught graduate and undergraduate level courses in spatial statistics, statistics applied to medicine, demography, sampling theory, and theory of epidemics. She was director of the Methodology Department of the Portuguese National Statistical Institute. She served five years on the European Statistical Advisory Committee (ESAC), appointed by the European Commission.

Irene Casas, PhD is a Professor in the School of History and Social Sciences at Louisiana Tech University where she has been a faculty member since 2009. Her interests are in the areas of transportation, health, and geographic information science (GIS). She has collaborated with researchers in different disciplines on problems relevant to these areas. She has served as vice-chair and chair of the Transportation Geography Specialty group from the Association of American Geographers and is part of the editorial board of several journals. She teaches a variety of geographic information science courses and serves as an adviser to students in the GIS major.

Annamaria Castrignanò, MS is Associate Senior Researcher at CNR-IRSA and full professor of Geostatistics at the G. D'Annunzio University of Chieti-Pescara (Italy). Her research interests mainly concern the processing of spatial and temporal data measured with different types of sensors. Methods used include multivariate geostatistics, mixed models, spatial data fusion, and stochastic simulation. She is a member of the editorial board of Precision Agriculture. She has been a member of the scientific committee of several national and international scientific conferences. She authored more than 400 scientific publications as articles in international journals, book chapters, and conference proceedings.

Howard H. Chang, PhD is an Associate Professor in the Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University. His primary research interest is in the development and application of statistical methods for analyzing complex spatial-temporal environmental exposure and health data. He conducts population health studies that leverage large administrative health databases, such as birth/death certificates, hospital billing records, electronic health records, and disease surveillance systems. His methodological research focuses on spatial epidemiology, analysis of adverse birth outcomes, statistical models for air quality exposure assessment, and uncertainty quantification in climate science.

Yanchao Cheng, MS is a researcher in the Department of Biogeography at the University of Bayreuth in Germany. She works in the field of biogeography, with a focus on vector-borne diseases. She is interested in interdisciplinary modeling research implementing diverse modeling approaches. She compares and integrates ecological niche modeling and epidemiological modeling approaches to better assess and improve the performance of diverse models.

Steven A. Cohen, DrPH with doctoral and master degrees in public health, is a social epidemiologist and applied public health demographer, studying the impacts of aging on health and health care delivery. His research agenda examines socioeconomic and demographic disparities in informal, family caregiving in the USA, and their impacts on caregiver health and quality of life, as well as social factors and place-based characteristics associated with health, health behaviors, and health care services utilization in older adults, with a primary focus on rural-urban and other sociodemographic disparities. He is a faculty member in the Department of Health Studies at the University of Rhode Island.

Sandra Aleixo, PhD has a bachelor's and master's degrees in Statistics and Operational Research from the Faculty of Sciences of the University of Lisbon, and a PhD in the scientific area of Statistics and Operational Research, in the specialty of Probability and Statistics, from the Faculty of Sciences of the University of Lisbon. She is a Coordinate Professor in the Departmental Area of Mathematics at the Instituto Superior de Engenharia de Lisboa of the Polytechnic Institute of Lisbon, teaching Curricular Units in Probabilities and Statistics. Her scientific interests are applied statistics, data modeling, and dynamic systems.

Eric Delmelle, PhD is an Associate Professor in the Department of Geography and Earth Sciences at the University of North Carolina at Charlotte. He received his PhD in Geography at SUNY Buffalo in 2005. His research interests include GIS, epidemiology, uncertainty, and spatial analysis and modeling. He has been funded by the March of Dimes Foundations, the Centers for Disease Control and Prevention (CDC), and the North Carolina Water Resources Research. He has been serving on the Medical Geography (Health) Group of the American Association of Geographers (AAG) since 2013. He has coedited 2 book volumes and coauthored over 60 papers.

Michael Desjardins, PhD is a postdoctoral fellow at Johns Hopkins Bloomberg School of Public Health in the Spatial Science for Public Health Center and Department of Epidemiology. His research interests include spatial epidemiology, spatial statistics and modeling, geovisualization, and mixed methods. His postdoc research mainly focuses on the development of spatiotemporal Bayesian models to understand the distribution of *Vibrio parahaemolyticus* in estuarine environments. He has published numerous peer-reviewed articles covering a variety of health-related topics, including COVID-19, vector-borne diseases, and sexually transmitted infections. He serves as chair (2021–2022) of the Health and Medical Geography Specialty Group of the American Association of Geographers.

Steven H. Douglas, MS received a BS in Environmental Science from Michigan State University and an MS in Environmental Science from the University of South Florida. His MS research focus was on modeling groundwater vulnerability to pathogens and pesticides. Following graduation, he worked at the USGS on modeling of pathogens in soil with Dr. Dale Griffin and with other USGS staff on coastal change, coastal hazards, and ocean acidification. Since joining Versar, he has worked on a wide variety of environmental projects, mainly focusing on sea-level rise, storm surge, and coastal resilience.

Christos Doulkeridis, PhD is an Associate Professor in the Department of Digital Systems at the University of Piraeus. He has been awarded both a Marie-Curie and an ERCIM Allain Bensoussan fellowship for postdoctoral studies at the Norwegian University of Science and Technology. Prior to this, he received his PhD from the Department of Informatics at the Athens University of Economics and Business. He has published in reputed international journals and presented at conferences in the areas of data management, knowledge discovery, and distributed systems. His research interests include Big Data, cloud-based data management, distributed query processing, mobility analytics, and spatial and spatiotemporal databases.

Elizabeth Erdman, MS is an Epidemiologist in the Special Analytic Project group within the Massachusetts Department of Public Health. She develops and uses complex data analysis techniques to guide public health practice in Massachusetts. Her team manages an innovative, linked public health dataset (PhD) which enables data-driven analysis of emerging public health issues such as opioid-related overdoses and maternal health. She is particularly interested in the intersection of environmental and public health and has extensively researched the health effects of the built environment, traffic-related air pollution, and green design to improve urban climates.

Fazlay S. Faruque, PhD is a Professor of Preventive Medicine at the University of Mississippi Medical Center (UMMC). He joined UMMC in 2000 as the founding Director of GIS and Remote Sensing program. For the last thirty years, he has been teaching and researching in the area of environmental health utilizing a variety of geospatial technology. As principal investigator, he managed extramural grants, including projects from NIH and NASA. His research projects are mainly within the areas of environmental health and the application of spatial methods in epidemiological and healthcare delivery-related studies. Since 2012, he has been serving as the Chair of the ISPRS Working Group on Environment and Health. By academic training, he is a geological engineer.

Sérgio Gomes, MS is a Supervising Nurse from the DGS and Coordinator of the Support Unit of the National Health Service. He holds a master's degree in Nursing Science from the Portuguese Catholic University, Specialist in Medical-Surgical Nursing – School of Military Health of Lisbon, and a postgraduate degree in Nursing Services Administration – ESE Nursing Maria Fernanda Resende, Lisbon; PhD student in Nursing Science at the Portuguese Catholic University. He is a member of the permanent team of the Public Health Emergency Unit and a member of the "Task Force" at the DGS for the conclusion of the work of the National Health Plan.

Mary L. Greaney, PhD is an Associate Professor and Chair of the Department of Health Studies at the University of Rhode Island. She received her PhD from the University of South Carolina and MPH from the University of Massachusetts (Amherst). Dr. Greaney has over two decades of experience in conducting and collaborating on public health research into health disparities and health promotion. She received her PhD from the University of South Carolina and MPH from the University of Massachusetts (Amherst).Her research has focused on identifying personal, social, and environmental factors associated with physical activity and other healthful behaviors; identifying sociodemographic, behavioral, and health-related factors associated with intervention engagement; and investigating the relationship between health behaviors and health-related quality of life among older adults.

Dale W. Griffin, PhD is an Environmental and Public Health Microbiologist at the US Geological Survey. He received a BS in Microbiology and a Master of Science in Public Health from the University of South Florida (USF). He received a PhD with a research focus on the use of molecular methods for the detection of water quality indicator microorganisms and pathogenic viruses in fresh and marine waters from USF. In his postdoctoral positions, he worked on human enterovirus detection assays, marine lysogeny and isolation of viruses lytic, and also on a NASA-funded grant to study microbiology and public health issues associated with transatlantic dust storms. He served as a Waksman Foundation Lecturer for ASM's Distinguished Lecturer Series. His current research projects include methods development, microbial water quality issues, pathogens in soils, the influence of aerosols on harmful algae, and aerobiology.

Laura A. Hempel, PhD is a Hydrologist with USGS, Colorado Water Science Center. She has a PhD in Geology from Oregon State University, where she studied the effects of the flow regime on sediment transport and river morphology. Previously, she worked on projects investigating hydrologic controls of mercury (Hg) deposition on floodplains and soil gas fluxes in a tropical montane forest. At USGS, her current work as a Hydrologist involves using novel techniques for geomorphic change detection, stream gaging and flood warning in post-wildfire landscapes, and using small Unmanned Aircraft Systems (sUAS) for mapping rivers and flow properties.

Alexander Hohl, PhD is an Assistant Professor in the Department of Geography at the University of Utah. He received his BS in geography from the University of Zurich, Switzerland, and his MA and PhD in Geographic Information Science from the University of North Carolina at Charlotte. His research focuses on computational aspects of spatial analysis with application to geographies of health and wellbeing. He has authored multiple articles in peer-reviewed scientific journals, including the *International Journal of Geographic Information Science*, *Spatial and Spatiotemporal Epidemiology*, *Cartography and Geographic Information Science*, and *Applied Geography*.

C. A. O. Jean-Babtiste, DrPH is a Public Health Analyst for the US Air Force. She is also an Adjunct Instructor of Public Health Policy. Cindy received her Doctor of Public Health from Loma Linda University. Cindy's passion is in health equity on a community, environmental, and systems approach. Cindy's work and interests are geared at using research and evaluation to influence Policy Systems and Environment (PSE).

Anja Jaeschke, PhD is a research assistant at the Bavarian Environment Agency. Previously, she has been a postdoctoral researcher at the Department of Biogeography at the University of Bayreuth in Germany. Her main focus was on species distribution modeling in nature conservation and vector-borne diseases. She mainly applied correlative modeling approaches and integrated ecological prerequisites like biotic interactions to increase the ecological relevance of the results.

Richard K. Kiang, PhD was a Group Leader before he retired from NASA Goddard Space Flight Center. His team collaborated with US, national, and international public health agencies and engaged in research concerning the outbreaks and propagations of vector-borne, influenza, and other respiratory diseases using satellite measurements. He codeveloped the Pandemic Prediction and Forecasting Science and Technology Plan for the US Office of Science and Technology Policy to broaden the use of remote sensing for health applications. Currently, he is a science advisor to the US National Institutes of Health's International Centers of Excellence for Malaria Research and other remote sensing projects in the academia.

Karen C. Kosinski, PhD is a Senior Lecturer in the School of Arts and Sciences at Tufts University. Her research interests focus on infectious disease and especially urogenital schistosomiasis, water infrastructure, engineering interventions, risk mapping, quantitative data analysis, and pedagogical approaches to training students to conduct Community-Based Participatory Research (CBPR). She has worked on the design, implementation, and evaluation of techniques to prevent urogenital schistosomiasis (UGS) and on risk mapping to understand the distribution of UGS in the Eastern Region of Ghana. She is the 2018 recipient of the Lerman-Neubauer Prize for Outstanding Teaching and Advising at Tufts University.

David J. Lary, PhD received a First- Class Double Honors BSc in Physics and Chemistry from King's College London with the Sambrooke Exhibition Prize in Natural Science, and a PhD in Photochemical Computer Modeling of Atmospheric Chemistry from the University of Cambridge. The thread running through all the research is holistic sensing in service of society through the use of machine learning and data-driven insights, using observation and automation to facilitate discovery, with a focus in the area of preemptive human protection, including autonomous robotic teams and comprehensive biometric sensing. David held positions at Cambridge University as a faculty member and a Royal Society University Research Fellow. He was awarded the first Alon Fellowship in the Department of Geophysics and Planetary Space Science at the University of Tel-Aviv. He was invited to join NASA for his work on data assimilation as the first distinguished Goddard fellow in Earth Science, receiving six NASA awards for his research and technology development. In 2018, David was appointed a United States Special Operations Command Fellow at SOFWERX by J5, the Futures Mission Directorate of USSOCOM.

Jay Lee, PhD received his PhD in Geography from the University of Western Ontario, Canada. His research stems from a broad interest in integrating analytics in Operations Research and Geographic Information Science/Systems, as applied to works in health geography, geography of crime, environmental studies, and business geography. Some of his publications and research grants involved digital elevation models, environmental conservation, GIS, web-based GIS, urban sprawl, and areal health disparities. He has been teaching GIS and related courses in the Department of Geography at Kent State University since 1991. Dr. Lee is the Editor-in-Chief for Papers in Applied Geography (Taylor & Francis).

Estelle Levetin, PhD is Professor Emerita of Biological Science at the University of Tulsa. She received her BS from Boston State College and PhD in Biological Science from the University of Rhode Island. She was a faculty member at the University of Tulsa for 48 years and also served as Department Chair of Biological Science. Her research has focused on airborne pollen and fungal spores, and she is continuing her studies in retirement. She has authored multiple articles in peer-reviewed journals and coauthored the textbook, *Plants and Society*, with eight editions. She is a Fellow of the American Academy of Allergy, Asthma and Immunology.

Huiyu Lin, MA is currently a PhD candidate in geography at Kent State University. Her work focuses on spatiotemporal analysis techniques, environmental criminology, and community development.

Juliana Maantay, PhD is a Professor of Urban Environmental Health Geography at City University of New York (CUNY), Lehman College since 1998. She founded and directs the graduate program in Geographic Information Science as well as the Urban GISc Lab, and has edited several compendia and written two widely used textbooks and numerous other publications on the urban environment and geospatial analysis. Her main research foci are environmental justice, health disparities, and exposure assessment, specifically in urban areas. She received a Fulbright Distinguished Chair Award to study health and the built environment. For 25 years prior to her academic career, she was an urban planner, environmental analyst, and architect.

Sabino Maggi, PhD is a Senior Researcher at CNR-IIA. His current research activities focus on the analysis of remote sensing and environmental data and on the development of software and hardware tools for environmental monitoring. His past research interests involved the development of tunnel devices for superconducting electronics and astrophysics. He is the author of more than 150 publications. He has been an adjunct professor at the University of Torino and is currently an associate professor at UniNettuno University, Rome. He is a permanent member of the IMKEKO Technical Committee for Environmental Measurements and of scientific and organizing committees of international conferences, workshops, and technical seminars.

Nabin Malakar, PhD received an MSc degree in physics from Tribhuvan University, Kirtipur, Nepal, in 2005, and a PhD degree in physics from the State University of New York, University at Albany, Albany, NY, USA, in 2011. He is currently an assistant professor at the Worcester State University, Massachusetts. He was a postdoctoral research scientist at the Jet Propulsion Laboratory, California Institute of Technology, USA, where he was involved in the development, validation, and evaluation of a new NASA land surface temperature and emissivity product (MOD21) for the thermal infrared sensors onboard MODIS Terra, MODIS Aqua, Landsat, and VIIRS satellites. His current research interests include the identification of relevant variables in a physical phenomenon to improve our understanding of atmospheric processes, as well as algorithm development for societal applications of remote sensing data.

Andrew Maroko, PhD is an Associate Professor of Environmental, Occupational, and Geospatial Health Sciences at the City University of New York's Graduate School of Public Health and Health Policy. He has developed and taught numerous environmental health and GISc courses, and has conducted and published his research on public health issues related to the urban environment and health disparities.

Rita Masciale, PhD is a Geologist with specialties in Geomorphology and Environmental Dynamics. She works as a researcher at CNR-IRSA. Her research topics and activities focus on hydrogeological characterization, monitoring, and treatment of environmental data finalized to groundwater state and aquifer vulnerability assessment; field tests for measuring and quantifying hydrogeological processes; sustainable management of groundwater in regions with water scarcity; implementation of GIS and management of spatial information; innovative approaches for the hydraulic characterization of porous and/or fractured rock material under variable saturation condition. She is the author of several publications, and she has been involved and is still involved in national and international projects.

Alina M. McIntyre, MHS is currently a PhD student in the Environmental Health Department at Boston University School of Public Health. Her research focuses on assessing human health impacts of the built environment through environmental epidemiology and exposure assessment. She has utilized geospatial methods to assess infectious disease risk, specifically urogenital schistosomiasis, in relation to environmental health factors. She is investigating the use of mapping techniques in combination with community-engaged research to evaluate heat exposure and air pollution risk in the Boston area. She received her MHS from the Johns Hopkins Bloomberg School of Public Health and BA from Tufts University.

Isabel Natário, PhD is an Assistant Professor in the Mathematics Department of the School of Science and Technology of the NOVA University of Lisbon and a researcher at the Centre of Mathematics and Application of the same University. She is a Fellow of the Royal Statistical Society. Her research areas of interest are spatial statistics (aggregated data and point processes), Bayesian statistics, epidemiology, and medical and environmental statistics (hierarchical models, generalized linear models, dynamic population models), areas where she has published scientific papers and a book.

Elena N. Naumova, PhD is a Professor of Mathematics, Chair of the Nutrition Epidemiology and Data Science at Friedman School of Nutrition Science and Policy and Director of the Tufts Initiative for the Forecasting and Modeling of Infectious Diseases (InForMID) at Tufts University. Her research focus is on the development of analytical and computational tools for understanding transient biological processes and outbreak forecasting. She is actively promoting the use of novel data sources, satellite imagery, computer-intense simulation techniques, and dynamic mapping in public health applications. She has over 30 years of experience in conducting international interdisciplinary research projects and educational programs and coauthored over 200 publications.

Tonya L. Nichols, PhD serves as a Senior Science Advisor to the US EPA. As a Senior Science Advisor, she represents the US EPA in the federal Homeland and National Security communities. She has conducted and managed chemical, biological, and radiological collaborative research programs for almost 20 years. Since 2005, she has actively participated in developing national policies and operational plans related to preparing for, responding to, and recovering from national CBR events, natural disasters, and accidental releases of hazards to the environment with potential cascading effects to human health.

Giuseppe Passarella, PhD hydraulic engineer and hydrogeologist, is a researcher at CNR-IRSA. His current research activities focus on methods and tools for environmental monitoring and characterization, risk assessment of groundwater degradation and coastal aquifer salinization, climate change impact assessment on water resources in Mediterranean environments, and use of geostatistical tools for environmental issues. He is a reviewer and editorial board member of several water and environmental topics related to scientific journals. He is a permanent member of the IMKEKO Technical Committee for Environmental Measurements and of scientific and organizing committees of international conferences and technical seminars. He has been involved in national and international research projects on integrated water resources management as principal investigator and work package leader. He authored about 150 peer-reviewed research papers.

Deryn Pressman-Mashin née Harbin, MBA was a hydrological technician at USGS, where she assisted in tracking soil samples on Google Earth and performed PCR reactions on soil samples, testing for Bacillus. She holds an MBA in Nonprofit Management from Brandeis University and currently serves as the Director of Community Engagement and Communications at Epstein Hillel School.

Qiang Qu, PhD is a Professor and the director of Guangdong Provincial R&D Center of Blockchain and Distributed IoT Security at Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS). He received his PhD at Aarhus University, supervised by Prof. Christian S. Jensen. His PhD research was supported by the GEOCrowd project under Marie Skłodowska-Curie Actions. He was a visiting scientist working with Prof. Gustavo Alonso at ETH Zurich in 2014–2015, a visiting scholar working with Prof. Christos Faloutsos at Carnegie Mellon University, a visiting scholar working with Prof. Gao Cong at Nanyang Technological University, and a research fellow at Singapore Management University. His current research interests are in data-intensive applications and systems, focusing on efficient and scalable algorithm design, blockchain, data sense-making, and mobility intelligence.

Abdallah M. Samy, PhD has expertise in vector-borne diseases, disease burden analysis, and health economics. His research addresses several questions at the interface of ecology, epidemiology, public health, and global health. He is broadly interested in studying zoonosis, arboviral epidemiology, and the climate change influences disease dynamic and spread. He uses an interdisciplinary and multifaceted approach to research questions, typically using a combination of field and lab experiments, geographic information systems, remote sensing, ecological modeling, and phylogeography. His current work on arboviruses and mosquito-borne diseases is focused on developing disease forecasts, understanding the major drivers of disease spread, and identifying the possible shifts at disease risk in response to global warming in the future.

Shikhar Shrestha, PhD is a postdoctoral fellow with the Department of Public Health and Community Medicine at the Tufts University School of Medicine. His research involves the study of substance use disorders and health outcomes using spatial epidemiological methods. He also has extensive research experience in the field of substance use disorder in pregnant women. He has published several peer-reviewed articles that examine the effects of prenatal opioid and alcohol exposure on early childhood growth and development.

Erin E. Silvestri, MHS has an MPH in Occupational and Environmental Epidemiology from the University of Michigan, School of Public Health and a certificate in Geographic Information Systems. She has worked for the US EPA for 14 years. As a public health biologist, she has been a key public health researcher for several EPA projects. Her current focus has been on the development of sampling methods for pathogens from environmental matrices and the development of a tool to document biological sampling and analysis plans.

Paula Simões, PhD is an Assistant Professor in the Department of Statistics and Operations Research, Faculty of Sciences, University of Lisbon. She is also an Assistant Professor in the Department of Exact Sciences and Engineering, Military Academy of the Military University Institute, and a researcher of the Mathematics and Applications Centre, of the New University of Lisbon. She holds a PhD in Statistics and Risk Management, and a master's degree in Probability and Statistics from the University of Lisbon. The main researcher areas include spatial statistics, spatial econometrics, and Bayesian statistics.

Radina P. Soebiyanto, PhD was a scientist at NASA Goddard Space Flight Center (Greenbelt, MD) working on the development of climate-based analytics for risk mapping and forecasting of infectious diseases, including chikungunya, Rift Valley fever, and seasonal influenza. She has experience in assessing the role of climate conditions in disease transmission such as cholera, dengue, malaria, hantavirus, and plague. Her previous work experience also includes developing large-scale mathematical models of the human immune system in the context of inflammatory diseases. He holds a PhD in systems and control engineering and a Master in Engineering and Management from Case Western Reserve University (Cleveland, Ohio). She is currently a senior data scientist at USAID President's Malaria Initiative (PMI) focusing on providing advanced analytics and tools as well as analytical technical assistance to PMI supported countries that are aimed at better data use to help in reducing malaria burden more efficiently.

David Sterling, PhD † PhD, CIH, ROH, FAIHA was a prodigious scholar and researcher in environmental and occupational health science. He was a Fellow of the American Industrial Hygiene Association. His research interests and publications included childhood lead exposure, air pollution and childhood asthma, asbestos exposure and disease progression, chemical exposures among agricultural workers, occupational safety in health care settings, and the epidemiology of parkinsonism in welders. His career spanned over three decades, during which he developed and directed environmental and occupational health graduate programs at Old Dominion University, Saint Louis University, and the University of North Texas Health Science. †Deceased 23 January 2020.

Thomas J. Stopka, PhD is an Associate Professor with the Department of Public Health and Community Medicine and the Clinical and Translational Science Institute at the Tufts University School of Medicine. He has contributed to and led numerous federally funded mixed methods, interdisciplinary, and translational studies focused on the intersection of opioid use disorder, overdose, and infectious disease since 1999. He has employed geographic information systems (GIS), and spatial epidemiological, qualitative, and biostatistical approaches in multisite studies and interventions to better understand and curb the ill effects of the opioid crisis. He also teaches courses in GIS, spatial epidemiology, and research methods.

Stephanie Margarete Thomas, PhD is a postdoctoral researcher at the Department of Biogeography at the University of Bayreuth. Her research provides detailed insights to better understand the relationships between biodiversity, climate change, and human and animal health. Using correlative and process-based modeling techniques, she assesses the spatial and temporal variability of distributional patterns of mosquitos and mosquito-borne diseases. The aim is to develop early warning systems that support the public health sector. She leads and participates in interdisciplinary, pan-European projects.

Chetan Tiwari, PhD is an Associate Professor and the Director for the Center for Disaster Informatics and Computational Epidemiology at Georgia State University. Before moving to GSU he was an Associate Professor of Geography at the University of North Texas. His research focuses on the development and application of spatial analysis methods for modeling and visualizing the spatial dimensions of disease risk. He has served as a coinvestigator on National Institute of Health (NIH) and National Science Foundation (NSF) grants on emergency preparedness and disaster response. In addition, he has served as a coinvestigator on environmental risk assessment projects funded by the Texas Commission on Environmental Quality (TCEQ) and Texas Environmental Health Institute. He served as the Chair of the Health and Medical Geography Specialty Group of the American Association of Geographers.

George Vouros, PhD received his BSc in Mathematics and PhD in Artificial Intelligence (1992), both from the University of Athens, Greece. Currently, he is a Professor in the Department of Digital Systems at the University of Piraeus and head of the Artificial Intelligence Lab (AI-Lab) in this Department. He has conducted research in the areas of expert systems, knowledge management, collaborative systems, ontologies, and agents and multiagent systems. He has been serving in leadership roles in numerous national and international conferences. He has been serving as a senior researcher in several European Union-funded and nationally funded projects. His recent works include the Data-driven AiRcraft Trajectory (DART) prediction research and the datACRON Big Data project (H2020 ICT-16). Professor George Vouros served three times as the chair on the Hellenic A.I. Society board.

Michele Vurro, PhD is a Senior Research Associate at CNR-IRSA where he held the task of a scientific coordinator the Integrated Water Resources Management. He has been involved in the following scientific topics: mathematical modeling of groundwater flow and transport in porous and fractured systems; methods for groundwater resources protection; artificial recharge of groundwater; databases and expert systems design and development for groundwater management; geostatistical methods for water resources management; integrated water resources management in water scarcity conditions; conflict analysis and community-based monitoring; adaptation strategies to the impact of climate change on water resources in Mediterranean area. He was an adjunct professor of hydrology and hydrogeology at the University of Basilicata. He is the author of about 100 papers published in international peer-reviewed journals and about 75 papers at international conferences.

Lance A. Waller, PhD is a Professor in the Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University. He is a member of the US National Academy of Sciences Board on Mathematical Sciences and Analytics, and cochair of the Committee on Applied and Theoretical Statistics. He is a Fellow of the American Statistical Association and a Fellow of the Royal Geographical Society. His research involves the development of statistical methods for geographic data including applications in epidemiology, disease surveillance, and disease ecology. He coauthored the textbook *Applied Spatial Statistics for Public Health Data* (2004, Wiley).

Angelika Winner, MS is a PhD candidate in Geography at the City University of New York's Graduate Center. She has been teaching undergraduate and graduate-level courses in GISc, Urban Geography, Population Geography, and Urban Health Geography at Lehman College for several years, and is an Urban Studies Core Fellow at CUNY. Her doctoral research focuses on health inequities surrounding food security and access to food in the urban environment.

Gebreab K. Zewdie, PhD is a postdoctoral researcher in the Department of Electrical and Computer Engineering, at Georgia Institute of Technology. He graduated with his PhD in Physics from the University of Texas at Dallas. Currently, he is working on space weather forecasting based on big data and advanced machine learning and deep learning methods.

Susheng Zhang, PhD is a PhD candidate at the University of Cambridge. She was a researcher at Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS). She received her master's and bachelor's degrees from Cambridge. She has very broad research interests, including biomedical engineering, machine learning, IT for health care, and bioinformatics.

The original version of this book was revised: Revised book has been uploaded to Springerlink. The correction to this book is available at https://doi.org/10.1007/978-3-030-71377-5_21.

Geospatial Technology for Human Well-Being and Health: An Overview

Fazlay S. Faruque

Introduction

There have been remarkable developments in the area of geospatial health over the last 30 years or so. Many of these advancements are discussed in detail throughout the chapters in this book. However, two important aspects of geospatial health have not received the required attention. These two aspects are (a) understanding the limitations of different spatial analytical tools and (b) advancements in making geospatial environmental health information available in a usable format for health science researchers for certain emerging areas and also for clinicians for their medical practice. The purpose of this chapter is to explore the contribution of geospatial technology in relation to (a) emerging health science research and (b) clinical practice of medicine. This chapter also provides background information on relevant concepts, terminologies, technologies, and organizations, which are often unfamiliar to the geospatial health early professionals.

The emergence of the geospatial concept for human well-being and health is inherited from 2500 years of philosophical premises of medicine, geography, and social science, although now, it may appear as a technology. When Hippocrates (460–375 BC) noted the importance of environmental exposure in medical investigation, in his classic work, *Airs, Waters, and Places* (*aëre, aquis et locis*), he logically signified the importance of location and environment in medical practice (Miller 1962). Unfortunately, Hippocrates' astute observations have not been translated into today's medical practice. That is why, today, obtaining a history of environmental exposure is not a routine practice, in fact it is an exception (ATSDR 2015; McClafferty et al. 2015; Hart 2017).

Advances in Earth observations, progresses in overall spatial data quality and availability, user-friendliness of the software, and affordability of required processing power have made it possible to apply geospatial tools in a wide range of health studies. Unfortunately, the use of such tools often occurs without the proper understanding of the theoretical developments of these tools and their strengths and weaknesses. Since the National Library of Medicine (NLM) first added "geographic information systems" as a MEDLINE indexing term in 2003, an extraordinary growth has been noticed. From 1994 to 2002, the growth in the number of GIS articles in MEDLINE was four times greater than articles in general health (Pickle et al. 2005). Whereas the increased emphasis on GIS is a recognition of its importance, the seemingly ease of use of the tools has also resulted in an increased number of poorly prepared papers. Journal editors often encounter surges of such papers and, unfortunately, some of those papers are published, which may hinder the sound progress of the discipline. To avoid the infiltration of poorly prepared manuscripts, more advanced discussions are needed about the limitations of spatial techniques. This book includes chapters on diverse geospatial health applications and also discussions on various limitations associated with applying geospatial techniques, including data, methodology, and available tools. This particular chapter highlights information that will further enhance the discussions on the potentials of geospatial technology in generating environmental exposure history for emerging health science research and for the use of disease investigation in clinical practice. To ensure that the readers are provided a relevant springboard, this chapter also discusses relevant concepts, terminologies, technologies, and organizations, with emphasis on contemporary and emerging issues. Hopefully, this chapter will provide readers a comprehensive idea of the newer potentials of geospatial technology

F. S. Faruque (✉)
Department of Preventive Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA
e-mail: ffaruque@umc.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_1

for our well-being and health irrespective of their level of experience.

Common diseases are the products of our genes, lifestyle behavior, and environment. Genes, lifestyle, and environment, in most cases, interact with each other causing different levels of disease risks (Fig. 1). In most cases, our genetic blueprint alone does not determine our health. In terms of disease risk factors, two different persons carrying two different types of gene will respond differently in response to the same environment and lifestyle. On the contrary, identical twins, having the same genetic blueprint, will respond differently to their different lifestyle and environmental exposures. A common saying "genetics load the gun, but environment (environmental exposure and lifestyle) pulls the trigger" makes it easy to understand the role of the environment in disease development. An individual's surrounding environmental factors are location specific, and that person's lifestyle also to some extent is location dependent. Thus, geospatial technology plays a critical role in disease risk factor analysis.

With the emerging interest in population health science, the use of geospatial technology may provide a unique opportunity to broaden our understanding of the multifactorial pathways that produce health and health disparities at the population level. In the case of the vulnerable population, geospatial technology is even more important to identify their surrounding environment. Subtle differences in genetic makeup can cause two individuals to respond differently to the same environmental exposure. As a result, some people may develop a disease after being exposed to certain environments and lifestyle exposures, while others may not. Such differences in response make it critical that vulnerable population and their healthcare providers are aware of the environmental conditions, particularly of the vulnerable people (NIEHS 2020).

Current studies on gene-lifestyle-environment interactions are shedding light on the causes of many diseases, their therapeutic solutions, and method of preventions. Characterizing these gene-lifestyle-environment interactions is particularly important to develop meaningful prevention strategies. Because of the importance of both gene (G) and environment (E), and their interactions, scientists are engaging in GxE studies. Generating environmental exposure information for the population (or even for an individual) is now possible through the proper use of geospatial technology, which, in fact, is an emerging area for geospatial professionals.

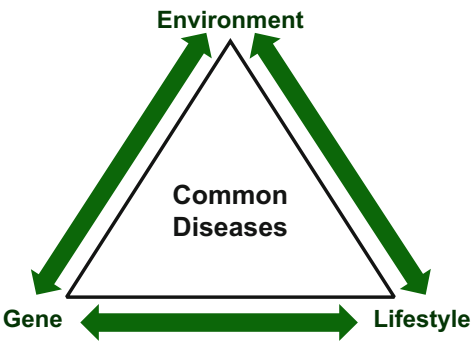


Fig. 1 Major causes of common diseases are known to be the environment, genes, and lifestyle. Interactions of these variables are responsible for a wide array of diseases

While medical science recognizes environmental exposure as one of the three major risk factors for common diseases (Fig. 1), the investigations on environmental exposure history is not a common practice during patient diagnosis in clinical settings. In most western countries, the doctor's office collects information from their patients during the first visit related to family history and habits, which could be grouped as gene and lifestyle categories. However, the other major risk factor, environment, is missing in this information collection practice.

Outstanding progress has been made in generating information on environmental exposures. Maantay and McLafferty (2011) eloquently discussed the role of geospatial technology in environmental health. Geospatial technology is now advanced enough to reveal complex associations between environmental exposure and health outcomes by incorporating multivariate and nonlinear spatial modeling. However, the full benefits of the advancements of geospatial technology have not reached the hands of all sectors of health science researchers and practicing clinicians, certainly not in a readily useable format. As the geospatial experts, who are working in the field of environmental health, are able to generate a "profile" of spatiotemporal environmental exposure, it is their responsibility to build the bridge with the health science and medical communities. Such engagement can guide the formulation of environmental health exposure information per the needs of the emerging health science research and healthcare practice. To harvest the results of the progress, the geospatial environmental health community needs to work with the healthcare providers to make environmental health information an essential part of the disease diagnosis system. This introductory chapter is expected to lead readers to comprehend the progress and potential of geospatial technology for human well-being and health.

Emergence of Geospatial Technology for Human Well-Being and Health

Over the last three decades, discussions on human well-being and health gained momentum, influencing the beginning of a shift from reactive care to proactive care, prevention, and precision medicine. Around the same time, applications of geospatial technology were developing more interdisciplinary approaches. Such approaches enabled geospatial

technology to explore the complex interrelationships between multidisciplinary variables related to human well-being and health.

Space technologies, the Earth observation (EO) in particular, have significantly contributed to the generation of environmental knowledge, shedding light on newer dimensions of well-being and health. Because of their position in orbit around the Earth, remote sensing satellites offer a unique vantage point and generate data streams that allow us to observe environmental changes, including the wide-ranging effects of climate change (ClimateAction 2020). The concepts of planetary approaches to study the health of the Earth and its living organisms clearly require the Earth observation (EO) data and geospatial technology. Planetary epidemiology, a term introduced by Colin Butler (2018), is a perfect example that requires applications of these data and technology to implement this sub-discipline of epidemiology.

Different space agencies have played critical roles in the evolution of geospatial environmental health by generating Earth observation data, inventing different analytical tools, and providing knowledgebase expertise. Since the 1970s, when Earth-observing satellites started collecting global environmental information, the geospatial community did not take long to engage in human well-being and health studies utilizing resources available from the space agencies. These studies got further momentum when Earth observation satellite data became easily available to academicians and researchers. Among the geospatial community, utilization of Earth observation (EO) of the environment for human well-being and health is now well-established (Hay et al. 1997; Herbreteau et al. 2007; Arvor et al. 2011; Wigbels 2011; Hamm et al. 2015; Viana et al. 2017; Parselia et al. 2019; Marti et al. 2020).

With increasing demands, some space agencies have even started specific programs to support public health studies. One such initiative by NASA was the Air Quality Applied Sciences Team (AQAST), which is still active. NASA also developed a set of tools, delivered online, known as Giovanni, which has a wealth of resources for public health research (Acker 2021). Other national/regional space agencies have also made progress in health applications. Notably, the Japan Aerospace Exploration Agency (JAXA) has developed a platform specifically for public health studies called public-health monitoring and analysis platform (Oyoshi et al. 2019).

Organizations Supporting Geospatial Applications for Human Well-Being and Health

Many of today's human well-being and health-related issues require international solutions. Some nongovernmental organizations play important roles directly or indirectly to shape the policy and research directions in the area of geospatial health. However, the structural relationships among these organizations may not be very apparent. Nevertheless, these organizations are important for researchers as they play roles in research prioritization. These organizations often provide funding support directly to researchers, but more importantly, they work closely with different national research and funding agencies and thus may have a larger impact. Below are few examples of such major organizations.

International Science Council (ISC)

The International Science Council (ISC) works at the global level to catalyze and convene scientific expertise, advice, and influence on issues of major concern to both science and society (ISC 2020). It is a nongovernmental organization with a global membership of 40 international scientific unions and associations and over 140 national and regional scientific organizations, including academies and research councils.

The ISC was created in 2018 after a merger between the International Council for Science (ICSU) and the International Social Science Council (ISSC). Now this organization brings together the natural and social sciences forming the largest global science organization of its type. Through its members and associates; its partnerships with other international scientific organizations, UN agencies, and intergovernmental bodies; and its wider networks of expertise, the Council is engaged in bringing together scientific excellence and science policy expertise from all fields of science and all regions of the world.

The unions and academy of sciences, under the umbrella of ISC, deal with science and its promotion as well as generate scientific knowledge and evidences for policymaking. However, the space of and interlinks between the individual ISC organizations are not necessarily very clear. Ismail-Zadeh (2016) attempted to clarify the common goals and activities of these professional societies and international unions and presented their differences as well. A special issue of the History of Geo- and Space Sciences on the "The International Union of Geodesy and Geophysics: from different spheres to a common globe" was published describing associated unions in further detail (Ismail-Zadeh and Joselyn 2019). It is important for the geospatial health community to keep up with these organizations as the breadth of wellness and health is very wide, and almost all of these organizations, comprising a diverse scientific community, directly or indirectly shape the research trend in this area. As mentioned earlier, these organizations are linked with the United Nations and thereby linked with the governments of different countries and also with the national agencies of those countries.

The following organizations and bodies within the ISC are directly or indirectly involved in geospatial health-related activities:

GeoUnions

One of the most effective international networks that benefit geospatial health studies is the International Geoscientific Unions or GeoUnions. With thousands of scientists from all over the world, the GeoUnions have been coordinating and promoting international efforts in Earth and space sciences since the beginning of the twentieth century (Ismail-Zadeh 2016).

In 2004, representatives of several of the GeoUnions met in Paris to establish a partnership to better promote the geosciences worldwide, to communicate and to coordinate the scientific activities of individual unions, and to gain recognition by ICSU (now ISC) bodies, the United Nations organizations, and other global stakeholders (Joselyn et al. 2019).

As a network of the International Science Council (ISC), the GeoUnions include those organizations who deal with Earth and space sciences, such as the International Geographical Union (IGU), the International Union of Geodesy and Geophysics (IUGG), the International Cartographic Association (ACA), the International Union of Geological Sciences (IUGS), and the International Society for Photogrammetry and Remote Sensing (ISPRS). It should be noted that ISPRS has its own Working Group on Environment and Health with one of its goals to bridge the geospatial, Earth, and health science communities for exploring interdisciplinary collaborations to improve our overall well-being and health.

GeoUnions are of special interest to the geospatial community who are working at the new frontiers of well-being and health as it requires a broad spectrum of understanding and collaboration. The roles that ISC and GeoUnions have been playing in human well-being and health are reflected in their reports and cited in other literature as well (Budge et al. 2009; ICSU 2011; Morain and Budge 2012; Bai et al. 2012).

Group on Earth Observations GEO

The Group on Earth Observations GEO is different from the ISC organizations. However, it has unique importance for the geospatial well-being and health community. GEO is an intergovernmental partnership working to improve the availability, access, and use of open Earth observations, including satellite imagery, remote sensing, and in situ data, to impact policy and decision-making in a wide range of sectors (GEO 2020). Currently, it has 112 member governments (https://www.earthobservations.org/members.php). GEO's member governments, participating organizations, and associates work together to develop and implement Earth observations projects and initiatives that address global environmental and societal challenges. GEO has a wide range of participating organizations (currently 133), including ISPRS, UN, and WHO. GEO participating organizations benefit from the global community of Earth observation experts to learn and share knowledge in the area of the GEO engagement priorities, namely, climate change, disaster risk reduction, and the UN Sustainable Development Goals.

For the geospatial community, GEO plays a special role through its Global Earth Observation System of Systems (GEOSS), which integrates observing systems and shares data by connecting existing infrastructures using common standards (GEOSS 2020). There are more than 400 million open data resources in GEOSS from more than 150 national and regional providers including NASA and ESA, international organizations such as World Meteorological Organization (WMO), and the commercial sector such as DigitalGlobe.

GEO has several initiatives related to health, such as its Earth observations for Public Health Surveillance, which utilizes Earth observations for public health alerts on air quality, outbreaks of disease carried by water-borne vectors, and assessments of access to healthcare and helps achieve Sustainable Development Goal (SDG) Goal 3 on Good Health and Well-being. Another initiative is the GEO Health Community of Practice (Geohealthcop 2020), a global network of governments, organizations, and observers, seeking to use environmental observations to improve health decision-making at the international, regional, country, and district levels. EO4HEALTH (EO4HEALTH 2020) is an element of the GEO Health Community of Practice (CoP), engaged in the development and elaboration of the CoP Work Plan. The CoP Work Plan will be aligned with the EO4HEALTH objectives and includes workgroups on seven specific topics: (1) heat; (2) infectious diseases; (3) air quality; (4) food security and safety; (5) healthcare infrastructure; (6) crosscutting issues; and (7) integrating EO data techniques.

GEO continues to focus on societal benefits, encouraging a diverge utilization of EO. To support the response and recovery actions related to the COVID-19 pandemic, the GEO Work Program activities, GEO Members, Participating Organizations, and Associates are using EO in diverse projects in many different countries.

Space Agencies During COVID-19 Crisis

During the crisis of the COVID-19 pandemic, space agencies came forward to play a responsible role utilizing their resources in a variety of ways to reveal different aspects of this unprecedented phenomenon. The notable areas of research are (a) environmental factors that affect survival and spread of SARS-CoV-2, (b) association between people's environmental exposures and COVID-19 outcome, and (c) impact of COVID-19 lockdown on the environment.

NASA made available some unique research opportunities during the early onset of this pandemic. One specific area was to model the epidemiological time series utilizing NASA

data. Experience from previously used infectious disease spreading models became surrogates for some of the SARS-CoV-2 spreading studies. NASA encouraged scientists to use information from NASA's Earth-observing satellites, on-the-ground sensors, and computer-based datasets to study the environmental, economic, and societal impacts of the COVID-19 pandemic. In addition, the agency's Earth Science Division sponsored new projects to examine the effects of shutdowns that brought changes to the environment, especially the atmosphere (NASA Earth Science 2020). Another topic of research of this initiative was the role of natural environmental phenomena that may impact the spread of the pandemic.

NASA's other initiative was to encourage citizen scientists around the world to solve challenges related to COVID-19 using NASA's open-source data in an all-virtual, global "hackathon" on the following four themes (NASA Earth Science 2020):

1. Learning about the virus and its spread using space-based data
2. Local response/change and solution
3. Impacts of COVID-19 on the Earth system/Earth system response
4. Economic opportunity, impact, and recovery during and following COVID-19

The Japan Aerospace Exploration Agency (JAXA) is another major national space agency that made several quick but major steps to respond to the COVID-19 pandemic. From early on, JAXA had started to study the effects of COVID-19 on economic activities by analyzing CO2 concentration variations using GOSAT data over major cities. JAXA also extended cooperation for international collaboration among other space agencies to use satellite data to contribute to the various measures taken against COVID-19.

As a regional body, the European Space Agency (ESA), in coordination with the European Commission, launched a special edition of the Custom Script Contest, focusing on the support of space assets during the COVID-19 crisis. The expectation of this initiative was to generate new ideas on how satellite data could help monitor and mitigate the situation for the upcoming months, while the world would organize to get back to business and would need to adapt from this crisis.

Different space agencies in China, including governmental and commercial, came forward to provide their remote sensing resources studying various aspects of COVID-19 at local, national, and international scales and in supporting the control of this pandemic.

In response to the COVID-19 pandemic, the Group on Earth Observations (GEO) has taken several initiatives, such as GEO Community Response to COVID-19, Teleconferences on COVID-19 Activities, and Webinars (http://www.geohealthcop.org/vision-and-goals).

It is worthy of including here a note from the ESA, written in collaboration with the European Commission and their other collaborators (Cheli 2020).

ESA Earth Observation Input
For the introductory chapter of the Springer publication:
"Application of Geospatial Technology in Prevention and Improvement of Human Health"

Space technologies, and Earth Observation (EO) in particular, provide key contributions to the generation and accuracy of geospatial knowledge and information. Because of their position in orbit around Earth, remote sensing satellites offer a unique vantage point and generate data streams that allow us to observe environmental changes, including the wide-ranging effects of climate change.

The European Space Agency (ESA) plays a key role in the European EO ecosystem. Its R&D programme prepares all future EO missions developed by ESA, in particular new science missions called Earth Explorers, and new capabilities for operational monitoring in meteorology and for use within the Copernicus programme. ESA also provides technology development for national and commercial missions of its Member States.

Copernicus is European Union's independent operational EO system, under the leadership of the European Commission, for which ESA develops and implements the space component: the Copernicus Sentinel missions and the related ground segment infrastructure. European EO also relies on a strong symbiosis between public and private entities, as the commercial sector and national missions feed high-resolution satellite data into the public programmes.

Thanks to Copernicus, the general public, downstream users and decision-makers have been able to witness some direct effects of the COVID-19 impacts from space. For instance, it showed how the levels of air pollution caused by NO2 emissions above global cities and industrial areas significantly dropped. This occurred as a consequence of lockdown measures: traffic was decimated and a lot of (fossil-fuel powered) industrial production was put temporarily on hold.

Copernicus also revealed information on the functioning of our economy, as it is capable of observing grounded planes at airport tarmacs, changes in large-scale agricultural productivity, traffic jams at

(continued)

closed or obstructed borders and ship and oil tankers waiting at sea to deliver their goods onshore. In short, Earth Observation can tell a great deal about human economic activity, and about the mobility of goods and people so crucial to the globalised economy of the twenty-first century.

In order to make optimal use of the current infrastructure and data streams of the various EO programmes, ESA has joined forces with the American National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA). Their coordinated action ensures that space agencies can maximise their contributions to the relief efforts and to help support the restart of the economy, sharing relevant EO data, developing joint methodologies and results on topics of common interest and on key selected supersites areas such as Tokyo, Los Angeles and France. ESA is also working together closely with the European Commission to support actions within a European context, for instance to monitor certain areas of particular strategic interest with greater attention thanks to updated satellite tasking[a].

Furthermore, EO provides contributions to our ability to improve human health in the future. In the specific case of viruses, for instance, scientific studies have confirmed that their prevalence and increasing frequency is linked to the accelerating destruction of ecosystems. In this sense, trends observed from space – such as the rate of disappearance of rainforests and wetlands – can help in our assessment of medium and long-term risks and, it can provide the information needed to take sensible decisions in terms of environmental and agricultural policy.

But the link between EO and health is in fact much broader. Human health is very much impacted by the overall health of our planet and the state of our environment. Climate change and other environmental changes are having an ever-stronger impact on local or regional level. EO systems therefore help us to increase societal resilience, to support rescue and disaster relief efforts and, to help predict how the overall biosphere will likely change in the coming decades.

Finally, Earth Observation also brings indisputable facts and increased transparency to the public debate. In fact, it is now even demonstrating that political actions can successfully lead to environmental repair. Last year, the hole in the ozone layer above Antarctica was measured the smallest since 1982. The first signs of its recovery show that environmental action does pay off, even if that only starts to show decades later. This will become increasingly important as society needs to grow the public support to tackle the key challenge of our time.

The Coronavirus pandemic has painfully illustrated that in times of a health crisis, it is vital to have a reliable testing and diagnosis capabilities at hand. Earth Observation infrastructure, and Copernicus in particular, is our planetary health monitoring devise. A better understanding of our Earth's complex climate system will be indispensable if we are to win the battles against climate change and biodiversity loss for the sake of current and future generations.

Written by Simonetta Cheli, Director of Earth Observation Programmes & Head of ESRIN, ESA - European Space Agency

[a]*This is the activity of updating instructions of a satellite's operations, which can be done i.a. to change the observation pattern of overflying satellites above a given area.*

(continued)

Discussions on Relevant Concepts and Terminologies

This section will discuss a few concepts and terminologies that may be useful to those new to the field of geospatial environmental health. The following discussions are particularly relevant when applying geospatial technology for human well-being and health.

Geospatial Technology

Geospatial technology is typically referred to as a suite of technology that can acquire, display, or analyze geospatial data, including geographic information systems (GIS), global positioning systems (GPS), remote sensing (RS), and others. These technologies can be applied to analyze, monitor, and forecast well-being and health and reveal the complex relationships of the interacting variables.

Human Well-Being

Human well-being is a very broad concept encompassing different aspects of our lives and requires a discussion rather than a definition(s). Not all aspects of human well-being can be observed nor measured. As the term "well-being" is an abstraction to refer to the evaluations of the state of life or "being," there are many different approaches to label that situation (McGillivray 2006). There are also several indicators of human well-being in different practices to measure

different aspects of human well-being, some of which are still evolving, and beyond the scope of this book.

While discussing human well-being, it is important to mention about the Millennium Ecosystem Assessment (MA). The MA was a 4-year long international collaboration, launched by the United Nations Secretary-General Kofi Annan in June 2001 (WHO 2005). From this initiative, a series of very comprehensive reports linking ecosystem and human well-being were published. One of the interesting features of MA was to include component assessments undertaken at multiple spatial scales – global, sub-global, regional, national, basin, and local levels – which clearly emphasize the significance of location while studying ecosystem and human well-being (Millennium Ecosystem Assessment 2005).

In the Ecosystems and Human Well-Being: A Framework for Assessment report, it is stated that "Human well-being has multiple constituents, including basic material for a good life, freedom and choice, health, good social relations, and security. Well-being is at the opposite end of a continuum from poverty, which has been defined as a "pronounced deprivation in well-being." This report also noted that the constituents of well-being, as experienced and perceived by people, are situation-dependent, reflecting local geography, culture, and ecological circumstances, further indicating the role of the geospatial technology in revealing the complex interrelationship of human well-being determinants. The five linked components of well-being and ill-being are (Butler et al. 2003):

1. Material sufficiency
2. Security
3. Good social relations
4. Freedom and choice
5. Health

The five dimensions should be viewed across a spectrum and their abilities to reinforce each other. These dimensions and their interactions produce a person's state of being (Fig. 2). On the negative end of the spectrum, the state of being can be referred to as ill-being, where the interaction of the dimensions results in negative experience including stress, pain, and anxiety. On the other end of the spectrum, the experience of a good life with peace of mind is the product of the interactions of its five dimensions, referred to as well-being. Ill-being is a term mostly used in the phycology literature along with well-being to group indicators of psychological conditions. However, it has a place in health and state of life to connote the opposite side of well-being and can also be used as situation-dependent reflections of factors affecting human life. Merriam-Webster defines ill-being as a condition of being deficient in health, happiness, or prosperity and claims that this term was first used in 1840. It could be noted that the components of well-being or ill-being could be characteristics of the population at various spatial and temporal scales.

The US Centers for Disease Control and Prevention (CDC) have discussed well-being at length (CDC 2020a). In this discussion, they mention that societies with higher well-being are those that are economically more developed, have effective governments with low levels of corruption, have high levels of trust, and can meet citizens' basic needs for food and health. It is noticeable that the spatial attributes of these conditions can make it possible to examine their internal characteristics, interaction patterns, and impact on well-being.

CDC (2020a) listed different aspects of well-being examined by researchers from different disciplines. This list includes:

- Physical well-being (health)
- Economic well-being
- Social well-being
- Development and activity
- Emotional well-being
- Psychological well-being
- Life satisfaction
- Domain specific satisfaction
- Engaging activities and work

CDC also discusses the measures of well-being collected with different instruments, which the geospatial community recognizes as variables that can be spatially analyzed and attributed to the local population for assessment of status and improvement of policy.

Health

The term health has been modified over the years. It is well-known that health is no longer recognized as being free from illness or injury. A well-accepted definition of health is defined by the WHO (2020a), "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." However, since the inception of this definition in 1948, there has not been any shortage of criticism (Callahan 1973; Saracci 1997). Indicating the shortfalls of the WHO definition, several new definitions of health have been proposed. Among those, Bircher (2005) and Huber et al. (2011) are worthy of mention here.

Bircher (2005) proposed health as a "dynamic state of well-being characterized by a physical, mental and social potential, which satisfies the demands of a life commensurate with age, culture, and personal responsibility. If the potential is insufficient to satisfy these demands the state is disease."

Claiming that the WHO definition of health as "complete well-being" is no longer fit for purpose given the rise of

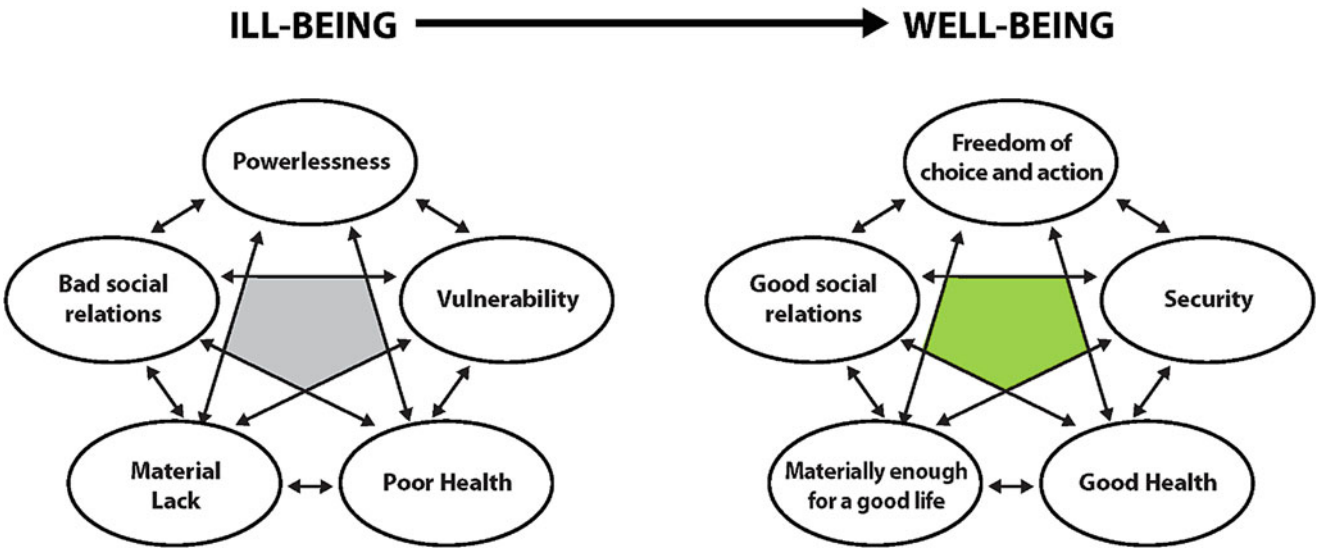


Fig. 2 Five dimensions of well-being reinforce each other, whether positively or negatively. A change in one often brings about changes in the others. The shaded space represents the experience of living and being – including stress, pain, and anxiety in the bad life and peace of mind and spiritual experience in the good life. (From *Ecosystems and Human Well-being: A Framework for Assessment* by the Millennium Ecosystem Assessment. Copyright © 2003 World Resource Institute. Reproduced by permission of Island Press, Washington, DC)

chronic disease, Huber et al. (2011) propose changing the emphasis on the ability of adapting and self-managing in the face of social, physical, and emotional challenges. They propose the formulation of health as "the ability to adapt and to self-manage."

Given the importance of the issue of the definition of health, the British Medical Journal (BMJ) published an editorial in December 2008, authored by Alex Jadad and Laura O'Grady which called for a "global conversation" about how to redefine health (BMJ 2008).

Readers are encouraged to carefully read the responses, published in the BMJ, to the proposed definition of health by Huber et al. (2011) to see how most of the responders indirectly pointed out the importance of social aspects of health beyond the control of an individual (BMJ 2011).

Irrespective of all these views, the fact is that there are inequalities in well-being and health due to the unequal distribution of the contributing factors. Therefore, geospatial technology can be used for investigating the patterns and interactions of these root contributing factors over space and sometimes overtime. Identifying the communities with inequalities is the first step to develop strategies for eliminating the causes.

Health is not a single phenomenon. Health is our daily life, our community, our country, our world. Health is our room, our home, our street, our river, our ecosystem, our planet. Geospatial technology is capable of dealing with the multidimensionality of all these contexts. The ability of geospatial technology to handle the location of events at multiple scales is its unique strength, and it can be coupled with other emerging technology for complex analysis of the nonlinear world.

Well-Being and Health

The definition of health by the WHO explicitly links well-being with health. Figure 2 presents the manner by which the expression and interactions of those dimensions produce a state of being. A significant driver of the expression of the dimensions is social determinants of health (SDOH) and conceptualizes health as a human right requiring physical and social resources to achieve and maintain.

Figure 3 depicts the importance of recognizing the need to appreciate how biomedical interventions need to be merged with socio-environmental strategies to improve well-being and, eventually, our health. The proactive roles of the key players who can make positive changes through required interventions are critical for the outcomes. Policy is certainly the driving force at different levels. However, along with other stakeholders, healthcare providers have critical roles too in improving our well-being and health. Interestingly, Allen et al. (2013), in their report "Working for Health Equity: The Role of Health Professionals," discussed how physicians could play roles even in further upstream, i.e., to the root causes by improving the SDOH, which eventually can improve well-being and health. It could be noted that England, under their Health and Social Care Act 2012, created statutory bodies called Health and Wellbeing Boards (HWBs) as a forum in which key leaders from the local

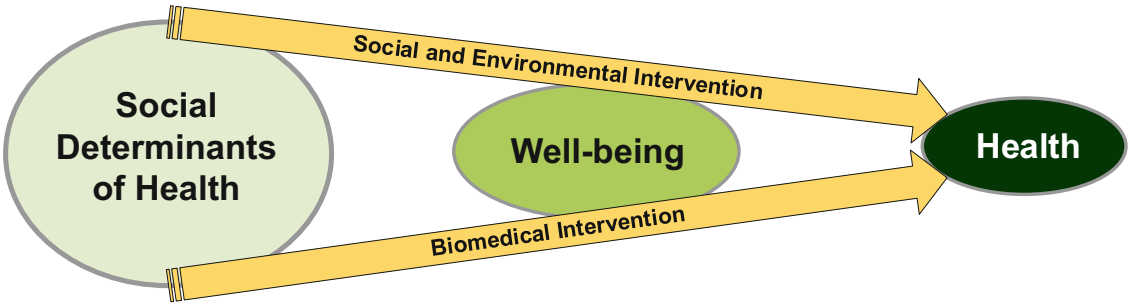


Fig. 3 Simplified linear relationships of SDOH, well-being, and health

care and health system could work together to improve the well-being and health of their local population (Greaves and McCafferty 2017; Coleman et al. 2016). The degree of success in the implementation of HWBs may have controversy, but the idea of working stakeholders together to address the root cause of health determinants was novel. In their report, Allen et al. (2013) described how physicians could also help to ensure taking into account a range of social, economic, and environmental factors to improve the health of the local population in the local health and well-being strategy. Although involving clinicians in the well-being and health of local people beyond clinical settings was a major step forward, this initiative was criticized for not addressing the academic training and opportunities in the current medical practice setup. The success of involving clinicians in "health and well-being" of local people will largely depend on the education of the clinicians in understanding the SDOH. While accessing and utilizing the SDOH data, clinicians will have a better appreciation of the utilization of geospatial technology in addressing the root causes of health determinants.

Social Determinants of Health

The CDC (2020b) states that social determinants of health (SDOH) are conditions in the *places* where people live, learn, work, and play that affect a wide range of health and quality-of-life risks and outcomes. Readers are encouraged to look at information resources from HP2020 (2020), HP2030 (2020), CDC (2020b), and WHO (2020b) to have a broader idea of how SDOH, human well-being, and health are related. Location is an integral part of all the seven topic areas of SDOH as listed in HP2020. Even if we take this simplified relationship of SDOH, well-being, and health (Fig. 3), it should be clear that each and every domain and the intervention phases have spatial attributes.

Of particular importance to this introductory chapter is the promotion of a place-based concept of five key areas of SDOH, i.e., economic stability, social and community context, education, neighborhood and built environment, and health and healthcare. Geospatial technology can assist in measuring, analyzing, and revealing the patterns and interactions of the variables of these domains and intervention phases. Through this place-based approach, each of the five domains of SDOH (Fig. 4) and their elements can be improved through proper planning and measurements with an objective to improve well-being and health. Such improvement initiatives must begin with an assessment of the pre-improvement status, which leads to proper intervention. However, the results and impacts of the improvement must be monitored to understand what worked and what did not, which can be referred to as three major steps: Assess, Intervene, and Monitor (AIM). The AIM steps can be performed using geospatial technology for the best possible specificity. Lack of specificity is often responsible for the failure of many population health improvement projects.

Assess, Intervene, and Monitor (AIM)
Assess: Assess the existing conditions of a community problems and resources
Intervene: Intervene strategically to improve the conditions
Monitor: Monitor the impacts qualitatively and quantitatively

AIM is a geospatial tool-based precise approach to improve the community health improvement program.

Community

The word "community" has different meanings for different professionals. However, to the geospatial health professionals, a community should be very specific and measurable. A group of people with a common characteristic or interest belongs to a community. On the other hand, a group of people living within a geographic boundary belongs to a community. In a geospatial context, communities are bounded by geography and are place based.

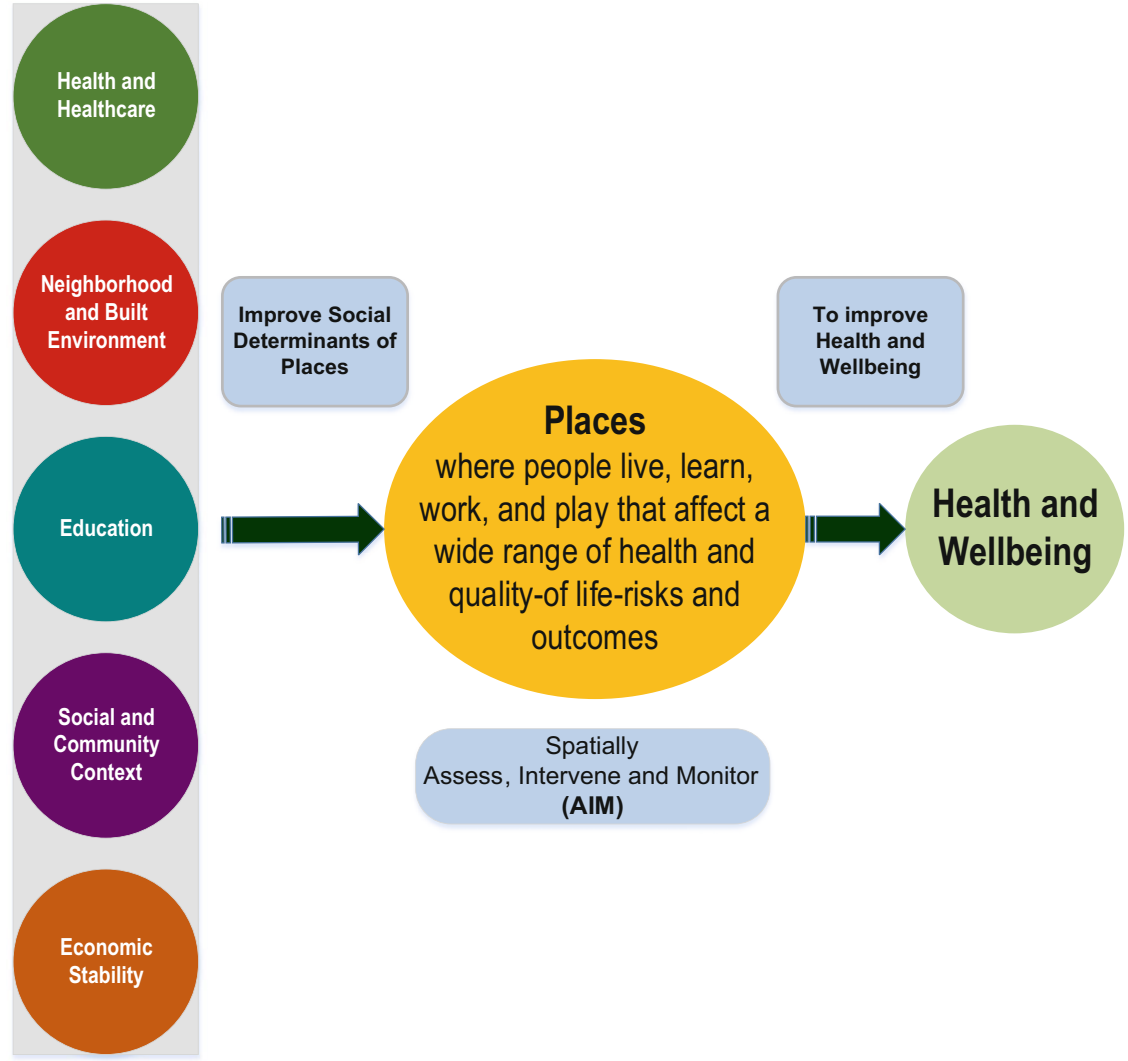


Fig. 4 A "place-based" framework, reflecting the concept of improving the five key areas of social determinants of health (SDOH) to improve our well-being and health

The community approach is critical because if every community becomes active to ensure its well-being and health, the entire nation improves. Why do the initiatives need to be at the community level instead of the national or regional or state level? Because the community itself knows its needs better than anybody else. Every community is somewhat unique with its strengths and weaknesses, which need to be incorporated into the action plan for well-being and health. Since community-based actions are place-based, geospatial technology becomes the obvious tool for the assessment, intervention, and monitoring (AIM) steps for community well-being and health improvement initiatives. Particularly, geospatial tools can be of great assistance in reducing waste and misuse, which is not uncommon, by monitoring the usage of resources and desired outcomes.

Neighborhood

A neighborhood is a place where people live near one another, usually having distinguishing characteristics, such as similarities in types of families, incomes, and education levels. In a geospatial health context, a neighborhood is the smallest geographic unit, where assessment, intervention, and monitoring (AIM) can take place to improve the well-being and health. Now, the questions are how big a neighborhood should be and how it can be brought into a measurable framework. There is no formal definition for the size of a neighborhood. While the availability of data and research designs may lead to different geographic units, in the USA, the most popular geographic unit used as a neighborhood is the Census Tract. When Earth observation data from satellites are incorporated in to a population health study, an artificial grid-based unit can be used as a proxy for the neighborhood

unit, and data from the Census or other sources are transposed to that grid. As a neighborhood is an area where people live and interact with one another, it has synergetic strengths improving its well-being and health.

Emerging Areas of Health Science Research and Geospatial Technology

Exposome

The concept of exposome refers to the totality of exposures from a variety of external and internal sources and how these exposures relate to health. About 15 years ago, shortly after the human genome was sequenced, Christopher Wild proposed the term exposome as an environmental complement to the genome in determining the risk of disease (Wild 2005; Dennis et al. 2017). When first introduced by Dr. Christopher Wild (2005), the term exposome seemed to be a wild idea (Miller and Jones 2014). Nevertheless, it immediately caught the attention of a diverse group of scientists who were studying the genome, environmental exposure, and health. The definition of the exposome has undergone several revisions; however, its underlying premise has remained the same.

The exposome comprises every exposure to which an individual is subjected, from conception to death (Wild 2005; Wild 2012; Miller 2020). It could be noted that the exposome concept considers the lifelong exposure history and therefore requires taking into account for exposure over time, or at least at more sensitive stages of life. The manner how the exposome interacts with a person's unique characteristics will, to a large extent, determine that person's trajectory of "successful" aging. Wild (2012) clarifies that in the context of the exposome, the environment comprises of "non-genetic," and the exposome complements the genome by providing a comprehensive description of lifelong exposure history.

While decoding of the human genome has helped explain the underlying causes of disease, it has left certain gaps in understanding the big picture. In fact, genetic factors are not the major causes of chronic diseases (Rappaport 2016). Without consideration recognizing the contributions of the environmental exposures, the picture of disease risk factors is incomplete. Accounting for the interactions of environmental factors with biological systems is providing a much greater understanding of disease etiology. The causal links among the genome, the environment, and human disease have made the exposome an integral part of modern health science (Vineis et al. 2020; Martin-Sanchez et al. 2020). Now, environmental exposures and genetic variation can both be considered in studying the causes of disease burden.

The increasing knowledge about the exposome is already causing a shift in the paradigms of studying environmental health, exposure science, and biomonitoring (Lioy and Rappaport 2011; Rappaport 2011, 2012, 2018; Rappaport et al. 2014; Vrijheid 2014; Siroux et al. 2016; Dennis et al. 2017; Stingone et al. 2017; Niedzwiecki and Miller 2017; Guloksuz et al. 2018; Steckling et al. 2018; Niedzwiecki et al. 2019; Sarigiannis 2019; Vermeulen et al. 2020). As a result, both clinical medicine and population health are expected to gain a significant increase in the utilization of the exposome approach (Rappaport 2011; Barouki et al. 2018; Niedzwiecki et al. 2019; Martin-Sanchez et al. 2020). While the technology to measure the exposome continues to advance, the horizon in environmental health is broadening by combining biomarkers and external exposures. Measurable links between environmental exposure and health or disease are biomarkers. Biomarkers are key molecular or cellular characteristics that can link a specific environmental exposure to a health outcome. These markers have been used to associate diseases with environmental exposures and are also useful to identify vulnerable people at increased disease risk. Biomarkers are better represented by omics-level measurements. Advancements in lab technology have led to the development of omics technologies, which is a huge breakthrough for exposome research. In a single experiment, large amounts of data about a specific type of molecules can be obtained and analyzed, using omics technologies (Quezada et al. 2017). Omics-based biomarkers are key for the new generation population health studies and clinical practices. Recent studies are providing key findings and new concepts that the combined use of data generated by omics and geospatial technology will lead to innovative solutions for next-generation medical science (Gulliver et al. 2018; Vineis et al. 2017; Juarez and Matthews-Juarez 2018; Vineis 2019; Canali 2020). However, the active collaboration will be the key determinant for progress in this multidisciplinary approach.

Wild (2012) outlined the exposome as three overlapping domains: (a) general external, (b) specific external, and (c) internal. The general external refers to exposures that people at the community level are subjected to a variety of environment, including chemical (e.g., lead, ozone), biological (e.g., bacteria, viruses, fungi), physical (e.g., noise, heat, cold, altitude), and social (e.g., crime, food insecurities) environments. Sillé et al. (2020) recognize the ability of geospatial technology to measure the general external exposome. Specific external refers to individual-level lifestyle-related exposures, such as diet, smoking, physical (in)activity, chemicals, occupational, etc. Our health status is shaped by the manner by which the specific external (lifestyle) and general external (unavoidable) environment interact with the internal environment (internal biochemical perturbations due to external exposures) (Sillé et al. 2020). Miller (2020) argues that the concept of the exposome should be integrated instead of categorizing it into different domains; otherwise, it may allude to as if there are different kinds of exosomes. Wild

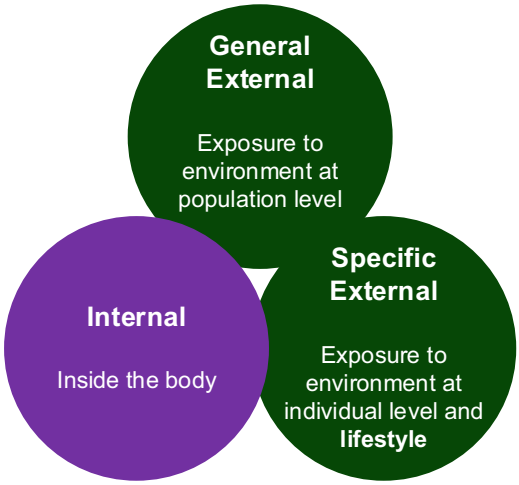


Fig. 5 Domains of exposome. (Modified from Wild 2012)

(2012)) himself mentions that there are overlaps in the three domains and sometimes may be difficult to place a particular exposure in one domain or another. However, for the sake of simplification and our discussions in this chapter, such categorization will suffice (Fig. 5).

(a) General external exposome (population-level exposure)
(b) Specific external exposome (individual-level exposure)
(c) Internal exposome (occurring within the body)

As shown in Fig. 5, both the general and specific exposomes reflect the exposure to the surrounding environment, respectively, at the population level and individual level, while specific external included individual lifestyle as well. The internal exposome is not an isolated entity because general and specific external exosomes also contribute to this domain. Individual genetic variety determines the nature of the internal response to the external exosome. Whereas individual genetic traits may determine the degree of the internal response to the external exosomes, genetic traits may also contribute to a person gravitating to certain lifestyles and community exposures. So, all three of these domains are impacted, to a varied degree, by location-based environmental factors, which ultimately contribute to producing a trajectory of health and aging. The purpose of the discussion on exposome in this chapter is to make geospatial health experts aware of their roles and responsibilities in the emerging environmental health science research.

For the sake of further discussion, let us try to simplify the relationship.

$$\text{Exposome} = \text{External Exposome (Specific} + \text{General)} + \text{Internal Exposome}$$

Pollutome

The Lancet Commission on environmental health (2018) coined a new term the "pollutome," which is the totality of all forms of pollution with the potential to harm human health (Landrigan et al. 2018). The pollutome is a nested subset of the exposome – the total amount of pollutants an individual is exposed to during the life course.

Contributing Factors of the Exposome

Environmental Exposure

We live in a wide range of diverse environments, which could be categorized as chemical, biological, physical, and social. We are constantly subjected to different magnitudes of these exposures. These environmental factors can be considered as contributors to the external exposome exposure, which in combination with lifestyle and internal exposures result in non-uniform health outcomes. The ability to measure environmental exposures is advancing significantly through geospatial technology, and its data can be incorporated into association analyses to understand or predict health outcomes. Place is directly related to environmental exposures.

Lifestyle

Wild (2012) recognizes that lifestyle serves as a major contributor to one's individual-level external exposure. Lifestyle example includes tobacco or alcohol use, physical inactivity, physical activity and exercise, diet, and other behavioral factors that affect health. Often lifestyle is related to the built environment that allows for access to the above factors. Arguably, the built environment is a component of the general external exposure indicating how the external environments are linked. Access to sidewalks, biking routes, parks, trails, exercise facilities, and crime-free conditions are direct contributing factors to lifestyle exposure. So, although lifestyle is within the specific external domain, it is very much shaped by the general external environment. Place plays a role in lifestyle exposure regarding the availability or unavailability of resources.

Internal Processes

Endogenous or internal biological processes, such as metabolism, hormones, inflammation, gut microflora, oxidative stress, and ageing, also contribute to the exposome of an individual. Although internal processes are driven by the genetic characteristics of the individual, the internal exposome is highly dependent on both the general and the specific external exposures. The impact of specific external exposures is well-documented as, for example, evidence indicates that exercise significantly modifies internal processes and the microbiome (Mailing et al. 2019) and, consequently, the exposome. The impact of

specific external exposures on the internal process is also evident because the internal exposome is related to the external exposures representing the individual's response to environmental stimuli or his/her physiologic and biologic responses needed for maintaining homeostasis (Louis et al. 2017). As the external environment influencing the internal processes is place-based, geospatial technology capturing those environmental factors is useful for predicting at least part of the internal exposome.

Role of Geospatial Technology in the Era of Exposome

Geospatial technology has been extensively used for the assessment of population-based environmental exposure but also has the potential for the evaluation of individual-level exposure (discussed later in this chapter). However, in the context of the exposome, the advantages of geospatial technology have not been fully explored. Figure 6 conceptualizes how geospatial technology can contribute to assessing the big picture of the exposome and assist in environmental health studies at population and individual levels. Geospatial technology could generate environmental exposure information across different time periods of life. Uniquely, geospatial technology can also generate a person's cumulative environmental exposure by incorporating the location history of an individual.

In order to determine environmental exposures, both the location of humans and their surrounding environmental conditions are required. Geospatial technology clearly plays a critical role in assessing and analyzing environment-related variables (Ward et al. 2000; Schmidt 2005; Yoo et al. 2015; Jerrett et al. 2017; Faruque 2019; Sogno et al. 2020), as well as lifestyle-related variables (Charreire et al. 2012; Tamura et al. 2014; Chai and Kwan 2015; Tung et al. 2017). However, the literature is very limited, connecting the application of geospatial technology to the exposome studies. The geospatial community is lagging behind in this effort. Nevertheless, recent research has started to suggest the importance of the use of geospatial technology in exposome studies (Stahler et al. 2013; Robinson and Vrijheid 2015; DeBord et al. 2016; Dennis et al. 2017; Vineis et al. 2017; Maitre et al. 2018; Juarez and Matthews-Juarez 2018; Prior et al. 2019).

The Lancet recently published a commentary about spatial lifecourse epidemiology as a conceptual model of epidemiological thinking in reference to exposome (Jia 2019). Spatial lifecourse epidemiology aims to utilize advanced spatial, location-aware, and artificial intelligence technologies to investigate the long-term effects of measurable biological, environmental, behavioral, and psychosocial factors on individual risk for chronic diseases (Jia et al. 2020a, 2020b).

Many consider that the right time for exposome research has arrived and it will bring about a significant paradigm shift in exposure-health outcome research (Louis and Sundaram 2012; Sarigiannis 2019; Jia 2019; Martin-Sanchez et al. 2020; Sillé et al. 2020). The geospatial community should now explore how geospatial technology can be best utilized to engage in this rapidly flourishing exposome-based research approach. The full implementation of the exposome concept in population health and medicine may require more time, but the time to embark on this journey is now. As technological advances are increasingly becoming more precise in identifying biomarkers associated with normal biological or pathogenic processes, advances in geospatial technology need to materialize to properly capture and deliver external exposure measures to the researchers and clinicians for population- and individual-level use. Geospatial technology can provide precise data that can shed light on how external exposures may influence or modify biomarkers in population- and/or individual-level research studies.

Biomarkers are useful for measuring exposome, which represents cumulative external exposures at that point plus the internal processes in response to those external exposures. However, biomarkers do not specify the history of environmental exposures, which is critical for disease prevention at the population or individual level.

As Fig. 6 shows, geospatial technology cannot capture the exposome (total) or the internal part of the exposome. However, it does capture several important parts of the external domains of the exposome. To determine exposome, advanced measures of biomarkers are becoming available. Biomarker measures are supposed to represent the total exposures plus the responses from the internal processes without describing any history of external exposures. On the contrary, geospatial technology can capture the Geospatial Individual Environmental Exposure (GIEE), which represents the external exposures.

Let us look at the relationships in a geospatial context.

$$\text{Exposome} = \text{Internal Exposome} + \text{External Exposome (Specific} + \text{General)}$$

$$\text{Exposome} = \text{Biomarker} = \text{Internal Exposome} + \text{External Exposome}$$

$$\text{Exposome} = \text{Biomarker} = \text{Internal Exposome} + \text{GIEE}$$

This, of course, is an oversimplified relationship. However, it does indicate the role of geospatial technology in this emerging field. Coordinated efforts should be made in the area of exposome-based research while ensuring further advancements in the fields of biomarkers, genetic, epigenetic, and GIEE.

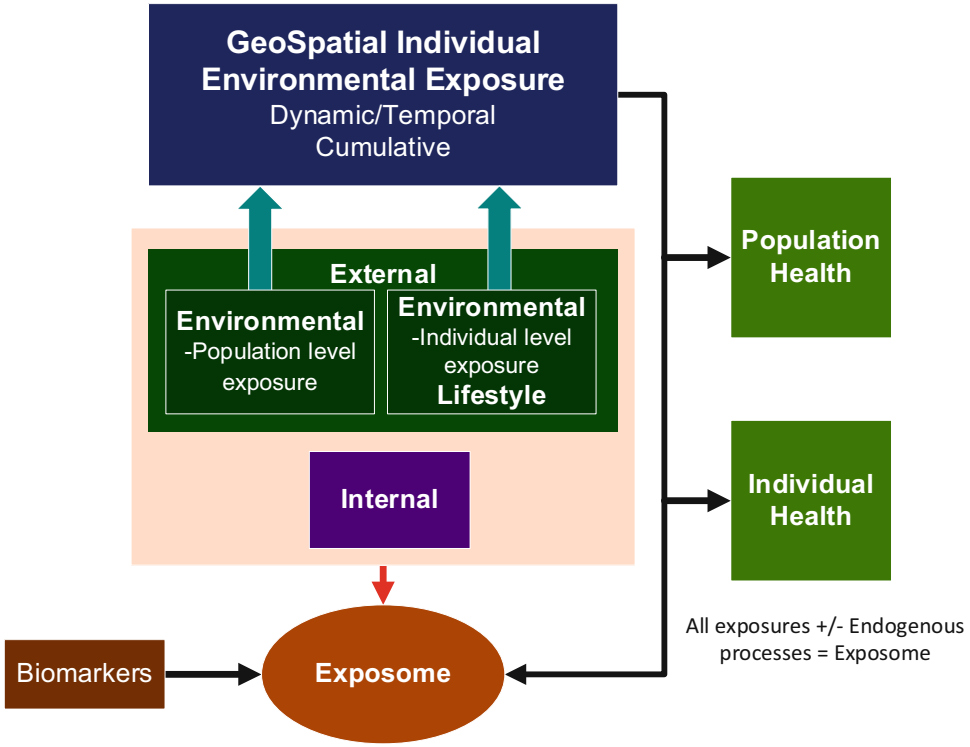


Fig. 6 Geospatial technology contributing to the paradigm of exposome

Advanced biomonitoring methods can provide improved exposome measures. However, these measures will not tell the exposure history. On the other hand, GIEE cannot give exposome measures, but it can give the exposure measures. GIEE can help to identify spatiotemporal environmental exposures that contribute to the exposome values. Hence, for population-based prevention research, GIEE can be instrumental. On the same token, GEE can generate environmental exposure history for individuals, which has the potential to be vital information in disease investigation in clinical settings. However, making environmental exposure history, which is one of the three major contributors to most of the diseases, available to the clinicians will take efforts from a diverse group of professionals, including IT, policymakers, legal experts, clinicians, and, of course, geospatial experts.

Epigenetics is the study of heritable changes in gene expression that occur without changes in DNA sequence (Wolffe and Guschin 2000). Epigenetics refers to mechanisms explaining how the environment impacts health. Exposome refers to measures of total environmental exposures and internal responses. Both acknowledge environmental factors that affect human well-being and health. Geospatial technology, which has been successfully generating information on environmental variables for human exposures, now has this unique opportunity to engage in this research. Data need to be captured on a finer scale, and the spatiotemporal dimension of environmental variables along individual dynamics needs to be captured to generate a personal exposure record.

In the exposome context, geospatial technology should be an integral part of environmental health studies. For example, the location of a mother during her pregnancy and the related physical and social environments of that place contribute to the long-term health outcome of the child. Therefore, time-stamped or spatiotemporal environmental information is required. Health scientists have already started to recognize how data generated through omics and geospatial technology can complement each other to interpret different health phenomena (Canali 2020). Biobanks around the world are becoming interested in omics-level data to support collaborative studies involving genetic, lifestyle, and environmental risk factors for diseases, which brings more opportunities for geospatial health professionals. Challenging but exciting tasks for the geospatial community are to generate geospatial environmental exposure information and to bridge with the medical professionals to properly support the study of diseases and treatment of patients.

Geospatial Individual Environmental Exposure (GIEE) can be defined as the geospatially tagged environmental exposure information for an individual. Such information should include cross-sectional as well as cumulative exposure of that person. GIEE should be useful for the exposome study and should have the potential for use in clinical investigation.

Geospatial Technology Contributing to Making Environmental Exposure Information Available to Clinicians in a Usable Format

The three major risk factors for common diseases are genes, behavior, and the environment. During a patient's visit to a physician, family history (genes) and lifestyle of the patient (behavior) are taken into account for disease investigation, but questions concerning possible environmental exposures are not included in the routine investigation. Ironically, the importance of the surrounding environment in disease development has been known since the very early history of medicine. Hippocrates, widely considered as the father of medicine, 25 centuries ago noted the importance of environmental exposure in the medical investigation (Miller 1962). Health scientists are now aware that certain environmental exposures can impact health conditions both through genetic and epigenetic mechanisms. Thus, without accounting for the environmental exposures, disease investigations using only hereditary and lifestyle information, therefore, can be incomplete. The major reasons for not considering environmental exposure in today's routine clinical practices are as follows:

(a) Medical education does not adequately prepare physicians to consider environmental history in disease investigation.
(b) Due to its dynamic and often obscured nature, environmental exposure history is difficult to obtain from the patients.
(c) Lack of access to environmental exposure information in a readily available format.

Since the 1990s, many studies have been advocating the needs for environmental health in medical education (Rall and Pope 1995; Goldman et al. 1999; Roberts and Reigart 2001; Kilpatrick et al. 2002; McCurdy et al. 2004; Gehle et al. 2011; Pelletier 2016; Walpole et al. 2017; Ihde et al. 2020; Brand et al. 2020). Similar calls have come from nursing education as well (Pope and Snyder 1995; Eddins 1998; Green 2000; McCurdy et al. 2004; Leffers et al. 2015). These calls are encouraging the inclusion of environmental health in medicine and nursing curricula. In the USA, medical schools are finding different ways to train students in environmental issues, including courses on environmental medicine. To address this gap, tutorials and other resources are also available from government agencies teaching how to take environmental exposure history (ATSDR 2015). The next-generation physicians are expected to have a much better understating of the importance of environmental exposure in disease investigation.

While the next-generation clinicians are getting prepared to incorporate environmental exposure in medical investigations, it is not clear whether the geospatial experts are fully aware and prepared to partake in this opportunity. Geospatial experts need to get prepared to engage in this new area of geospatial health of generating environmental exposure history for clinical practice.

In order to generate the patient's environmental medical history, two types of information are necessary: (a) relevant spatiotemporal environmental variables and (b) location of the individual in that environment. Advanced geospatial technology has been implemented to estimate common environmental agents such as pollutants, mold spores, pesticides, etc. Until recently, the other component, the location of an individual, was limited to a static representation such as residential or workplace location. Now, with the development of mobile technology, dynamics in an individual's location can be tracked even in real time. Technological advancements in both the areas, estimating environmental agents and identifying locations of individuals at flexible spatiotemporal scales, now present the potential of a paradigm shift in clinical practices by incorporating environmental exposure history into determining disease risk factors.

Needs for Environmental Exposure Information

Nearly all human diseases result from the interaction of genetic susceptibility factors and modifiable environmental factors (CDC 2000). Environmental factors include pollutants, which can be measured or estimated to assess the risk factors for the population or for the individual. Research on environmental epigenetics suggests a stronger impact of pollutants on individual health than previously known (Baccarelli and Ghosh 2012; Hou et al. 2012; Bollati and Baccarelli 2010; Tarantini et al. 2009).

Earth Observations and Geospatial Technology

In the early days of Earth observations, the resolutions of satellite images were much coarser. The utilization of satellite data was mostly restricted at regional ecosystem and landscape levels. These data were suitable for epidemiological health studies. Recent satellite images provide much higher resolution data suitable to be considered for community- and even individual-level analysis (Fig. 7). At the same time, the quality of image-derived products, even from the older satellites, is getting improved by applying advanced retrieval algorithms. Now advanced geospatial technology, by incorporating Earth observation data with data from multiple other sources, can generate information, which was not

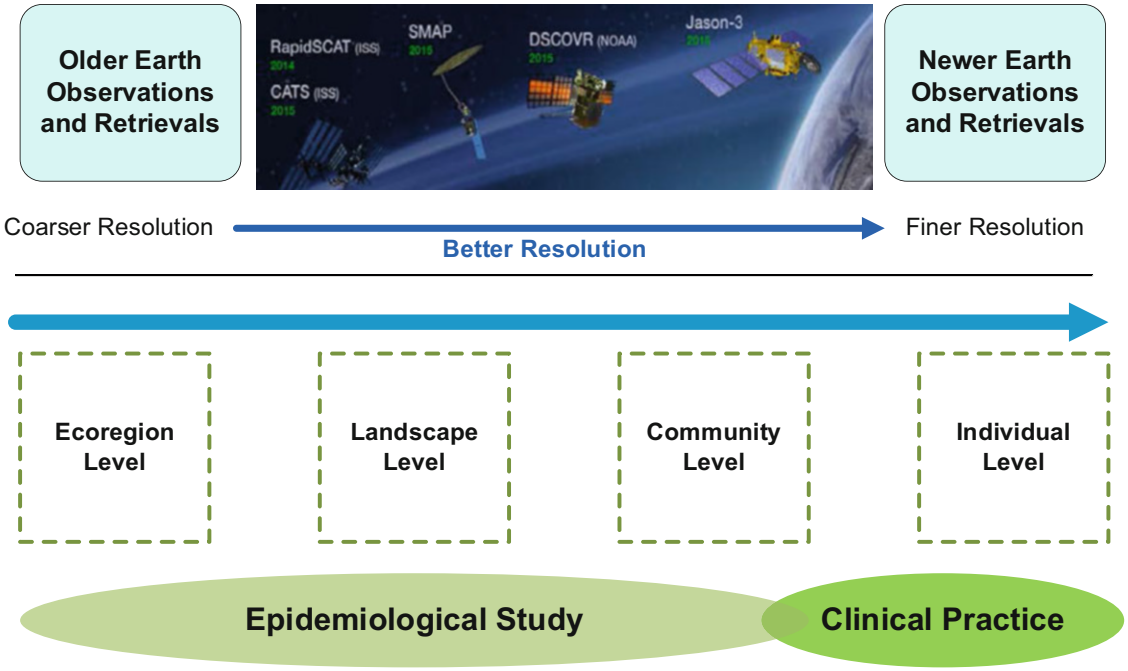


Fig. 7 Generalized scenario of the advancement of Earth observations and geospatial technology generating environmental data suitable for clinical practice. (From Faruque and Finley 2016)

possible earlier, yielding unprecedented new opportunities. This includes environmental exposure information suitable for determining environmental health risk profiles, useful in clinical practice.

Scale

Finer scale data can accommodate more detailed studies. However, the scale is determined during the data acquisition and preparation. Space agencies have been relentlessly trying to improve their data to better understand the integrated Earth system globally to benefit human health and as well as the health of the ecosystem (Space Studies Board 2015). The scale is a major factor determining the usefulness of the measurements, whether for individual human health or for the ecosystem. A unique strength of geospatial technology is its ability to handle multiple scales at the same time when the basic data can support finer scale.

A Millennium Ecosystem Assessment report on Health Synthesis by the WHO (WHO 2005) states that as the interactions and changes that affect human well-being can take place at more than one scale and also across the scales, a multiscale approach that simultaneously uses larger- and smaller-scale assessments can help identify important dynamics of the system that might otherwise be overlooked. While human health is influenced even at the ecosystem level, incorporating environmental exposure for clinical practice for individual health requires measurements at a much finer scale.

Figure 7 shows how finer scale data are becoming available due to the gradual advancements in Earth observation capabilities. The utilization of Earth observation data is commonly applicable in population-level health studies. Now, these EO data, in combination with data from other sources, can attain a finer scale and can generate environmental exposure history for individuals with potentials for individual-level application, such as for exposome studies and for clinical practices.

Earth Observations in Tracking Air Pollutants

Earth-observing satellite systems have been playing a major role in tracking a variety of air pollutants that are harmful to human health (Griffin et al. 2012). Particulate matters are among the worst air pollutants to cause multiple health hazards, and these pollutants have been estimated at various scales.

Particulate Matter (PM)

It is known that among the air pollutants, PM2.5 affects more people than any other pollutant. This air pollutant is responsible for a wide variety of adverse health conditions, including respiratory problems (Dominici et al. 2006), cardiovascular disease (Brook et al. 2010), cancer (Andersen et al. 2017; Pun et al. 2017), birth defects (Vinikoor-Imler et al. 2013; Guo et al. 2018; Alman et al. 2019; Huang et al. 2019), and neurological disorders (Kioumourtzoglou et al. 2016; Fu et al. 2019; Shi et al. 2020). Globally, two to four million annual deaths, more than malaria and HIV-AIDS combined, are associated with these fine inhalable particles (Anenberg et al. 2010; Lim et al. 2012; Lozano et al. 2012; Silva et al.

2013; Apte et al. 2015; Cohen et al. 2017). A comprehensive list of health hazards due to PM2.5 exposure is yet to be completed. During the current pandemic, evidences have emerged suggesting exposure to PM2.5 is responsible for higher mortality and morbidity due to COVID-19 (Wu et al. 2020a; Chakrabarty et al. 2020; Hendryx and Luo 2020; Borro et al. 2020; Becchetti et al. 2020).

Some satellite sensors have been successfully providing the total column atmospheric particulate matter. However, estimating ground-level particles (GLP) from satellite data is still evolving with a varied success. Techniques, such as machine learning, that can account for multivariate and nonlinear relationships show potential for generating more reliable ground-level PM2.5 from satellite-derived aerosols. Several researchers, including Lary et al. (2014), generated daily global PM2.5 estimates using a suite of remote sensing and meteorological data products validated using ground-based PM2.5 data. More recently, researchers from Duke University demonstrated how high-resolution microsatellite imagery using a machine learning algorithm could generate 200 m resolution PM2.5 (Zheng et al. 2020). Such studies show potential for estimating more reliable PM2.5 at better resolution utilizing newer satellite data and techniques. Scientists are continuing their efforts to generate improved GLP measurements from satellite data by incorporating other useful data and applying newer techniques (Chowdhury et al. 2019; Hu et al. 2019; Mhawish et al. 2020; Singh et al. 2020). Improved air quality indices for health hazards can be developed with such efforts when better data from Earth observation-based sources can be utilized. The level of variable detail and fineness of scale of environmental data required for providing adequate information for individual health exposure is a matter of discussion among geospatial experts, medical scientists, and clinicians. Nevertheless, this discussion will open up new challenges and opportunities for geospatial experts.

Harmful Airborne Fungal Spores (HAFS)

Airborne fungal spores impose significant health risks, particularly for vulnerable people. Exposure to harmful airborne fungal spores (HAFS) is known to cause a wide range of adverse health effects, mild to severe. While the impact of mold spores on health has been well-documented (Dales et al. 2000; Burney et al. 2008; Simon-Nobbe et al. 2008; Bousquet et al. 2009), information on the outdoor abundance of these spores is not available on a local scale because of the very sparse distribution of monitoring facilities. Currently available daily nationwide fungal spore abundance maps (Fig. 8, left) rely on a very limited amount of actual data, which are extrapolated over large areas often across multiple states. In a pilot study, spatiotemporal surface models were generated for six clinically significant spore types for the Central Mississippi region in the USA at 10 km resolution utilizing Earth observation data (Fig. 8, right) (Faruque and Finley 2016). Newer technology and data can support generating such daily estimates even at a much finer spatial resolution, which may allow physicians and vulnerable people to establish timely preventative strategies.

Technological Advancements

Satellite Data

Over the years, space agencies across the globe have achieved significant improvements in data quality, including improved spatial and spectral resolutions. In addition, there have been developments in specialized instrumentation necessary for health applications. For example, NASA funded a project to put new instruments in low Earth orbit to track the abundance and types of particulate matter at 1 km resolution (NASA 2016). Such initiatives can bring major breakthroughs in determining pollutant abundance using satellite data. This type of initiative is bringing the potential of using environmental exposures closer to medical practice.

Modern Technology

With the technological advent of wireless communication, low-cost air pollution sensors, and increased computational power, there is a potential for a paradigm shift in pollution monitoring. Personal monitoring devices and mobility trackers are capable of providing information about environmental exposures for individuals, which can be integrated with satellite-generated pollution data. Regardless of whether these types of measures are integrated with Earth observation measures or other monitoring networks, there will be a role for geospatial experts engaged in well-being and health research to make these measures readily available and usable to healthcare practitioners.

Internet of Things

The Internet of Things (IoT) is a rapidly advancing network of physical objects embedded with sensors, software, and other technologies that enables communication between electronic devices and sensors through the internet, bringing a wealth of information useful for our lives. Through IoT, air pollution monitoring systems could be improved to address some of the current limitations, such as low geographic coverage, low precision, and high cost, of the existing monitoring systems (Mokrani et al. 2019). For personal air quality monitoring, real-time air quality can be transmitted to the user, and even the pollution level can be predicted if the route is known (Dhingra et al. 2019). Recently, researchers are reporting the utilization of the IoT also for indoor air quality (Saini et al. 2020). Smart architectures and algorithms are being developed to capture and share these data through the IoT with potentials for well-being and health applications. The

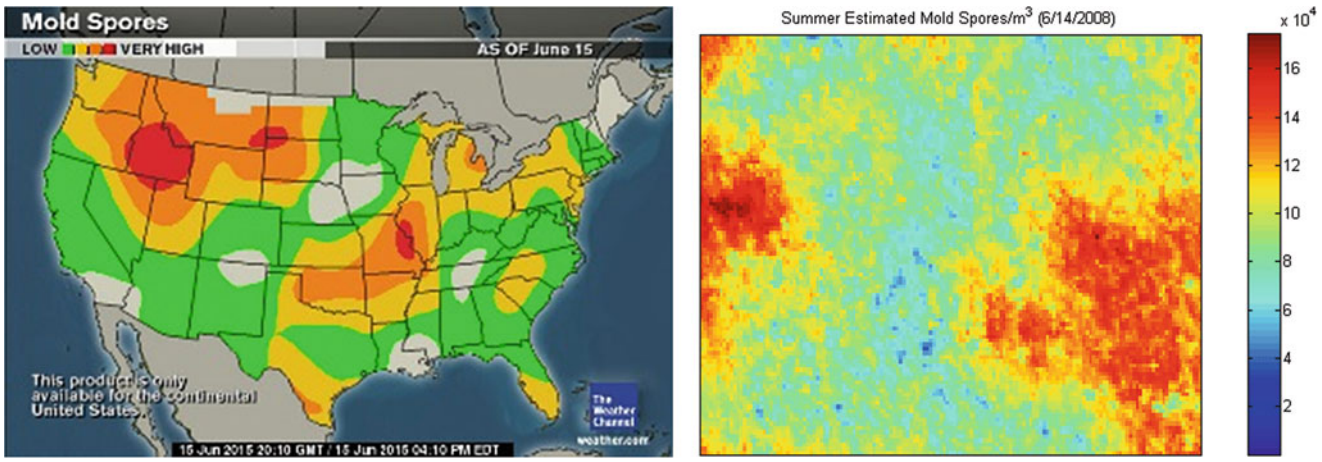


Fig. 8 Example of currently available relative mold spore frequency map (http://www.weather.com/maps/health) (left). Example of the geospatial surface model of estimated mold spore abundance at 10 km resolution for central Mississippi, USA (right). (From Faruque and Finley 2016)

IoT is providing a very useful platform for the application of mobile health technology, which is being successfully used in tracking individual mobility to combat COVID-19 (Wu et al. 2020b). The same platform is useful for individual mobility tracking for collecting environmental exposure history.

Next-Generation Air Quality Measurement Technologies

The Next-Generation Air Quality Measurement Technologies are emerging so fast that the nomenclature of its components has not been standardized yet. Most well-known components include different types of sensors, monitoring systems, wireless sensor networks, and IoT. These technologies are capable of providing personal exposure as well as capable of contributing to satellite-based estimates.

Ground-monitoring stations are considered the gold standard for air pollution data. Data from these conventional stationary monitoring facilities are also used for the validation of estimated data using other sources. However, the limitations of these conventional stations are well-known because of their sparsity over large areas (Evans et al. 2013; Faruque et al. 2014). For example, even in the USA, more than 80 percent of counties do not contain a single PM2.5 monitor (Fowlie et al. 2019). Considering the critical importance of air pollutions impacting health, satellite imagery with innovative approaches has been essential in generating estimates of air pollutants for a broader coverage (Gupta et al. 2006; Anderson et al. 2012; Evans et al. 2013; Lary et al. 2014; Van Donkelaar et al. 2015; Lasko et al. 2018; Fowlie et al. 2019; Jin 2020). However, since such estimates also depend on the utilization of ground-monitored data, estimates are less reliable where ground-monitored data are not available. Again, the sparsity of conventional ground-monitoring stations imposes a significant problem in estimating reliable data. With the increasing availability of low-cost portable sensors and the opportunity of using the wireless sensor network (WSN), the next-generation air pollution monitoring system (TNGAPMS) has emerged (Yi et al. 2015; Hagler 2016; Morawska et al. 2018; Arano et al. 2019). With such a network comes the possibility of compensating for the sparsity of conventional ground monitors and generating better estimates of air pollutants by incorporating satellite imagery and other data.

Different government agencies are becoming proactive in combining their network data with individual-level pollution data derived from other sources. US EPA has an Air Measuring and Monitoring Research initiative, which includes the development of innovative air sensor technology and analysis tools to improve the availability and accessibility of air quality measurement technology for communities and citizen scientists EPA (2020). EPA also had a funding opportunity announcement in 2010 "Developing the Next Generation of Air Quality Measurement Technology" (EPA 2010), which funded three grants. In a review of low-cost air monitoring technologies for exposure assessment, Morawskaa et al. (Morawska et al. 2018) state that current low-cost sensing technologies are able to (1) supplement routine ambient air monitoring networks and (2) expand the conversations with communities. In the area of epigenetic and exposome research, improved data and coverage will play an important role.

Patient Location

Environmental exposure information for a person can be generated only when the environmental condition data and that person's concurrent location data both are available. A person's home address, workplace address, or even approximate locality, such as zip code, is used as that person's location data to assess the environmental exposure of that person. For most epidemiological studies, such locational

data are adequate to examine the association between health outcomes and environmental exposures. However, in order to incorporate environmental exposure data into clinical practices, much more precise personal location data are required. A wide range of technologies are evolving to collect personal location data with the required precision; some are only for location and some can collect environmental exposure data as well (Phillips et al. 2001; Fang and Lu 2012; De Nazelle et al. 2013; Su et al. 2015; Chatzidiakou et al. 2019). The space-time cube has become a popular approach to show a person's movement at high resolution (Kraak 2003; Adams et al. 2009; Wagner Filho et al. 2019; Bach et al. 2014), which can be adopted to examine personal exposure measures (Kwan 2009; Fang and Lu 2011; Lu and Fang 2015; Jing et al. 2017; Ma et al. 2020).

Tracking Personal Movements

As individual-level spatiotemporal mobility can now be easily tracked, such as by GPS-enabled phone devices, the estimation of personal exposure is also possible when pollution data around that person are known.

Based on the reality that, whether we always know it or not, our movements can be tracked and stored, there are initiatives to make this information useful for healthcare. When environmental pollution data can be generated at a useful resolution, tracking an individual's movement can generate cumulative environmental exposure information over space and time. There are vendors developing applications useful at clinicians' offices to upload a patient's location history to examine that person's crossing paths for tracking transmission risk factors (Faruque and Finley 2016). The same application can generate an environmental exposure profile when the environmental factors along the way of that person are known (Fig. 9).

Crisis often speeds up technological developments. During the COVID-19 pandemic, applications of GPS-enabled mobile technology for the use of contact tracing have sprouted. The number of published articles demonstrating the applications and addressing different opportunities, technological challenges, and privacy issues continues to grow (Akarturk 2020; Buchanan et al. 2020; Dar et al. 2020; Ekong et al. 2020; Frith and Saker 2020; Garg et al. 2020; Gupta et al. 2020; Kretzschmar et al. 2020; Liang 2020; Mbunge 2020; Pan 2020; Prabu et al. 2020; Wu et al. 2020b; Ye et al. 2020). This event may promote serious thoughts about tracking individual patient location for exposure assessment, which is a critical component to assess Geospatial Individual Environmental Exposure (GIEE), potentially useful in disease investigation during regular patient visits to doctors.

Privacy and Confidentiality

While the data gathering on individual mobility is becoming common, whether for commercial purposes, law and order, or health, the issues related to the privacy of individuals and confidentiality of data remain to be the critical concerns. Some legal frameworks are already in place that could guide the development of effective tools for protecting individual privacy. Notable legal frameworks in this respect include the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the USA, the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada (Department of Justice Canada, 2000) (Act, P. 2000), and the EU General Data Protection Regulation (GDPR) in the European Union (EU) (Voigt and Von dem Bussche 2017).

It is encouraging that significant technological development as well as the number of articles discussing different aspects of mobility tracking, from accuracy to ethical issues,

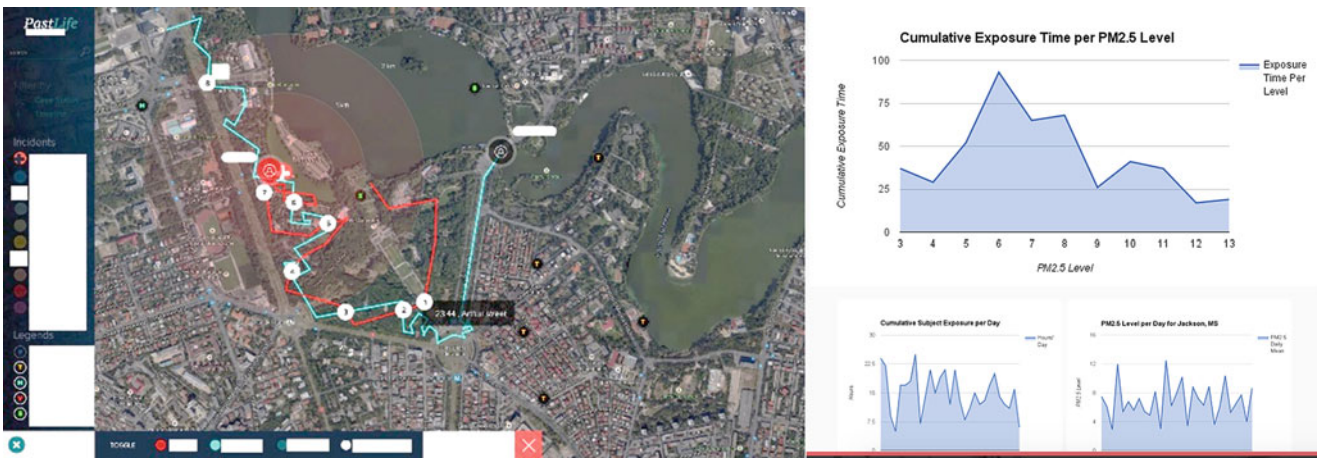


Fig. 9 Left: Representation of a patient's crossing paths. When a patient with a communicable disease visits a clinician, the patient's location history can be uploaded to see when and where the patient shares a common location history with another individual with the same disease. This information can help the clinician in deciding the necessary tests and treatments plan. Right: Environmental exposure profile along the pathway of the patient. (From Faruque and Finley 2016)

has continued to grow (Sohraby et al. 2007; Lee et al. 2008; Cao et al. 2009; Abouchar et al. 2015; Birenboim and Shoval 2016; Kargl et al. 2019; Apte et al. 2019; Kim et al. 2020). Specific interest in health applications is noticeable (Goldenholz et al. 2018; Çakmak and Eroğlu 2019; Fraccaro et al. 2019; Breslin et al. 2019; Ulrich et al. 2020). While addressing the COVID-19 pandemic crisis, breakthroughs in many areas are occurring, often without fully considering the legal aspects (De Carli et al. 2020; McLachlan et al. 2020a; b; Fenton et al. 2020; Ayres et al. 2020; McLachlan et al. 2020a; b; Klar and Lanzerath 2020). Initiatives are absolutely necessary, whether at national, regional, or even global levels, to ensure the protection against the possibility of abusing individual privacy while making scientific use of these data specifically for individual well-being and health.

Conclusions

There are certain areas in environmental health that are explored very little by the geospatial health community. This chapter discusses the areas where geospatial technology is not yet fully implemented, namely, the area of exposome-based environmental health research and the area of clinical health practice. This chapter also discusses some of the common terminologies that may be unfamiliar to the newcomers in this field. Paradigms are shifting in health research and in health practice. The geospatial health community needs to engage more intensely in both of these areas.

Health, health and well-being, social determinants of health, and population health – none of these reside in isolated domains. From the viewpoint of the geospatial approach, location is common to all. However, the application of geospatial technology to explain these phenomena is often overlooked. It is the responsibility of the geospatial community to establish the practical needs of geospatial technology for quantitative or qualitative analyses of well-being and health-related data. The new generation of geospatial experts, who will be engaged in well-being and health studies, is expected to focus on the new arena of generating environmental exposure information by taking advantage of improved data, computational power, and analytical tools.

Approximately 70% to 90% of chronic diseases are attributed to environmental exposures, much more than genetic risks. To better understand the phenomena of environmental exposures, particularly, in the context of genetic factors, the current trend of research is encouraging the exposome approach for health outcome studies. As exposome is a resultant of total exposure plus endogenous processes, it cannot represent the history of exposures. On the contrary, geospatial technology can capture the exposure history when the location of a person and the surrounding environment are known. Geospatial technology is playing a critical role in generating both types of data – a person's location and the surrounding environment of that person. In the era of the exposome, geospatial experts should be knowledgeable in gene-environment interactions. In reality, the huge number of genes and environmental variables and their interactions make experimental health research very challenging. By generating useful environmental exposure information with different dimensions and at different scales, geospatial technology can make a significant contribution to the advancement in this area.

It is true that environmental and occupational exposure information rarely enters into the clinician's history taking or diagnosis process (Marshall et al. 2002). Medical education has been blamed for this lack of enthusiasm in utilizing environmental exposure history in disease diagnosis (ATSTDR 2015). Education in environmental medicine is largely omitted in the continuum of US medical education, leaving current practitioners and future physicians without expertise in environmental medicine (Gehle et al. 2011). Collecting environmental exposure history from the patient may be incomplete as, in many instances, the patient may not be aware of the surrounding harmful environment, particularly if the exposure is not dramatic or at a low level. Without knowing the patient's environmental exposure history, physicians are limited in providing or facilitating environmental preventive or curative patient care.

Mapping the local-level environmental conditions in conjunction with the spatiotemporal positional history of a person can generate individual-level environmental exposure data. Geospatial technology is instrumental for mapping environmental pollutants as well as connecting individual's location with their surrounding environment. When individual level environmental exposure data is provided to the clinicians in a readily usable format that can contribute to the better diagnosis and prevention for a significant portion of all global diseases impacted by environmental factors.

Figure 10 summarizes the concept presented in this chapter. As shown in Fig. 10, environmental exposure information over time and space at a finer resolution can provide useful environmental exposure history both for population studies and for clinical practices. In terms of gathering data, satellites can cover larger areas, but inherently, the resolution in most cases is not adequate. Standard ground-monitoring stations, such as the EPA air pollution monitoring stations, are very sparse, and while generating national scale data, the resolution becomes poor. On the contrary, by utilizing next-generation measurement technologies and personal monitoring devices, better resolution data can be generated, but of course, the coverage will be very small. A combination of these methods can provide environmental exposure data for a wide range of applications for both population- and

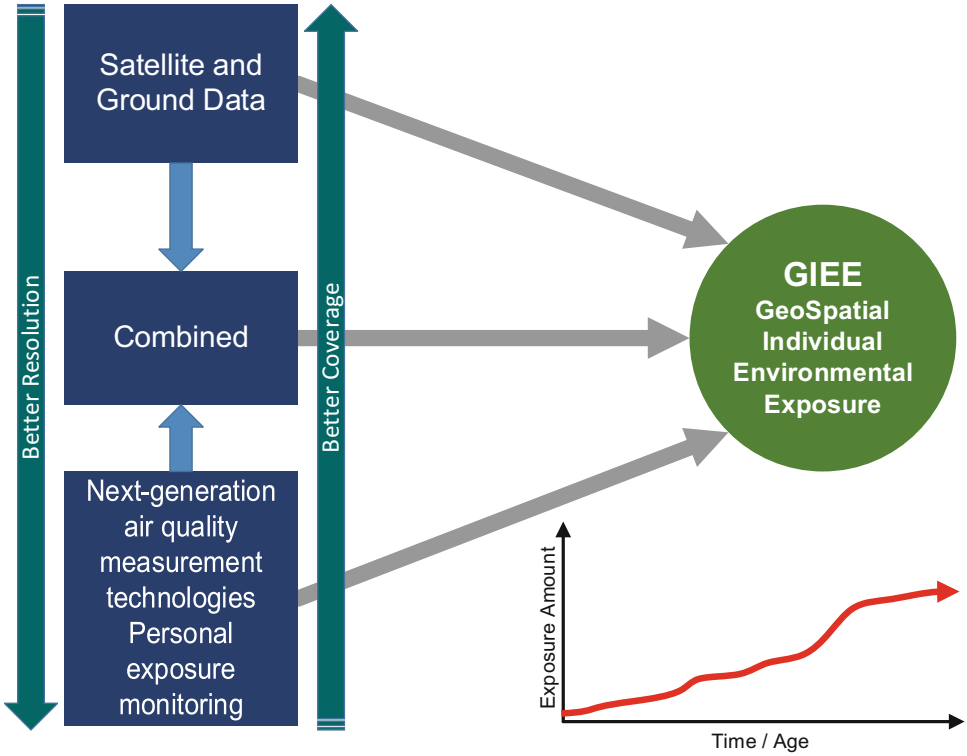


Fig. 10 Geospatial technology generating individual environmental exposure history for (**a**) exposome research and (**b**) clinical practice

individual-level well-being and health. The chart on the right in Fig. 10 illustrates the exposure measurement over time, which can be utilized as cumulative lifecourse exposure or, for certain times, particularly for the key stages of human development. A person's exposure record can be stored in a secured national database, which can be accessed only with the permission of that individual.

The European Union funded a project, EXPOsOMICS, to develop a novel approach to the assessment of exposure to high-priority environmental pollutants by characterizing the external and the internal components of the exposome (Vineis et al. 2017). It is interesting that some of the features of this project very much aligns with the concept of GIEE presented in this chapter. The project EXPOsOMICS ultimately addresses two things: (1) exposure assessment at the personal and population levels and (2) multiple "omic" technologies for the analysis of biological samples (internal markers of external exposures). To detect pollutants, provide accurate and instant estimates of changes in human exposures and estimate physical activity; this study has integrated personal exposure monitoring (PEM) with satellite-based exposure assessment and has included GPS-based techniques, smartphones, and accelerometers. This project is a perfect example of formally implementing the concept of GIEE.

Preventive measures and therapeutic regimes require an explicit understanding of the links between external exposures and health outcomes. Simultaneous analysis of external exposures, biological responses, and genetic susceptibility can help revealing such complex links. Advancements in technologies have made it possible to reveal the complex relationships among multidisciplinary and multiscale variables of health. Geospatial technology is one of many that have experienced immense advancements in recent years. It is the responsibility of the geospatial health community to be proactive in making this technology useful to its full potential for our well-being and health. The new era of health science research and medical practice is going to need multidisciplinary collaboration using very diverse sets of data at diverse scales.

References

Abouchar, T.S., M.J. Biernat, A.C. Guinn, D.E. Gura, M. Lakshmanaperumal, and R.B. Robbins. 2015. *Uses of location tracking in mobile devices. U.S. Patent No. 8,944,916*. Washington, DC: U.S. Patent and Trademark Office.

Acker, J.G. 2021. Using the NASA Giovanni system to assess and evaluate re-motely-sensed and model data variables relevant to public health issues. In *Geospatial Technology for Human Well-being and Health*, ed. F.S. Faruque. Springer.

Act, P. 2000. Personal information protection and electronic documents Act. Department of Justice, Canada. Retrieved 24 December 2020, from http://laws.justice.gc.ca/en/P-8.6/text.html.

Adams, C., P. Riggs, and J. Volckens. 2009. Development of a method for personal, spatiotemporal exposure assessment. *Journal of Environmental Monitoring* 11 (7): 1331–1339.

Akarturk, B. 2020. The role and challenges of using digital tools for COVID-19 contact tracing. *The European Journal of Social & Behavioural Sciences* 29 (3): 3241–3248.

Allen, M., J. Allen, S. Hogarth, and M. Marmot. 2013. *Working for health equity: The role of health professionals*. London: UCL Insti-

tute of Health Equity.
Alman, B.L., J.A. Stingone, M. Yazdy, L.D. Botto, T.A. Desrosiers, S. Pruitt, et al. 2019. Associations between PM2.5 and risk of preterm birth among liveborn infants. *Annals of Epidemiology* 39: 46–53.
Andersen, Z.J., M. Stafoggia, G. Weinmayr, M. Pedersen, C. Galassi, J.T. Jørgensen, et al. 2017. Long-term exposure to ambient air pollution and incidence of postmenopausal breast cancer in 15 European cohorts within the ESCAPE project. *Environmental Health Perspectives* 125 (10): 107005.
Anderson, H.R., B.K. Butland, A. van Donkelaar, M. Brauer, D.P. Strachan, T. Clayton, et al. 2012. Satellite-based estimates of ambient air pollution and global variations in childhood asthma prevalence. *Environmental Health Perspectives* 120 (9): 1333–1339.
Anenberg, S.C., L.W. Horowitz, D.Q. Tong, and J.J. West. 2010. An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environmental Health Perspectives* 118 (9): 1189–1195.
Apte, J.S., J.D. Marshall, A.J. Cohen, and M. Brauer. 2015. Addressing global mortality from ambient PM2. 5. *Environmental Science & Technology* 49 (13): 8057–8066.
Apte, A., V. Ingole, P. Lele, A. Marsh, T. Bhattacharjee, S. Hirve, et al. 2019. Ethical considerations in the use of GPS-based movement tracking in health research–lessons from a care-seeking study in rural West India. *Journal of Global Health* 9 (1).
Arano, K.A.G., S. Sun, and J. Ordieres-Mere. 2019. The use of the internet of things for estimating personal pollution exposure. *International Journal of Environmental Research and Public Health* 16 (17): 3130.
Arvor, D., N. Stelling, M. Van der Merwe, S. Richter, A. Richter, G. Neumann, et al. 2011, April. Identification of earth observation data for health-environment studies. In *In 34th international symposium on remote sensing of environment*. Sydney: Australia.
ATSDR. 2015. Taking an exposure history. Retrieved 24 Dec 2020, from https://www.atsdr.cdc.gov/csem/exphistory/docs/exposure_history.pdf
Ayres, I., A. Romano, & C. Sotis. 2020. How to make COVID-19 contact tracing apps work: insights from behavioral economics. Available at SSRN 3689805.
Baccarelli, A., and S. Ghosh. 2012. Environmental exposures, epigenetics and cardiovascular disease. *Current Opinion in Clinical Nutrition and Metabolic Care* 15 (4): 323.
Bach, B., P. Dragicevic, D. Archambault, C. Hurter, & S. Carpendale. 2014, June. A review of temporal data visualizations based on space-time cube operations. Eurographics Conference on Visualization, Jun 2014, Swansea, Wales, United Kingdom. hal-01006140.
Bai, X., I. Nath, A. Capon, N. Hasan, and D. Jaron. 2012. Health and wellbeing in the changing urban environment: Complex challenges, scientific responses, and the way forward. *Current Opinion in Environmental Sustainability* 4 (4): 465–472.
Barouki, R., K. Audouze, X. Coumoul, F. Demenais, and D. Gauguier. 2018. Integration of the human exposome with the human genome to advance medicine. *Biochimie* 152: 155–158.
Becchetti, L., G. Conzo, P. Conzo, & F. Salustri. 2020. Understanding the heterogeneity of adverse COVID-19 outcomes: The role of poor quality of air and lockdown decisions. Available at SSRN 3572548.
Bircher, J. 2005. Towards a dynamic definition of health and disease. *Medicine, Health Care and Philosophy* 8 (3): 335–341.
Birenboim, A., and N. Shoval. 2016. Mobility research in the age of the smartphone. *Annals of the American Association of Geographers* 106 (2): 283–291.
BMJ. 2008. Editorials, How should health be defined? *BMJ*: 337. https://doi.org/10.1136/bmj.a2900. (Published 10 December 2008). Cite this as: BMJ 2008;337:a2900/.
———. 2011. How should we define health?- Responses. *BMJ* 343: d4163. https://www.bmj.com/content/343/bmj.d4163/rapid-responses.
Bollati, V., and A. Baccarelli. 2010. Environmental epigenetics. *Heredity* 105 (1): 105–112.
Borro, M., P. Di Girolamo, G. Gentile, O. De Luca, R. Preissner, A. Marcolongo, et al. 2020. Evidence-based considerations exploring relations between SARS-CoV-2 pandemic and air pollution: involvement of PM2. 5-mediated up-regulation of the viral receptor ACE-2. *International Journal of Environmental Research and Public Health* 17 (15): 5573.
Bousquet, J., P.G. Burney, T. Zuberbier, P.V. Cauwenberge, C.A. Akdis, C. Bindslev-Jensen, et al. 2009. GA2LEN (Global Allergy and Asthma European Network) addresses the allergy and asthma 'epidemic'. *Allergy* 64 (7): 969–977.
Brand, G., Collins, J., Bedi, G., Bonnamy, J., Barbour, L., Ilangakoon, C., . . . Nayna Schwerdtle, P. 2020. 'I Teach It Because It Is the Biggest Threat to Health': Integrating a Planetary Health Perspective into Health Professions Education. Available at SSRN 3566173.
Breslin, S., M. Shareck, and D. Fuller. 2019. Research ethics for mobile sensing device use by vulnerable populations. *Social Science & Medicine* 232: 50–57.
Brook, R.D., S. Rajagopalan, C.A. Pope III, J.R. Brook, A. Bhatnagar, A.V. Diez-Roux, et al. 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* 121 (21): 2331–2378.
Buchanan, W.J., M.A. Imran, M. Ur-Rehman, L. Zhang, Q.H. Abbasi, C. Chrysoulas, et al. 2020. Review and Critical Analysis of Privacy-preserving Infection Tracking and Contact Tracing. *arXiv preprint arXiv* 2009: 05126.
Budge, A.M., Grobicki, A.M., Rosenberg, M., Selinus, O., Steinnes, E., and Enow, A. 2009. Mapping GeoUnions to the ICSU Framework for Sustainable Health and Wellbeing: Focus on sub-Saharan African Cities. Joint Science Project Team for Health (JSPT-H). Contractor Report for ICSU Committee on Scientific Planning and Review.
Burney, P.G.J., R.B. Newson, M.S. Burrows, and D.M. Wheeler. 2008. The effects of allergens in outdoor air on both atopic and nonatopic subjects with airway disease. *Allergy* 63 (5): 542–546.
Butler, C.D. 2018. Planetary epidemiology: Towards first principles. *Current Environmental Health Reports* 5 (4): 418–429.
Butler, C., R. Chambers, K. Chopra, P. Dasgupta, A. K. Duraiappah, P. Kumar, . . . W.-Y. Niu 2003. Ecosystems and human well-being. Ecosystems and human well-being a framework for assessment, 71–84.
Çakmak, T., & Ş. Eroğlu. 2019. User privacy in mobile health applications. HEALTHINFO 2019 : The Fourth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing.
Callahan, D. 1973. The WHO definition of 'health'. Hastings Center Studies, 77–87.
Canali, S. 2020. Making evidential claims in epidemiology: Three strategies for the study of the exposome. Studies in history and philosophy of science Part C: Studies in history and philosophy of biological and biomedical sciences. 101248.
Cao, H., V. Leung, C. Chow, and H. Chan. 2009. Enabling technologies for wireless body area networks: A survey and outlook. *IEEE Communications Magazine* 47 (12): 84–93.
CDC. 2000. Gene-Environment Interaction Fact Sheet. Retrieved 24 December 2020, from https://advancedmedicine.ca/wp-content/uploads/2013/09/The-Gene-Environment-Interaction-Centre-for-Disease-Control.pdf
———. 2020a. Health-Related Quality of Life (HRQOL). Retrieved 24 December 2020, from https://www.cdc.gov/hrqol/wellbeing.htm
———. 2020b. Social determinants of health: know what affects health. Retrieved 24 December 2020, from https://www.cdc.gov/socialdeterminants/index.htm
Chai, Y., and M.P. Kwan. 2015. Suburbanization, daily lifestyle and space-behavior interaction in Beijing. *Dili Xuebao/Acta Geographica Sinica* 70 (8): 1271–1280.

Chakrabarty, R.K., P. Beeler, P. Liu, S. Goswami, R.D. Harvey, S. Pervez, et al. 2020. Ambient PM2. 5 exposure and rapid spread of COVID-19 in the United State. *Science of the Total Environment* 760: 143391.

Charreire, H., C. Weber, B. Chaix, P. Salze, R. Casey, A. Banos, et al. 2012. Identifying built environmental patterns using cluster analysis and GIS: Relationships with walking, cycling and body mass index in French adults. *International Journal of Behavioral Nutrition and Physical Activity* 9 (1): 59.

Chatzidiakou, L., A. Krause, O.A. Popoola, A. Di Antonio, M. Kellaway, Y. Han, et al. 2019. Characterising low-cost sensors in highly portable platforms to quantify personal exposure in diverse environments. *Atmospheric Measurement Techniques* 12 (8): 4643.

Cheli. 2020. Personal communication, November 2020. Simonetta Cheli, Head of Strategy, Programme & Coordination Office, Directorate of Earth Observation Programmes. ESA - European Space Agency Headquarters.

Chowdhury, S., S. Dey, L. Di Girolamo, K.R. Smith, A. Pillarisetti, and A. Lyapustin. 2019. Tracking ambient PM2. 5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset. *Atmospheric Environment* 204: 142–150.

ClimateAction. 2020. European Space Agency on how their technology can be used to combat climate change. Retrieved 24 December 2020, from http://www.climateaction.org/climate-leader-interviews/european-space-agency-on-how-their-technology-can-be-used-to-combat-climate

Cohen, A.J., M. Brauer, R. Burnett, H.R. Anderson, J. Frostad, K. Estep, et al. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet* 389 (10082): 1907–1918.

Coleman, A., S. Dhesi, and S. Peckham. 2016. Health and wellbeing boards: The new system stewards. *Dismantling the NHS*: 279–300.

Dales, R.E., S. Cakmak, R.T. Burnett, S.T.A.N. Judek, F. Coates, and J.R. Brook. 2000. Influence of ambient fungal spores on emergency visits for asthma to a regional children's hospital. *American Journal of Respiratory and Critical Care Medicine* 162 (6): 2087–2090.

Dar, A.B., A.H. Lone, S. Zahoor, A.A. Khan, and R. Naaz. 2020. Applicability of mobile contact tracing in fighting pandemic (covid-19): Issues, challenges and solutions. *Computer Science Review* 38: 100307.

De Carli, A., M. Franco, A. Gassmann, C. Killer, B. Rodrigues, E. Scheid, et al. 2020. WeTrace–A Privacy-preserving Mobile COVID-19 Tracing Approach and Application. *arXiv preprint arXiv* 2004: 08812.

De Nazelle, A., E. Seto, D. Donaire-Gonzalez, M. Mendez, J. Matamala, M.J. Nieuwenhuijsen, and M. Jerrett. 2013. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution* 176: 92–99.

DeBord, D.G., T. Carreón, T.J. Lentz, P.J. Middendorf, M.D. Hoover, and P.A. Schulte. 2016. Use of the "exposome" in the practice of epidemiology: A primer on-omic technologies. *American Journal of Epidemiology* 184 (4): 302–314.

Dennis, K.K., E. Marder, D.M. Balshaw, Y. Cui, M.A. Lynes, G.J. Patti, et al. 2017. Biomonitoring in the era of the exposome. *Environmental Health Perspectives* 125 (4): 502–510.

Dhingra, S., R.B. Madda, A.H. Gandomi, R. Patan, and M. Daneshmand. 2019. Internet of things mobile–air pollution monitoring system (IoT-Mobair). *IEEE Internet of Things Journal* 6 (3): 5577–5584.

Dominici, F., R.D. Peng, M.L. Bell, L. Pham, A. McDermott, S.L. Zeger, and J.M. Samet. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* 295 (10): 1127–1134.

Eddins, E.A.R. 1998. Nursing, health, and the environment: Strengthening the relationship to improve the Public's health. *Nursing and Health Care Perspectives* 19 (1): 43–45.

Ekong, I., E. Chukwu, and M. Chukwu. 2020. COVID-19 Mobile positioning data contact tracing and patient privacy regulations: Exploratory search of global response strategies and the use of digital tools in Nigeria. *JMIR mHealth and uHealth* 8 (4): e19139.

EO4HEALTH. 2020. Earth Observations for Health (EO4HEALTH). Retrieved 24 December 2020, from http://www.geohealthcop.org/eo4health

EPA 2010. Developing the Next Generation of Air Quality Measurement Technology, initial announcement of this funding opportunity. Retrieved 24 December 2020, from https://cfpub.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.rfatext/rfa_id/540

———. 2020. Air Measuring and Monitoring Research. Retrieved 24 December 2020, from https://www.epa.gov/air-research/air-measuring-and-monitoring-research

Evans, J., A. van Donkelaar, R.V. Martin, R. Burnett, D.G. Rainham, N.J. Birkett, and D. Krewski. 2013. Estimates of global mortality attributable to particulate air pollution using satellite imagery. *Environmental Research* 120: 33–42.

Fang, T.B., and Y. Lu. 2011. Constructing a near real-time space-time cube to depict urban ambient air pollution scenario. *Transactions in GIS* 15 (5): 635–649.

———. 2012. Personal real-time air pollution exposure assessment methods promoted by information technological advances. *Annals of GIS* 18 (4): 279–288.

Faruque, F.S. 2019. Geospatial technology in environmental health applications. *Environmental Monitoring and Assessment* 191 (2): 333.

Faruque, F.S., and R.W. Finley. 2016. Geographic Medical History: Advances in Geospatial Technology Present New Potentials in Medical Practice. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* XLI-B8: 191–195.

Faruque, F.S., H. Li, W.B. Williams, L.A. Waller, B.T. Brackin, L. Zhang, et al. 2014. GeoMedStat: an integrated spatial surveillance system to track air pollution and associated healthcare events. *Geospatial Health* 8: S631–S646.

N. Fenton, S. McLachlan, P. Lucas, K. Dube, G. Hitman, M. Osman, ... and M. Neil. 2020. A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing. medRxiv.

Fowlie, M., E. Rubin, & R. Walker. 2019. Bringing satellite-based air quality estimates down to earth. Paper presented at the AEA Papers and Proceedings.

Fraccaro, P., A. Beukenhorst, M. Sperrin, S. Harper, J. Palmier-Claus, S. Lewis, et al. 2019. Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: A systematic review. *Journal of the American Medical Informatics Association* 26 (11): 1412–1420.

Frith, J., and M. Saker. 2020. It is all about location: Smartphones and tracking the spread of COVID-19. *Social Media+ Society* 6 (3): 2056305120948257.

Fu, P., X. Guo, F.M.H. Cheung, and K.K.L. Yung. 2019. The association between PM2.5 exposure and neurological disorders: A systematic review and meta-analysis. *Science of the Total Environment* 655: 1240–1248.

Garg, Lalit, E. Chukwu, N. Nasser, C. Chakraborty, and G. Garg. 2020. Anonymity preserving IoT-based COVID-19 and other infectious disease contact tracing model. *IEEE Access*.

Gehle, K.S., J.L. Crawford, and M.T. Hatcher. 2011. Integrating environmental health into medical education. *American Journal of Preventive Medicine* 41 (4): S296–S301.

GEO. 2020. Group on Earth Observations. Retrieved 24 December 2020, from http://www.earthobservations.org/geo_community.php

Geohealthcop. 2020. GEO Health Community of Practice. Retrieved 24 December 2020, from http://www.geohealthcop.org/

GEOSS. 2020. Global Earth Observation System of Systems (GEOSS). Retrieved 24 December 2020, from https://earthobservations.org/geoss.php

Goldenholz, D.M., S.R. Goldenholz, K.B. Krishnamurthy, J. Halamka, B. Karp, M. Tyburski, et al. 2018. Using mobile location data in biomedical research while preserving privacy. *Journal of the American Medical Informatics Association* 25 (10): 1402–1406.

Goldman, R.H., S. Rosenwasser, and E. Armstrong. 1999. Incorporating an environmental/occupational medicine theme into the medical school curriculum. *Journal of Occupational and Environmental Medicine* 41 (1): 47–52.

Greaves, Z., and S. McCafferty. 2017. Health and wellbeing boards: Public health decision making bodies or political pawns? *Public Health* 143: 78–84.

Green, P.M. 2000. Taking environmental health education seriously. *Nursing and Health Care Perspectives* 21 (5): 234–234.

Griffin, D.W., E.N. Naumova, J.C. McEntee, D. Castronovo, J.L. Durant, M.L. Lyles, F. Faruque, and D. Lary. 2012. Chapter 4: Air quality and human health, (Chapter 4), pages 129–185. In *Environmental Tracking for Public Health Surveillance, International Society for Photogrammetry and Remote Sensing (ISPRS) Commission VIII/WG-2*, ed. S. Morain and A. Budge. Leiden: CRC Press Taylor & Francis. ISBN 9780415584715.

Gulliver, J., D. Morley, C. Dunster, A. McCrea, E. van Nunen, M.Y. Tsai, et al. 2018. Land use regression models for the oxidative potential of fine particles (PM2. 5) in five European areas. *Environmental Research* 160: 247–255.

Guloksuz, S., J. van Os, and B.P. Rutten. 2018. The exposome paradigm and the complexities of environmental research in psychiatry. *JAMA Psychiatry* 75 (10): 985–986.

Guo, T., Y. Wang, H. Zhang, Y. Zhang, J. Zhao, Q. Wang, et al. 2018. The association between ambient PM2.5 exposure and the risk of preterm birth in China: A retrospective cohort study. *Science of the Total Environment* 633: 1453–1459.

Gupta, P., S.A. Christopher, J. Wang, R. Gehrig, Y. Lee, and N. Kumar. 2006. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment* 40 (30): 5880–5892.

Gupta, R., M. Bedi, P. Goyal, S. Wadhera, and V. Verma. 2020. Analysis of COVID-19 tracking tool in India: Case Study of Aarogya Setu Mobile Application. *Digital Government: Research and Practice* 1 (4): 1–8.

Hagler, G. 2016. Next-generation air measurement technologies. EPA Office of Research and Development. https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=527435

Hamm, N.A., R.J. Soares Magalhães, and A.C. Clements. 2015. Earth observation, spatial data quality, and neglected tropical diseases. *PLoS Neglected Tropical Diseases* 9 (12): e0004164.

Hart, J. 2017. Taking an environmental exposure history. *Alternative and Complementary Therapies* 23 (2): 64–65.

Hay, S.I., M.J. Packer, and D.J. Rogers. 1997. Review article the impact of remote sensing on the study and control of invertebrate intermediate hosts and vectors for disease. *International Journal of Remote Sensing* 18 (14): 2899–2930.

Hendryx, M., and J. Luo. 2020. COVID-19 prevalence and fatality rates in association with air pollution emission concentrations and emission sources. *Environmental Pollution* 265: 115126.

Herbreteau, V., G. Salem, M. Souris, J.P. Hugot, and J.P. Gonzalez. 2007. Thirty years of use and improvement of remote sensing, applied to epidemiology: From early promises to lasting frustration. *Health & Place* 13 (2): 400–403.

Hou, L., X. Zhang, D. Wang, and A. Baccarelli. 2012. Environmental chemical exposures and human epigenetics. *International Journal of Epidemiology* 41 (1): 79–105.

HP2020. 2020. Social determinants of health, Healthy People 2020. Retrieved 24 December 2020, from https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

HP2030. 2020. Social determinants of health, Healthy People 2030. Retrieved 24 December 2020, from https://health.gov/healthypeople/objectives-and-data/social-determinants-health

Hu, H., Z. Hu, K. Zhong, J. Xu, F. Zhang, Y. Zhao, and P. Wu. 2019. Satellite-based high-resolution mapping of ground-level PM2. 5 concentrations over East China using a spatiotemporal regression kriging model. *Science of the Total Environment* 672: 479–490.

Huang, C.C., B.Y. Chen, S.C. Pan, Y.L. Ho, and Y.L. Guo. 2019. Prenatal exposure to PM2.5 and congenital heart diseases in Taiwan. *Science of the Total Environment* 655: 880–886.

Huber, M., J.A. Knottnerus, L. Green, H. van der Horst, A.R. Jadad, D. Kromhout, et al. 2011. How should we define health? *BMJ* 343: d4163.

ICSU. 2011. *Report of the ICSU planning group on health and wellbeing in the changing urban environment: A systems analysis approach*. Paris: International Council for Science.

Ihde, E., B. Kligler, G.P. Zipp, & C. Rocchetti. 2020. The impact of integrating environmental health into medical school curricula: A survey-based study.

ISC. 2020. The International Science Council (ISC). Retrieved 24 December 2020, from https://council.science/about-us/

Ismail-Zadeh, A. 2016. Geoscience international: The role of scientific unions. *History of Geo-and Space Sciences* 7 (2): 103–123.

Ismail-Zadeh and Joselyn. 2019. Editors, special issue-The International Union of Geodesy and Geophysics: From different spheres to a common globe, History of Geo- and Space Sciences.

Jerrett, M., M.C. Turner, B.S. Beckerman, C.A. Pope III, A. Van Donkelaar, R.V. Martin, et al. 2017. Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates. *Environmental Health Perspectives* 125 (4): 552–559.

Jia, P. 2019. Spatial lifecourse epidemiology. *The Lancet Planetary Health* 3 (2): e57–e59.

Jia, P., W. Dong, S. Yang, Z. Zhan, L. Tu, and S. Lai. 2020a. Spatial lifecourse epidemiology and infectious disease research. *Trends in Parasitology* 36 (3): 235–238.

Jia, P., C. Yu, J.V. Remais, A. Stein, Y. Liu, R.C. Brownson, et al. 2020b. Spatial lifecourse epidemiology reporting standards (ISLE-ReSt) statement. *Health & Place* 61: 102243.

Jin, X. 2020. *Observing the distributions and chemistry of major air pollutants (O 3 and PM 2.5) from space: Trends, uncertainties, and health implications*. Columbia University: Doctoral dissertation.

Jing, M., C. Yanwei, and F. Tingting. 2017. Progress of research on the health impact of people's space-time behavior and environmental pollution exposure. *Progress in Geography* 36 (10): 1260–1269.

Joselyn, J.A., A. Ismail-Zadeh, T. Beer, H. Gupta, M. Kono, U. Shamir, et al. 2019. IUGG in the 21st century. *History of Geo-and Space Sciences* 10 (1): 73–95.

Juarez, P.D., and P. Matthews-Juarez. 2018. Applying an exposome-wide (ExWAS) approach to cancer research. *Frontiers in Oncology* 8: 313.

Kargl, F., R.W. van der Heijden, B. Erb, and C. Bösch. 2019. Privacy in mobile sensing. In *Digital Phenotyping and Mobile sensing*, 3–12. Cham: Springer.

Kilpatrick, N., H. Frumkin, J. Trowbridge, C. Escoffery, R. Geller, L. Rubin, et al. 2002. The environmental history in pediatric practice: A study of pediatricians' attitudes, beliefs, and practices. *Environmental Health Perspectives* 110 (8): 823–827.

Kim, J., M.P. Kwan, M.C. Levenstein, and D.B. Richardson. 2020. How do people perceive the disclosure risk of maps? Examining the perceived disclosure risk of maps and its implications for geoprivacy protection. *Cartography and Geographic Information Science*: 1–19.

Kioumourtzoglou, M.A., J.D. Schwartz, M.G. Weisskopf, S.J. Melly, Y. Wang, F. Dominici, and A. Zanobetti. 2016. Long-term PM2.5 exposure and neurological hospital admissions in the northeastern United States. *Environmental Health Perspectives* 124 (1): 23–29.

Klar, R., and D. Lanzerath. 2020. The ethics of COVID-19 tracking apps–challenges and voluntariness. *Research Ethics* 16 (3–4): 1–9.

Kraak, M. J. (2003, August). The space-time cube revisited from a geovisualization perspective. In Proc. 21st international cartographic conference (pp. 1988-1996). Citeseer.

Kretzschmar, M.E., G. Rozhnova, M.C. Bootsma, M. van Boven, J.H. van de Wijgert, and M.J. Bonten. 2020. Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study. *The Lancet Public Health* 5 (8): e452–e459.

Kwan, M.P. 2009. From place-based to people-based exposure measures. *Social Science & Medicine* 69 (9): 1311–1313.

Landrigan, P.J., R. Fuller, N.J. Acosta, O. Adeyi, R. Arnold, A.B. Baldé, et al. 2018. The lancet commission on pollution and health. *The Lancet* 391 (10119): 462–512.

Lary, D.J., F.S. Faruque, N. Malakar, A. Moore, B. Roscoe, Z.L. Adams, and Y. Eggelston. 2014. Estimating the global abundance of ground level presence of particulate matter (PM2.5). *Geospatial Health* 8: S611–S630.

Lasko, K., K.P. Vadrevu, and T.T.N. Nguyen. 2018. Analysis of air pollution over Hanoi, Vietnam using multi-satellite and MERRA reanalysis datasets. *PLoS One* 13 (5): e0196629.

Lee, H. H., I. K. Park, & K. S. Hong. (2008, September). Design and implementation of a mobile devices-based real-time location tracking. In 2008 The Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (pp. 178–183). IEEE.

Leffers, J.M., C.M. Smith, R. McDermott-Levy, L.K. Resick, M.J. Hanson, L.C. Jordan, et al. 2015. Developing curriculum recommendations for environmental health in nursing. *Nurse Educator* 40 (3): 139–143.

Liang, F. 2020. COVID-19 and health code: How digital platforms tackle the pandemic in China. *Social Media+ Society* 6 (3): 2056305120947657.

Lim, S.S., T. Vos, A.D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, et al. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380 (9859): 2224–2260.

Lioy, P.J., and S.M. Rappaport. 2011. Exposure science and the exposome: An opportunity for coherence in the environmental health sciences [Editorial]. *Environmental Health Perspectives* 119 (11): A466–A467.

Louis, G.M.B., and R. Sundaram. 2012. Exposome: Time for transformative research. *Statistics in Medicine* 31 (22).

Louis, G.M.B., M.M. Smarr, and C.J. Patel. 2017. The exposome research paradigm: An opportunity to understand the environmental basis for human health and disease. *Current Environmental Health Reports* 4 (1): 89–98.

Lozano, R., M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, et al. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380 (9859): 2095–2128.

Lu, Y., and T.B. Fang. 2015. Examining personal air pollution exposure, intake, and health danger zone using time geography and 3D geovisualization. *ISPRS International Journal of Geo-Information* 4 (1): 32–46.

Ma, J., C. Li, M.P. Kwan, L. Kou, and Y. Chai. 2020. Assessing personal noise exposure and its relationship with mental health in Beijing based on individuals' space-time behavior. *Environment International* 139: 105737.

Maantay, J.A., and S. McLafferty. 2011. Environmental health and geospatial analysis: An overview. In *Geospatial analysis of environmental health*, 3–37. Dordrecht: Springer.

Mailing, L.J., J.M. Allen, T.W. Buford, C.J. Fields, and J.A. Woods. 2019. Exercise and the gut microbiome: A review of the evidence, potential mechanisms, and implications for human health. *Exercise and Sport Sciences Reviews* 47 (2): 75–85.

Maitre, L., J. De Bont, M. Casas, O. Robinson, G.M. Aasvang, L. Agier, et al. 2018. Human early life Exposome (HELIX) study: A European population-based exposome cohort. *BMJ Open* 8 (9): e021311.

Marshall, L., E. Weir, A. Abelsohn, and M.D. Sanborn. 2002. Identifying and managing adverse environmental health effects: 1. Taking an exposure history. *CMAJ* 166 (8): 1049–1055.

Marti, R., Z. Li, T. Catry, E. Roux, M. Mangeas, P. Handschumacher, et al. 2020. A mapping review on urban landscape factors of dengue retrieved from earth observation data, GIS techniques, and survey questionnaires. *Remote Sensing* 12 (6): 932.

Martin-Sanchez, F., R. Bellazzi, V. Casella, W. Dixon, G. Lopez-Campos, and N. Peek. 2020. Progress in Characterizing the Human Exposome: a Key Step for Precision Medicine. *Yearbook of Medical Informatics* 29 (1): 115.

Mbunge, E. 2020. Integrating emerging technologies into COVID-19 contact tracing: opportunities, challenges and pitfalls. diabetes & metabolic syndrome: Clinical Research & Reviews.

McClafferty, H., A. Brooks, S. Dodds, and V. Maizes. 2015. Environmental health: Evaluating an online educational curriculum for healthcare workers. *Journal of Preventive Medicine* 1: 1–8.

McCurdy, L.E., J. Roberts, B. Rogers, R. Love, R. Etzel, J. Paulson, et al. 2004. Incorporating environmental health into pediatric medical and nursing education. *Environmental Health Perspectives* 112 (17): 1755–1760.

McGillivray, M. 2006. Human well-being: Concept and measurement. Springer.

McLachlan, S., P. Lucas, K. Dube, G. S. McLachlan, G. A. Hitman, M. Osman, and N. E. Fenton. 2020a. The fundamental limitations of COVID-19 contact tracing methods and how to resolve them with a Bayesian network approach.

McLachlan, S., P. Lucas, K. Dube, G. A. Hitman, M. Osman, E. Kyrimi, ... and N. E. Fenton. 2020b. Bluetooth smartphone apps: Are they the most private and effective solution for COVID-19 contact tracing?. arXiv preprint arXiv:2005.06621.

Mhawish, A., T. Banerjee, M. Sorek-Hamer, M. Bilal, A.I. Lyapustin, R. Chatfield, and D.M. Broday. 2020. Estimation of high-resolution PM2. 5 over the Indo-Gangetic Plain by fusion of satellite data, meteorology, and land use variables. *Environmental Science & Technology* 54 (13): 7891–7900.

Millennium Ecosystem Assessment. 2005. *Ecosystems and Human Well-Being: Wetlands and Water Synthesis*. Washington, DC: World Resources Institute.

Miller, G. 1962. "Airs, waters, and places" in history. *Journal of the History of Medicine and Allied Sciences* XVII (1): 129–140. https://doi.org/10.1093/jhmas/XVII.1.129.

Miller, G.W. 2020. *The Exposome: A new paradigm for the environment and health*. Academic Press.

Miller, G.W., and D.P. Jones. 2014. The nature of nurture: Refining the definition of the exposome. *Toxicological Sciences* 137 (1): 1–2.

Mokrani, H., R. Lounas, M. T. Bennai, D. E.Salhi, & R. Djerbi. 2019. Air quality monitoring using iot: A survey. Paper presented at the 2019 IEEE International Conference on Smart Internet of Things (SmartIoT).

Morain, S.A., and A.M. Budge. 2012. Earth observing data for health applications. In *Environmental Tracking for Public Health Surveillance*, ed. S.A. Morain and A.M. Budge. London: CRC Press, ISBN-10 X, 41558471, 3-18.

Morawska, L., P.K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, et al. 2018. Applications of low-cost sensing technologies

for air quality monitoring and exposure assessment: How far have they gone? *Environment International* 116: 286–299.
NASA. 2016. NASA selects instruments to study air pollution, Tropical Cyclones. RELEASE 16–025, March 10, 2016. Retrieved 24 December 2020, from http://www.nasa.gov/press-release/nasa-selects-instruments-to-study-air-pollution-tropical-cyclones
NASA Earth Science. 2020. NASA Probes Environment, Covid-19 Impacts, Possible Links. Retrieved 24 December 2020, from https://www.nasa.gov/feature/nasa-probes-environment-covid-19-impacts-possible-links
Niedzwiecki, M.M., and G.W. Miller. 2017. The exposome paradigm in human health: Lessons from the emory exposome summer course. *Environmental Health Perspectives* 125 (6): 064502.
Niedzwiecki, M.M., D.I. Walker, R. Vermeulen, M. Chadeau-Hyam, D.P. Jones, and G.W. Miller. 2019. The exposome: Molecules to populations. *Annual Review of Pharmacology and Toxicology* 59: 107–127.
NIEHS. 2020. Gene and environment interaction. Retrieved 24 December 2020, from https://www.niehs.nih.gov/health/topics/science/gene-env/index.cfm
Oyoshi, K., Y. Mizukami, R. Kakuda, Y. Kobayashi, H. Kai, and T. Tadono. 2019. Japan Aerospace Exploration Agency's public-health monitoring and analysis platform: A satellite-derived environmental information system supporting epidemiological study. *Geospatial Health* 14 (1).
Pan, X.B. 2020. Application of personal-oriented digital technology in preventing transmission of COVID-19, China. *Irish Journal of Medical Science* 1971: 1–2.
Parselia, E., C. Kontoes, A. Tsouni, C. Hadjichristodoulou, I. Kioutsioukis, G. Magiorkinis, and N.I. Stilianakis. 2019. Satellite Earth observation data in epidemiological modeling of malaria, dengue and West Nile virus: A scoping review. *Remote Sensing* 11 (16): 1862.
Pelletier, S. 2016. Experts see growing importance of adding environmental health content to medical school curricula. AAMC News.
Phillips, M.L., T.A. Hall, N.A. Esmen, R. Lynch, and D.L. Johnson. 2001. Use of global positioning system technology to track subject's location during environmental exposure sampling. *Journal of Exposure Science & Environmental Epidemiology* 11 (3): 207–215.
Pickle, L.W., L.A. Waller, and A.B. Lawson. 2005. Current practices in cancer spatial data analysis: A call for guidance. *International Journal of Health Geographics* 4 (1): 3.
Pope, A., Snyder, M., & Mood, A. Committee on enhancing environmental health content in nursing practice, division of health promotion and disease prevention, Institute of Medicine, .(Eds.).(1995). Nursing, health & the environment: Strengthening the relationship to improve the public's health. In: Washington, DC: National Academy Press.
Prabu, S., B. Velan, F.V. Jayasudha, P. Visu, and K. Janarthanan. 2020. Mobile technologies for contact tracing and prevention of COVID-19 positive cases: A cross-sectional study. *International Journal of Pervasive Computing and Communications*.
Prior, L., D. Manley, and C.E. Sabel. 2019. Biosocial health geography: New 'exposomic' geographies of health and place. *Progress in Human Geography* 43 (3): 531–552.
Pun, V.C., F. Kazemiparkouhi, J. Manjourides, and H.H. Suh. 2017. Long-term PM2. 5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *American Journal of Epidemiology* 186 (8): 961–969.
Quezada, H., A.L. Guzmán-Ortiz, H. Díaz-Sánchez, R. Valle-Rios, and J. Aguirre-Hernández. 2017. Omics-based biomarkers: Current status and potential use in the clinic. *Boletín Médico Del Hospital Infantil de México (English Edition)* 74 (3): 219–226.
Rall, D.P., and A.M. Pope. 1995. *Environmental medicine: Integrating a missing element into medical education*. Washington, D.C: National Academies Press.
Rappaport, S.M. 2011. Implications of the exposome for exposure science. *Journal of Exposure Science & Environmental Epidemiology* 21 (1): 5–9.
———. 2012. Biomarkers intersect with the exposome. *Biomarkers* 17 (6): 483–489.
———. 2016. Genetic factors are not the major causes of chronic diseases. *PLoS One* 11 (4): e0154387.
———. 2018. Redefining environmental exposure for disease etiology. *NPJ Systems Biology and Applications* 4 (1): 1–6.
Rappaport, S.M., D.K. Barupal, D. Wishart, P. Vineis, and A. Scalbert. 2014. The blood exposome and its role in discovering causes of disease. *Environmental Health Perspectives* 122 (8): 769–774.
Roberts, J.R., and J.R. Reigart. 2001. Environmental health education in the medical school curriculum. *Ambulatory Pediatrics* 1 (2): 108–111.
Robinson, O., and M. Vrijheid. 2015. The pregnancy exposome. *Current Environmental Health Reports* 2 (2): 204–213.
Saini, J., M. Dutta, and G. Marques. 2020. Indoor air quality monitoring systems based on internet of things: A systematic review. *International Journal of Environmental Research and Public Health* 17 (14): 4942.
Saracci, R. 1997. The World Health Organization needs to reconsider its definition of health. *BMJ* 314: 1409–1410.
Sarigiannis, D.A. 2019. The exposome paradigm in environmental health. In *Environmental exposures and human health challenges*, 1–29. Hershey: IGI Global.
Schmidt, C.W. 2005. Global Earth observations for health. *Environmental Health Perspectives* 113 (11): 738–740.
Shi, L., X. Wu, M.D. Yazdi, D. Braun, Y.A. Awad, Y. Wei, et al. 2020. Long-term effects of PM2.5 on neurological disorders in the American Medicare population: a longitudinal cohort study. *The Lancet Planetary Health* 4 (12): e557–e565.
Sillé, F.C., S. Karakitsios, A. Kleensang, K. Koehler, A. Maertens, G.W. Miller, et al. 2020. The exposome–a new approach for risk assessment. *ALTEX-Alternatives to animal experimentation* 37 (1): 3–23.
Silva, R.A., J.J. West, Y. Zhang, S.C. Anenberg, J.F. Lamarque, D.T. Shindell, et al. 2013. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environmental Research Letters* 8 (3): 034005.
Simon-Nobbe, B., U. Denk, V. Pöll, R. Rid, and M. Breitenbach. 2008. The spectrum of fungal allergy. *International Archives of Allergy and Immunology* 145 (1): 58–86.
Singh, N., T. Banerjee, V. Murari, K. Deboudt, M.F. Khan, R.S. Singh, and M.T. Latif. 2020. Insights into size-segregated particulate chemistry and sources in urban environment over central Indo-Gangetic Plain. *Chemosphere* 263: 128030.
Siroux, V., L. Agier, and R. Slama. 2016. The exposome concept: A challenge and a potential driver for environmental health research. *European Respiratory Review* 25 (140): 124–129.
Sogno, P., C. Traidl-Hoffmann, and C. Kuenzer. 2020. Earth observation data supporting non-communicable disease research: A review. *Remote Sensing* 12 (16): 2541.
Sohraby, K., D. Minoli, and T. Znati. 2007. *Wireless sensor networks: Technology, protocols, and applications*. Hoboken: John Wiley & Sons.
Space Studies Board. 2015. Continuity of NASA earth observations from space: A value framework: National Academies Press. ISBN 978–0–309-37743-0, DOI: https://doi.org/10.17226/21789
Stahler, G.J., J. Mennis, and D.A. Baron. 2013. Geospatial technology and the "exposome": New perspectives on addiction. *American Journal of Public Health* 103 (8): 1354–1356. https://doi.org/10.2105/AJPH.2013.301306.
Steckling, N., A. Gotti, S. Bose-O'Reilly, D. Chapizanis, D. Costopoulou, F. De Vocht, et al. 2018. Biomarkers of exposure in environment-wide association studies–opportunities to decode

the exposome using human biomonitoring data. *Environmental Research* 164: 597–624.

Stingone, J.A., G.M. Buck Louis, S.F. Nakayama, R.C. Vermeulen, R.K. Kwok, Y. Cui, et al. 2017. Toward greater implementation of the exposome research paradigm within environmental epidemiology. *Annual Review of Public Health* 38: 315–327.

Su, J.G., M. Jerrett, Y.Y. Meng, M. Pickett, and B. Ritz. 2015. Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. *Science of the Total Environment* 506: 518–526.

Tamura, K., R.C. Puett, J.E. Hart, H.A. Starnes, F. Laden, and P.J. Troped. 2014. Spatial clustering of physical activity and obesity in relation to built environment factors among older women in three US states. *BMC Public Health* 14 (1): 1–16.

Tarantini, L., M. Bonzini, P. Apostoli, V. Pegoraro, V. Bollati, B. Marinelli, et al. 2009. Effects of particulate matter on genomic DNA methylation content and iNOS promoter methylation. *Environmental Health Perspectives* 117 (2): 217–222.

Tung, E.L., K.A. Cagney, M.E. Peek, and M.H. Chin. 2017. Spatial context and health inequity: Reconfiguring race, place, and poverty. *Journal of Urban Health* 94 (6): 757–763.

Ulrich, C. M., G. Demiris, R. Kennedy, and E. Rothwell. 2020. The ethics of sensor technology use in clinical research. Nursing Outlook

Van Donkelaar, A., R.V. Martin, M. Brauer, and B.L. Boys. 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental Health Perspectives* 123 (2): 135–143.

Vermeulen, R., E.L. Schymanski, A.L. Barabási, and G.W. Miller. 2020. The exposome and health: Where chemistry meets biology. *Science* 367 (6476): 392–396.

Viana, J., J.V. Santos, R.M. Neiva, J. Souza, L. Duarte, A.C. Teodoro, and A. Freitas. 2017. Remote sensing in human health: A 10-year bibliometric analysis. *Remote Sensing* 9 (12): 1225.

Vineis, P. 2019. What is the Exposome and how it can help research on air pollution. *Emission Control Science and Technology* 5 (1): 31–36.

Vineis, P., M. Chadeau-Hyam, H. Gmuender, J. Gulliver, Z. Herceg, J. Kleinjans, et al. 2017. The exposome in practice: Design of the EXPOsOMICS project. *International Journal of Hygiene and Environmental Health* 220 (2): 142–151.

Vineis, P., O. Robinson, M. Chadeau-Hyam, A. Dehghan, I. Mudway, and S. Dagnino. 2020. What is new in the exposome? *Environment International* 143: 105887.

Vinikoor-Imler, L.C., J.A. Davis, R.E. Meyer, and T.J. Luben. 2013. Early prenatal exposure to air pollution and its associations with birth defects in a state-wide birth cohort from North Carolina. *Birth Defects Research Part A: Clinical and Molecular Teratology* 97 (10): 696–701.

Voigt, P., and A. Von dem Bussche. 2017. *The EU general data protection regulation (gdpr). A Practical Guide*. 1st ed. Cham: Springer International Publishing.

Vrijheid, M. 2014. The exposome: A new paradigm to study the impact of environment on health. *Thorax* 69 (9): 876–878.

Wagner Filho, J.A., W. Stuerzlinger, and L. Nedel. 2019. Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration. *IEEE Transactions on Visualization and Computer Graphics* 26 (1): 514–524.

Walpole, S.C., A. Vyas, J. Maxwell, B.J. Canny, R. Woollard, C. Wellbery, et al. 2017. Building an environmentally accountable medical curriculum through international collaboration. *Medical Teacher* 39 (10): 1040–1050.

Ward, M.H., J.R. Nuckols, S.J. Weigel, S.K. Maxwell, K.P. Cantor, and R.S. Miller. 2000. Identifying populations potentially exposed to agricultural pesticides using remote sensing and a geographic information system. *Environmental Health Perspectives* 108 (1): 5–12.

WHO. 2005. *Ecosystems and human well-being: health synthesis: A Report of the Millennium Ecosystem Assessment*. Geneva: WHO Press.

———. 2020a. *Basic documents: forty-ninth edition (including amendments adopted up to 31 May 2019)*. Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO. ISBN 978–92–4-000052-0. https://apps.who.int/gb/bd/.

———. 2020b. Social determinants of health. Retrieved 24 December 2020, from https://www.who.int/gender-equity-rights/understanding/sdh-definition/en/

Wigbels, L. 2011. Using Earth observation data to improve health in the United States: Accomplishments and future challenges. Center for Strategic & International Studies.

Wild, C.P. 2005. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention* 14 (8): 1847–1850.

———. 2012. The exposome: From concept to utility. *International Journal of Epidemiology* 41 (1): 24–32.

Wolffe, A.P., and D. Guschin. 2000. Chromatin structural features and targets that regulate transcription. *Journal of Structural Biology* 129 (2–3): 102–122.

Wu, X., R.C. Nethery, M.B. Sabath, D. Braun, and F. Dominici. 2020a. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances* 6 (45): eabd4049.

Wu, J., X. Xie, L. Yang, X. Xu, Y. Cai, T. Wang, and X. Xie. 2020b. Mobile health technology combats COVID-19 in China. *Journal of Infection*.

Ye, Q., J. Zhou, and H. Wu. 2020. Using information technology to manage the COVID-19 pandemic: Development of a technical framework based on practical experience in China. *JMIR Medical Informatics* 8 (6): e19515.

Yi, W.Y., K.M. Lo, T. Mak, K.S. Leung, Y. Leung, and M.L. Meng. 2015. A survey of wireless sensor network based air pollution monitoring systems. *Sensors* 15 (12): 31392–31427.

Yoo, E., C. Rudra, M. Glasgow, and L. Mu. 2015. Geospatial estimation of individual exposure to air pollutants: Moving from static monitoring to activity-based dynamic exposure assessment. *Annals of the Association of American Geographers* 105 (5): 915–926.

Zheng, T., M.H. Bergin, S. Hu, J. Miller, and D.E. Carlson. 2020. Estimating ground-level PM2. 5 using micro-satellite images by a convolutional neural network and random forest approach. *Atmospheric Environment* 230: 117451.

Building the Analytic Toolbox: From Spatial Analytics to Spatial Statistical Inference with Geospatial Data

Lance A. Waller

Introduction

Beginning with early maps of yellow fever in New York City in the late 1700s and Dr. John Snow's famous maps of cholera in London in 1854, maps have played an important role in public health for more than 200 years (Waller 2017). The early twenty-first century has seen a transition to data-intensive science where health studies make use of multiple data sets from heterogeneous sources to gain insight into associations and relations with a goal of moving toward understanding underlying disease processes and causal relationships with putative risk and protective factors. These foundational developments in data availability and analytic approaches transition from the past setting where analytic methods were defined in order to gain as much information as possible from expensive (high cost, limited content) data sets, to the emergence of Data Science approaches seeking to learn from expansive and easily accessible (very large, potentially high content) data sets, often arising from multiple sources. This conceptual shift occurs (and is occurring) in all branches of science, including those intersecting with geographic information systems, spatial epidemiology, and spatial statistics, resulting in unique and profound influences on current and future directions of development, application, and interpretation of geospatial analysis. For georeferenced data, these general shifts toward data-intensive science impact and expand the intersection of three interrelated areas of science: Geographic Information Science (Goodchild 2010), Statistical Science, and the emerging discipline of Data Science. While each area has its own history and highlights, they each also provide complementary as well as intersecting insights into the future of analysis of georeferenced spatial and spatiotemporal data sets, particularly so in health-related fields. In the sections below, we provide a geographic perspective on Data Science, a brief history of the intersections of Geographic Information Science and Statistical Science, and an outline of methods for spatial analysis in health noting transitions from each of the three domains into their intersection and how these transitions define new approaches within the analytic toolbox for geospatial analysis and health. We also consider two sets of methods and applications that illustrate evolution of thought, methodological development, and application across all three areas of science.

From a geospatial analysis perspective, it is clear that the so-called data revolution referenced above is occurring at the intersection of Geographic Information Science, Statistical Science, and Data Science. Specifically, geospatially aware data science requires *spatial thinking* (National Research Council 2006) wherein location and geography provide essential insight into patterns and processes; *statistical thinking* wherein probabilistic models of uncertainty provide inferential frameworks for estimation and prediction (Chance 2002); and *spatial statistical thinking* (Waller 2014) wherein statistical results are not only constructed via geographic relationships but also evaluated and interpreted in a geographic context as well. This mutually beneficial intersection of the Geographic Information, Statistical, and Data Sciences and associated types of thinking is necessary to link concepts, tools, assumptions, problems, and solutions spanning the geographical, statistical, and data worlds to further expand and harmonize developments often occurring in one discipline into an integrated set of concepts, tools, and knowledge spanning all three.

In many ways, the Geographic Information Science community predates the rise of Data Science, not only in the coining of the terms but also in its appreciation and use of georeferenced data sets from multiple sources, creatively

L. A. Waller (✉)
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA
e-mail: lwaller@emory.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_2

linked to provide novel insight unavailable from any single data component. The general data management, linkage, and query tools available in geographic information systems and the layered data storage of Google Earth and other global scale data systems (Goodchild et al. 2012) provide a framework for working with big data in general and big spatial data in particular. More recently, the use of distributed data and cloud implementations extend popular frameworks to the geographic setting. All told, we find modern geospatial analyses benefiting from Data Science developments and contributing to specific spatial and geographic dimensions to the future of Data Science.

The sections below consider three key elements of the geospatial analytic toolbox, namely: (1) geographic information system data management, (2) geospatial analytics within spatial analysis, and (3) spatial and spatiotemporal statistics, particularly those applied to epidemiologic applications. Fig. 1 illustrates several examples of how these three elements build on and reinforce each other to provide an essential and expanding set of tools to interact with georeferenced data, to summarize and display spatial and spatiotemporal patterns and relationships, and to estimate, predict, and infer associations and observations within an interconnected geographic space. The arrows in Fig. 1 illustrate a sequence of analytic topics moving from discipline-specific topics toward integrated concepts and tools spanning two and three disciplines in order to move toward a general geoanalytic perspective.

Spatial Data Tools in GIS: Disparate Data Linked by Location

A central tenet of geospatial analysis is that location matters. Location links different types of measurements taken near to one another, and location predicts new observations of measured variables taken nearby in space or time. Geographic information systems (GIS) use location as a central reference point for measured and observed attribute values. Location provides a key for data matching, linking, and layering, and location provides a searchable reference for defining attributions from one data set that fall within a given distance and/or direction of observations in another. Since their inception, GIS have dealt with *uncomfortably large* data sets (data sets pushing current storage and/or processing limits), a good, rule-of-thumb working definition of "big data" (i.e., more data than you know what to do with).

Historical uncomfortably large geographic data include satellite imaging data (Goodchild 2016), small area data from the US Census, and myriad now-familiar GIS layers (rivers and streams, road networks, building-specific maps). While these represent now-familiar data sets to GIS users, all geospatial analysts have had the experience of slow rendering times, system crashes, and common but confusing incompatibilities associated with large georeferenced data from different sources. While such traditional (and popular) data sets may seem small by today's standards, the GIS and GIScience communities have a history of pushing the envelope on wanting more data, wanting more detailed data, and working creatively on the edge of what current computing will allow.

Modern challenges at the interface of GIScience and Data Science include distributed georeferenced data across multiple platforms, divide-and-conquer approaches using distributed cloud computing (Goldberg et al. 2014), machine/deep learning for georeferenced data, and analysis of location-based services. Each of these raises technical and algorithmic challenges but also can generate new ethical issues relating to privacy (how I feel about my data) and confidentiality (protections I am required to provide for data in my possession). To draw from the basic questions of journalism, geospatial analysis often builds on a premise that *where* and *when* you are can provide insight on *what*, *how*, and *why* you experience/observe/measure. Taken together, the increasing availability and use of location-based services relating to *where* and *when* you are also can provide quite accurate assessments of *who* you are, especially when combining information across multiple data sets (Rocher et al. 2019).

In addition to the technical, algorithmic, and ethical challenges, GIS also generates challenges to the application of traditional statistical methods. While the by-now-familiar notion of spatial correlation motivates and permeates spatial statistical analyses, GIS also provides additional challenges by linking data from multiple sources each exhibiting different levels of accuracy and uncertainty. Tracking multiple sources and magnitudes of uncertainty across each data layer can be complicated and may not fit neatly into traditional statistical techniques, motivating the development of novel analytic methods in the chapters of this volume.

Spatial Analytics: Defining Where to Take Action

In Fig. 1, at the intersection of Data Science and Statistical Science, we find the rise of "analytics," i.e., general purpose methods and sometimes quite sophisticated data summaries (and summaries of data summaries) that scale up familiar calculations to application within and between massive data sets. While there is no single definition of "analytics" versus, say, "statistics," generally the term refers to clearly defined statistical and analytic tools that can be computationally scaled up to apply to very large data sets and provide actionable insight from results (Cooper 2012). That is, the term "analytics" tends to focus on providing tools for data-

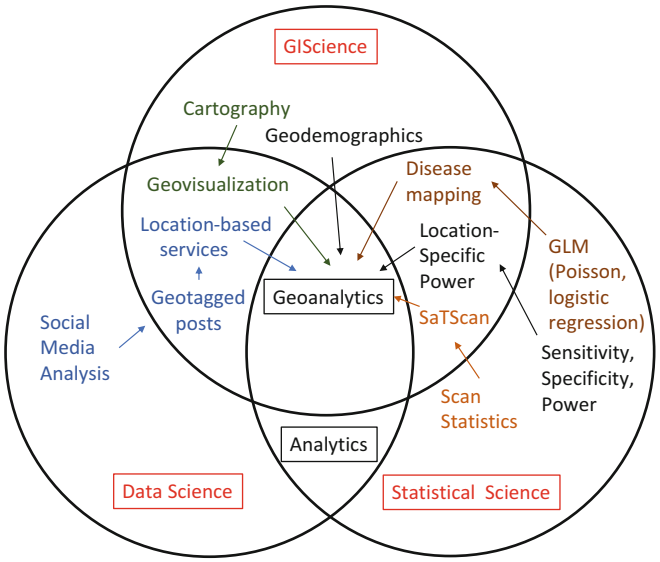


Fig. 1 Illustration of system approach of GIScience, Data Science, and Statistical Science and their components to achieve data-driven goals. Arrows indicate related areas of research moving from discipline-specific topics toward general geoanalytic concepts and tools but do not necessarily represent a sequence of approaches that need to be conducted in order for any single analysis

driven decision-making versus data-informed understanding of underlying processes. This distinction is subtle: analytics involve statistical calculations but tend to focus more on decision outcomes rather than on the properties of the statistical estimates themselves or on the properties of the underlying epidemiologic and/or biologic processes associated with the outcome of interest.

Analytics provide insight into patterns and variation in observations with a particular goal of, say, influencing future observations (e.g., reducing disease burden in an area or placing police patrols during a festival weekend). Analytics often involve tools such as leave-one-out cross-validation, bootstrapping, and more sophisticated divide-and-conquer approaches wherein calculations on data subsamples or subsets provide descriptive (and actionable) insight into distributions within and between data sets without relying on classical statistical parametric families for more advanced analysis.

Bridging the framework of analytics between Statistical Science, Data Science, and GIScience expands the definition to include cartographic aspects of data visualization. To illustrate, in Fig. 1, we begin with cartography within the GIScience framework, often building on Bertin's visual variables to best display and distinguish local quality, direction, differences, and magnitude (cf., classic references such as MacEachren 1995, Monmonier 2018, Slocum et al. 2004). To date, the literature relating to data visualization (e.g., Chen et al. 2008, Kerren et al. 2008) and that relating to cartographic visualization (e.g., Andrienko et al. 2011) remains relatively separate. However, as illustrated in Fig. 1 by the intersection of GIScience and Data Science, novel collaborations in these areas can and will provide fertile ground for expansion in the continued development of geovisualization tools drawing from both GIScience and Data Science (Andrienko et al. 2011).

In the setting of human well-being and health, actionable questions of interest include (but are not limited to) the detection of clusters or clustering of disease (Thun and Sinks 2004, Waller 2015), the detection of local concentrations of risk factors (e.g., environmental pollution or concentrations of social determinants of disease such as poverty or illegal drug use), the siting and staffing of health clinics, and the location and evaluation of health information campaigns. As noted, the distinction between an analytics-based focus on actionable outcomes (e.g., identifying locations that have the highest concentrations of disease and/or pollution) may differ from overall interest in estimating associations between exposures and disease incidence and/or prevalence. In some cases, we seek assessment of whether the concentrations of disease are statistically unusual (since some location will have the highest rate, but is it too high?), and in others we may simply wish to know where the highest concentrations of patients are regardless of the statistical significance (e.g., for determining clinic locations). While epidemiologic studies seeking to understand causes and drivers of local rates are important, they are not the only geospatial analyses of interest in the assessment of local human health and well-being.

In addition to geovisualization tools mapping local rates of disease, local values of pollution, and local summaries of risk factors, other specific tools often used as analytics for spatial data include global (e.g., Moran's I and Geary's c statistics) and local measures of spatial association ((i.e., LISAs), cf.

Lloyd (2010)). Such measures identify the overall level of similarity between neighboring values (for global statistics) and local hot/cold spots of association where particular regions are very similar/dissimilar from their neighbors. While measures of statistical significance are often associated with measures of association, their primary purpose is often to assess if there is spatial correlation in the observations and where this local correlation might be highest in magnitude.

Another area of research interest involves the analysis of social media posts, a very active area of Data Science research. As noted in Fig. 1, the addition of geotags (locations) to social media posts allows linkage to location-based services within GIScience, another pathway of development for present and future geoanalytics. Challenges include the relatively low (but growing) fraction of social media data with linked location information. (All data-centric analytics require solid support of both location *and* health data in order to fully realize their full potential!)

Adapting Analytic Tools to the Geospatial Setting for Public Health Analysis

We next turn to the evolution of analysis tools from Geographic Information Science and Statistical Science toward automated, actionable use as geoanalytics. This pathway is often slow and multidisciplinary, involving a series of developments rather than a single landmark publication or proposal. To illustrate this process of development, we review two specific areas of analytic tool development drawing from both Geographic Information Science and Statistical Science.

As noted above, many (if not most) geographic public health applications maintain an epidemiologic perspective, seeking to better understand causes and drivers of observed incidence and prevalence of disease. In this setting, analysts seek to detect deviations from a setting where the risk of disease is the same for individuals everywhere (i.e., a hypothesis of no clusters/clustering) or, more generally, where risk is higher than expected based on known or suspected local risk factors. Identification of geographic patterns or outliers can be used in an analytics setting (i.e., act here versus there) or in more of a statistical/epidemiologic manner (i.e., why are rates high here?).

To see the influence of Geographic Information, Statistical, and Data Science more clearly, we outline contributions to the development of methodological thinking around the detection of disease clusters.

Example 1: Detecting Clusters of Disease

An unexpected "cluster" or "hot spot" of disease cases is an evocative image in public health, often framed as beginning with Dr. Snow's investigation of cholera deaths in London in 1854. The image captures the imagination of scientists, policymakers, and the general public and generates a strong desire for discovery of hidden drivers of risk based on the geographic pattern observed in cases.

In 1990, the US Centers for Disease Control and Prevention hosted a workshop bringing together public health officials, epidemiologists, statisticians, and others to discuss how best to seek out clusters and how best to respond to reports of clusters by concerned groups. Beginning around the same time, several analytic methods were proposed drawing on advances in geographic data processing, advances in statistical methodology, and advances in data availability and access. The initial guidelines for analysis focused on traditional epidemiologic summaries such as standardized mortality ratios and standardized incidence ratios to describe observed local excess cases and risk. The next decade witnessed a rapid expansion in proposed analytic methods, but application and interpretation typically required customized development and programming by analysts embedded in research groups, advocacy groups, or health agencies. From 2000 to 2010, textbooks (e.g., Waller and Gotway 2004, Lawson 2006) provided collective descriptions and open-source software with spatial analytic libraries provided broad access to novel analytic methods. The most recent decade has seen further expansion of computing power, open-source tools, freely distributed software, and rapid access to vast quantities of georeferenced data. Recent revisions to guidelines for understanding disease clusters now anticipate broadly sophisticated analyses from all quarters, and responsible responses to reports from analysts, advocates, and the public now require familiarity with tools that have moved rapidly from their origins in Geographic Information Science, Data Science, or Statistical Science toward implementation as geoanalytic tools.

To see this point more clearly, we note that, immediately preceding the three-decade time period outlined above, Geographic Information Science, building on digitized maps of disease incidence and prevalence, explored automated detection approaches, most notably the Geographical Analysis Machine (GAM) of Openshaw et al. (1987). While the GAM predates the coining of term "Geographic Information Science" by a few years, and the term "Data Science" by approximately two decades, it is very much in the spirit of coupling geographic concepts and spatial relationships with computational power to scale up simple tasks to address complex, spatial problems. The approach considered a large number of potential clusters (locally defined collections of observed cases) and assigned a statistical significance value to each potential cluster, plotting the boundaries of those which exceeded a user-specified threshold. Due to the very large number of overlapping potential clusters, each with its own p-value, formal statistical inference presented a challenge. However, the graphical output identified areas on the map where greater than expected rates of cases were ob-

served. Investigators from Statistical Science provided some early formalization of the GAM structure by limiting potential clusters to collections of either a fixed number of cases (Besag and Newell 1991) or a fixed number of individuals at risk (Turnbull et al. 1990). Such approaches provided more interpretable evaluations of statistical significance for putative clusters but were not as comprehensive or automatic as the original GAM. Further research led to the now-popular approach of the space-time scan statistic (SaTScan, Kulldorff et al. 2005, Kulldorff 2009) which reframed the question to avoid providing significance levels for every potential cluster and instead provide focused and accurate statistical significance relating to the most likely cluster. The approach maintains the large-scale search aspect of the GAM but provides sound inference for the potential cluster of greatest concern. (A thorough and growing bibliography of analyses using SaTScan across many different disciplines appears at www.satscan.org.)

While the GAM-to-SaTScan path illustrates a historical example of moving from one of the three fields through others and toward the center node of geoanalytics in Fig. 1, the example also illustrates that this path typically involves the work of multiple individuals from multiple fields and multiple perspectives to fully navigate the transition. In addition, it is important to note that such explorations rarely end in the only possible approach to a problem. For example, in addition to scan statistics, many other investigators have developed statistically based analytic methods for the detection of spatial or spatiotemporal clusters. Tango (2010) provides a catalog of many such methods, and Waller and Gotway (2004, Chaps. 6 and 7) provide discussion of interpretation of such hypothesis tests. With one path to geoanalytics in place, many others often quickly follow providing analysts with a broad collection of tools.

In addition to the historical development of cluster detection tools, Fig. 1 also illustrates the development pathway of small area estimation and disease mapping models, beginning in Statistical Science with generalized linear models of small area rates and counts based on independent observations (McCullagh and Nelder 1989) to the incorporation of spatial correlation (GIScience) through the inclusion of random effects (Clayton and Kaldor 1987, Besag et al. 1991). The statistical properties of such approaches are well understood (Banerjee et al. 2014), and recent advances in computing (Blangiardo and Cameletti 2015) offer potential for data science-based distributed computing to allow application to very large-scale data sets. The basic framework is widely used by spatial analysts, and many extensions to the basic model have been proposed and developed. One area of ongoing research involves adjustments to allow associations between an outcome variable and particular covariates to vary across space, i.e., the strength of association between a risk factor and a health effect may be stronger in some areas than others, perhaps due to unobserved confounders. For example, if one were exploring the association between illegal drug activity (measured by local arrest counts) and the rate of violent crime, one might expect a stronger association at the border of two rival distributors (say, due to competition) than one might expect within areas largely covered by a single distributor. A brief history of these developments provides a second illustrative example of the move from one of the three Sciences toward the definition of geoanalytic tools.

Example 2: Spatial Variation in Associations

For almost two decades, two different approaches have been proposed for estimating spatial variation in outcome-covariate associations, one originating in Geographic Information Science, the other from Statistical Science, and both benefiting from developments in Data Science.

Tobler's First Law of Geography, paraphrased as: all things are related but things closer together are more related, is central to Geographic Information Science, as are measures of spatial association. Such measures (e.g., Moran's I, Geary's c) often draw on a matrix of spatial "weights" associated with every pair of observations giving higher weights given to closer pairs of observation locations. Fotheringham et al. (2002) linked the Geographic Information Science idea of weighting nearby observations to the Statistical Science idea of using weights to increasing influence of certain observations to provide local statistical estimation of associations between outcomes and covariates within a regression setting. While in Statistical Science local regressions provide smooth curves based on data with similar values of covariates, Fotheringham et al. (2002) proposed estimating smooth relationships based on data from nearby locations. The shift in perspective from covariate space to geographic space provides smoothly varying surfaces describing the estimated association between a covariate and outcome. The results are visually appealing and descriptive of the varying associations. With available software, "geographically weighted regression" (GWR) quickly became a popular analytic tool with many applications in many different areas of application. As with the GAM, some statistical challenges remained, namely, calculation of local estimates of the variability of the spatially varying estimates remains difficult since this variance is entangled with the weights and variance of nearby observations in a complicated manner. That is, it is difficult to see if the spatial variations induced by the method are significantly different from a model with a single value of the association everywhere.

From the Statistical Science perspective, other researchers have proposed extensions to disease mapping models to allow spatially correlated random slopes in a mixed effects framework. While such "spatially varying coefficient" (SVC) models are cleaner statistically, the approach is not exactly

the same as GWR, and direct comparisons between the two approaches remain a challenge (Waller et al. 2007). Output from SVC models provides model-based estimates of local rates that are smoother than values based on local data alone by "borrowing information" from neighboring observations. Such neighbors are often defined by a spatial weight associated to pairs (as in GWR); however, GWR and SVC use the weights quite differently. SVC weights define spatial correlation between observations, while GWR weights define the strength of influence of each observation on association estimates across the study area. Typically, GWR estimates are smoother (largely by definition), and SVC estimates retain some residual statistical noise yielding less smooth maps of the spatially varying associations.

With respect to Fig. 1, GWR begins in Geographic Information Science and uses ideas from Statistical Science without providing a full statistical assessment of estimation and uncertainty, while SVC begins in Statistical Science via Bayesian hierarchical models and then uses ideas from Geographic Information Science, but its results are less clear geographically. With respect to Data Science, current implementations of GWR are much faster to compute and closer to automation than are Markov chain Monte Carlo implementations of SVC. (Markov chain Monte Carlo algorithms estimate model parameters through (often lengthy) simulations of potential values based on the observed data and a probability model relating each parameter with other model parameters.) Both sets of approaches continue to move toward providing geoanalytic capability, i.e., actionable insight, but both still require care in implementation and interpretation and likely require more refinements before they can be viewed as robust, automatic, general purpose tools within the geoanalytic toolbox.

Pulling It All Together

As illustrated in Fig. 1 and the discussion above, the three fields of GIScience, Data Science, and Statistical Science all offer unique but complementary contributions to the future development, application, and interpretation of geoanalytic methods in studies of health and well-being. We stress that no single field serves as the sole source of development, nor does any single field serve as the final arbiter of successful development of geoanalytic strategies. All solutions contain elements of computation, geography, and statistics/epidemiology, and the best solutions will borrow from all three areas. In addition to the development of the methods, we also note that the evaluation of their accuracy, precision, and overall performance should also be viewed through the composite lens of the intersecting fields.

For example, Waller et al. (2006) and Waller (2014) note that the statistically familiar concept of power, the probability of detecting a feature (e.g., a cluster of disease) when that feature is really present, has a geographic as well as a statistical dimension. That is, the probability of detecting a cluster of disease in a given location depends critically on the size of the population at risk in that area. This intersection of Statistical Science and GIScience offers novel geographic insight into current discussions of false-positive rates in Data Science-based detection algorithms, but such cross-fertilization is still developing and will likely yield much promise for further development.

Finally, while our discussion above primarily focuses on the spatial aspect of geoanalytics, incorporating time will allow the expansion of geoanalytics for spatiotemporal analyses. Such research enables a dynamic assessment of spatial patterns allowing analysts to explore the emergence of outbreaks, the effectiveness of intervention policies, the impact of season on spatial patterns of disease and health, and many other aspects that vary by location *and* time (Cressie and Wikle 2011).

Conclusions

In summary, Fig. 1 and the examples above illustrate the valuable contributions offered by the viewpoints of GIScience, Data Science, and Statistical Science in the development, application, interpretation, and assessment of geoanalytics, especially for their application to studies of health and well-being. Such hybrid thinking identifies the connection of tools and concepts across all three settings in order to provide accurate, reliable, and actionable conclusions as well as to extend established tools from each area into a more robust analytic toolbox for spatial analyses in public health and biomedicine.

Future directions include further expansion of ideas from each of the three areas into more integrated tools and training that draw from the strengths of the others. Such work should focus attention on the development of geoanalytic tools incorporating the best ideas in visualization, geography, statistics, epidemiology, and data science. This is necessarily interdisciplinary work and will benefit greatly from expanded team science collaborations across the disciplines with a central focus on creating better tools for the broader application of spatial and spatiotemporal concepts and analytics across the biomedical and public health sciences.

Acknowledgments This research is supported in part by grant R01AI125842 from the National Institute of Allergy and Infectious Diseases and grant R01HD092580 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The thoughts and opinions expressed above reflect those of the author and should not be construed to represent those of NIAID or NICHD.

References

Andrienko, G., N. Andrienko, D. Keim, A.M. MacEachren, and S. Wrobel. 2011. Challenging problems of geospatial visual analytics. *Journal of Visual Languages & Computing* 22: 251–256.

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.

Besag, J., and J. Newell. 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A* 154: 143–144.

Besag, J., J. York, and A. Mollié. 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43: 1–20.

Blangiardo, M., and M. Cameletti. 2015. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Chichester: Wiley.

Chance, B. L. 2002. Components of statistical thinking and implications for instruction and assessment. *Journal of Statistical Education* 10. http://www.amstat.org/publications/jse/v10n3/chance.html.

Chen, C.-H., W. Härdle, and A. Unwin, eds. 2008. *Handbook of Data Visualization*. New York: Springer.

Clayton, D., and J. Kaldor. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43: 671–681.

Cooper, A. 2012. *CETIS Analytics Series Volume 1, Number 5: What is Analytics? Definition and Essential Characteristics.* Centre for Educational Technology and Interoperability Standards Series ISSN 2051-9214.

Cressie, N., and C. Wikle. 2011. *Statistics for spatio-temporal data*. Hoboken, NJ: Wiley.

Fotheringham, A.S., C. Brunsdon, and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley and Sons.

Goldberg, D., M. Olivares, Z. Li, and A.G. Klein. 2014. Maps & GIS libraries in the era of Big Data and cloud computing. *Journal of Map & Geography Libraries* 10: 100–122.

Goodchild, M.F. 2016. GIS in the era of big data. *Cybergeo: European Journal of Geography* (online). http://journals.openedition.org/cybergeo/27647.

———. 2010. Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science* 1: 3–20.

Goodchild, M.F., H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, M. Ehlers, J. van Genderen, D. Jackson, A.J. Lewis, M. Pesaresi, G. Remety-Fülöpp, R. Simpson, A. Skidmore, C. Wang, and P. Woodgate. 2012. Next-generation Digital Earth. *Proceedings of the National Academy of Science USA* 109: 11088–11094.

Kerren, A., J.T. Stasko, J.-D. Fekete, and C. North, eds. 2008. *Information Visualization: Human-centered Issues and Perspectives*. Berlin: Springer.

Kulldorff, M and Information Management Services, Inc. 2009. SaTScan™ v8.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/.

Kulldorff, M., R. Heffernan, J. Hartman, R.M. Assunção, and F. Mostashari. 2005. A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine* 2: 216–224.

Lawson, A.B. 2006. *Statistical Methods for Spatial Epidemiology*. 2nd ed. CRC Press: Boca Raton, FL.

Lloyd, C.D. 2010. *Local Models for Spatial Analysis*. 2nd ed. Boca Raton, FL: CRC Press.

McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC: Boca Raton, FL.

MacEachren, A.M. 1995. *How Maps Work: Representation, Visualization, and Design*. New York: The Guilford Press.

Monmonier, M. 2018. *How to Lie with Maps*. 3rd ed. University of Chicago Press: Chicago.

National Research Council, Committee on the Support for Thinking Spatially: The Incorporation of Geographic Information Science Across the K-12 Curriculum, Committee on Geography. 2006. *Learning to Think Spatially*. Washington DC: National Academies Press.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft. 1987. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1 (4): 335–358. https://doi.org/10.1080/02693798708927821.

Rocher, L., J.M. Hendrickx, and Y.-A. de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10: 3069. https://doi.org/10.1038/s41467-019-10933-3.

Slocum, T.A., R.B. McMaster, F.C. Kessler, and H.H. Howard. 2004. *Thematic Cartography and Geographic Visualization*. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall.

Tango, T. 2010. *Statistical Methods for Disease Clustering*. New York: Springer.

Thun, M.J., and T. Sinks. 2004. Understanding Cancer Clusters. *CA: A Cancer Journal for Clinicians.* 54: 273–280.

Turnbull, B.W., E.J. Iwano, W.S. Burnett, H.L. Howe, and L.C. Clark. 1990. Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 132 (supplement): S136–S143.

Waller, L.A. 2014. Putting spatial statistics (back) on the map. *Spatial Statistics* 9: 4–19.

———. 2015. Discussion: Statistical cluster detection, epidemiologic interpretation, and public health policy. *Statistics and Public Policy* 2 (1): 1–8. https://doi.org/10.1080/2330443X.2015.1026621.

———. 2017. Mapping in Public Health. In *Mapping Across Academia*, ed. S.D. Brunn and M. Dodge. Dordrecht: Springer.

Waller, L.A., and C.A. Gotway. 2004. *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley and Sons.

Waller, L.A., E.G. Hill, and R.A. Rudd. 2006. The geography of power: Statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine* 25: 853–865.

Waller, L.A., L. Zhu, C.A. Gotway, D.M. Gorman, and P.J. Gruenewald. 2007. Quantifying geographic variations in associations between alcohol distribution and violence: A comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment* 21: 573–588.

Geostatistical Methods for Modeling Environmental Exposures with Applications to Ambient Air Pollution

Howard H. Chang

Introduction

Advancing our understanding of how environmental exposures affect human health continues to be an important area of public health research. Environmental exposures include toxic contaminants from anthropogenic sources (e.g., traffic-related air pollution, agricultural pesticides) and risk factors such as extreme temperature and ecological changes. Pollutants, especially those released into the atmosphere and the water system, can affect large populations. One particular challenge with environmental health studies is that they are predominantly observational. This is because exposures to environmental pollutants are often involuntary, and studies are often conducted in response to emerging concerns (e.g., hydraulic fracturing) or disasters (e.g., hurricanes and oil spills). It is infeasible or unethical for researchers to assign individuals to different levels of environmental exposures. Hence, to protect public health, environmental regulatory standards and policies have relied on findings from large population-based epidemiologic studies.

Being able to accurately link environmental exposures to health data in space and time is a crucial, but difficult, task in any environmental health study. It is generally not possible to measure exposures continuously for all participants, especially in studies of long-term health effects. Moreover, many environmental health studies are conducted retrospectively where exposure levels are not collected with the health data. For example, health data can be obtained from large databases developed for administrative purposes (e.g., vital certificates, hospital billing records) or disease surveillance systems and registries (e.g., for cancers, birth defects, and neurological diseases). In these study designs, environmental exposures need to be estimated from different data sources.

Past environmental epidemiology studies have routinely utilized measurements from monitoring networks that are set up by government agencies to perform exposure assessment. However, reliance on monitoring networks can lead to several well-recognized analytic challenges. First, monitoring networks are spatially sparse and temporally incomplete due to maintenance costs. This is a particular concern for pollution fields that exhibit high spatial variability. Second, monitors in networks designed for regulatory purposes are often preferentially located in areas with large populations and high pollution levels. Hence, when linking health data to monitoring measurements, the complex spatial-temporal missing data pattern will not only restrict the study population, but it can also result in exposure measurement error that impacts the accuracy of health studies (Levy et al. 2019).

The increasing availability of geo-referenced health data is accompanied by increasing interest in estimating environmental exposures at fine spatial scales with complete spatial-temporal coverage to support health studies. In order to improve the availability and resolution of environmental pollution data, one approach is to supplement monitoring measurements with additional data sources that can reflect pollution levels. In this chapter, these data sources are referred to as *proxy data*. For example, land use variables (e.g., elevation, roadway density, distance to pollution sources) and meteorological conditions may be highly predictive of pollution levels. Statistical models can be used to exploit observed relationships between the pollutant of interest and these predictors in space and in time.

Recently, satellite imagery and numerical model simulations are two proxy data sources that have received particular attention. Satellite imagery has been used to measure environmental processes such as temperature, wildfire, and ambient air pollution. Advantages of remotely sensed data

H. H. Chang (✉)
Department of Biostatistics and Bioinformations, Emory University, Atlanta, GA, USA
e-mail: howard.chang@emory.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_3

include their fine spatial resolutions, public and near real-time availability, and excellent geographical coverage. In contrast, numerical models aim to simulate a pollutant's creation and dispersion using information on pollutant sources and state-of-the-art knowledge on chemical and physical processes. Moreover, numerical models can provide three-dimensional deterministic outputs that have complete spatial-temporal coverage for the study domain. Numerical models have been utilized extensively for weather forecast and climate research. Advances in geographical information system, remote-sensing technology, and numerical model simulation have contributed to the proliferation of modeling approaches to estimate environmental exposures over the past decade.

While satellite imagery and numerical model simulation can provide useful information on environmental exposures, these proxy data cannot directly replace monitoring measurements in health analyses. Specifically, associations between pollutant levels and remotely sensed parameters often vary across meteorological condition, land cover, and pollution composition. Satellite data are also subject to retrieval errors and informative missingness. The main disadvantage of numerical models is their high computational demand. Numerical model outputs are also not based on observations, and errors can arise from incorrect input data on sources, incorrect representation of the underlying complex processes with partial differential equations, and discretization of the continuous environmental field in space and in time.

This chapter provides a review of *geostatistical methods* for modeling environmental pollution fields. These statistical models aim to combine monitoring measurements and proxy data to improve exposure assessment while accounting for the errors in proxy data. Geostatistical modeling, particularly, aims to borrow information spatially from nearby observations to perform interpolation. In contrast to many machine learning approaches, statistical models can also provide interpretable parameters and uncertainty quantification in a principled way.

Recent approaches in geostatistical methods for modeling environmental exposures can be categorized into two broad groups: *melding* and *calibration*. In melding, both measurements and proxy data are viewed as error-prone realizations of an unobserved (latent) true pollution field, whereas in calibration, the proxy serves as a predictor for the observed measurements. Both melding and calibration encounter several common modeling challenges. First, satellite images and numerical model outputs represent areal spatial data over contiguous grid cells. When linked to point-reference monitoring locations, a spatial change of support is encountered. Second, the bias between monitoring measurements and proxy data can exhibit complex spatial and temporal structures.

The modeling approaches presented in this chapter are largely drawn from air quality research because of the rich literature on exposure modeling and health effect estimation. In air quality research, recent work has focused predominantly on two pollutants: fine particulate pollution $PM_{2.5}$ (particulate matter less than 2.5 μm in aerodynamic diameter) and ground-level ozone. Both $PM_{2.5}$ and ozone have been linked to health outcomes such as premature mortality, asthma exacerbation, cardiorespiratory morbidity, and adverse birth outcomes. As a remotely sensed proxy, satellite-derived aerosol optical depths (AOD) have been used to monitor $PM_{2.5}$ concentrations. AOD measures light extinction due to airborne particles in the atmospheric column, and previous studies have found positive associations between $PM_{2.5}$ level and AOD at different spatial and temporal scales. Similarly, several global and regional numerical models have been developed to simulate $PM_{2.5}$ and ozone concentrations. Examples include the Community Multiscale Air Quality Model (CMAQ) and GEOS-Chem.

The rest of this chapter is organized as follows. First, we will introduce geostatistical methods often used to model environmental processes with land use and meteorological variables. We will then review the general framework of Bayesian melding, followed by statistical calibration. To simplify notation, the spatial version of the approach will be presented with noted spatial-temporal extensions. Several recent advances such as multipollutant models, multiscale fusion, and quantile calibration will be discussed.

Geostatistical Models for Environmental Exposures

The goal of a geostatistical model is to use observed point-level exposure measurements at sparse locations to estimate the unobserved spatial exposure surface. Let $Y(\mathbf{s}_i)$ denote the pollutant measurement observed at point-location $\mathbf{s}_i$ with coordinate (s_{1i}, s_{2i}), which is often the projected x–y coordinate. We assume $Y(\mathbf{s}_i)$ is an error-prone version of an unobserved true exposure $W(\mathbf{s}_i)$. For a set of monitoring locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, a typical spatial regression model is given by

$$Y(\mathbf{s}_i) = W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \tag{1}$$

$$W(\mathbf{s}_i) = \mathbf{Z}(\mathbf{s}_i)^T \boldsymbol{\alpha} + v(\mathbf{s}_i), \tag{2}$$

where $\mathbf{Z}(\mathbf{s}_i)$ is a $p \times 1$ vector of covariates that are useful for predicting of the exposure and $\boldsymbol{\alpha}$ is the corresponding $p \times 1$ vector of regression coefficients. Component $v(\mathbf{s}_i)$ in Eq. (2) represents *spatially-dependent* residuals not explained by the covariate $\mathbf{Z}(\mathbf{s}_i)$, and component $\epsilon(\mathbf{s}_i)$ represents independent

residuals due to instrumental error or fine-scale spatial variation not captured by $v(\mathbf{s}_i)$ and $\mathbf{Z}(\mathbf{s}_i)^T\boldsymbol{\alpha}$.

Assuming the exposure measurement $Y(\mathbf{s}_i)$ is normally distributed or suitably transformed to be normal, the latent variable $v(\mathbf{s}_i)$ is typically modeled as a Gaussian process. A mean-zero Gaussian process assumes that the joint distribution of $v(\mathbf{s}_i)$ at any finite set of locations is multivariate normal:

$$\mathbf{v} = [\, v(\mathbf{s}_1), v(\mathbf{s}_2), \ldots, v(\mathbf{s}_n)\,]^T \sim N\,(\,\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\,)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is an $n \times n$ covariance matrix parameterized by $\boldsymbol{\theta}$. One can also think of $v(\mathbf{s}_i)$ as spatially dependent random effects in a mixed model or hierarchical model framework. The entries of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are determined by a covariance function $C(\cdot|\boldsymbol{\theta})$. Let $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ denote the Euclidean distance between two locations $\mathbf{s}_i$ and $\mathbf{s}_j$. One popular spatial covariance function is the Matérn:

$$C(d_{ij}|\boldsymbol{\theta}) = \theta_1 \frac{1}{\Gamma(\theta_3)2^{\theta_3-1}} \left(\frac{d_{ij}}{\theta_2}\right)^{\theta_3} K_{\theta_3}\left(\frac{d_{ij}}{\theta_2}\right)$$

where $\theta_1, \theta_2, \theta_3 > 0$, and $K_{\theta_3}(\cdot)$ are the modified Bessel function of the second kind. The Matérn covariance function contains two important special cases. First, when $\theta_3 = 0.5$, it corresponds to the exponential covariance function: $C(d_{ij}|\boldsymbol{\theta}) = \theta_1 e^{-d_{ij}/\theta_2}$; second, as $\theta_3 \to \infty$, it gives the double-exponential covariance function $C(d_{ij}|\boldsymbol{\theta}) = \theta_1 e^{-d_{ij}^2/\theta_2^2}$.

In the above covariance functions, θ_1 corresponds to the value when $d_{ij} = 0$. Hence, θ_1 is known as the *marginal variance* which describes the variability at any location across independent realizations of the Gaussian process. Parameter θ_2 is known as the *range* parameter and describes the correlation as a function of distance. Finally, θ_3 determines the *smoothness* of the covariance function. The Matérn family is an example of *isotropic* covariance function because its value only depends on the distance between locations, regardless of direction and regions of the modeling domain.

Modeling the spatial dependence in $v(\mathbf{s}_i)$ is what allows us to perform spatial interpolation (kriging) using the set of observations $\mathbf{W} = [W(\mathbf{s}_1), W(\mathbf{s}_2), \ldots, W(\mathbf{s}_n)]^T$ to predict at locations without measurements. Similarly, let $\mathbf{W}^*$ denote a vector of n^* exposures to be interpolated and $\mathbf{Z}^*$ the corresponding matrix of covariate predictors. The joint distribution of $\mathbf{W}$ and $\mathbf{W}^*$ is given by the multivariate Gaussian distribution

$$\begin{bmatrix}\mathbf{W}\\ \mathbf{W}^*\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{Z}\boldsymbol{\alpha}\\ \mathbf{Z}^*\boldsymbol{\alpha}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta},11} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},12}\\ \boldsymbol{\Sigma}_{\boldsymbol{\theta},21} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},22}\end{bmatrix}\right),$$

where the block covariance matrix is governed by the same covariance function of $\mathbf{v}$. Assuming the residual error is given by $\epsilon(\mathbf{s}_i) \sim N(0, \sigma_\epsilon^2)$, the multivariate Gaussian conditional distribution of $\mathbf{W}^*$ can be derived with respect to observations $\mathbf{Y}$:

$$[\mathbf{W}^*|\mathbf{Y}] \sim N\Big[\, \mathbf{Z}^*\boldsymbol{\alpha} + \boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} + \sigma_\epsilon^2\mathbf{I}_n)^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}),$$
$$\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} + \sigma_\epsilon^2\mathbf{I}_n)^{-1}\boldsymbol{\Sigma}_{12}\Big], \qquad (3)$$

suppressing the $\boldsymbol{\theta}$ subscript in $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ above for notational ease. Finally, in most studies, the exposure estimate $\widehat{\mathbf{W}}^*$ is then defined as the conditional mean:

$$\widehat{\mathbf{W}}^* = \mathbf{Z}^*\boldsymbol{\alpha} + \boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} + \sigma_\epsilon^2\mathbf{I}_n)^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}). \qquad (4)$$

The above is similar to what is known as *universal Kriging*. In the special case of no spatial dependence, i.e., $\boldsymbol{\Sigma}_{21} = \mathbf{0}$ in Eq. (3), the estimate becomes the mean trend $\widehat{\mathbf{W}}^* = \mathbf{Z}^*\boldsymbol{\alpha}$ and does not utilize information from observations $\mathbf{Y}$. Finally, if we have no covariate, the estimate is given by

$$\widehat{\mathbf{W}}^* = \boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{Y}\,,$$

which can be viewed as a weighted average of observed measurements $\mathbf{Y}$, also known as *simple kriging*.

We now review several commonly used regression approaches for modeling environmental exposures using different forms of Eq. (2). First, if we assume $v(\mathbf{s}_i) = 0$ for all $\mathbf{s}$, then the model becomes a standard multiple regression model and $\widehat{\mathbf{W}}^* = \mathbf{Z}^*\boldsymbol{\alpha}$. This is also known as a *land use regression* model where the exposure surface is driven only by spatially varying predictors, such as elevation, roadway density, and distance to pollution sources (Ryan and LeMasters 2007; Hoek et al. 2008). One challenge with land use regression models is that predictions can only capture spatial variation represented by the selected land use variables. Often additional measurement campaigns are conducted to enrich the observation dataset and the distribution of predictor variables.

Instead of specifying the residual spatial trend using a Gaussian process, it can also be modeled parametrically using functions of coordinates directly, for example, with quadratic terms s_{1i}, s_{2i}, s_{1i}^2, and s_{2i}^2 in the mean. More generally, one can use spatial basis functions:

$$v(\mathbf{s}_i) = \mathbf{S}(\mathbf{s}_i)^T\boldsymbol{\theta}\,,$$

where $\mathbf{S}(\mathbf{s}_i) = [S_1(\mathbf{s}_i), \ldots, S_K(\mathbf{s}_i)]^T$ is a vector of K basis functions evaluated at location $\mathbf{s}_i$. If the coefficient vector $\boldsymbol{\gamma}$ is assumed to be mean-zero Gaussian random effects with covariance matrix $\boldsymbol{\Omega}$, then $v(\mathbf{s}_i)$ is Gaussian with the covariance between $\mathbf{s}_i$ and $\mathbf{s}_j$ given by $\mathbf{S}(\mathbf{s}_i)^T\boldsymbol{\Omega}\mathbf{S}(\mathbf{s}_j)$. The above formulation has two main advantages. First, the choice

of basis functions and covariance matrix $\mathbf{\Omega}$ can allow for flexible structures. Second, when the number of basis functions K is smaller than the number of spatial locations, one can avoid large matrix inversions in the kriging solution, which can be computational burdensome for large datasets. This approach is also known as *fixed-rank* kriging (Cressie and Johannesson 2008). Different basis functions have been proposed using pre-specified covariance function (Kammann and Wand 2003), thin-plate splines (Wood 2003), and compact kernels (Nychka et al. 2015).

Spatial-temporal models are used for exposures that are measured across both space and discrete time points, indexed by t. Spatial-temporal covariates $Z(\mathbf{s}_i, t_i)$ often include short-term meteorological variables such as temperature and precipitation. For exposures with smooth temporal trends, including them in the mean via temporal splines or cyclic functions may be sufficient; otherwise, additional temporal autocorrelation can built into the random effect $v(\mathbf{s}_i, t_i)$. There is a rich literature on how to model spatial-temporal data with different covariance functions (Gneiting et al. 2006). The simplest covariance function assumes that correlation due to spatial and temporal proximity is *separable*, i.e., $\text{Corr}[v(\mathbf{s}_i, t_i), v(\mathbf{s}_j, t_j)] = h_1(||\mathbf{s}_i - \mathbf{s}_j||) \times h_2(|t_i - t_j|)$, where $h_1(\cdot)$ and $h_2(\cdot)$ are correlation functions in space and time, respectively. A dynamic model is another popular formulation where the spatial process evolves through time. For example, let $\rho \in [0, 1]$, we can assume $v(\mathbf{s}_i, t_i) = \rho \times v(\mathbf{s}_i, t_i - 1) + \eta(\mathbf{s}_i, t_i)$, where $\eta(\mathbf{s}_i, t_i)$ is a second GP that is independent across time points. Alternatively, one might consider modeling spatial-temporal data as time-series data with spatial dependence; for example, see Lindström et al. (2014).

Application: Modeling Daily $PM_{2.5}$ Concentration

Here we describe an application of geostatistical models for estimating $PM_{2.5}$ concentrations in Southeastern USA. Figure 1 shows the observed 24-h average $PM_{2.5}$ concentrations at 78 sites on a particular day. We first construct a 12×12 km grid over the modeling domain. Then for each grid cell, three additional spatial predictors are obtained: elevation (m), percent forest cover, and simulated $PM_{2.5}$ levels from the Community Multiscale Air Quality (CMAQ) modeling system from the US Environmental Protection Agency. CMAQ version 5.0.2 was run at a 12×12 km resolution with 35 vertical layers that span till the top of the free troposphere. In this analysis, we only used the surface layer, which is nominally 19 m tall. Figure 2 shows the CMAQ $PM_{2.5}$ simulations.

We evaluate the spatial prediction performance of linear regression models and spatial kriging with different set of predictors. This is accomplished using cross-validation (CV) experiments where we split the data repeatedly into a training set and a validation set. The training set is used to fit various models and predictions made from the training set compared to the left-out observations in the validation set. Here we implement at leave-one-site CV where each observation is treated as a validation data point, while the other 77 observations are used to fit different models. The process is repeated 78 times, until each observation has served as a validation data point.

Let $\hat{Y}_i$ be the prediction for observation i when i is treated as the validation data and Y_i the actual observation. Each prediction is also associated with a Kriging variance $\text{Var}(\hat{y}_t)$.

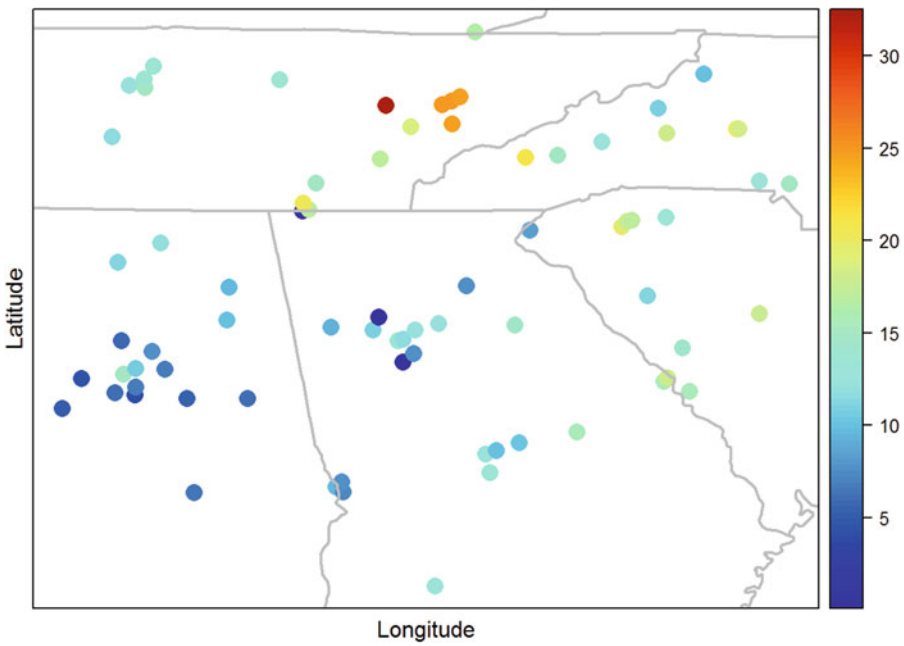


Fig. 1 Daily 24-h average $PM_{2.5}$ concentration ($\mu g/m^3$) in the Southeastern US modeling domain

Table 1 Prediction performance for $PM_{2.5}$ using linear regression or Kriging with different covariates: elevation (Elev), percent forest cover (Forest), and simulations from numerical model (CMAQ). Performance criteria include root mean square error (RMSE), mean absolute error (MAE), empirical 95% prediction interval coverage (Cov95), and the average prediction standard error (SE)

Model	Covariates	RMSE	MAE	Cov95	Avg SE
Linear regression	Elev+Forest	5.80	4.33	0.94	6.78
Linear regression	CMAQ	5.67	4.05	0.95	6.48
Linear regression	Elev+Forest+CMAQ	5.52	3.87	0.96	6.62
Kriging	None	4.26	3.22	0.94	4.22
Kriging	Elev+Forest	4.07	2.79	0.94	4.26
Kriging	CMAQ	4.07	2.87	0.94	4.05
Kriging	Elev+Forest+CMAQ	4.16	2.86	0.92	4.23

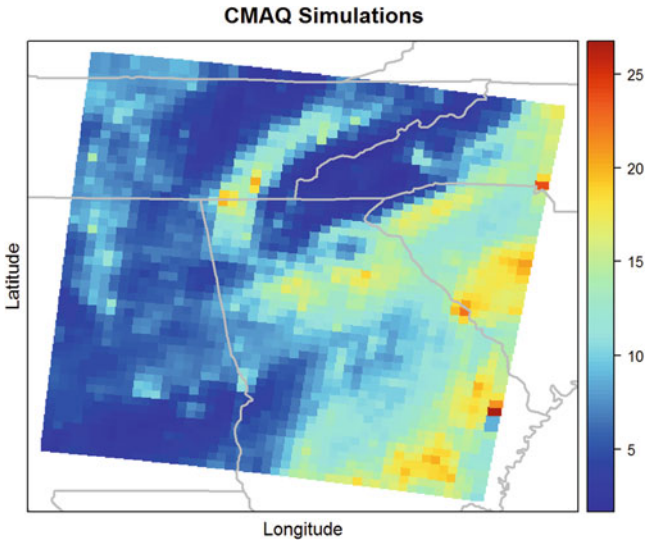


Fig. 2 Simulated $PM_{2.5}$ concentrations ($\mu g/m^3$) in Southeastern USA

We use the following criteria for prediction performance to compare models:

- Root mean squared error (RMSE):

$$\sqrt{\frac{1}{78}\sum_{i=1}^{78}(\hat{Y}_i - Y_i)^2}$$

- Mean absolute error (MAE):

$$\frac{1}{78}\sum_{i=1}^{78}|\hat{Y}_i - Y_i|$$

- Average prediction standard error (SE):

$$\frac{1}{78}\sum_{i=1}^{78}\sqrt{\text{Var}(\hat{y}_i)}$$

- Empirical coverage of the 95% prediction interval:

$$\frac{1}{78}\sum_{i=1}^{78}1_{Y_i} \in \hat{Y}_i \pm 1.96 \times \sqrt{\text{Var}(\hat{y}_i)}$$

RMSE and MAE measure the overall error between out-of-sample prediction and observations. Average SE describes the precision of the prediction. The empirical coverage probability assesses whether the prediction intervals constructed have the desired property. Specifically, a 95% prediction interval should include the observations 95% of the time.

Table 1 summarizes the CV results, where we see that Kriging has better prediction performance (lower RMSE, lower MAE, lower Avg SE) compared to linear regression that does not incorporate spatial correlation in the prediction. For Kriging models, we assumed an exponential spatial correlation structure, but other correlation structures are also possible. For all models, the prediction intervals have good empirical coverage probability. We also find that Kriging model with the means having CMAQ as the predictive gives the smaller RMSE and average SE among the models examined.

Bayesian Melding

Bayesian melding is a data integration approach developed specifically to combine point-level monitoring measurements with gridded proxies. Again, let $Y(\mathbf{s})$ denote the pollutant concentration measurement from an air quality monitor at point-location $\mathbf{s}$ with coordinates (s_1, s_2). For notational ease, we now suppress the subscript i for locations $\mathbf{s}_i$. The gridded proxy data are denoted by $X(B_\mathbf{s})$, where $B_\mathbf{s}$ indexes the contiguous grid cell that includes point location $\mathbf{s}$. The spatial resolution of the proxy data varies based on satellite retrieval algorithms and whether the numerical simulation is performed on a global or a regional scale. In air quality applications, the spatial resolution of satellite-derived AOD ranges from 10 km to 1 km; the spatial resolution ranges from 100 km to 4 km for numerical model simulations.

In Bayesian melding, we assume observations $Y(\mathbf{s})$ and $X(B_\mathbf{s})$ arise from a common unobserved latent process $W(\mathbf{s})$, representing the true pollutant field. The model is formulated as follows:

$$Y(\mathbf{s}) = W(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{5}$$

$$W(\mathbf{s}) = \mu(\mathbf{s}) + v(\mathbf{s}) \tag{6}$$

$$X(B_\mathbf{s}) = \frac{1}{|B_\mathbf{s}|}\int_{B_\mathbf{s}} \widetilde{X}(\mathbf{s})\, d\mathbf{s} \tag{7}$$

$$\widetilde{X}(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})W(\mathbf{s}) + \delta(\mathbf{s})\ . \tag{8}$$

Equation (5) treats the observed monitoring data $Y(\mathbf{s})$ as an error-prone realization of the latent process $W(\mathbf{s})$ with independent measurement error $\epsilon(\mathbf{s})$. The latent process $W(\mathbf{s})$ has a spatial trend $\mu(s)$ and a spatially dependent residual component $v(\mathbf{s})$ that is a Gaussian process (GP). Equation (7) introduces a conceptual point-referenced proxy data $\widetilde{X}(\mathbf{s})$ and its spatial average over grid cell $B_\mathbf{s}$ resulting in the observed grid-level proxy value. Finally, the point-level proxy $\widetilde{X}(\mathbf{s})$ is linked to the latent true pollutant field $W(\mathbf{s})$ via a linear regression model. Coefficients $a(\mathbf{s})$ and $b(\mathbf{s})$ are often interpreted as the additive and multiplicative calibration parameters for the proxy, and component $\delta(\mathbf{s})$ represents random proxy error.

The above framework was first described by Fuentes and Raftery (2005) for assessing spatial bias in CMAQ. Subsequent applications in predicting pollution fields have found that Bayesian melding consistently outperforms kriging (Berrocal et al. 2010b; Liu et al. 2011). Several features are worth noting. The main advantage of melding is the use of a latent continuous field that allows the spatially misaligned measurements and proxy data to jointly provide information on $W(\mathbf{s})$, which is the quantity of interest. Conceptually, the latent variable approach offers straightforward extensions to multiple proxies (Crooks and Isakov 2013) and multiple pollutants (Sahu et al. 2010). However, the model is highly parameterized, and identifiability is a frequent concern. Specifically, we first need to decompose the residual variation in $Y(\mathbf{s})$ into two error components $v(\mathbf{s})$ and $\epsilon(\mathbf{s})$. The structures of the spatial calibration parameters $a(\mathbf{s})$ and $b(\mathbf{s})$ also need to be selected with care as it determines how much variation in the proxy can be attributed to the true pollutant field versus output bias. Specifically, $a(\mathbf{s})$ and $b(\mathbf{s})$ are usually parametrized as fixed effects, instead of spatial random fields, to avoid identifiability problems with estimating $W(\mathbf{s})$.

Using CMAQ proxy to estimate weekly SO_2 concentration in eastern USA, Fuentes and Raftery (2005) have the following parameterizations for the three independent variance components: $\epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \sigma_e^2)$, $v(\mathbf{s})$ is a Gaussian process with nonstationary covariance function, and $\delta(\mathbf{s}) \overset{iid}{\sim} N(0, \sigma_\delta^2)$. For the structural components, $\mu(\mathbf{s})$ and $a(\mathbf{s})$ are polynomial functions of $\mathbf{s}$, and $b(\mathbf{s})$ is assumed to be an unknown constant. For modeling ozone concentration and numerical model outputs, a similar model is employed by Berrocal et al. (2010b) and Liu et al. (2011), with the exception that the covariance function of $v(\mathbf{s})$ is assumed to be exponential. One important observation from Berrocal et al. (2010b) is that the predictive surface obtained from melding closely follows the spatial gradients of the proxy, with values closer to CMAQ especially in unmonitored areas. In melding, the proxy can dominate for two reasons. First, there is considerably more proxy data than monitoring measurements. Second, by specifying $a(\mathbf{s})$ as a smooth spatial trend, fine-scale spatial variation in the proxy is assumed to reflect the true pollutant field. The issue of disentangling spatial scales between $a(\mathbf{s})$ and $W(\mathbf{s})$ is further investigated by Paciorek (2012). Through a simulation study and an application of predicting $PM_{2.5}$ levels using CMAQ or AOD, Paciorek finds that modeling $a(\mathbf{s})$ flexibly significantly reduces the usefulness of the proxy.

Bayesian melding often involves considerable computational effort because of the change-of-support integral in Eq. (7), and a large number of spatial points need to be evaluated for $W(\mathbf{s})$. Typically, the integral is approximated using Monte Carlo integration. For example, Berrocal et al. (2010b) use a systematic sample of 4 points for each CMAQ grid cell (12 km resolution). Liu et al. (2011) consider randomly selecting a fixed number of points to avoid ill-conditioned spatial covariance matrix. Several approaches to decrease computational burden have been proposed. First, Sahu et al. (2010) introduce an areal true pollutant process $\widetilde{W}(B_\mathbf{s})$ that has the same grid as the proxy. This latent discrete spatial variation can be efficiently estimated using a conditionally autoregressive (CAR) model (Besag 1974). A measurement error model, Eq. (9), is then used to resolve the mismatch between point-referenced $W(\mathbf{s})$ and areal true pollutant processes:

$$\begin{aligned} Y(\mathbf{s}) &= W(\mathbf{s}) + \epsilon(\mathbf{s}) \\ W(\mathbf{s}) &= \widetilde{W}(B_\mathbf{s}) + v(\mathbf{s}) \\ X(B_\mathbf{s}) &= \gamma_0 + \gamma_1 \widetilde{W}(B_\mathbf{s}) + \psi(B_\mathbf{s}) \end{aligned} \tag{9}$$

where $\epsilon(\mathbf{s})$, $v(\mathbf{s})$), and $\psi(B)$ are normal independent errors. An additional simplification is taken by McMillan et al. (2010) to model $PM_{2.5}$ levels and CMAQ outputs by eliminating the latent variable, $W(\mathbf{s})$, for all $\mathbf{s}$. Under this model, monitoring measurements are linked to the areal latent process directly as in Eq. (6):

$$\begin{aligned} Y(\mathbf{s}) &= \widetilde{W}(B_\mathbf{s}) + \epsilon(\mathbf{s}) \\ X(B_\mathbf{s}) &= \gamma_0 + \gamma_1 \widetilde{W}(B_\mathbf{s}) + \psi(B_\mathbf{s}). \end{aligned} \tag{10}$$

Here $\epsilon(\mathbf{s})$ incorporates both measurement error in monitoring data and error due to spatial misalignment. One important consequence of Eq. (10) is that the model can only provide gridded pollution predictions because $\widetilde{Z}(B_\mathbf{s})$ is modeled as a discrete spatial process. This is often referred to as *upscaling* as the original point-referenced monitoring data has been coarsened to areal level.

Because of the gain in computational speed, both Sahu et al. (2010) and McMillan et al. (2010) are able to perform Bayesian melding in a spatio-temporal setting with an additional dynamic time series model for the latent process. Their models provide daily pollution predictions that are useful for health studies that require short-term exposure assessment. Choi et al. (2009) also extend the melding framework of Fuentes and Raftery (2005) to a temporal setting by modeling the latent pollution field on day t as $Z(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)$. The trend $\mu(\mathbf{s}, t)$ includes time-varying meteorological variables, and $\epsilon(\mathbf{s}, t)$ is modeled as an autoregressive Gaussian process. More recently, Gilani et al. (2016) used Bayesian melding to model near-roadway pollution with monitoring measurements and dispersion model output. They used a non-stationary covariance function that is dependent on wind direction.

Statistical Calibration

Motivated by the increasing amount of proxy data being generated and the limitations of Bayesian melding, Berrocal et al. (2010b) develops a statistical calibration approach for using CMAQ data. Under a calibration framework, the proxy is viewed as a predictor for measurements. Let $X(B_\mathbf{s})$ denote the proxy grid cell linked to a monitor at point-location $\mathbf{s}$. The regression model is given by:

$$Y(\mathbf{s}) = \alpha_0(\mathbf{s}) + \alpha_1(\mathbf{s})X(B_\mathbf{s}) + \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \sim N(0, \sigma^2_\epsilon)\,. \tag{11}$$

Here $\alpha_0(\mathbf{s})$ and $\alpha_1(\mathbf{s})$ can be interpreted as the spatially varying additive and multiplicative calibration parameters of the error-prone proxy data. Parameters $\alpha_0(\mathbf{s})$ and $\alpha_1(\mathbf{s})$ are modeled as a bivariate continuous spatial process via linear coregionalization model (LMC) (Gelfand et al. 2004). Briefly, two independent mean-zero, unit-variance Gaussian processes $U_1(\mathbf{s})$ and $U_2(\mathbf{s})$ are introduced. Correlation between $\alpha_0(\mathbf{s})$ and $\alpha_1(\mathbf{s})$ is induced by letting $\alpha_0(\mathbf{s}) = c_1U_1(\mathbf{s})$ and $\alpha_1(\mathbf{s}) = c_2U_1(\mathbf{s}) + c_3U_2(\mathbf{s})$, where c_1, c_2, c_3 are constants that determine the marginal variance for each random effect and their correlation.

There are several advantages that statistical calibration offers compared to Bayesian melding. First, note that the only proxy data used for model fitting are those linked to a monitor. Proxy data not linked to a monitor are only used for predictions. This reduces the computational effort considerably as the number of monitoring locations is usually much smaller than the number of proxy grid cells. Second, by treating $\alpha_0(\mathbf{s})$ and $\alpha_0(\mathbf{s})$ as continuous processes, they can be interpolated smoothly in space. This allows point-level predictions even though the proxy predictor represents an areal average. This feature is often referred to as *downscaling*, in contrast to the upscaler by McMillan et al. (2010) in Eq. (10). Statistical downscaling has been used to combine CMAQ data (Berrocal et al. 2010b), as well as AOD for predicting $PM_{2.5}$ fields (Chang et al. 2014). When compared to Bayesian melding, Berrocal et al. (2010b) found that downscaling CMAQ for daily ozone level results in smaller spatial prediction error in cross-validation experiments. Similar findings are observed by Paciorek (2012) in a simulation study.

Outputs from dispersion models that provide air quality simulations at point level have also been considered (Lindström et al. 2014; Pirani et al. 2014). Finally, in the recent annual Global Burden of Disease Study published by the World Health Organization, global ambient $PM_{2.5}$ estimates were derived from monitoring measurements, satellite-derived AOD, and numerical model simulations (Shaddick et al. 2018), and data integration was conducted under the statistical calibration framework.

One major limitation for using the proxy as a predictor is that it cannot contain missing values. For example, satellite-retrieved AOD can be missing due to cloud cover and highly reflective surfaces such as snow. Several methods have been proposed to account for missing satellite information. For example, Grantham et al. (2018) developed a spatial regression framework that accounts for informative spatial missingness. Murray et al. (2019) modeled $PM_{2.5}$ using an ensemble approach such that when AOD is not available, one can use estimates from other models (e.g., a model with CMAQ as the proxy). Finally, the calibration framework assumes that the observed measurements $Y(\mathbf{s})$ are the gold standard, even though instrumental error is likely to be present.

Extension to spatio-temporal data is straightforward by allowing the calibration parameters in Eq. (11) to be time-varying: $\alpha_0(\mathbf{s}, t)$ and $\alpha_1(\mathbf{s}, t)$. For computational efficiency, the space-time processes can be decomposed into additive components, i.e., $\alpha_j(\mathbf{s}, t) = \alpha_j(\mathbf{s}) + \alpha_j(t)$, for $j = 1, 2$. The temporal component is then assumed to evolve dynamically in time via an autoregressive model. In combining $PM_{2.5}$ and satellite-derived AOD, Chang et al. (2014) demonstrate the importance of considering temporal dependence in the calibration parameters because missing AOD data can result in days with no linked AOD-measurement pair. With sufficient monitoring locations, one can assume that $\alpha_j(\mathbf{s}, t)$ evolves dynamically over time. However, for combining ozone and CMAQ outputs in the eastern USA, Berrocal et al. (2010b) find that the model allowing the spatial calibration parameters to be independent across days has the best prediction

performance. Finally, in Zidek et al. (2012), the authors assume spatially varying calibration parameters but model the residuals $\epsilon(\mathbf{s}, t)$ to be autoregressive temporal processes with spatially varying autoregressive parameters. The motivation is to model the temporal variation in the measurements via the residuals, instead of that inherent in the proxy data.

Model Extensions

Multivariate Exposure Modeling

Humans are exposed to multiple pollutants simultaneously, and different pollutants can share the same sources. Consequently, there is growing interest in developing data fusion methods for multiple pollutants to support health research. A multipollutant approach may also improve prediction performance as we can exploit the dependence between pollutants. This is particularly advantageous when the pollutant monitors are not co-located or have different measurement schedules.

Choi et al. (2009) present an interesting application of Bayesian melding for five constituents of $PM_{2.5}$. Here the proxy data are measurements from another network that only provides the sum of the five constituents. The objective is to model individual pollutant concentration at location $\mathbf{s}$ as a function of a latent pollutant sum field. Let $Y_l(\mathbf{s})$ denote the measured lth pollutant's concentration, and let $S(\mathbf{s}) = \sum_{l=1}^{5} Y_l(\mathbf{s})$ denote the unobservedlatent sum. Similarly, let $X(\mathbf{s})$ be the sum measured from the proxy network. The hierarchical model is given by:

$$\begin{aligned} Y_l(\mathbf{s}) &= \theta_l(\mathbf{s})\, S(\mathbf{s}) + \epsilon_l(\mathbf{s}, t) \\ S(\mathbf{s}) &= \mu(\mathbf{s}) + \epsilon(\mathbf{s}) \\ X(\mathbf{s}) &= a(\mathbf{s}) + S(\mathbf{s}) + \delta(\mathbf{s})\ . \end{aligned} \tag{12}$$

Equation (12) expresses the observed pollutant as a proportion of the latent sum. The pollutant-specific proportion $\theta_k(\mathbf{s})$ is allowed to vary spatially and is specified as

$$\theta_k(\mathbf{s}) = \frac{\exp(\alpha_k(\mathbf{s}))}{\sum_{j=1}^{5} \exp(\alpha_j(\mathbf{s}))}$$

where for $j = 1, \ldots, 5$, $\alpha_j(\mathbf{s})$ is a Gaussian process, independent across j. Parameter $\alpha_5(\mathbf{s})$ is set to 0 for all $\mathbf{s}$ for identifiability purposes. To account for the correlated measurement error, $\epsilon_l(\mathbf{s})$ is modeled jointly using LMC. Note that here a change-of-support calculation is not needed because the proxy is available at the point level. Spatial-temporal data are accommodated by replacing $\theta_l(\mathbf{s})$ with a dynamic Gaussian process.

Multipollutant approaches have also been proposed under the downscaling framework. First, Berrocal et al. (2010a) extend their original model to a bi-pollutant setting for ozone and $PM_{2.5}$ using CMAQ outputs as the proxy. Following previous notation in Eq. (11), the bi-pollutant model is given by:

$$\begin{aligned} Y_1(\mathbf{s}) &= \alpha_{10}(\mathbf{s}) + \alpha_{11}(\mathbf{s})X_1(B_\mathbf{s}) + \alpha_{12}(\mathbf{s})X_2(B_\mathbf{s}) + e_1(\mathbf{s}), \quad e(\mathbf{s}) \sim N(0, \sigma^2_{\epsilon_1}) \\ Y_2(\mathbf{s}) &= \alpha_{20}(\mathbf{s}) + \alpha_{21}(\mathbf{s})X_2(B_\mathbf{s}) + \alpha_{22}(\mathbf{s})X_1(B_\mathbf{s}) + e_2(\mathbf{s}), \quad e(\mathbf{s}) \sim N(0, \sigma^2_{\epsilon_2}). \end{aligned}$$

We note that the proxy variables $X_1(B_\mathbf{s})$ and $X_2(B_\mathbf{s})$ are used to model each outcome, maximizing potential information in the proxy data. Again, the six calibration parameters are modeled jointly using LMC where various between-pollutant and/or between-proxy dependence structures can be investigated. To handle sites where only one of the pollutants is observed, a data augmentation step for the LMC latent variables is included in the Bayesian estimation algorithm. The above model involves numerous parameters, and extension to more than two pollutants has yet to be examined.

Finally, we describe a calibration approach by Crooks and Özkaynak (2014) that includes a sum constraint for modeling $PM_{2.5}$ constituents. The motivating application entails simultaneously combining monitoring data and CMAQ outputs for five $PM_{2.5}$ constituents and the total $PM_{2.5}$ mass. The sum constraint is accomplished by modeling the klth pollutant concentration using a Gamma distribution:

$$Y_l(\mathbf{s}) \sim \text{Gamma}\left(\tau^{-1} \times [\,\alpha_{k0}(\mathbf{s}) + \alpha_1(\mathbf{s})X_l(B_\mathbf{s})\,],\ \tau^{-1}\right).$$

Note that the Gamma rate parameter τ and the multiplicative calibration parameter $\alpha_1(\mathbf{s})$ do not vary across pollutants. Assuming the pollutants are independent, the observed sum also follows a Gamma distribution. This allows a mass conservation requirement using the observed total $PM_{2.5}$ mass. Despite these distributional assumptions, in cross-validation experiments, the authors find that both mass conservation and the multipollutant approach improve prediction accuracy.

Multiple-Scale Calibration

In calibration, at each monitoring location, only the single linked proxy grid cell is used to provide information on the observed measurements. Because of the spatial dependence in pollutant field, it is reasonable to also consider whether neighboring grid cells are useful for prediction. Berrocal et al. (2012) propose two approaches to borrow proxy information across multiple grid cells. First, they replace the single grid cell predictor $X(B_{\mathbf{s}})$ in Eq. (11) by a *smoothed* version, $\tilde{X}(B_{\mathbf{s}})$:

$$Y(\mathbf{s}) = \alpha_0(\mathbf{s}) + \alpha_1 \tilde{X}(B_{\mathbf{s}}) + \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon^2) \quad (13)$$

The new predictor $\tilde{X}(B_{\mathbf{s}})$ is taken as the spatial grid-level random effect with a CAR structure: $X(B_{\mathbf{s}}) = \tilde{X}(B_{\mathbf{s}}) + \psi(\mathbf{s})$, where $\psi(\mathbf{s})$ is an independent normal residual error. Note that the above multiplicative calibration parameter α_1 is constant in space to avoid identifiability problems. The use of smoothed proxy data offers two advantages. First this approach avoids the spatial misalignment due to a monitor's location within a proxy grid cell. Specifically, if a monitor is near the boundary of a grid cell, then it's natural to also consider proxy values at the closer grid cell. Second, it provides a flexible framework to utilize all of the proxy data in predicting $Y(\mathbf{s})$. This mimics the Bayesian melding approach where all the proxy data are used to estimate the latent pollutant field. However,the downscaler enables the measurement data to decide how much local smoothing is required to achieve optimal prediction.

Berrocal et al. (2012) also consider an alternative smoothing approach by deriving a point-referenced proxy that represents a weighted average across all proxy grid cells:

$$\tilde{X}(\mathbf{s}) = \sum_{g=1}^{G} w_g(\mathbf{s}) X(B_k).$$

Let $\mathbf{r}_g$ be the centroid of grid cell g; the weights are defined as:

$$w_g(\mathbf{s}) = \frac{K(\mathbf{s} - \mathbf{r}_g)\ \exp(Q(\mathbf{r}_g))}{\sum_{l=1}^{G} K(\mathbf{s} - \mathbf{r}_l)\ \exp(Q(\mathbf{r}_l))}$$

where $K(\cdot)$ is a Gaussian kernel with bandwidth covering three grid cells in each direction, and $Q(\cdot)$ is a mean-zero Gaussian process approximated using predictive process (Banerjee et al. 2008). By including $Q(\cdot)$, the weight $w_g(\mathbf{s})$ is allowed be to asymmetrical among the grid cells around location $\mathbf{s}$. In their application to daily ozone concentration, the authors find that the use of smoothed CMAQ proxy provides better prediction power, especially at locations farther from the rest of the monitors.

Reich et al. (2014) propose a *spectral* downscaler that extends Berrocal et al. (2012) further by using multiple smoothed proxies at different spatial scales. The conceptual framework begins by considering the spectral representation of the continuous processes associated with the measurement and the proxy:

$$Y(\mathbf{s}) = \int \exp(-i\boldsymbol{\omega}^T\mathbf{s}) H_1(\boldsymbol{\omega})\, d\boldsymbol{\omega} \quad \text{and} \quad X(\mathbf{s}) = \int \exp(-i\boldsymbol{\omega}^T\mathbf{s}) H_2(\boldsymbol{\omega})\, d\boldsymbol{\omega}$$

where $H_1(\boldsymbol{\omega})$ and $H_2(\boldsymbol{\omega})$ are mean-zero Gaussian processes. The correlation between $H_1(\boldsymbol{\omega})$ and $H_2(\boldsymbol{\omega})$ is assumed to vary across frequency $\boldsymbol{\omega}$. Assuming $X(\mathbf{s})$ is observed everywhere, the conditional distribution of $Y(\mathbf{s})$ is given by:

$$E[Y(\mathbf{s}) \mid X(\mathbf{s}')\text{forall}\mathbf{s}'] = \int \exp(-i\boldsymbol{\omega}^T\mathbf{s})\, \alpha(\boldsymbol{\omega}) H_2(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

The above describes a scenario where the usefulness of the proxy to predict $Y(s)$ differs across spatial scales as captured by parameter $a(\boldsymbol{\omega})$. To estimate $a(\boldsymbol{\omega})$, Reich et al. (2014) parameterized it using basis expansion where $a(\boldsymbol{\omega}) = \sum_{l=1}^{L} A_l(\boldsymbol{\omega})\theta_l$. The standard downscaler in Eq. (11) now takes the form

$$Y(\mathbf{s}) = \alpha_0(\mathbf{s}) + \sum_{l=1}^{L} \theta_l \tilde{X}_l(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \sim N(0, \sigma_e^2) \quad (14)$$

where

$$\tilde{X}_l(\mathbf{s}) = \int A_l(\boldsymbol{\omega}) \exp(-i\boldsymbol{\omega}^T\mathbf{s}) H_2(\boldsymbol{\omega})\, d\boldsymbol{\omega}\,.$$

Since the proxy data are observed completely over a grid, $\tilde{X}_l(\mathbf{s})$ can be constructed efficiently using fast Fourier transform.

The spectral downscaler is similar to using a smoothed proxy as predictor because the decomposed proxy signal at each frequency is driven by more than one individual grid cell linked to the monitor. By considering the entire range of frequencies, the spectral downscaler also provides unique insights into the utility of the proxy at different scales. When applied to CMAQ ozone simulations, the spectral downscaler showed that CMAQ outputs at 12 km resolution have low correlation for features with a period less than 24 km. This

may highlight the deterministic model's limited resolution at this scale due to the coarse meteorological inputs. However, the associations between CMAQ outputs and measurements increase with increasing period, likely because ozone concentration tends to exhibit strong regional trends.

Rank and Quantile-Based Calibration

All the models we have presented focus on modeling the mean of the pollutant fields. However, environmental exposures can exhibit extreme tails, and these extreme values are often more detrimental to health (e.g., extreme heat). In the USA, the air quality standards for ozone pollution use the annual fourth-highest daily 8-h maximum concentration as the metric to determine non-attainment status. Berrocal et al. (2014) consider a downscaler that models the annual largest kth order statistic based on the generalized extreme value distribution which contains three parameters: location, scale, and shape. The same calibration approach in Eq. (11) is then applied to the location parameter.

Zhou et al. (2011) propose an alternative approach that aims to characterize the entire distribution of the pollutant field at each location. This is accomplished by estimating a one-to-one mapping between the quantile functions of the measurements and the proxy data using monotonic splines. This approach also falls under the framework of non-parameteric density estimation. Unlike Berrocal et al. (2014), to address the tails of the distribution, Zhou et al. (2011) assume the 10% tail of the pollutant distribution at each extreme follow a generalized Pareto distribution, and the central 80% of the distribution is determined flexibly by the splines.

Concluding Remarks

Current modeling approaches to integrate different data sources can be classified into two broad paradigms: machine learning (e.g., random forest and neural network) and advanced geostatistical modeling (e.g., Bayesian hierarchical model). This chapter focuses on geostatistical approaches that aim to optimally borrow information from nearby observations to perform interpolation; this model-based approach can also provide more interpretable parameters and uncertainty quantification in a principled way. However, machine learning methods offer several advantages including the ability to handle a number of highly correlated predictors and the ability to construct complex predictive algorithms that are non-additive and nonlinear. There have been very limited cross-paradigm comparisons (Adam-Poupart et al. 2014), likely due to the analytic effort and expertise required to carry out the different approaches.

Improved exposure assessment methods will continue to be valuable for epidemiological research and health impact studies. The prospect of combining different sources of data to assess environmental pollution is well recognized. As data products of environmental exposures become more readily available, researchers face the challenge of how to utilize them for epidemiological research.

References

Adam-Poupart, A., A. Brand, M. Fournier, M. Jerrett, and A. Smargiassi. 2014. Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy–LUR approaches. *Environmental Health Perspectives* 122(9): 970–976.

Banerjee, S., A.E. Gelfand, A.O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(4): 825–848.

Berrocal, V.J., A.E. Gelfand, and D.M. Holland. (2010a). A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics* 4(4): 1942.

Berrocal, V.J., A.E. Gelfand, and D.M. Holland. (2010b). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* 15(2): 176–197.

Berrocal, V.J., A.E. Gelfand, and D.M. Holland. (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics* 68(3): 837–848.

Berrocal, V.J., A.E. Gelfand, and D.M. Holland. (2014). Assessing exceedance of ozone standards: a space-time downscaler for fourth highest ozone concentrations. *Environmetrics* 25(4): 279–291.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2): 192–225.

Chang, H.H., X. Hu, and Y. Liu. 2014. Calibrating MODIS aerosol optical depth for predicting daily pm 2.5 concentrations via statistical downscaling. *Journal of Exposure Science and Environmental Epidemiology* 24(4): 398.

Choi, J., B.J. Reich, M. Fuentes, and J.M. Davis. 2009. Multivariate spatial-temporal modeling and prediction of speciated fine particles. *Journal of Statistical Theory and Practice* 3(2): 407–418.

Cressie, N., and G. Johannesson. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(1): 209–226.

Crooks, J., and V. Isakov. 2013. A wavelet-based approach to blending observations with deterministic computer models to resolve the intraurban air pollution field. *Journal of the Air & Waste Management Association* 63(12): 1369–1385.

Crooks, J.L., and H. Özkaynak. 2014. Simultaneous statistical bias correction of multiple pm2. 5 species from a regional photochemical grid model. *Atmospheric Environment* 95: 126–141.

Fuentes, M., and A.E. Raftery. 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* 61(1): 36–45.

Gelfand, A.E., A.M. Schmidt, S. Banerjee, and C. Sirmans. 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test* 13(2): 263–312.

Gilani, O., V.J. Berrocal, and S.A. Batterman. 2016. Non-stationary spatio-temporal modeling of traffic-related pollutants in near-road environments. *Spatial and Spatio-Temporal Epidemiology* 18: 24–37.

Gneiting, T., M.G. Genton, and P. Guttorp. 2006. Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs on Statistics and Applied Probability*, vol. 107, 151.

Grantham, N.S., B.J. Reich, Y. Liu, and H.H. Chang. 2018. Spatial regression with an informatively missing covariate: Application to mapping fine particulate matter. *Environmetrics* 29(4): e2499.

Hoek, G., R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* 42(33): 7561–7578.

Kammann, E., and M.P. Wand. 2003. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52(1): 1–18.

Levy, M.C., P.A. Collender, E.J. Carlton, H.H. Chang, M.J. Strickland, J.N. Eisenberg, and J.V. Remais. 2019. Spatiotemporal error in rainfall data: consequences for epidemiologic analysis of waterborne diseases. *American Journal of Epidemiology* 188(5): 950–959.

Lindström, J., A.A. Szpiro, P.D. Sampson, A.P. Oron, M. Richards, T.V. Larson, and L. Sheppard. 2014. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics* 21(3): 411–433.

Liu, Z., N.D. Le, and J.V. Zidek. 2011. An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics* 22(3): 340–353.

McMillan, N.J., D.M. Holland, M. Morara, and J. Feng. 2010. Combining numerical model output and particulate data using Bayesian space–time modeling. *Environmetrics* 21(1): 48–65.

Murray, N.L., H.A. Holmes, Y. Liu, and H.H. Chang. 2019. A Bayesian ensemble approach to combine pm2. 5 estimates from statistical models using satellite imagery and numerical model simulation. *Environmental Research*: 178: 108601.

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics* 24(2): 579–599.

Paciorek, C. J. (2012). Combining spatial information sources while accounting for systematic errors in proxies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61(3): 429–451.

Pirani, M., J. Gulliver, G.W. Fuller, and M. Blangiardo. 2014. Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science and Environmental Epidemiology* 24(3): 319.

Reich, B.J., H.H. Chang, and K.M. Foley. 2014. A spectral method for spatial downscaling. *Biometrics* 70(4): 932–942.

Ryan, P.H., and G.K. LeMasters. 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicology* 19(sup1): 127–133.

Sahu, S.K., A.E. Gelfand, and D.M. Holland. 2010. Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59(1): 77–103.

Shaddick, G., M.L. Thomas, A. Green, M. Brauer, A. van Donkelaar, R. Burnett, H.H. Chang, A. Cohen, R. Van Dingenen, C. Dora, et al. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(1): 231–253.

Wood, S.N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(1): 95–114.

Zhou, J., M. Fuentes, and J. Davis. 2011. Calibration of numerical model output using nonparametric spatial density functions. *Journal of Agricultural, Biological, and Environmental Statistics* 16(4): 531–553.

Zidek, J.V., N.D. Le, and Z. Liu. 2012. Combining data and simulated data for space–time fields: application to ozone. *Environmental and Ecological Statistics* 19(1): 37–56.

Spatial Epidemiology and Public Health

Shikhar Shrestha and Thomas J. Stopka

Introduction: Geographic and Spatial Health

The World Health Organization (WHO) defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO 1995). Health and the context where such a "state" exists are critical in understanding the pathways that lead there. Health is a product of one's body and its complex interaction with the surrounding environment. The surrounding environment could be as local as our workstation, the physical structures in our immediate surroundings, and the people that surround us, or as broad as the geographic, economic, and socio-political environment. Environments can determine the food we consume, the air we breathe in, the light that shines on us, and the sounds that we hear. As society has progressed, humans have gained the privilege and ability to modify our surroundings for our well-being. We control a broad range of decisions that contribute to our health, yet we are unlikely to be able to extricate ourselves from the environmental and social contexts tied to the geography within which we exist. For example, we may try to lead ideal lives with proper nutrition, exercise, and healthy habits, but there is little we can do about air pollution, population density, and traffic-related stress in the city or town within which we live. We may go further and remove ourselves from the polluted, high-density, stressful city, and move to another, but the impacts of those previous health stresses will have already taken their toll on our lives. Furthermore, our lives and state of health will still be tied to the social norms, governmental policies, and environmental factors, in one form or the other, in our new surroundings. Thus, the "environment," taken broadly to mean the natural, social, and risk environments that surround us, has a significant role in determining our individual health and broadly determining the trajectory of the public's health.

Epidemiology is the study of health and diseases across populations, whether it be neighborhoods, cities, towns, counties, states, nations, or the globe (Celentano and Mehta 2018). In public health, we consider the health and well-being of the community, as opposed to looking at the health of an individual patient. In fact, Schools of Public Health view public health as an opportunity to save lives – millions at a time (Thomas 2016). Spatial epidemiology reminds us that place matters when it comes to health (Cummins et al. 2007; Macintyre et al. 2002; Cubbin et al. 2008; Newburger et al. 2011). Where we are born and where we grow up, go to school, hang out, work, and recreate over the life course have an impact on our health status. Throughout our lives, we are exposed to a wide range of health-promoting (e.g., fresh air, healthy food, and exercise outlets) and health-inhibiting factors (e.g., pollution, unhealthy food, disease vectors, stress, carcinogens), and a majority of these factors are geographically based. When epidemiologists study risk factors for a certain disease, they need to account not only for the very proximal causes of disease, say exposure to *Mycobacterium tuberculosis*, the pathogen that causes tuberculosis (TB), but also for a broader context of how an individual got exposed to the bacteria (e.g., crowded living spaces) or what biological factor predisposed one individual to get the infection (e.g., a compromised immune system) while others were able to avoid infection. A major

I've always been fascinated by maps and cartography. A map tells you where you've been, where you are, and where you're going – in a sense it's three tenses in one.– Peter Greenaway

S. Shrestha
Department of Public Health and Community Medicine, Tufts University School of Medicine, Boston, MA, USA
e-mail: Shikhar.Shrestha@tufts.edu

T. J. Stopka (✉)
Department of Public Health and Community Medicine, Clinical and Translational Science Institute, Tufts University School of Medicine, Boston, MA, USA
e-mail: Thomas.Stopka@tufts.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_4

predisposing factor for contracting TB is the built and natural environment surrounding the individual. Therefore, it is critical for a researcher to incorporate location or environment as a part of the overall study of the disease.

Geographic Information Systems (GIS) for Public Health

Recognizing the important role of geography in health, maps were used to understand where diseases occurred as early as in the 1700s. These efforts in "Medical Geography" strove to help improve our understanding of the distribution of disease within the local context of environmental conditions. Some of the first known works in the field of medical geography came from Leonhard Ludwig Finke in the late 1700s, when he complied information on the location of human diseases (Barrett 2000). While there has been a long debate on who compiled the first world map of diseases (Barrett 2000; Barkhuus 1945; Schnurrer 1827; Light 1944), these efforts ushered in a new age of research into the field of medical geography and spatial health. Perhaps one of the better known disease maps is attributed to the "Father of Modern Epidemiology," Dr. John Snow, who mapped caskets in London in the midst of a cholera epidemic (Fig. 1) (Cameron and Jones 1983). Snow understood that there was a geographic pattern in cholera cases and in using "shoe leather epidemiology"(Koo and Thacker 2010), moving about local neighborhoods and talking to residents and family members who had lost loved ones to cholera, he was ultimately able to pinpoint the geographic location of the implicated disease source, the Broad Street pump, which was downstream from sewage outflows that ultimately found their way into the water supply that was pumped through the Broad Street pump and into local community members' water receptacles (Snow 1855).

In a similar medical geographical effort, Henry L. Bowditch developed a statewide map for Massachusetts that highlighted soil types across municipalities in order to assess potential spatial correlations between soil types and "consumption," or what we know better today as tuberculosis (Bowditch 1862). Importantly, through the work of Snow and Bowditch, as well as others, the medical and public health fields began to recognize not only the importance of mapping the diseases but also the importance of mapping the key exposures that were thought to be associated with disease outcomes in local populations. Recognizing the integral role of geography and spatial associations between exposures and outcomes, the field of spatial epidemiology evolved to integrate spatial attributes into epidemiologic analysis (Elliott and Wartenberg 2004). These fields, and related studies, recognized the importance of (1) the *location of diseases or medical conditions, (2) the relationship between disease outcomes and the environment in local populations, (3) the importance of access to resources to help manage these diseases, and (4) the key role of policies to mitigate a condition in the right place and at the right time (Elliott and Wartenberg* 2004).

The value of mapping diseases, as well as their relationship with underlying factors and outcomes, gained prominence in the last few decades as new tools were developed that enabled researchers to incorporate spatially-oriented data into their studies. Geographic information systems (GIS) formally came into being in the 1960s when Roger Tomlinson used the term in his paper "A Geographic Information System for Regional Planning." GIS came to be known as a system to create, store, analyze, and present geographical data (Clarke 1995). GIS evolved as a tool to visualize and map spatially-oriented health data (as well as other types of data) by facilitating the portrayal of health event locations, counts, rates, densities, and clusters across geographic and population landscapes. GIS began to facilitate an enhanced understanding of the spatial distribution of health phenomena within a specific geographic area of interest, allowing us to study the location of key features and health promotion resources in a local community and to document the geolocation of important events tied to public health outcomes. It also helped to facilitate the mapping of change over time across geographies. In the last three decades, GIS has been used for a wide range of public health and epidemiological initiatives (Moore and Carpenter 1999; Kohli et al. 1995; Kohli et al. 2000) to portray local realities tied to health and disease across geographic space and time. Whether focused on infectious diseases such as cholera or tuberculosis (Snow 1855; Vindenes et al. 2018), diseases correlated with environmental pollutants, such as respiratory illness (Sakai et al. 2004), diseases of addiction and related comorbidities (Stopka et al. 2017a, 2019a, b; Stahler et al. 2013; Brownstein et al. 2010; Wangia and Shireman 2013), or non-communicable diseases such as cancer (Openshaw et al. 1988; Kulldorff et al. 1997), GIS use in research has become commonplace.

Let us further consider the application of GIS to the study of infectious diseases. The viability of the vectors responsible for disease transmission (e.g., ticks, mosquitoes, and contaminated syringes), the virulence and survival of the pathogens tied to the diseases, and the spatiotemporal components of infectiousness (e.g., proximity between the person living with the infection and the person(s) susceptible to the infection) can all affect disease transmission, and all can vary in geographic space. Effective measures to combat infectious disease rely on methods that can locate these infections and their many related characteristics (some noted above) in space and time. GIS could be used to identify the location where these infectious diseases were transmitted along with data on the factors that fostered the transmission. When the spatial information is fed into a data processor

Fig. 1 Viewing John Snow's cholera map with modern GIS tools. Panel A represents John Snow's cholera map. Panel B is a modern rendition of the original map where the red circles represent the location of cholera deaths, with larger red circles (proportional symbols) representing larger clusters of cholera deaths, and the blue dots representing the water pumps in the area. Panel C is a density map, or a "heat map," created using the original data. Panel D highlights the location of cholera deaths as red points, juxtaposed with the blue points for water pumps, and a thematic multi-colored map layer that consists of Thiessen polygons. Thiessen polygons divide the area such that the space that the polygon encloses is closer to the point of origin of that polygon (in this case the water pumps) than to any other points. This panel shows that most of the deaths were in close proximity to one of the water pumps (i.e., The Broad Street water pump)

and linked with other non-spatial data, we could gain a greater understanding of "where" the disease is occurring, which brings us closer to explaining "why" the disease is occurring in local populations. Our ability to identify and characterize the distribution of such diseases and their causal agents is vital to successful evaluation of the risk of disease outbreaks and development of interventions to prevent or manage them. As evidenced by the published literature, GIS has been instrumental in carrying out such analyses. A systematic review published in 2015 listed 80 peer reviewed articles that studied infectious diseases using GIS ranging from respiratory infections (such as SARS and influenza)

to intestinal infections (e.g., cholera and salmonellosis) to sexually transmitted infections (Smith et al. 2015). The authors highlighted the importance of the spatial context tied to the infectious disease outbreaks, including connections to the origins of key resources and risk factors (e.g., water, food, vectors) in local risk landscapes. While most of the studies presented some form of visualization for disease outbreaks, multiple studies used spatial exploration, cluster analysis, and advanced spatial modeling to predict the distribution of diseases.

Moving away from the more traditional uses of GIS in epidemiologic studies (e.g., risk mapping, resource mapping), this tool is also being used increasingly to conduct more complex analyses ranging from assessment of access to public health and healthcare services to examining spatial distribution of the social determinants of health and policies that govern healthcare utilization. As greater focus is put on health disparities, health officials and local policymakers are increasingly utilizing GIS to understand where disparities in access and utilization exist (Evans et al. 1994). Early efforts were focused on studying variations in local delivery of health services and medical care by David Wennberg (Wennberg 1973). Based on years of study following the initial discovery of variations in healthcare delivery and outcomes, Wennberg famously mentioned that "*Geography is destiny*" for medical care, demonstrating growing disparities in health service access across local communities (Wennberg 1998). This recognition of the role of geography in defining and influencing community health forced public health officials and healthcare providers to consider new and enhanced ways of studying the role of geography and spatiotemporal relationships in health outcomes. Nowadays, studies employ complex spatiotemporal and geostatistical analyses to better understand the geographic dynamics of community health. Recent advances with data collection and the advent of "Big Data" have made it easier for more epidemiologists to incorporate spatial data into their research. In fact, more than 80% of health data have a spatial component, whether tied to a specific address, ZIP code, or community of residence. While, in public health, we believe that geography should not be destiny (Nunn et al. 2014) and we strive to minimize disparities in healthcare across multiple dimensions, the reality is that many disease outcomes are strongly tied to our local surroundings.

GIS and spatial analyses that enhance our ability to incorporate these factors and determinants at the community level are congruent with the shift in research paradigm from a proximal "individual health" model to a distal "population health" model. Population health refers to the "*health of a population as measured by health status indicators and as influenced by social, economic, and physical environments, personal health practices, individual capacity and coping skills, human biology, early childhood development, and health services"* (Dunn and Hayes 1999). The methodological implementation of population health theory necessitates five broad steps: (a) identification of the population, (b) assessment of the health of the population along with the environment surrounding it, (c) descriptive evaluation of current services (utilization and distribution), (d) evaluation of gaps and overlapping services, and (e) evaluation of outcomes (Barnard and Hu 2005). The identification of patterns of health outcomes in large groups of people supported by the evidence of an underlying environmental, socioeconomic, and geopolitical pathology can help support development of policies to promote the well-being of the population. GIS is well placed to study these factors at multiple levels (individual, community, and geopolitical) and allows us to better understand and address the issues that govern the health of the masses.

As we move through the contents of this chapter, we aim to give readers a brief glimpse of how GIS and spatial epidemiology can be applied in public health research. GIS tools and applications can span a wide range of topics, which can extend well beyond the confines of this publication. But through simple examples, and building up on some of the topics covered, we hope that readers can gain a solid understanding of GIS as it applies to public health research. Furthermore, we present details on a number of ways in which public health leaders, researchers, policymakers, and community members can use GIS and spatial analyses to better understand public health. We begin with a brief introduction to the development and use of descriptive maps and risk maps that are commonly used in public health research to help generate hypotheses. Next, we describe GIS and spatial epidemiologic tools that are commonly used to create new measures, based on geographical information and calculations of spatial relationships, which can be mapped on their own or can be used in statistical models as covariates and outcomes. Finally, we discuss spatial epidemiological and geostatistical analyses and modeling that can be used to test hypotheses regarding the distribution of events, assess spatial statistical associations (e.g., Is proximity to a health clinic associated with better health outcomes?), and identify and characterize spatiotemporal clustering (e.g., hotspots). We then provide a case study, along with an overview of studies that elucidate the range of GIS and spatial analytical tools and approaches that are currently available, to help us better understand one of the most complex public health challenges of our times (i.e., the opioid crisis). We close with considerations for GIS and spatiotemporal analytical applications for public health in the years ahead, highlighting promising opportunities to bolster our understanding of and responses to geographically influenced destinies across local, state, and national communities.

Overarching Domains for GIS and Spatial Analyses

In this section, we present three broad general uses of GIS and spatial analyses in public health: (1) risk mapping/descriptive mapping, (2) calculation of variables in a GIS, and (3) spatial epidemiological and geostatistical analyses. While these three categories may not encompass all applications of GIS and spatial analyses tied to health, we believe that they incorporate a broad spectrum of uses and applications that are relevant for a wide range of health and public health issues.

An In-Depth Look at Methods – Descriptive Mapping

Descriptive mapping, which is done widely today on static and online, interactive maps, allows us to obtain a general understanding of the lay of the land as it pertains to a specific phenomenon of interest related to health, whether we focus on disease outcomes, risk exposures, or health and social service resources that are available in a local community or region. Taken together, the descriptive elements of these disease outcome and resource maps or map layers can help local public health officials and community members to understand the risk environment as it relates to public health. These descriptive maps typically depict "push pin" and thematic polygon maps that are hypothesis generating, allowing us to observe potential spatial associations that need to be tested through more complex approaches.

Choropleth Maps A choropleth ("Khora" = region and "plethos" = multitude) map is a thematic map in which geographic regions (i.e., polygons) are displayed in relation to a value. They are useful when visualizing a variable and how it changes across defined regions or geopolitical areas (e.g., counties, towns, ZIP codes); data are typically aggregated over these predefined areal units. They are best used for standardized data (e.g., rates), discrete variables (e.g., counts), and measures that are evenly distributed. In public health, mapping counts for disease on the polygon level within thematic maps can help policymakers to garner an initial understanding of the burden of disease across geographic regions. Understanding such burdens can provide an initial understanding of the extent of a problem and potential public health needs (e.g., enhanced vaccination coverage, disease treatment outlets, and policy-based interventions to limit access to vaping equipment). Meanwhile, mapping rates, which take population denominators into consideration (i.e., "normalizing by population"), allow for a better understanding of disparities in the spatial distribution of health effects and disease outcomes across diverse communities. While choropleth maps provide a good visual overview of differences over space, they have a few disadvantages. Choropleths are not appropriate to show total values as they are likely to be correlated with total population. In choropleth maps, values are presented over bounded regions, giving a false impression that values abruptly change at the border of man-made political boundaries. Additionally, it could be difficult to separate different color shades in a choropleth map, especially when there are multiple bounded regions in proximity within small areas.

More extreme colors (darker or lighter) in choropleth maps typically indicate increased or decreased rates for the outcome highlighted in the choropleth maps, respectively. In Fig. 2, for instance, higher death rates associated with the hepatitis C virus (HCV) are depicted by darker shades of orange. Note that to address smaller areas, an inset map has been used, which is an effective way of focusing on small areas that have high attribute variability.

Push-Pin or Point-Vector Maps Push-pin maps are descriptive maps that rely on geocoded data at the address level. By "geocoding" the data, we use GIS tools to obtain the "X" (longitude) and "Y" (latitude) measures for a specific address in a local community. Push-pins for specific addresses, based on the XY data, may represent public health resources (e.g., health centers, disease testing sites, pharmacies), sources of risk (e.g., smoke stacks, liquor stores, contaminated water sources), or the homes of individuals who are at risk for or living with a specific disease. It is important to note that varying sensitivities and guidelines come into play with the portrayal of push-pin data. It is not typically problematic, for instance, to map the location of hospitals in a local community, but it is not typically possible to map point-vector data for people living with sensitive and stigmatized diseases (e.g., HIV), as it would compromise the confidentiality of local community members. To map such sensitive data, "geomasking" approaches are needed to "jitter" the location of disease cases, aggregate disease outcomes at the polygon (e.g., county) level, or calculate densities that can be depicted in heat maps that portray locations with higher and lower risk, without displaying precise addresses for people living with a specific disease.

Proportional Symbol Maps A proportional symbol map is another form of thematic map that builds from typical push-pin or polygon-level thematic maps. In proportional symbol maps, the values referring to a specific attribute in a geographical area (i.e., a polygon) or at a specified XY location are represented by a symbol (e.g., a dot on a map). The map follows a simple concept: "the larger the value, the more something exists at a location," i.e., change in the absolute value of the attribute is shown as change in the size

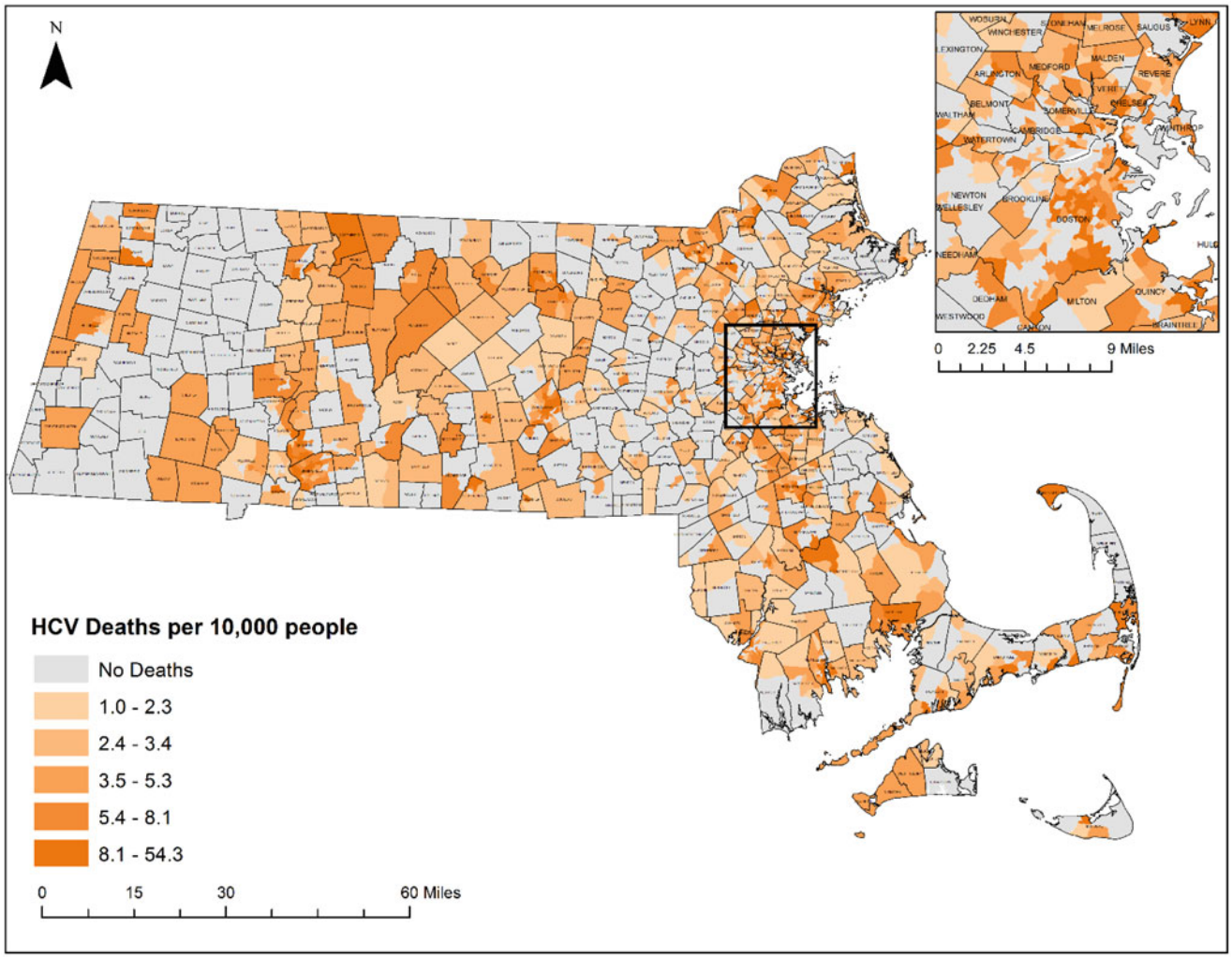


Fig. 2 A descriptive risk or thematic map depicting the rate of hepatitis C virus (HCV)-related deaths per 10,000 population in the state of Massachusetts (2002–2011) (Meyers et al. 2014)

(length, area, or volume) of the symbol, which is also termed "absolute scaling" or "apparent magnitude scaling." While proportional symbols are commonly used for both raw and standardized data, they are best used when there is substantial variation within a specific measure across geographic space. They are commonly used to obtain an overview of localized incidence rates as the larger size of a symbol in a polygon surrounded by smaller sized symbol (for low incidence rates or counts) is likely to stand out, thereby enabling the "map audience" to better assess variations in the phenomena of interest. Proportional symbol maps can also be useful when there is substantial overlap in point vector data in underlying layers, such that it can be difficult to observe how many cases are collocated in a small area. Further, they offer an additional advantage over choropleth maps, particularly when it comes to visualizing small areas in a large map, which helps to more clearly depict differing values at a glance. Regardless of the value associated with a polygon, there is a larger likelihood that information and variation can be missed or underestimated in a choropleth map as larger areas tend to attract more attention from the map audience, regardless of the color, whereas a larger symbol within a smaller polygon will be able to stand out more prominently. Additionally, flexibility regarding the type of data (one can use both raw and standardized data) and specification (attributes related to geographic points or areas) make proportional symbol maps very popular among cartographers and public health researchers. Proportional maps are however not without limitations. If there is a large array of values with small differences, the differences between symbols become indistinguishable. In other cases, the symbols for large values can obscure other symbols and the underlying map.

Figure 3 is an example of a proportional symbol map where the size of the circle represents the number of people diagnosed with diabetes in states across the United States.

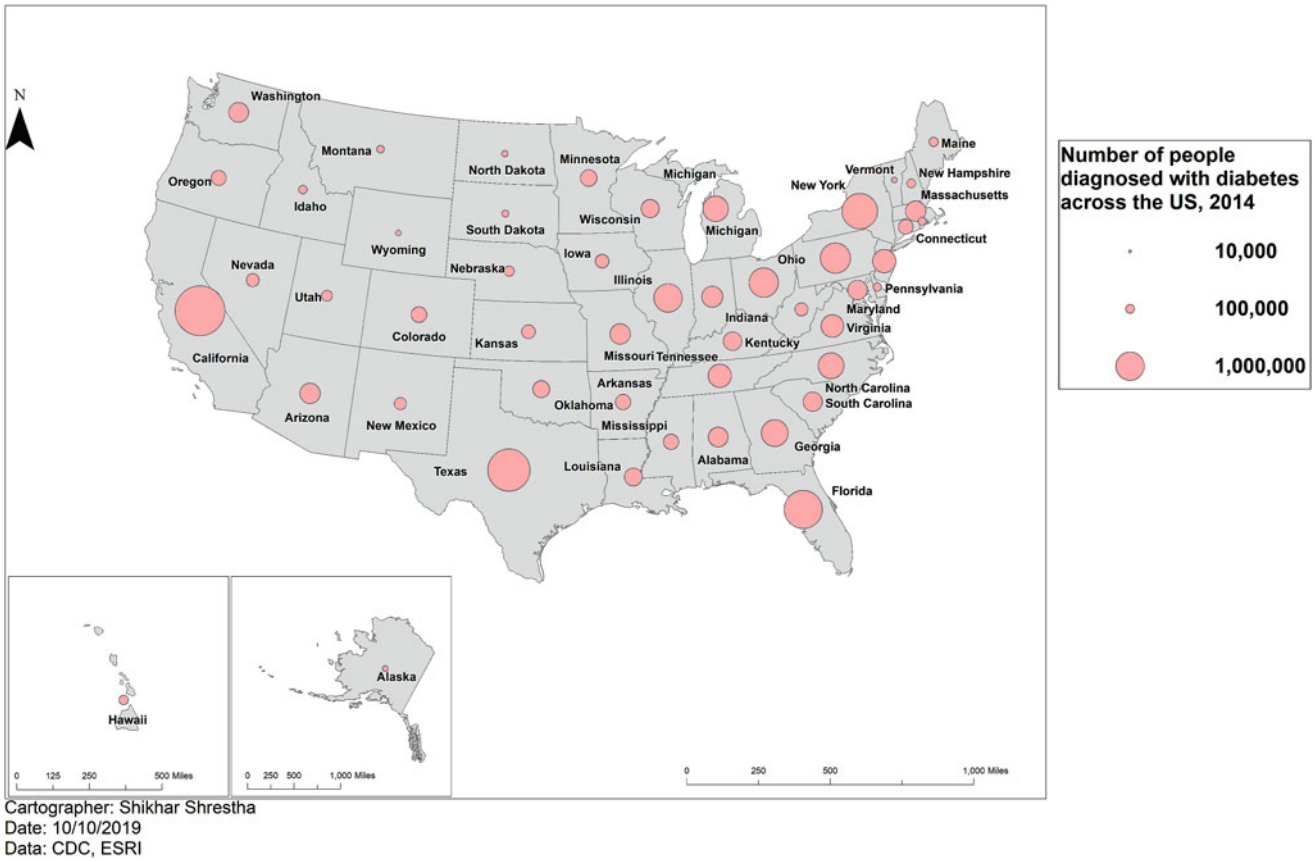


Fig. 3 Prevalence of diabetes in the United States (proportional symbol map), 2014

While this figure is effective in showing the difference on a state-by-state level, representation of some of the values may be hindered by larger symbols (notice the overlapping of symbols in the New England region). Further, the presentation of numbers, rather than rates, fails to normalize by population.

Graduated Symbol Maps A graduated symbol map is similar to a proportional symbol map. However, in this map, the values of the attribute are divided into discrete categories and then denoted by a symbol for which size is relative to the value or ranking of the category it represents (Fig. 4). This process is termed "range grading," which simplifies the matching of a symbol to its proportional value. Graduated symbol maps are often preferred over proportional symbol maps, especially in cases where the values of an attribute do not provide any additional information once they hit a certain value and discrete categories provide a more sensible representation of data compared to a continuous range of symbol sizes. This process, however, leads to loss of actual values of the observation. Therefore, in instances where actual values of the attributes are considered important (e.g., a map trying to show maximum or minimum values), proportional symbol maps should be used.

Dot Density Maps Dot density maps utilize dots or points to show the presence of a feature within a bounded location (Kimerling 2009). Instead of larger symbols meaning more, dot density maps portray more dots in locations where there are higher counts. They can make it very easy to visualize data, with higher densities of dots representing higher values. The distribution of dots shows spatial patterns and relative densities of occurrences. It is important to note, however, that the dots are placed randomly within the specified polygon within which they fall, so they do not represent exact addresses or XY data for disease cases. The individual dots within a specified area could represent single or multiple occurrences. We can have each dot represent a different number of occurrences (e.g., one dot = 5 heart attacks). Dot density map symbology is best used to show count data. One of the disadvantages of dot density maps is that it is difficult to extract quantities or values from them, and high dot densities may block important background features or boundaries on a map. This may be especially difficult

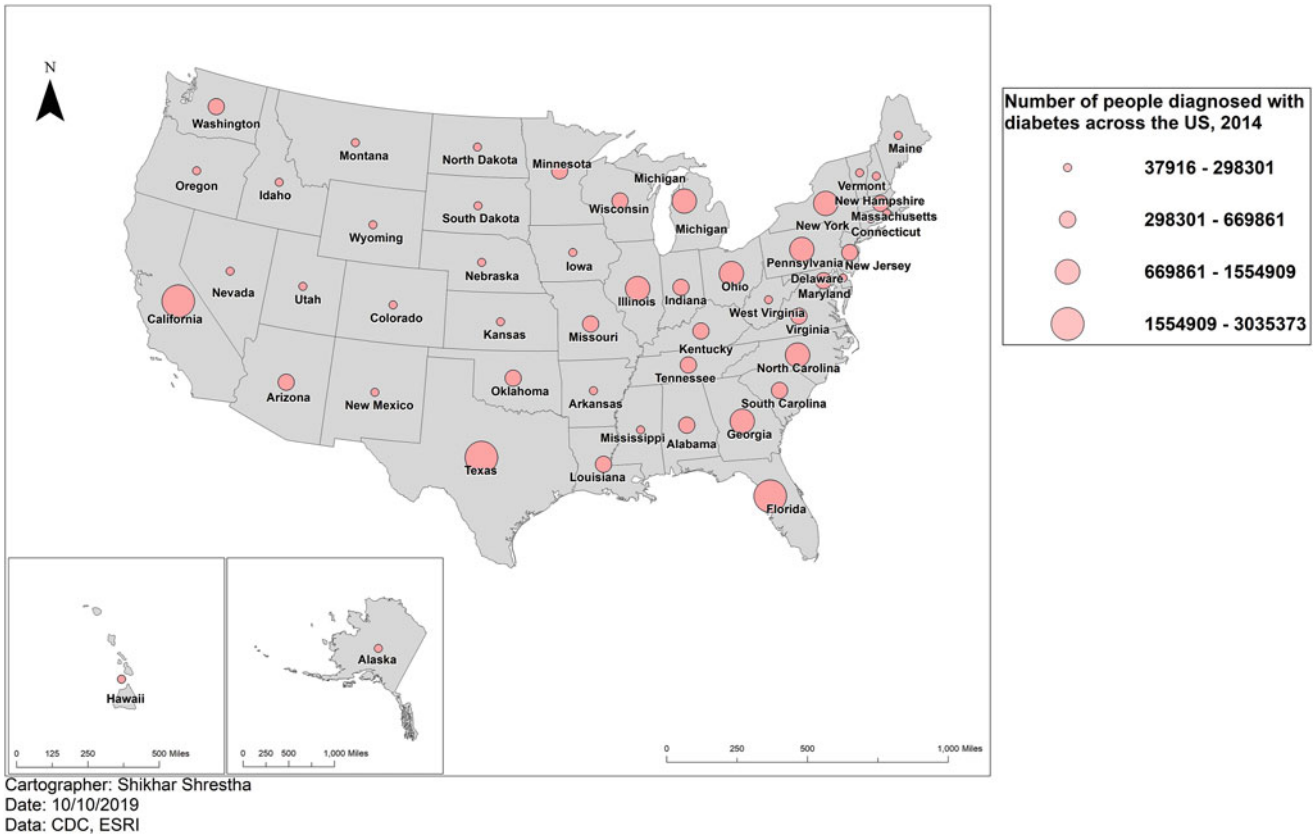


Fig. 4 Prevalence of diabetes in the United State (graduated symbol map), 2014

when the polygons are small in size and have high counts (Fig. 5).

An In-Depth Look at Methods – Calculation of Variables

Through the calculation of variables within a GIS, we can better understand spatial distributions and relationships tied to exposures and outcomes of interest. Examples of calculated variables or measures include distances from place of residence to health centers, drive-time or walk-time distances to public health resources, small area estimates for risk and disease counts and rates on the macro (e.g., state) or micro (e.g., census tract) levels, distance buffers around putative exposures or preventative and curative health centers to understand proximity to risk and access to services, respectively, and Kernel density estimates or "heat maps" that portray areas with higher and lower densities of our health or disease outcome of interest. It is important to note that the variables calculated within a GIS may subsequently be used in maps that are developed and displayed through a GIS, and they may also be exported from a GIS for use in statistical models and more complex analyses. It is not uncommon, for instance,for investigators to use distance calculations as a continuous or categorical measure in subsequent multivariable models to determine whether, for instance, the distance from one's place of residence to a disease prevention program is associated with disease outcomes (e.g., disease progression, cure, transmission), while controlling for other measures.

Proximity Analysis Following are some of the commonly asked GIS questions in relation to location of a specific feature: "What is nearby? What is farther away?" Additionally, questions could also be framed to focus on the population of people who live within 10 miles of a specific healthcare resource or source of risk (e.g., hospital, nuclear power plant). A proximity map can be an efficient tool to answer these types of questions. A proximity map can be used to calculate and depict the distance between features, and between features that are within a certain distance from a point, line or a polygon. There are multiple ways of creating proximity maps depending upon the context of study. In the following paragraphs, we describe buffering, Thiessen polygons (also known as Voronoi diagrams), and distance calculations as some of the methods to conduct proximity analysis.

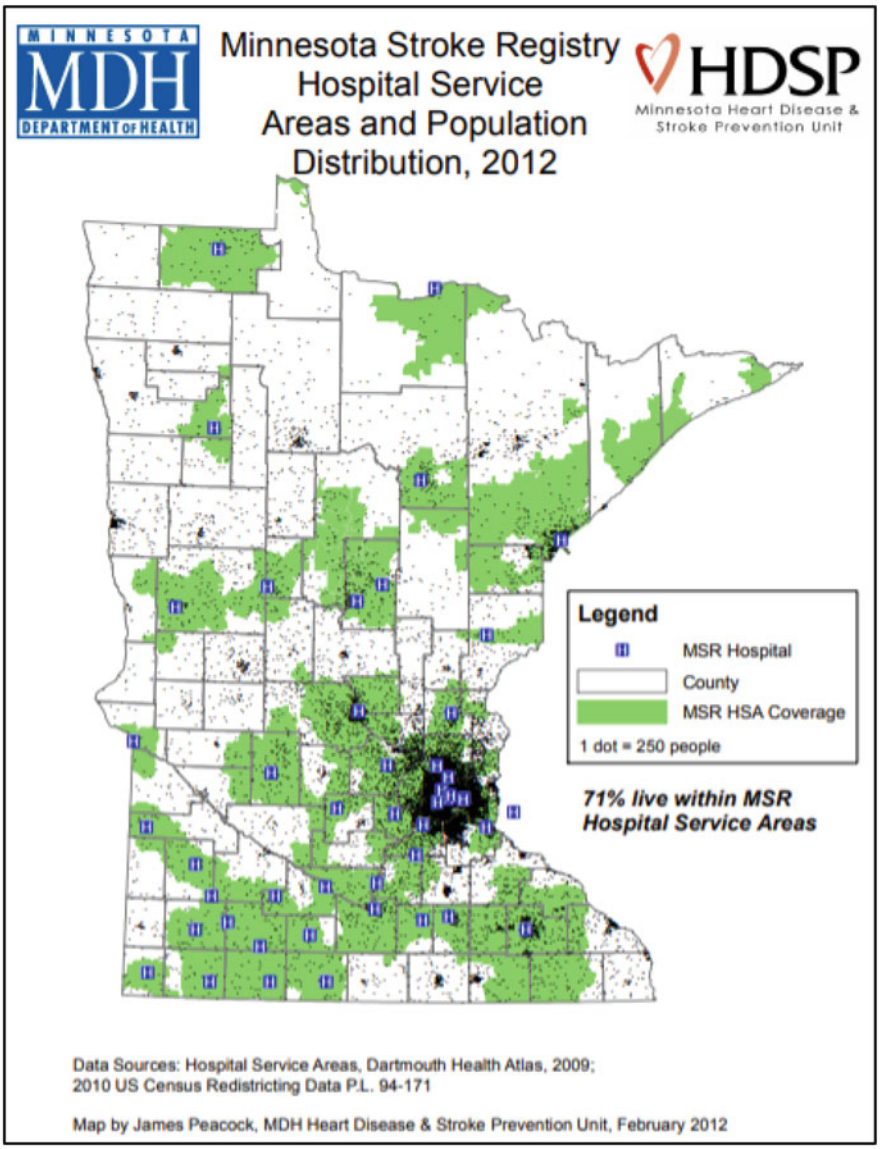


Fig. 5 Dot density map for the distribution of the population (black dots; 1 dot = 250 people) in the State of Minnesota. Dot densities for the population are juxtaposed with Minnesota stroke registry hospitals (represented by the blue H symbols), and the green areas, which represent local stroke health service coverage, or the Minnesota stroke registry hospital health service area (Peacock 2012)

Buffering A buffer is an area with a specified distance around a spatial object. The object may either be a point, line, or a polygon. Buffering is commonly used for proximity analysis to determine if an object is within a certain distance from a geographical object. For example, 1000-foot buffer rules are commonly used in planning to demarcate areas around specific locations (such as schools) where other businesses may not operate (e.g., liquor store, marijuana dispensary). Buffer areas are also commonly used to determine flood zones or contamination zones around specific locations (e.g., hazardous areas around a damaged nuclear reactor). Similarly, buffer maps in conjunction with other measures of proximity analysis have been used in the study of built and natural environments and their association with health outcomes. These analyses have, for instance, focused on the presence of built environment "assets," such as parks and sidewalks, and their association with outcomes such as physical activity, and obesity (Brown et al. 2014; Jago et al. 2006; Browning and Lee 2017; Thornton et al. 2011). Many factors need to be considered when pondering information that should be included in a buffer map. We need to consider whether straight-line Euclidian distances, "as the crow flies," or Manhattan distances, that take street networks into consideration, would be more relevant in developing buffers (i.e., ring buffers vs. walk-time or drive-time buffers). When considering risky substances that are air-borne, like smoke or poisonous gases from a factory, for instance, wind direction can drastically influence the direction and area where smoke and poisonous gases travel from the factory to local communities. In this example, street and road networks are inconsequential when it comes to considering exposure to the toxic air-borne source. Similarly, when considering distance from the epicenter of an earthquake, multi-ring buffers with straight line

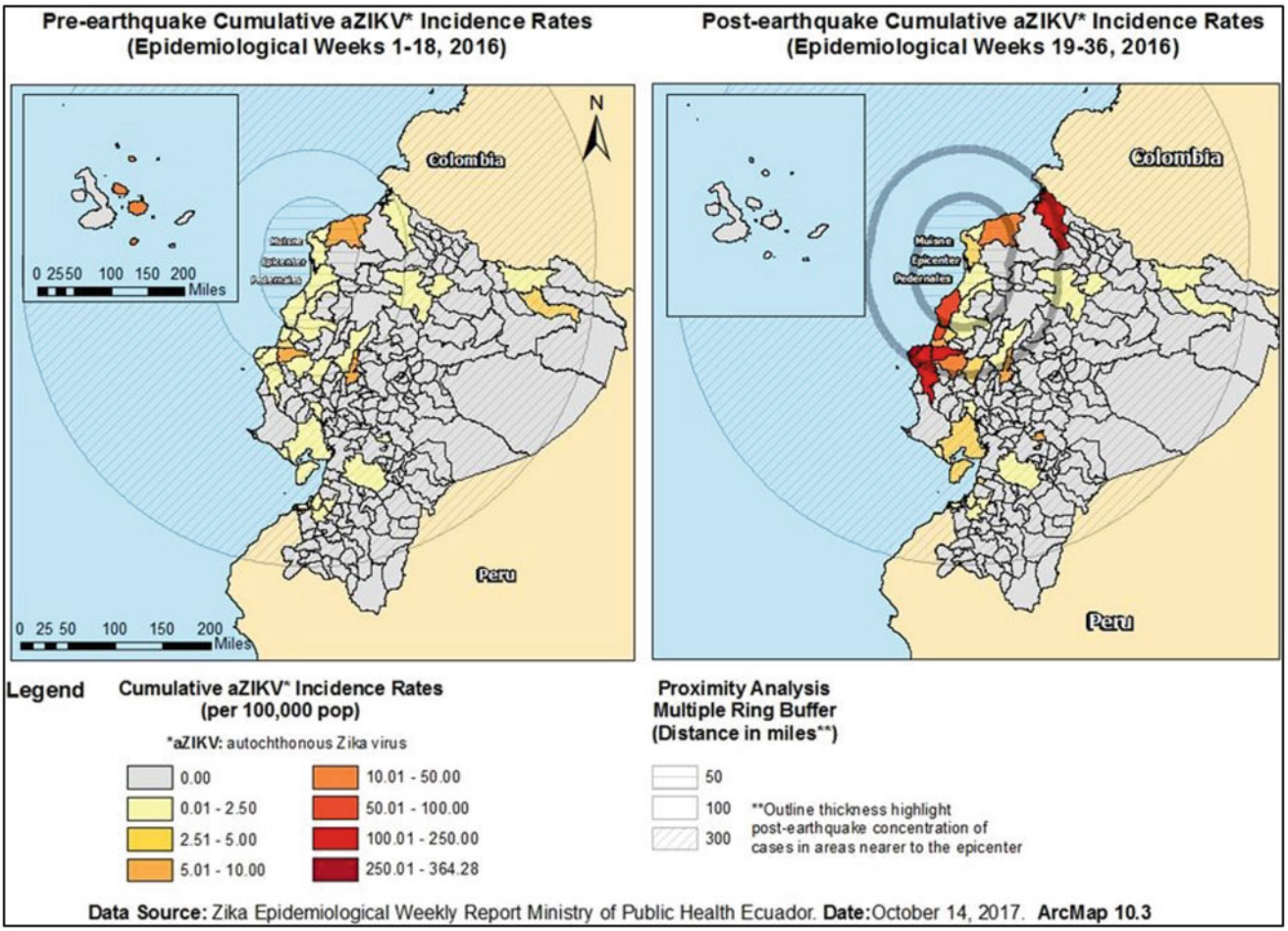


Fig. 6 Pre- and post-earthquake incidence of Zika Virus in Ecuador with buffer rings showing proximity to the epicenter of the quakes (Ortiz et al. 2017)

radii can allow us to define "regions of risk" denoted by concentric circles around the epicenter (Fig. 6). If we are interested in assessing access to disease prevention resources like syringe services programs, on the other hand, buffers that take street networks into consideration can help us to develop walk-time and drive-time buffers, providing a more precise understanding of local levels of access to such harm reduction services that can reduce syringe-mediated HIV transmission risk. It is, however, important to consider the type of buffer or proximity analysis utilized for a given study. For example, studies have indicated that a circular buffer may not be accurate in reflecting spatial features that influence walking particularly in locations where the natural or built environment restrict free movement around the point of origin (Oliver et al. 2007).

Thiessen Polygons Thiessen polygons are created from a given set of points in an area such that each polygon represents an area of influence around a point of interest (e.g., a source of health risk; for example, the contaminated Broad Street water pump from John Snow's Cholera map). In a Thiessen polygon, any point within the polygon is closer to the point of origin compared to other points in the map. We can refer back to the modern rendition of John Snow's Cholera maps where Thiessen polygons were created using local water pumps as the points of origin (see Fig. 1). Thiessen polygons are also used in geosciences and can be used in modeling catchment areas for shops, health centers, or public transport stations.

Distance Calculations A variety of tools can be employed within a GIS to calculate distances for a variety of applications. One can calculate distance from one location to another (e.g., one's home to a medical clinic) or distance from one's place of residence or workplace to a putative source of a dangerous exposure (e.g., place of work to the nearest chemical factory). We can also calculate a number of different types of distances: (1) Euclidean ("as the crow flies"), (2) Manhattan (which takes street networks into consideration), (3) walk-time distances, and (4) drive-time distances (which can take

traffic patterns into consideration). There are pros and cons to each of these different distance calculations, and some may be more appropriate in some situations than others. If you are interested, for instance, in the local proximity between a community member's home and the nearest green space, walk-time calculations, which take street networks and sidewalks into consideration, may be most appropriate. If, on the other hand, you are interested in the distance from the site of a nuclear meltdown and the homes in the local area, Euclidean distances may be most appropriate given that the straight-line distance that radiation travels is of utmost importance. We can use tools within a GIS (e.g., "near tool") to calculate these distances. Distance calculations can provide a rough measure of access to a specified location or health service. One of the shortcomings of these tools, however, is that they do not always take population density and other population dynamics into consideration. For accurate estimates of proximity or access to services, we need to take the distribution of underlying population and the geographical and man-made structures into account. Many tools and approaches have been developed to take these variables into account.

The Two-Step Floating Catchment Area (2SFCA) Approach The 2SFCA approach takes the distribution of underlying population (and several other parameters) into consideration when assessing access to health care services (as well as other services). Individuals are free to choose care, up to a certain extent, from whomever they wish. We must take into consideration that an individual can have multiple health services that they can access, and these options are not bounded by arbitrary boundaries. The complex nature of human movement within and across arbitrary boundary lines (e.g., census tracts or ZIP codes) to receive health care or use other services makes the assessment of "access" to a service a challenging task. The 2SFCA method was developed by Lou and Wang to overcome the shortcomings of previous measures of health care access, such as the physician-population ratio (PPR, calculated by dividing the number of physicians by the number of patients in a given area) (Guagliardo 2004; Wang and Luo 2005; Luo and Wang 2003; Luo 2004). PPR can be inaccurate in determining access as it does not acknowledge the possibility of crossing a boundary line for accessing treatment services. To overcome such limitation, the 2SFCA approach utilizes floating catchment areas, which can overlap each other enabling the estimation of "access" that closely models movement of individuals within and across arbitrary borders in real life. The 2SFCA, as its name specifies, is calculated in two steps. The first step is the calculation of PPR at each provider location (e.g., given a certain drive-time or distance from the provider and the approximate number of people who have access to the provider). The second step is the calculation of a spatial accessibility index, which is a summation of the PPR (e.g., the summation of the PPR of the providers that are located within the given drive-time or distance).

While the 2SFCA has several advantages over PPR, it does have certain limitations. First, the 2SFCA approach assumes equal access within a catchment area, i.e., there is no difference between a 10-minute drive-time and a 25-minute drive-time. Other limitations include its limited application in rural areas (catchment areas need to be expanded greatly to capture sparse population) and only inclusion of one measure (drive time) to determine the catchment area. In the last few years, several improvements have been made to the original 2SFCA approach to overcome these limitations. One of the improvements was the development of an enhanced two-step floating catchment area (E2SFCA) method, which integrates a distance decay function within the catchment area (Luo and Qi 2009) (Fig 7).

Heat Maps A heat map shows the intensity of an occurrence or an attribute in a dataset, hence they are also called "intensity maps." The concentration or density of the occurrence is represented by the "heat." A heat map utilizes color gradients to represent intensity. A major difference compared to a choropleth or thematic map is that a heat map does not use specific boundaries to group data. Usually, creation of a heat map requires use of point vectors, based on the precise longitude (X) and latitude (Y) measures, to create a continuous surface area known as a density surface, which maps the frequency of occurrences at a specific point. When creating a density surface, two parameters are usually specified, which facilitate calculations that culminate in the final surface, taking the form of a raster or pixelated image file. The first parameter that needs to be specified is the cell size of the raster output map, which determines the degree of detail of the density surface. Second, the bandwidth or search radius needs to be specified. This affects the area of restriction around the points or features. If too narrow a bandwidth is chosen, the density patterns are restricted to points that are very close to it. On the other hand, using a wide bandwidth can lead to overgeneralization of the "heat" in the density surface. Heat maps are commonly used to visualize locations of higher occurrence of events such as crime and traffic accidents, or newly reported disease cases. This information can be used to assess the proximity of areas with high density of events to certain geographic or manmade features (Fig. 8).

Small Area Estimates Small area estimates pertain to estimates that are calculated for local communities, and they are typically calculated for small area polygons tied to political boundaries (e.g., counties, municipalities, census tracts). They can, however, also be calculated for polygons that are created by the researcher (e.g., fishnet matrices, buffer

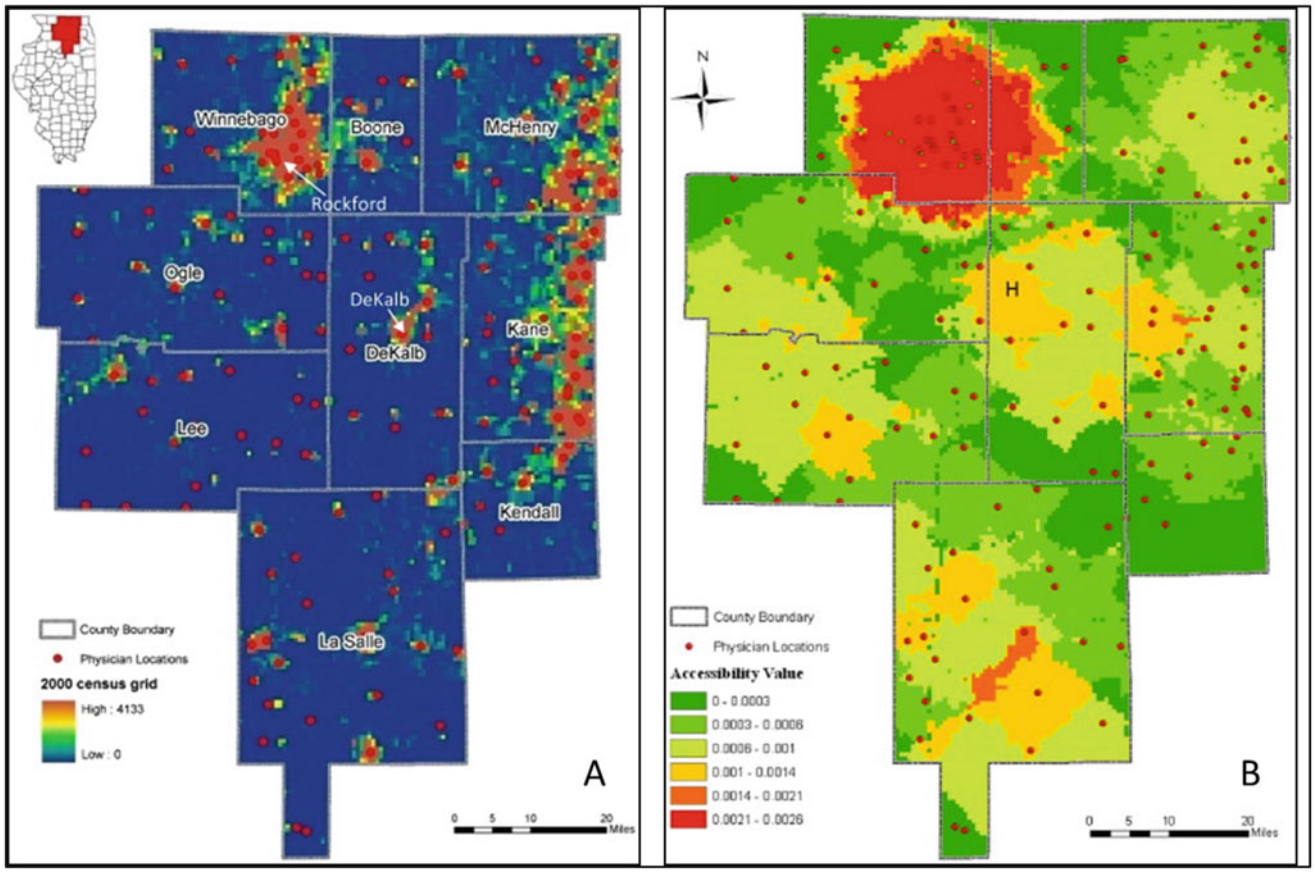


Fig. 7 Measuring physician access in rural and suburban Chicago. The first panel (**a**) shows the location of the physicians and the population density (based on 2000 Census data). The second panel (**b**) shows accessibility value generated using two-step floating catchment area (2SFCA) method. The areas colored in red (panel B) denote high accessibility and the areas colored in green have low physician accessibility. (Luo and Qi 2009)

spaces around locations of interest). Estimates can be as simple as calculating rates based on disease counts within the specified area (i.e., polygons) in the numerator and the population at risk in the denominator. Calculations can get more complex in terms of the measures we are able to calculate (e.g., incidence rates, densities per square mile, smoothed or spatially-lagged rates that take neighboring polygons into consideration) and the tools available to us in a GIS (e.g., areal interpolation, geographically weighted regressions).

A More In-Depth Look at Methods – Spatial and Geostatistical Approaches

Spatial epidemiology and geostatistical analyses represent more complex uses of GIS for public health. When using such analyses in GIS, we are typically testing hypotheses to determine whether, in fact, there is a statistically significant spatial pattern or clustering of an outcome of interest. Several tools are available to analyze the patterns of events of interest over a geographical area. We present a few examples here.

Average Nearest Neighbor analyses are used to assess whether the distance between events that are observed is similar to what would be expected if the events were randomly distributed in geographic space (e.g., studying patterns of dengue distribution in Kuala Lumpur (Aziz et al. 2012)). The Moran's *I* test statistic is used to detect spatial autocorrelation of an attribute, indicating agglomeration or dispersion of features with similar attributes. The Getis-Ord Gi* statistic (Getis and Ord 2010, 1992) is used to identify clustering of events (e.g., identification of healthcare hotspots in Taiwan (Tsai et al. 2009)). Bayesian spatiotemporal models are used to predict the distribution of attributes or events over time within a geographical area (e.g., forecasting life expectancy in England and Wales (Bennett et al. 2015)). These and many other tools and approaches exist within a wide range of GIS software and freeware programs, as well statistical analysis programs that increasingly include packages, tools, and commands for geostatistical analyses. Software and freeware such as QGIS, ArcGIS, GeoDA, SatScan, CrimeStat, GRASS, and gvSIG provide users of different programming

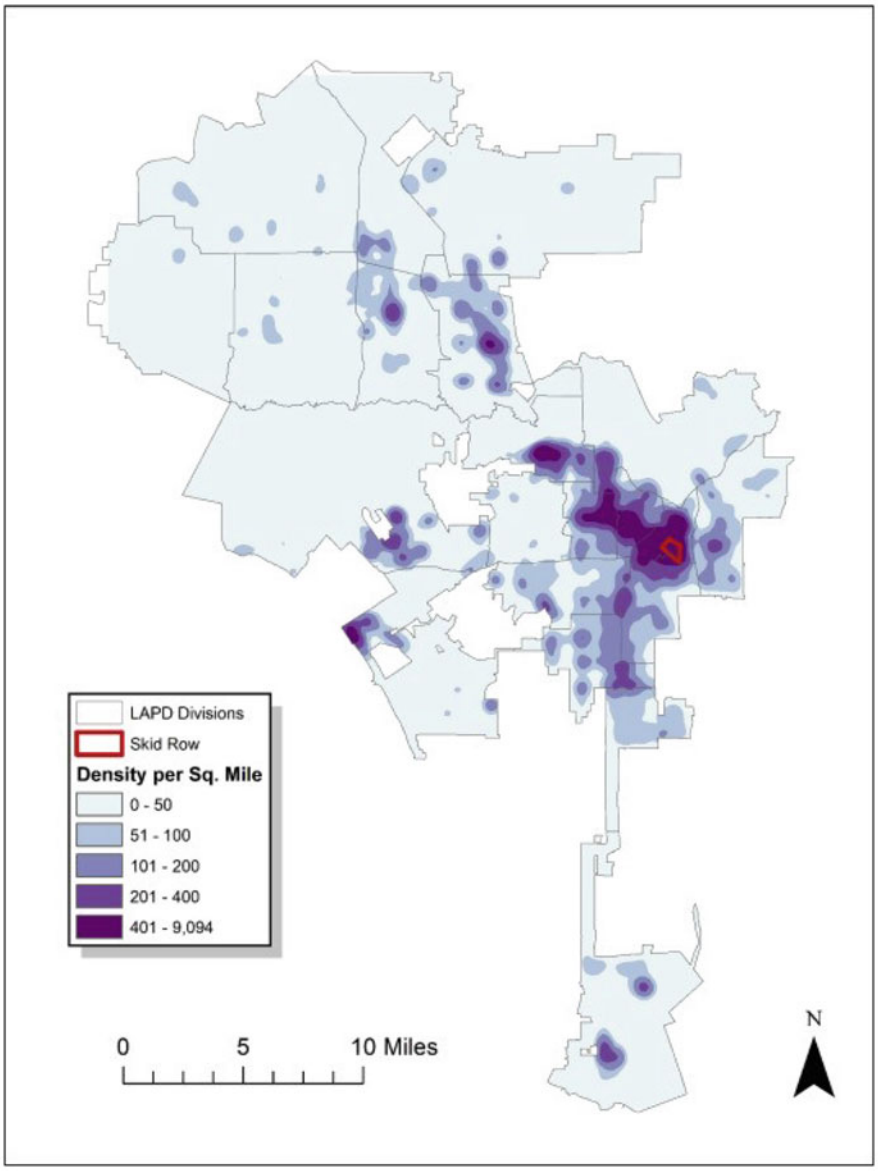


Fig. 8 Crime related to homelessness in Los Angeles, California. Map generated using Kernel density estimation (Yoo and Wheeler 2019)

capabilities and backgrounds with the ability to integrate spatially-oriented data into epidemiologic research.

Spatial Interpolation Spatial interpolation is a method used to predict values of a certain area or point based on other values in the study area. In a GIS, interpolation is used to predict values for cells in a raster format from limited sample points. Spatial interpolation is used to predict unknown values in locations where data points were not collected, due to time constraints, costs, and access considerations. The product is a continuous surface with modeled estimates across an entire study area, providing a "topographic" map layer of sorts where we can observe higher and lower estimated values of interest, even in locations where data points were not originally available. Based on the type and amount of data collected (also known as "control points"), different techniques can be applied to interpolate data. In general, spatial interpolation can be global or local, and deterministic (output is fully determined by the parameters supplied into a model) or stochastic (some random component is present in the model whereby the same set of parameters supplied can lead to a different output). Trend surface interpolation is considered global and deterministic, where as a regression-based interpolator is global and stochastic. Similarly, inverse distance weighting is considered a local deterministic interpolator, while Kriging is a local stochastic interpolating process. Trend surface analysis uses a polynomial function across known values to create an interpolated surface. In a regression model, a set of predictors is used to fit a linear model, which can be used to estimate the values of an attribute at a location where the observation is missing. One of the local measures of interpolation, Thiessen polygons, has

been previously described in this chapter. Thiessen polygons assume that any point within the polygon is the closest to the polygon's point of origin than any other known point of origin. Inverse distance weight interpolation assumes that each point has a local influence that diminishes with increasing distance. Kriging is a statistical method used for spatial interpolation which considers both degree of variation and distance when estimating values in unknown areas. Kriging uses a semivariogram for spatial interpolation. A semivariogram is a visual depiction of the spatial autocorrelation of points that are measured. Once a semivariogram has been modeled for a given set of observed data, it can be used for prediction of unmeasured data which is similar in nature to inverse distance weighting.

Cluster Analysis Cluster analysis is a spatial analysis method which is used to identify statistically significant clusters of a spatial event. It is not to be confused with a heat map, which relies on relatively straightforward calculations, whereas the hotspot cluster analysis relies on a multi-step process that allows us to examine the probability of the occurrence of events under the assumption that our event of interest is randomly distributed in space and/or time. A hotspot is an area that has a higher incidence/prevalence of an event compared to what would be generally expected if the events were randomly distributed in a spatial area. By identifying clusters, we can determine, with more certainty, whether locations may be unsafe if, for instance, we are studying public safety and wish to identify the locations where crimes cluster. Similarly, these techniques can also be used to observe if two different events are occurring in close proximity, which could help in examining the causal chain of events (e.g., an increase in the clustering of incidence of Lyme disease with a simultaneous increase in the clustering of *Ixodes scapularis* tick populations in Michigan between the years 2000 and 2014 (Lantos et al. 2017)).

Multiple tools have been developed to answer nuanced questions regarding the assessment of clusters in a geographical area. Among the spatial patterns and clustering tests, the nearest neighbor index (NNI) can be used as an indicator for clustering of point data. It is calculated by observing the distribution of events against an expected random distribution of these events in a geographic area. Spatial autocorrelation analysis allows us to look at similarities in values that are in close proximity to each other. Measures of spatial autocorrelation can be categorized as global or local indicators of spatial association (LISA). Moran's *I* and Geary's *C* (Getis et al. 1973) are examples of global spatial autocorrelation statistics. Measures of global spatial autocorrelation assess relationships between neighboring polygons within a dataset on a macro level (e.g., state or county level). Moran's *I* is akin to a coefficient of correlation in that it measures relationship between two variables; however, the second variable is the "spatial lag" of the first variable.

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}\left(x_i - \overline{X}\right)\left(x_j - \overline{X}\right)}{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}\right)\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}$$

Where,

N = number of observations (points/polygons)
$\overline{X}$ = mean of the variable
x_i = variable at a particular location i
x_j = variable at a particular location j
W_{ij} = weighting index

Moran's I can be classified into positive (when similar values are clustered; Moran's $0 < I \leq 1$), negative (when dissimilar values are clustered; Moran's $I = -1 \leq I < 0$), and non-spatially correlated (Moran's $I = 0$) categories. Global tests for spatial autocorrelation do not point to the location of clusters. Rather, they provide evidence of the presence of spatial autocorrelation, and a spatial sphere of influence, that merits further attention. Like Moran's *I*, Geary's *C* (Getis et al. 1973) is also used as a measure of global spatial autocorrelation. The main difference in the calculation of Geary's *C* is that the cross product used for the calculation of the *C* statistic is based on deviation from the actual values themselves rather than the mean location which is used in Moran's *I*. The Geary's *C* ranges from 0 to 2 where, 0 indicates perfect positive autocorrelation, 1 indicates no autocorrelation, and 2 indicates perfect negative autocorrelation. While Moran's *I* is a simple method to measure spatial autocorrelation, its use is limited by the fact that it tends to average the local variations in spatial autocorrelation. To overcome this limitation, statisticians developed local indicators of spatial autocorrelation (LISA) (Anselin 1995).

LISA is widely used to assess the significance of clustering on the local level. Local spatial autocorrelation analysis can help identify hotspots (areas that have higher numbers of events than the estimated average number of events), coldspots (areas that have fewer numbers of events than the estimated average number of events), and outliers (very high or very low values in relation to all the values in adjacent polygons) that exist spatially. The Getis-Ord Gi* statistic, which is calculated for each feature in a dataset, allows us to locate clusters (high or low) of events.

$$G_i^* = \frac{\sum_{j=1}^{N} W_{ij}x_j - \overline{X}\sum_{j=1}^{N} W_{ij}}{S\sqrt{\frac{\left[N\sum_{j=1}^{N} W_{ij}^2 - \left(\sum_{j=1}^{N} W_{ij}\right)^2\right]}{N-1}}}$$

Where,

N = number of features
$\overline{X}$ = mean of the variable

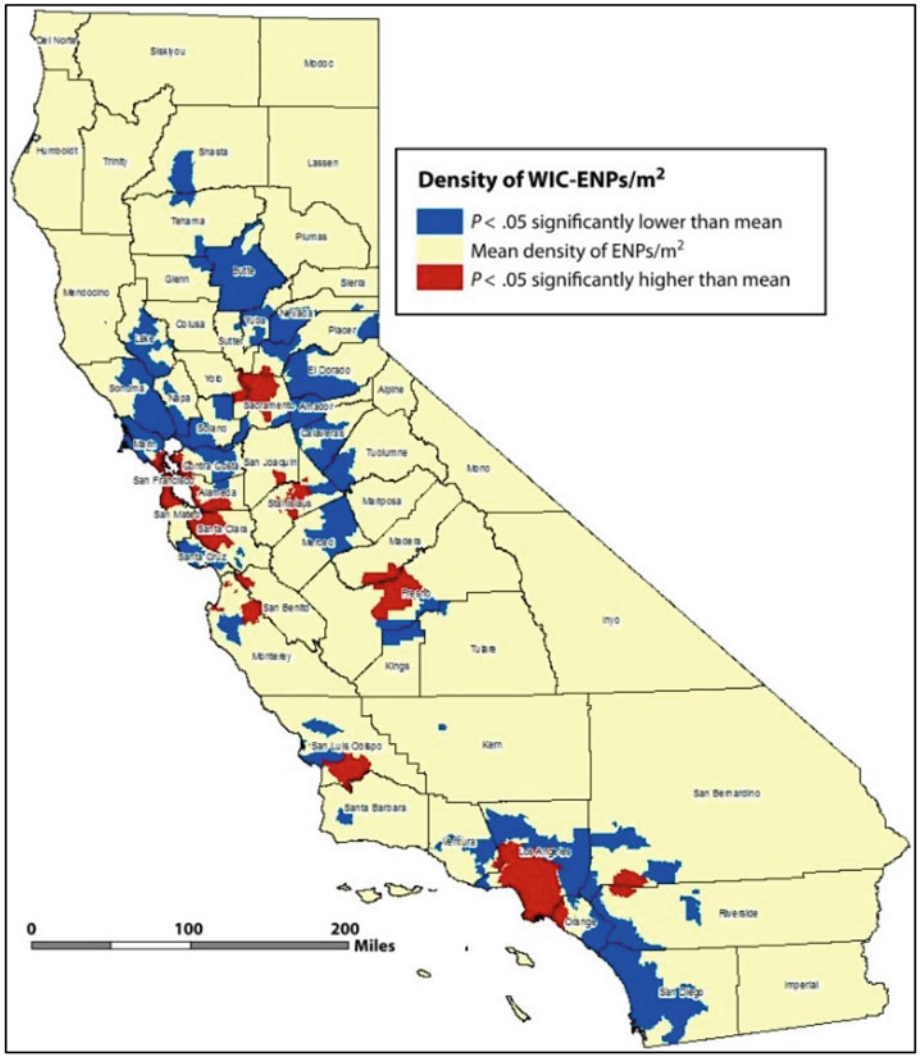


Fig. 9 Clusters of census tracts with higher densities of women who were eligible for but not receiving (i.e., eligible non-participants = ENPs) the services of Special Supplemental Nutrition Program for Women, Infants and Children (WIC), California, 2010 (Stopka et al. 2014a)

x_j = attribute value for feature j
W_{ij} = weighting index
S = standard deviation (modified)

The Gi* statistic returns a z-score for each feature used in the analysis. Significant positive z-scores are related to clustering of hotspots and significant negative z-scores are indicators of clustering of coldspots. Figure 9 shows the clusters of Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) eligible nonparticipants based on WIC eligible nonparticipants per square mile per census tract. It is important to note that cluster assessment tools are often used in tandem with one another, working from the macro to the micro level, to determine the influence of specific measures in locations or interest, and taking into consideration neighboring areas. It can be useful, for instance, to first assess the spatial relationships between polygons (e.g., census tracts) and their neighboring polygons, then conduct incremental spatial autocorrelations to determine the appropriate spatial sphere of influence for our phenomenon of interest, and ultimately to assess clustering at the local level with tests that rely on the Getis-Ord Gi* statistic (Stopka et al. 2014a).

Spatiotemporal Analyses Spatiotemporal analyses allow us to assess statistically significant clustering in areal and time dimensions – to answer the questions regarding "Where and when clustering is taking place?" Space-time cluster analyses can be thought of an extension of a simple time-based regression model. Consider a simple time series model where an event at time t is associated with an event at time t-x where x is greater than zero. The addition of geographical measures (e.g., location of event, distance between event and relevant health resources) to the time series data creates a general spatiotemporal model. Spatiotemporal models are used to analyze data that span the dimensions of both time (in terms of minutes, weeks, years) and space (in terms of linear distances, neighborhoods, and political boundaries). Analysis of spatiotemporal data not only allows researchers to study the geographical clustering of an outcome but also facilitates evaluation of changes in clustering patterns over

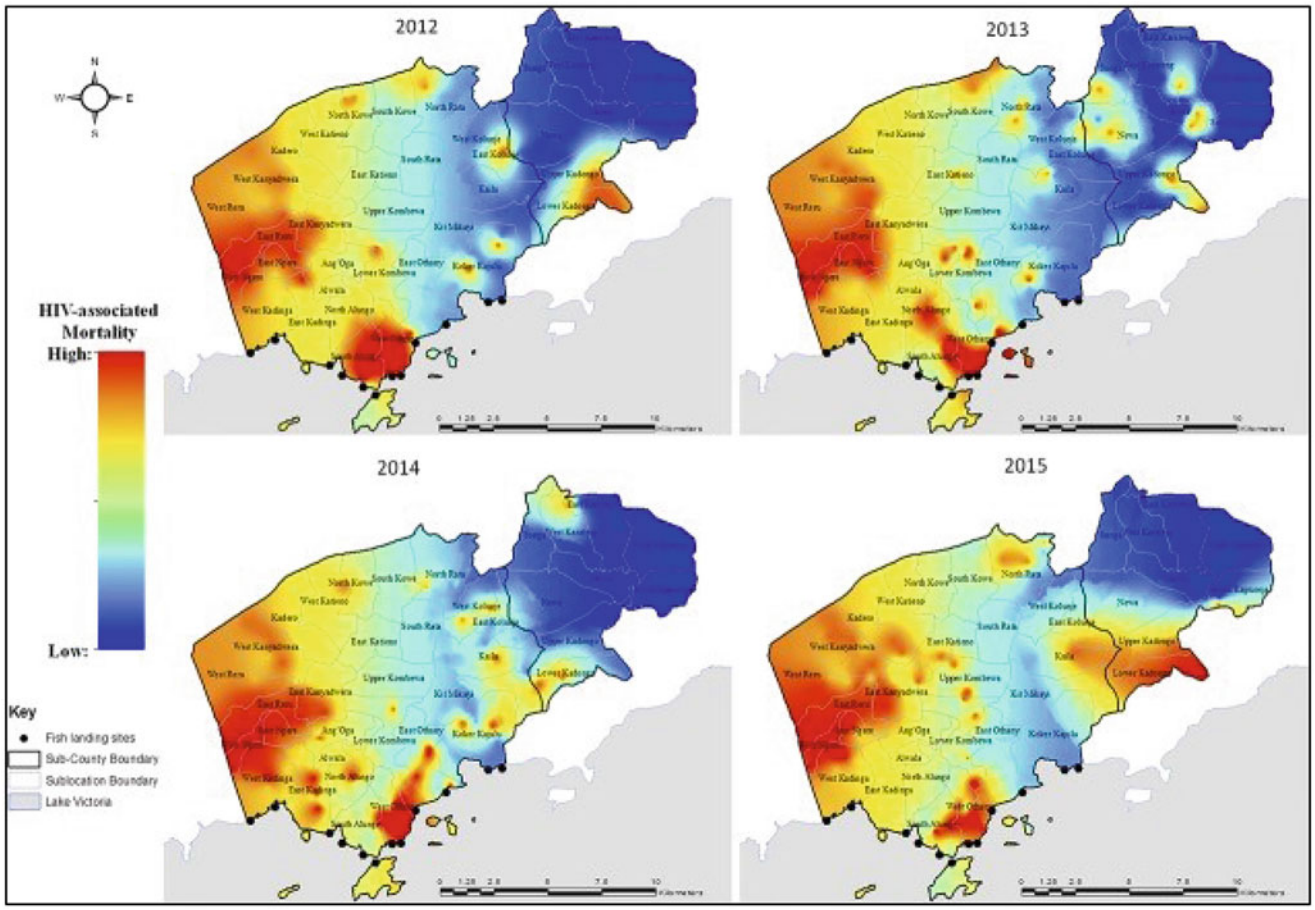


Fig. 10 Trends of HIV-related death in Kombewa health and demographic surveillance system (HDSS) between the years 2012 and 2015. In the figure above, spatiotemporal trends of deaths related to HIV in Kombewa HDSS between the year 2012 and 2015 are mapped using interpolation (Sifuna et al. 2018)

time. The complexity of a spatiotemporal analysis emanates from the unlimited directionality within the dimension of space and time. Consider a time-series dataset that consists of a sequence of observations that measures the same thing over a period of time, i.e., a specific variable is measured at a uniform time interval. A time-series dataset can be visualized by plotting time on the x-axis and the value of the observation on the y-axis. An ARIMA (Autoregressive, Integrated, Moving Average) model is a commonly used method to model and forecast time-series data based on different orders of its own past values (e.g., modeling hospital bed occupancy during a SARS outbreak (Earnest et al. 2005)). Similar to the ARIMA model for generalized time series data, techniques such as conditional autoregression, space-time ARIMA models are used for analyzing spatiotemporal data. Bayesian spatiotemporal models are also used increasingly to assess and predict clustering of events in specific locations and for specific timeframes. In the same way that weather forecasters can determine where the next hurricane is predicted to land, informing public safety advisories and responses, it is possible to employ Bayesian spatiotemporal prediction models to forecast disease outbreaks (da Costa et al. 2018; Yu et al. 2013) that can inform pre-emptive public health interventions (e.g., flu vaccine campaigns) (Fig. 10).

Case Study – The Opioid Crisis

In this section, we present a case study for one of the most pressing public health challenges of our times – the opioid crisis in the US (Scholl et al. 2018, Florence et al. 2016, Ciccarone 2019). In 2016, health care providers across the United States wrote more than 214 million prescriptions for opioid pain medication (CDC 2019). Drug overdoses claimed approximately 70,000 American lives in 2017; a majority (>60%) involved an opioid prescription or illicit opioids (Scholl et al. 2018), and the economic burden of the opioid crisis was estimated, in 2013, to be approximately $78 billion (Florence et al. 2016). In more recent years, synthetic opioids such as fentanyl, which is 50–100 times more potent than morphine, have been responsible for more than 80% of fatal opioid overdoses (Ciccarone 2019; Somerville et al.

2017). While there have been several national substance-use epidemics in the United States and internationally over the past few decades (Manchikanti et al. 2012; Cornish and O'Brien 1996; Pacurucu-Castillo et al. 2019), the opioid crisis of the first two decades of the 2000's is, perhaps, the most widespread from a geographic perspective and from a community perspective – affecting people from all socioeconomic strata, all racial and ethnic communities, urban, suburban, and rural locations, and people of all ages. In response to the immense social, human, and economic impact of the opioid crisis, significant clinical and public health interventions have been developed and implemented. Along with the traditional tools of epidemiologic research, GIS has become integral in the surveillance of the opioid epidemic, as well as the targeting of responses. It is being utilized to help local officials, community members, and health policy members better understand the extent of local, state, and national opioid epidemics, assess the vulnerability for related infectious disease outbreaks, and to monitor local responses. In reviewing this crisis from a GIS for public health and spatial epidemiological perspective, we take into consideration the risk factors (e.g., socioeconomic status, place of residence, family history, local exposures) and disease outcomes, including opioid use disorder (OUD), non-fatal and fatal overdose, the hepatitis C virus (HCV), the human immunodeficiency virus (HIV), infectious endocarditis, and sexually transmitted infections associated with opioid use and misuse.

Descriptive Mapping of the Opioid Crisis

Public health departments have increasingly relied on thematic maps to gain a general understanding of disease landscapes within their public health jurisdiction (MDPH 2017). Such risk maps have been widely used to monitor opioid use and misuse, nonfatal and fatal opioid overdoses, related comorbidities, and emergency medical service utilization (Rossen et al. 2014). While many techniques can be used to visualize information such as mortality data, a simple illustration is shown in Fig. 11 where opioid overdose death rates per 100,000 people are depicted by US county using a choropleth map. We can easily discern from the map that the rate of opioid overdose is high in the Northeastern, Central, and Southwestern regions of the United States. Such descriptive maps can assist health officials in assessing disparate outcomes across US geography, help to foster hypothesis generation with regard to causes, and inform targeted public health responses.

In addition to static maps that are often presented in reports and articles, web-based, dynamic, interactive maps provide opportunities to place spatially-oriented descriptive maps and data dashboards in the hands of key stakeholders, including community members, field staff, program managers, researchers, and policymakers (Overdose Prevention and Intervention Taskforce 2019). The advantages of online maps include opportunities to juxtapose a wide range of map layers, overlaying point data for health services on polygon-level thematic or choropleth maps for socioeconomic or health measures (Fig. 12).

Similarly, descriptive GIS maps have also been used to describe the risk environment related to opioid use. The study of individual-level factors associated with opioid use and misuse are important in tailoring specific interventions. However, the study of the broader socioeconomic and cultural factors that affect opioid use disorder is warranted as the presence of opioids is not a "sufficient" condition to cause such a crisis. In recent years, public health researchers have linked the opioid crisis to "economic and social upheaval" (Dasgupta et al. 2018). Studies have found that opioid use and misuse are concentrated in areas where there are higher rates of poverty and unemployment (Spiller et al. 2009), and it has plagued communities with high proportions of blue collar workers (Harduar Morano et al. 2018). County-level drug-related mortality data shows that between 2006 and 2015, mortality rates were higher in counties with economic and social challenges (Monnat 2018). An Assistant Secretary for Planning Evaluation (ASPE) report indicated that poverty and unemployment were correlated with higher rates of adverse opioid-related outcomes (Ghertner and Groves 2018). A bivariate choropleth map can be very useful in exploring the association between two variables (e.g., poverty and overdose deaths). As shown in Fig. 13, a color matrix can be used to denote incremental changes in poverty and opioid overdose deaths which can help researchers understand associations between these two variables and identify regions that may need a thorough evaluation to understand other risk factors related to opioid overdose deaths.

Proximity and Cluster Analysis: Applications for Studying Clusters of Opioid Use, Surveillance of Disease Outbreaks, and Access to Harm Reduction Services

With the development of advanced geostatistical methodologies and complementary software, analytic research to identify opioid use hotspots has garnered increasing attention over the years. In New Mexico, patient-level data from addiction treatment facilities was used to map opioid misuse hotspots (Brownstein et al. 2010). The study showed clustered opioid use around Albuquerque (the largest metro area in the state) and Las-Cruces (the southernmost city bordering Mexico). Furthermore, the integration of multiple sources of data into spatial analyses has also benefitted these studies. Analysis of opioid overdose in the state of Massachusetts linked 16 different administrative datasets and was able to

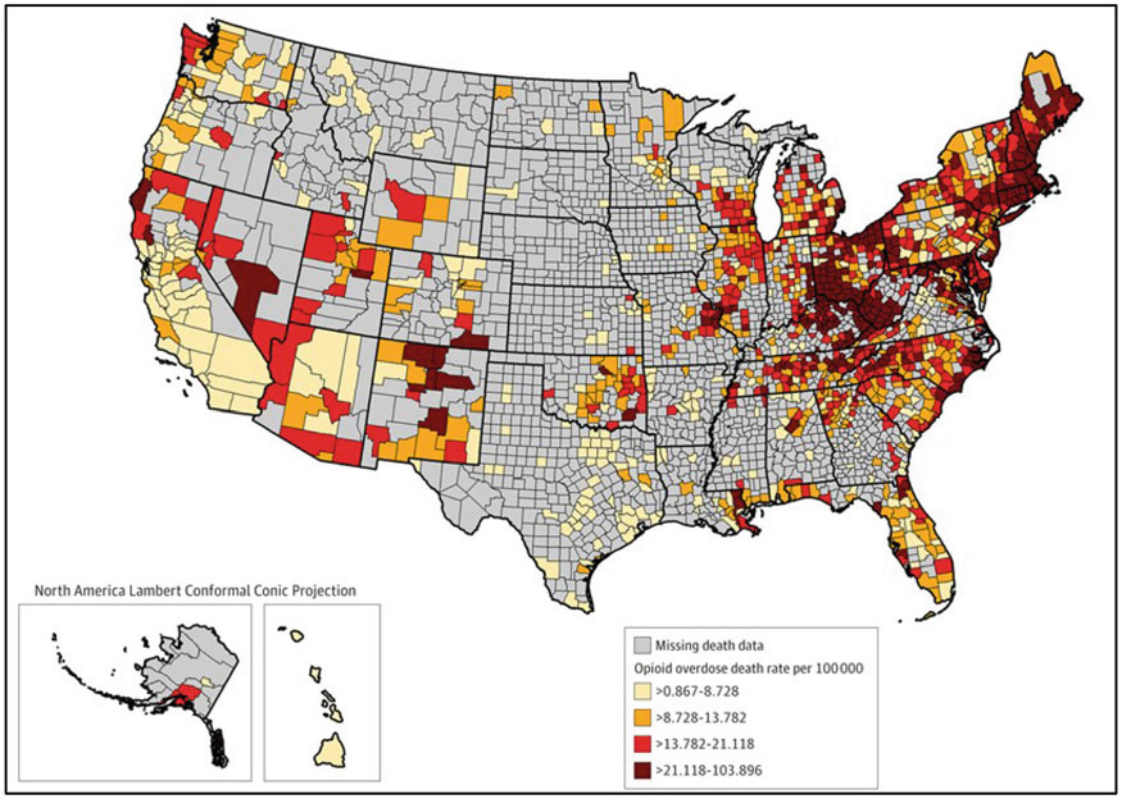


Fig. 11 Fatal opioid overdose rates per 100,000 people by US county, 2015–2017 (Haffajee et al. 2019)

identify clusters of potentially inappropriate opioid prescribing (PIP) practices and fatal opioid overdose, as well as overlapping clusters in specific regions (Stopka et al. 2019a). In Orange County, Florida, three different sources of heroin surveillance were used to assess heroin-related emergency department visits and deaths between the years 2010 and 2014 (Hudson et al. 2017). A study in Indianapolis, Indiana, leveraged EMS, coroners toxicology, and crime data to study the spatial distribution of opioid-related death (Carter et al. 2019). The study acknowledged that opioid hotpots overlapped within regions of higher crime activities. Another study conducted in Northeast Boston was able to identify a specific census tract that was associated with over 45% of emergency department visits between the years 2012 and 2015 (Dworkis et al. 2017). The figure below shows the density of non-fatal opioid overdoses in Cambridge, Massachusetts, for the years 2016–2017, with high density of these events around subway stations in Cambridge and East Cambridge (Fig. 14).

A growing body of research has utilized GIS maps and spatial epidemiological analyses to assess the risk environment related to the opioid crisis. Opioid-use disorder, opioid-related overdose, HCV, HIV, and other blood-borne infections have increased on similar trajectories between 2000 and 2019. While harm reduction strategies put in place during the early years of the HIV epidemic were successful in curbing infection rates among people who inject drugs (PWID) (Drucker et al. 1998; MacDonald et al. 2003; Wodak and McLeod 2008), increases in injection frequency (Broz et al. 2018) tied to prescription opioids in Scott County, Indiana (Peters et al. 2016; Conrad et al. 2015), and synthetic opioids such as fentanyl (Cranston et al. 2019), in recent years, have led to increases in HIV infections, once again, among PWID in the United States. Further, the exceptionally high virulence of HCV and high co-infection susceptibility have placed PWID at a greater risk of HIV and HCV infection (Cranston et al. 2019). Cooper and colleagues employed spatial techniques to develop neighborhood-level measures for access to prevention services – namely syringe exchange programs as well as drug-related crime measures – to assess community-level risk (Cooper et al. 2009a). In more recent years, Davidson et al. employed GIS to assess the need for targeted syringe service interventions to reduce injection-mediated risks (Davidson et al. 2011), and Brouwer employed spatial epidemiological analyses to assess HIV risk among people who inject drugs (PWID), down to the neighborhood level, to identify hotspots for HIV transmission risk in Tijuana, Mexico (Brouwer et al. 2012). In Southern and Northern California, Stopka and colleagues assessed the risk environment by measuring access to pharmacies that sold sterile syringes in Los Angeles (Stopka et al. 2013) and San Francisco (Stopka et al. 2012), respectively, and determined

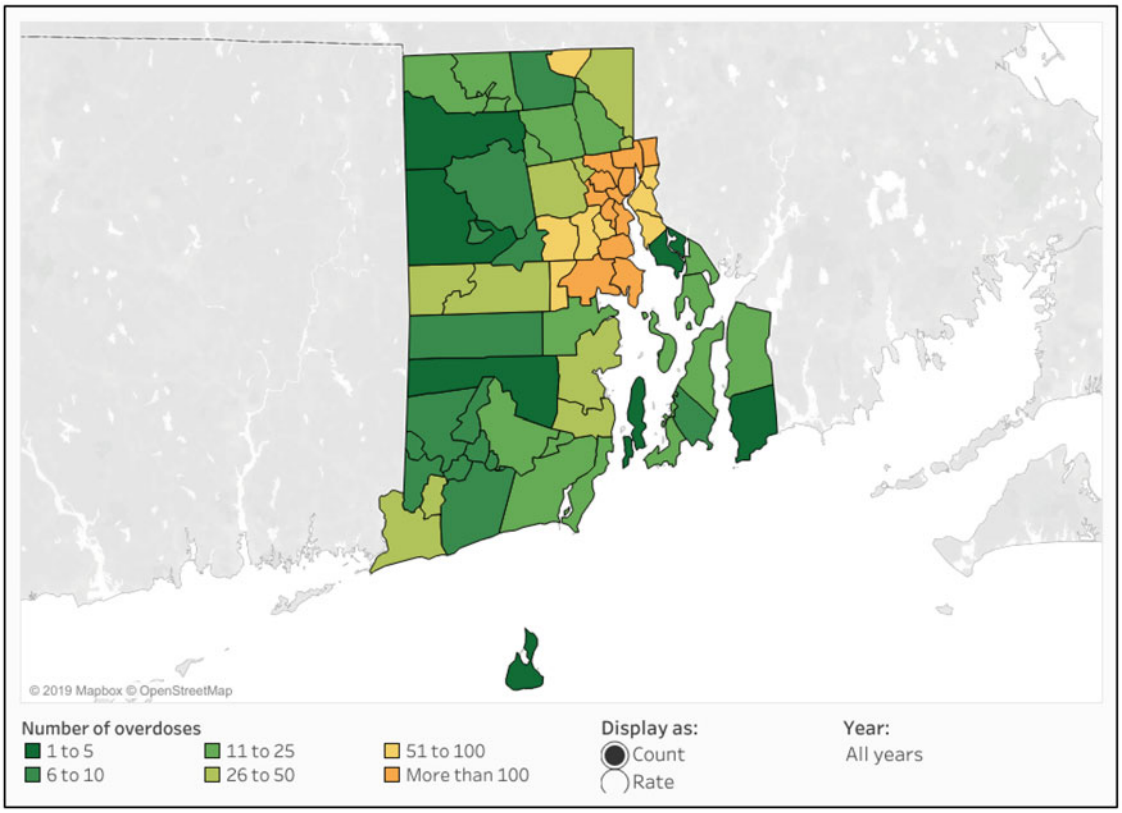


Fig. 12 Overdose deaths by municipality in Rhode Island, 2014–2018 (Overdose Prevention and Intervention Taskforce 2019). This online interactive map allows health policy experts and community members to assess overdose burden across communities throughout Rhode Island, allowing the viewer to turn on and off different layers and observe different variables of interest across geography and time (Overdose Prevention and Intervention Taskforce 2019).

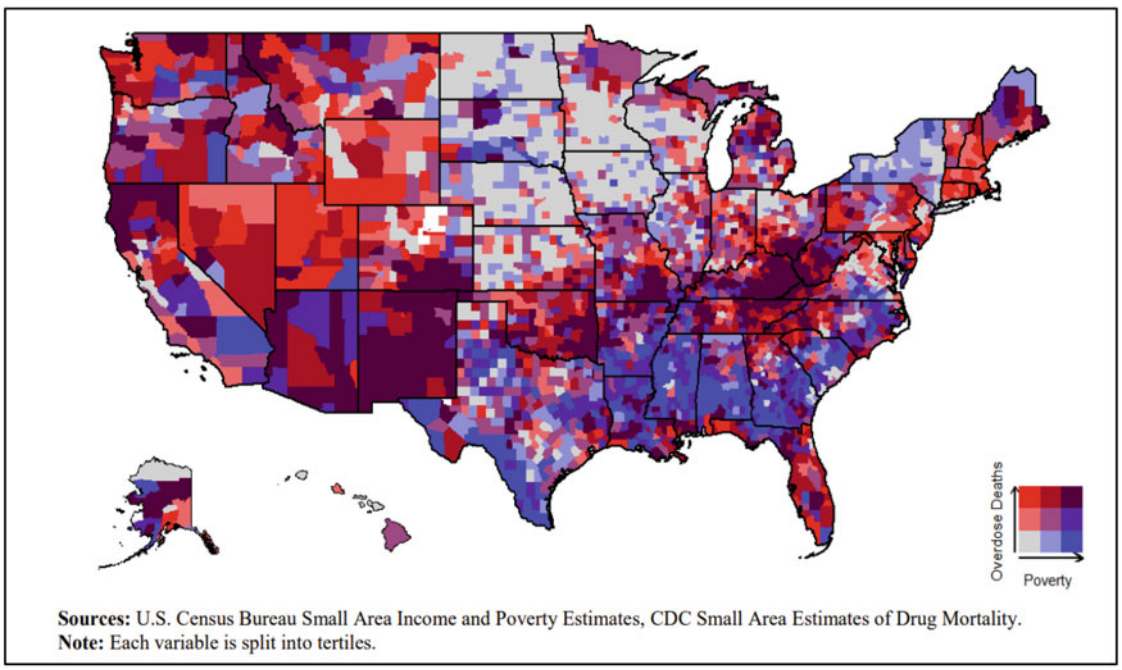


Fig. 13 Poverty rates and drug overdose deaths in the United States, 2016. A color matrix of all possible combination of poverty and overdose rate tertiles has been used in this figure to display the relationship between these variables across the country (Ghertner and Groves 2018).

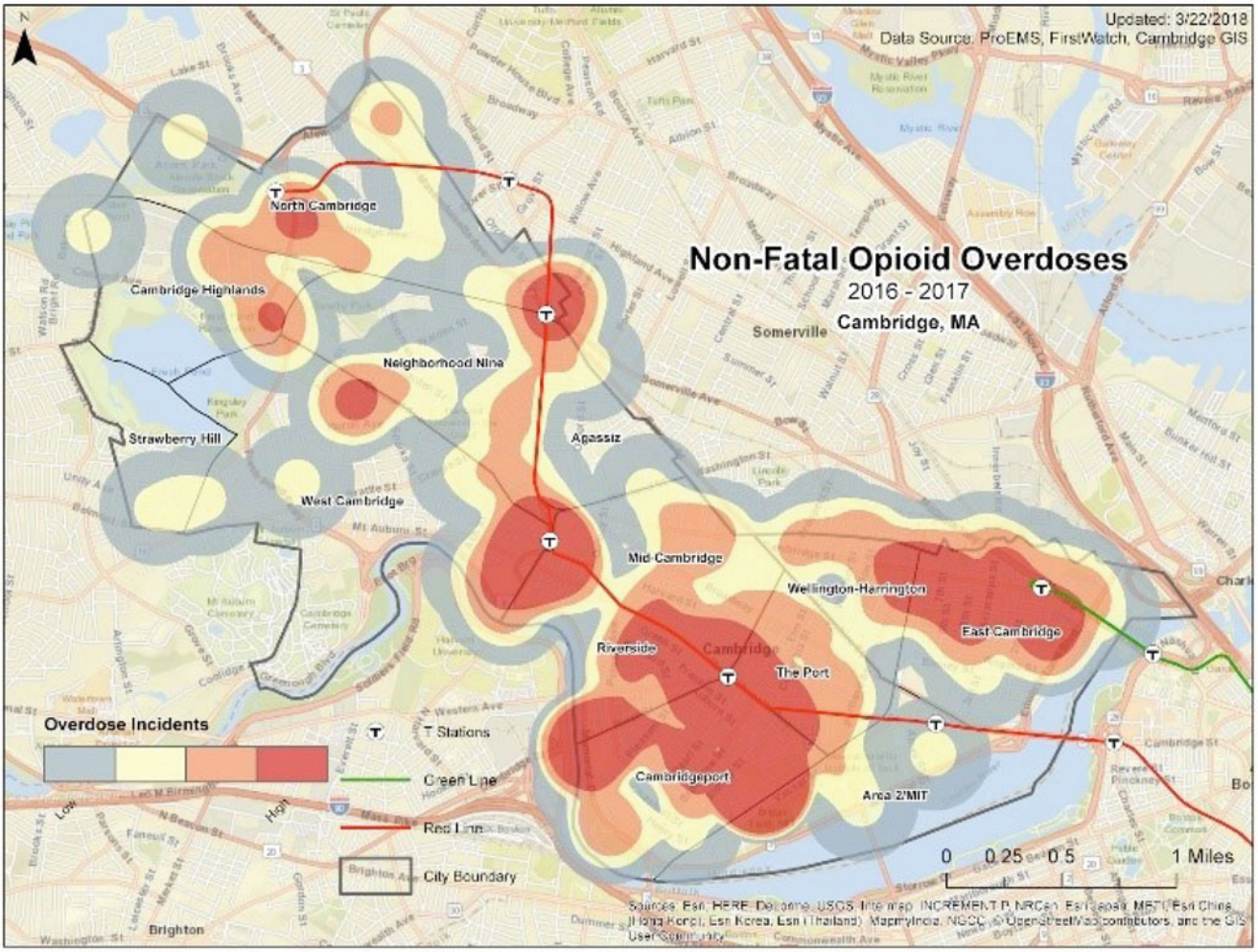


Fig. 14 Heat map representing density of non-fatal opioid overdoses in Cambridge, Massachusetts, 2016–2017 (Cambridge Health Alliance 2019).

whether over-the-counter syringe sales in Los Angeles were associated with increased crime rates in local communities (they were not) (Stopka et al. 2014b). In a comprehensive study of the opioid epidemic in New England, researchers studied the structural factors (race/ethnicity, income, unemployment, insurance rates), opioid-related outcomes (HCV and HIV infections, overdose-related deaths), and access (drive time) to harm reduction strategies such as syringe services programs, naloxone access, and opioid treatment services in the region (Fig. 15) (Stopka et al. 2019b).

GIS has been employed to map statistically significant clusters of HIV and HCV with a particular focus on PWID (Meyers et al. 2014; Stopka et al. 2017b; Stopka et al. 2018). Trooskin et al. identified six clusters of HCV in Connecticut and suggested a link to injection drug use and poor socioeconomic conditions (Trooskin et al. 2005). HIV clusters in injection drug users were also identified in San Francisco between 1987 and 2005 using spatial analysis, where the clusters mapped closely to higher poverty levels (Martinez et al. 2014). The geographic distribution of HCV infection among young PWID in two cities in the United States between 2002 and 2004 showed a different picture and helped to identify a significant difference in spatial patterns (Boodram et al. 2010). Notably, regression analysis ultimately demonstrated that the difference in HCV prevalence and spatial distributions was tied to specific city locations rather than specific socioeconomic or demographic factors. More recent analyses in San Francisco employed GIS, spatial, and statistical analyses to assess injection-mediated disease transmission risks by overlaying heat maps for HIV and HCV on maps that depicted access to pharmacies selling syringes, which could help reduce disease transmission risks (Fig. 16) (Stopka et al. 2012).

As evidenced by the studies presented above, results generated from spatial epidemiologic analysis have been widely used to plan, implement, and evaluate harm reduction strategies to enhance sterile syringe access, naloxone distribution, and adherence to medication for opioid use disorders (e.g.,

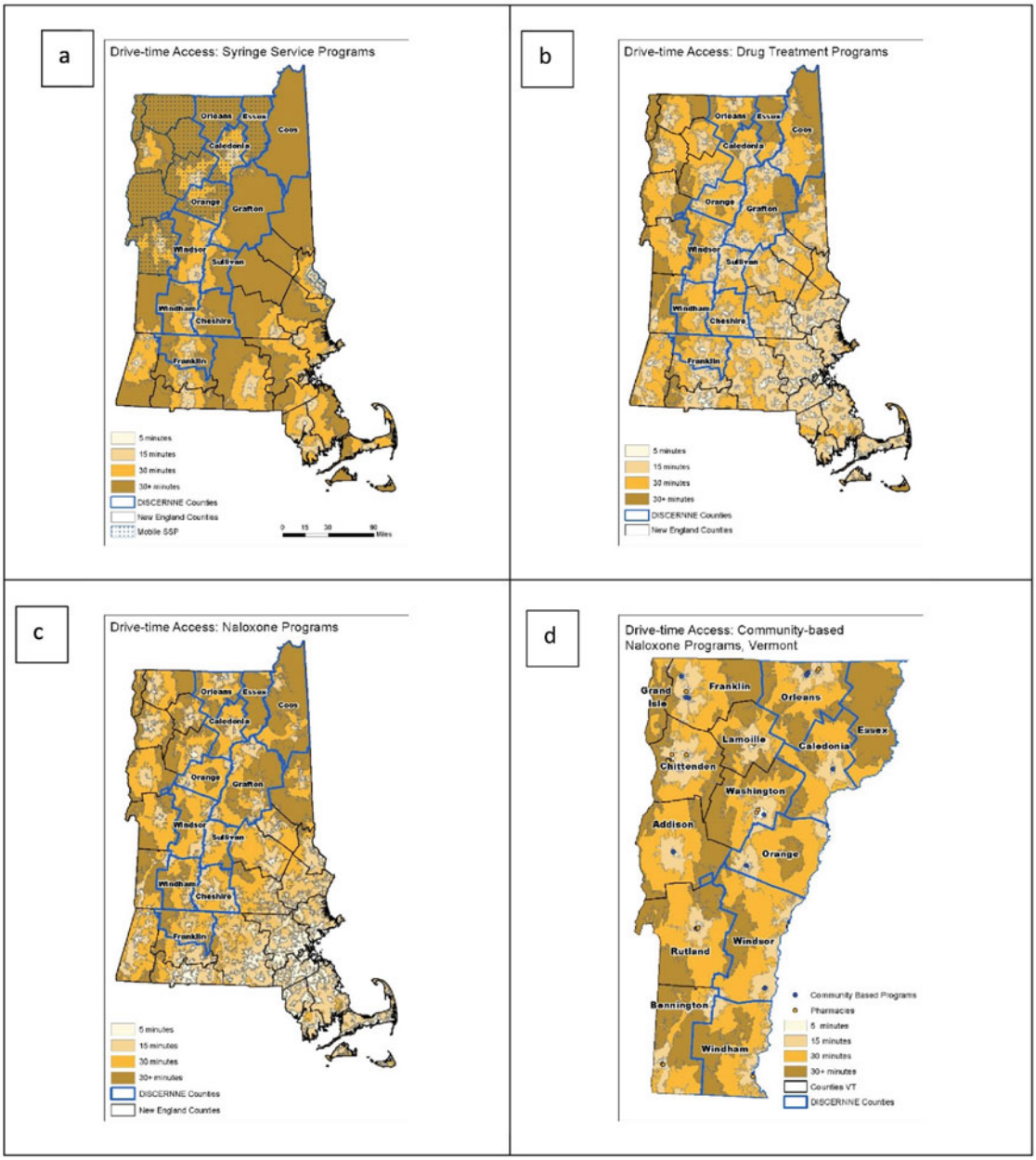


Fig. 15 Drive-time buffers to assess access to (**a**) syringe services programs, (**b**) drug treatment programs, (**c**) naloxone distribution programs, and (**d**) community-based naloxone programs in rural Northern New England, 2018

methadone, buprenorphine, naltrexone) (Stopka et al. 2017a; Amiri et al. 2018; Amiri et al. 2019; Cooper et al. 2009b; Fedorova et al. 2013; Rowe et al. 2016). In Massachusetts, data from phone surveys with retail pharmacies were used to identify areas where pharmacies sold non-prescription naloxone and syringes (Stopka et al. 2017a). Through the study, and the use of GIS maps and spatial analyses, investigators detected limited access to naloxone from a spatial perspective, prompting the need to improve naloxone distribution through pharmacies. A similar study was conducted in California to assess access to non-prescription syringes based on drive times to pharmacies that sold syringes (Pollini et al. 2015). In Fresno and Kern counties, for instance, Pollini and colleagues found that 80% of the residents lived within a 5-minute drive to a retail pharmacy; however, only half of the residents were within 5-minute drive of a pharmacy that sold syringes. Spatial epidemiological analyses were also used to identify hotspots where injection drug users resorted to syringe sharing behaviors to inform targeting of syringe services programs (Davidson et al. 2011). Spatial analysis also has been helpful in improving our understanding of geographical factors that affect adherence to medication for

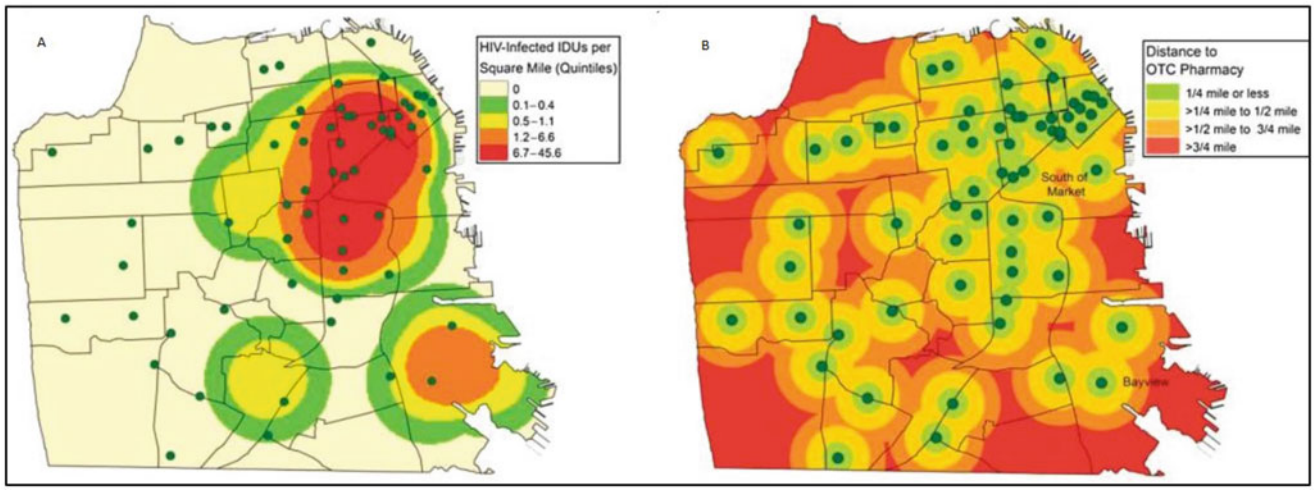


Fig. 16 Evaluating access to pharmacies that sold syringes over the counter (i.e., without a prescription) in San Francisco, California, 2008. Panel A represents the density of people who injected drugs, or injection drug users (IDUs), who were infected with HIV, juxtaposed with pharmacies that were registered with the local Department of Public Health to sell syringes (green dots) over the counter. Panel B highlights multiple ring buffers depicting distance to the nearest pharmacy that sold over-the-counter syringes, providing an initial understanding of access to risk reduction supplies (Stopka et al. 2012)

opioid-use disorders. Studies have reported, for instance, that increases in distance to treatment programs were associated with lower adherence (Amiri et al. 2018; Amiri et al. 2019). In a recently published study, Haffajee et al. combined county-level mortality data with substance-use treatment provider data to identify areas with lower access to treatment services, which had higher opioid overdose rates (Haffajee et al. 2019). This type of spatially-relevant information can be used for effective mobilization of resources to provide treatment services for people with opioid-use disorder, thereby reducing their chances for subsequent overdose. These studies demonstrate the critical role of GIS and spatial analyses in identifying the distribution of risk, highlighting the concerning clusters of disease outcomes, and evaluating access to existing prevention and treatment services to inform enhanced targeting of resources to minimize opioid-related adverse outcomes.

Spatial and Geostatistical Approaches to Understand and Predict Opioid-Related Adverse Events

Opioid use and misuse patterns, and the related adverse events such as opioid overdose and infectious disease transmission, are not static nor are they isolated. As mentioned previously, increases in sales of prescription opioids paralleled increasing trends in opioid overdose deaths between 2000 and 2015 (Scholl et al. 2018; CDC Wonder n.d.; Rudd et al. 2016). However, the main cause of opioid-related death has changed over time, from prescription opioids to heroin and ultimately to fentanyl, in what has been termed "the triple wave of the opioid epidemic" (Ciccarone 2019). Various studies have traced the changing nature of the opioid crisis employing various statistical and geospatial techniques (Rudd et al. 2016; Schoenfeld et al. 2019; Chen et al. 2019) (Hoffman et al. 2017). These studies have been instrumental in predicting changes in patterns of opioid use and misuse, in an attempt to provide adequate time to implement and augment needed services to mitigate opioid-related adverse events.

Various spatial and temporal analytic methods have been used to visually demonstrate the change in risk patterns for opioid-related adverse events. A recent study examined temporal and spatial patterns of heroin-related overdose events and identified the demographic and built environment factors that were associated with these overdoses (Li et al. 2019) (Fig. 17).

Increasingly, Bayesian spatiotemporal models are being developed to assess local opioid crises. Kline and colleagues used a joint spatial effects model of opioid-associated death and drug treatment using a generalized common spatial factor model (Kline et al. 2019). In addition to covariate effects, they were able to estimate a spatial factor for each county that characterized structural factors not accounted for by other covariates in the model that are associated with deaths and treatment counts. They ultimately identified associations between social and structural covariates (e.g., health professional shortages, people on disability, and single female households) and opioid-associated deaths and treatment counts. They were also able to characterize counties with latent risk that could help to guide future research to identify potential community risks (Fig. 18).

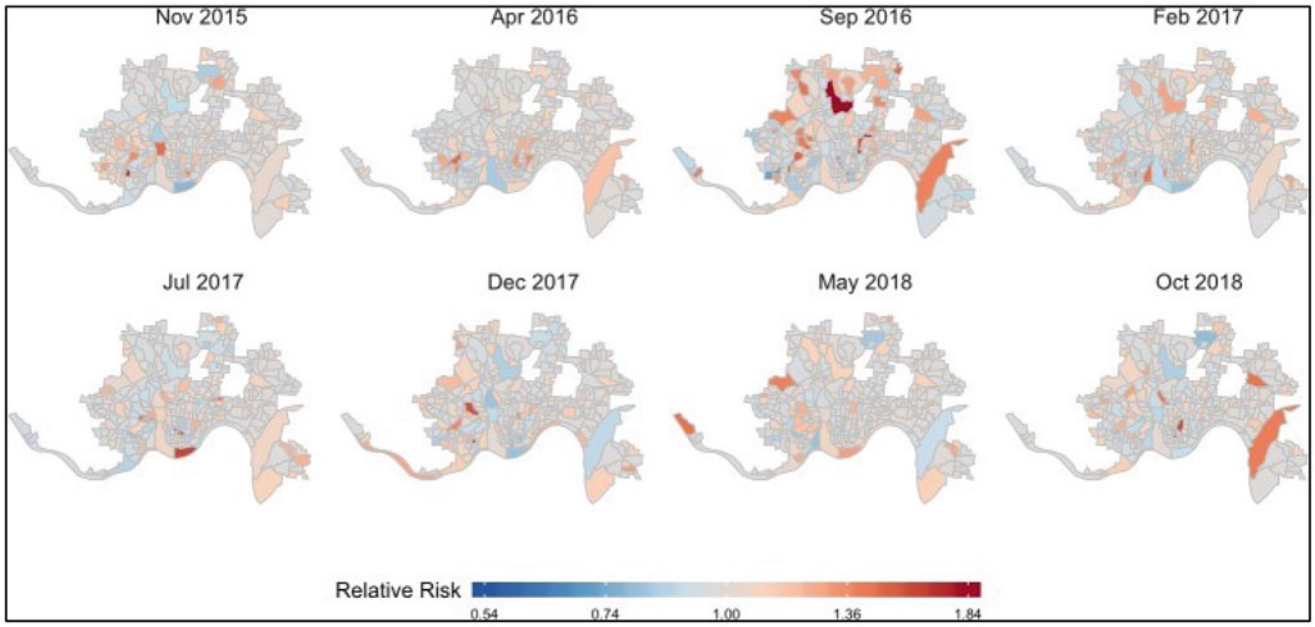


Fig. 17 Figure showing relative risk of heroin-related adverse event compared to the baseline trend over time and space (areas highlighted in red have higher risk for adverse events), Cincinnati, Ohio, August 2015 through January 2019 (Li et al. 2019)

These studies depict the changing nature of the opioid crisis over time and space. The results of such analyses can help inform local responses by public health officials, first responders, clinicians, and harm reduction personnel. As the threat of fentanyl, for instance, has grown across the United States in recent years, many harm reduction programs have begun to offer fentanyl test strips in key locations that have allowed PWID to examine whether their drug supply had been contaminated with fentanyl, providing opportunities for behavior change to reduce overdose risks (e.g., by carrying naloxone and avoiding use while alone). In locations where increases in opioid overdoses have been observed, enhanced opioid education and naloxone distribution programs and increased access to medication for opioid-use disorder can be bolstered. And, in areas where HIV clusters tied to opioid use and misuse are discovered, access to syringe services programs and other harm reduction services could be increased to curb local outbreaks.

Summary of the Case Study

In this case study, we provided examples of studies that have employed GIS and spatial epidemiology to provide an enhanced understanding of the risk environment with regard to the opioid crisis and to help inform targeted responses in the locations that are most in need. As with any disease, there are a myriad of contributing factors, some more specific and proximal while others are more distal. We focused on studies that evaluated a broad range of factors that contributed to the opioid crisis and its many challenging health outcomes. Some focused on risk factors such as potentially inappropriate opioid prescription rates (Haffajee et al. 2019; Spiller et al. 2009; Basak et al. 2019), while others focused on the broader contexts, such as socioeconomic, environmental factors, as well as health policies (Stopka et al. 2019a; Stopka et al. 2019b). Direct outcomes, such as fatal overdoses, and infectious diseases related to opioid injection, were also a topic of focus in studies that utilized GIS (Stopka et al. 2019a; Haffajee et al. 2019; Carter et al. 2019; Basak et al. 2019; Scholl et al. 2019). Sound epidemiological inquiries, which consider broader socio-cultural and economic contexts, supported by accurate spatiotemporal data can be used to guide public health responses. It is then the task of public health researchers to communicate those results effectively. As illustrated throughout this chapter, GIS can contribute to these communication efforts, which helps produce descriptive maps that can be used to inform community members and public health officials alike about the risk landscape. Spatial epidemiological and geostatistical approaches can facilitate analyses that allow researchers, public health officials, and health policy experts to monitor and forecast health outcomes across geographic regions, which helps inform targeted public health responses to curb local epidemics.

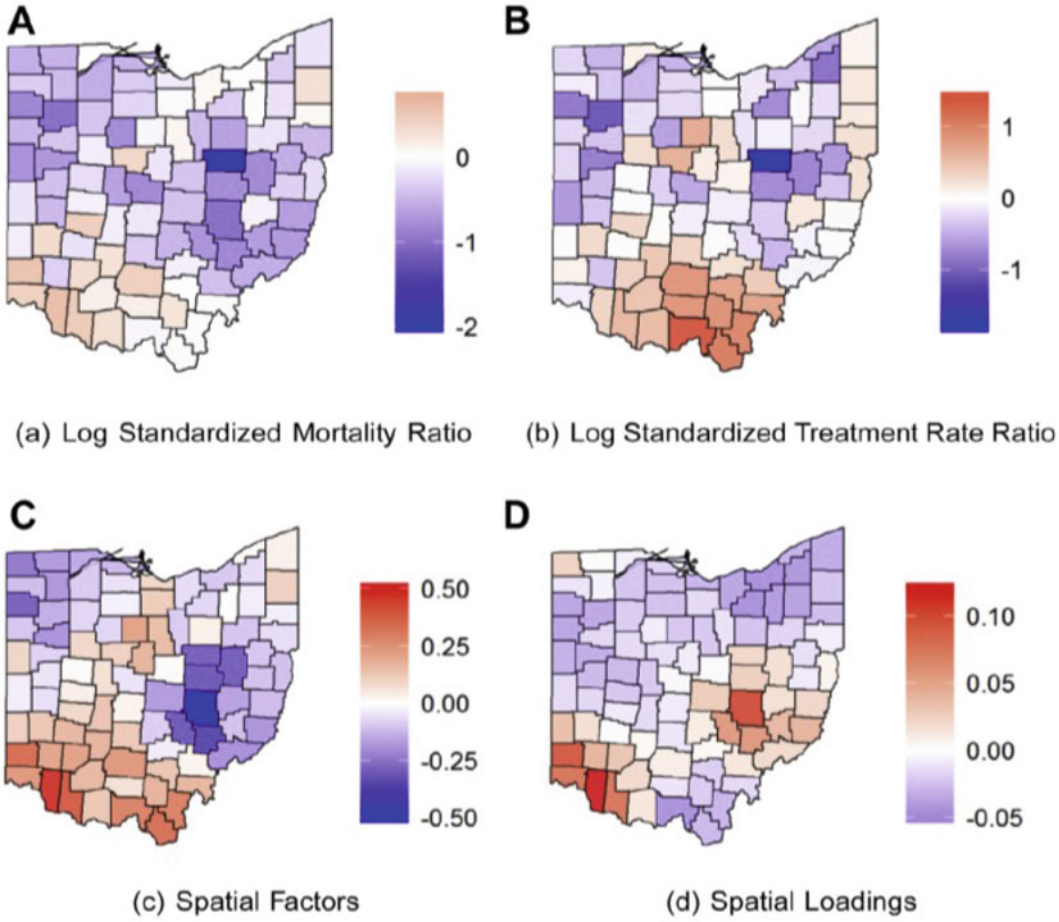


Fig. 18 Results of joint effects models focus on opioid-associated deaths (**a**) and opioid treatment admissions (**b**) in Ohio counties. Models facilitated identification of spatial factors associated with latent risk (**c**), as well as calculation of spatial loadings associated with key factors (Kline et al. *Annals of Epi*, 2019)

Discussion

Challenges and Limitations of GIS and Spatial Epidemiology

Despite the many opportunities and advantages of mapping and analyzing spatially-oriented data within a GIS, several challenges exist that can limit effective use and approaches.

The scope of primary data collection for spatial epidemiological studies is massive. Many resources, including funding, time, and trained personnel, are required to foster collection of information required to perform many of these analyses. The absence of an underlying data collection infrastructure can remain as a significant barrier to conducting spatial epidemiologic studies. Specifically, in low-income communities, limited GIS infrastructure can make incorporation of spatial epidemiology in public health difficult. Poor infrastructure, as related to GIS, can be tied to inadequate access to licensed software, a lack of trained professionals, limited access to needed instruments, and paucity of databases that can be spatially linked and analyzed (Boulos 2004; Bergquist and Rinaldi 2010; McLafferty 2003). Other issues arise in the form of non-uniform measures of spatial data, problems with privacy, lack of data sharing between agencies, and restricted access to national databases. In response to these limitations, Fletcher-Lartey and Caprarelli have put forth a summary of possible ways to address these issues with GIS (Fletcher-Lartey and Caprarelli 2016). The authors have suggested utilization of free software, creating a community of researchers and developers who champion open source platforms. They have identified the importance of educating locals to sustain GIS research. Finally, they also recommend creation of data sharing protocols for efficient data collection and development of a governance system in GIS research to safeguard privacy and promote ethical research.

Apart from the issues arising from an infrastructure perspective, GIS has other limitations that originate from the underlying theory of spatial epidemiology itself. Spatial analysis is largely conducted at an ecologic level. This creates two sets of problem. First, it leads to data loss as the information is aggregated to the unit of analysis. Second, the results are only generalizable to the unit of analysis and any interpretation of the results at an individual level can lead

to ecological fallacy. Additionally, improper sampling can lead to underrepresentation and thereby case selection bias. A well-cited example is that of positive association between prostate cancer incidence and socioeconomic factors (Oliver et al. 2005). Median household income and urban status had a positive association with incidence of prostate cancer, while poverty and lower education were associated with lower incidence (among whites only). These associations manifested only at the census tract level and were absent when the unit of analysis was at the county level. A detailed analysis showed that missing data in the study were linked to unbalanced measures causing "cartographic confounding."

Geographic boundaries in GIS and spatial analyses present another major challenge while using GIS. Boundary problems are tied to issues arising from the loss of neighboring polygons in spatial analyses. For many analyses, researchers use arbitrary boundaries to focus on the area under study. However, spatial processes continue beyond these arbitrary boundaries. For example, in a hypothetical study of disease transmission within different counties in a state, there might be considerable spatial effects on the edges of the study area based on the disease patterns in the surrounding states. If data from surrounding states are not included in spatial analyses, results on or near local boundaries may be less stable. The modifiable areal unit problem (MAUP) is another common issue with spatial analyses where the results of data aggregation are associated with the cartographer's choice of areal unit of analysis (Openshaw 1979; Openshaw 1984). Other issues that need to be considered while using GIS and conducting spatial analyses include the limited reliability of some census databases, which may not account for population migration, as well as challenges in estimating rare events.

The issues and limitations highlighted here are far from insurmountable, and with adequate investment in infrastructure, training, data systems, and development of new tools, GIS for public health and spatial epidemiology will continue to expand.

The Future of GIS and Spatial Epidemiology

There is a growing cadre of researchers and students in training who will be the GIS mapping and spatial epidemiological analysts of the future. The number and scope of GIS and spatial analysis courses on college campuses are rapidly expanding, and new scientific journals and conferences continue to appear with a focus on the geography of health and spatial epidemiology. Ongoing and future investments are needed to support continued growth, development, and innovation in the field to support sound, evidence-based, and geographically targeted public health interventions.

As big data and health informatics systems continue to expand, facilitating access to exposure and outcome data tied to health in near real time, the future is ripe for GIS, spatial epidemiology, and spatially-focused research. Vivid depictions of static and interactive online maps will allow us to pinpoint locations at highest risk for health and disease outcomes, better assess access to health services, and inspire community- and data-driven responses to local health disparities. Our increasing access to accurate and spatially-granular big data will also help to pave the way for growing opportunities to develop spatiotemporal prediction models that can inform pre-emptive public health and clinic responses, which helps to further decrease morbidity and mortality in local communities.

Conclusion

Throughout this chapter, we have highlighted the role of GIS and spatial analyses in public health and epidemiology. We have shared details on a number of different tools and methodologies that have been and that can be employed in GIS mapping and spatial analyses to better understand risks tied to detrimental public health outcomes and access to services through a geographic lens. While the field of GIS for public health and spatial epidemiology has taken major steps forward during the past two decades, there is much room for improvement in the years ahead. In the era of "Big Data," health informatics, and data science and analytics, we have an increasing array of spatially-oriented data with which to work, and we have a broadening spectrum of spatial analytical tools at our fingertips. Areas for growth include the following:

- Workforce development and capacity building within public health departments
- Spatiotemporal forecasting of disease to guide pre-emptive public health and clinical responses
- Imputation approaches to account for missing and censored data, to avoid gaps or "holes" in maps and spatial analyses, and to provide a more complete picture of local public health landscapes
- Vulnerability analyses that take a wide range of socioeconomic, social determinants of health, and disease outcomes into consideration, leading to calculation, mapping, and spatial analysis of composite measures of risk.

As public health experts, epidemiologists, researchers, students, and community members embrace maps, spatially-oriented health data, and geographically focused data science in the years to come, GIS for public health and spatial epidemiology will continue to expand, further enhancing our understanding of the spatially-oriented risk factors and

outcomes that surround us and inform better targeting of sound public health and clinical resources to improve the public's health.

References

Alliance C.,H. *Addressing the overdose epidemic in Cambridge.* 2019.

Amiri, S., R. Lutz, M.E. Socías, M.G. McDonell, J.M. Roll, and O. Amram. 2018. Increased distance was associated with lower daily attendance to an opioid treatment program in Spokane County Washington. *Journal of Substance Abuse Treatment* 93: 26–30.

Amiri, S., R.B. Lutz, M.G. McDonell, J.M. Roll, and O. Amram. 2019. Spatial access to opioid treatment program and alcohol and cannabis outlets: Analysis of missed doses of methadone during the first, second, and third 90 days of treatment. *The American Journal of Drug and Alcohol Abuse*: 1–10.

Anselin L. 1995. Local Indicators of spatial association–LISA. *Geographical Analysis* 27(2): 93–115.

Aziz, S., R. Ngui, Y.A.L. Lim, et al. 2012. Spatial pattern of 2009 dengue distribution in Kuala Lumpur using GIS application. *Tropical Biomedicine* 29 (1): 113–120.

Barkhuus A. Medical geographies. 1945.

Barnard, D.K., and W. Hu. 2005. The population health approach: Health GIS as a bridge from theory to practice. *International Journal of Health Geographics* 4: 23.

Barrett, F.A. 2000. Finke's 1792 map of human diseases: The first world disease map? *Social Science & Medicine* 50 (7–8): 915–921.

Basak, A., J. Cadena, A. Marathe, and A. Vullikanti. 2019. Detection of spatiotemporal prescription opioid hot spots with network scan statistics: Multistate analysis. *JMIR Public Health and Surveillance* 5 (2): e12110.

Bennett, J.E., G. Li, K. Foreman, et al. 2015. The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. *The Lancet* 386 (9989): 163–170.

Bergquist, R., and L. Rinaldi. 2010. Health research based on geospatial tools: A timely approach in a changing environment. *Journal of Helminthology* 84 (1): 1–11.

Boodram, B., E.T. Golub, and L.J. Ouellet. 2010. Socio-behavioral and geographic correlates of prevalent hepatitis C virus infection among young injection drug users in metropolitan Baltimore and Chicago. *Drug and Alcohol Dependence* 111 (1–2): 136–145.

Boulos, M.N. 2004. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *International Journal of Health Geographics* 3 (1): 1.

Bowditch H.I. *Consumption in new england: or, locality one of its chief causes. an address delivered before the Massachusetts Medical Society.* Ticknor & Fields. 1862.

Brouwer, K.C., M.L. Rusch, J.R. Weeks, et al. 2012. Spatial epidemiology of HIV among injection drug users in Tijuana, Mexico. *Annals of the Association of American Geographers* 102 (5): 1190–1199.

Brown, G., M.F. Schebella, and D. Weber. 2014. Using participatory GIS to measure physical activity and urban park benefits. *Landscape and Urban Planning* 121: 34–44.

Browning, M., and K. Lee. 2017. Within what distance does "greenness" best predict physical health? A systematic review of articles with GIS buffer analyses across the lifespan. *International Journal of Environmental Research and Public Health* 14 (7): 675.

Brownstein, J.S., T.C. Green, T.A. Cassidy, and S.F. Butler. 2010. Geographic information systems and pharmacoepidemiology: Using spatial cluster detection to monitor local patterns of prescription opioid abuse. *Pharmacoepidemiology and Drug Safety* 19 (6): 627–637.

Broz, D., J. Zibbell, C. Foote, et al. 2018. Multiple injections per injection episode: High-risk injection practice among people who injected pills during the 2015 HIV outbreak in Indiana. *The International Journal on Drug Policy* 52: 97–101.

Cameron, D., and I.G. Jones. 1983. John Snow, the broad street pump and modern epidemiology. *International Journal of Epidemiology* 12 (4): 393–396.

Carter, J.G., G. Mohler, and B. Ray. 2019. Spatial concentration of opioid overdose deaths in Indianapolis: An application of the law of crime concentration at place to a public health epidemic. *Journal of Contemporary Criminal Justice* 35 (2): 161–185.

CDC. U.S. Opioid Prescribing Rate Maps. https://www.cdc.gov/drugoverdose/maps/rxrate-maps.html. Accessed August 31, 2019.

CDC Wonder. http://wonder.cdc.gov/.

Celentano, D.D., and S. Mhs. 2018. *Gordis epidemiology*. Elsevier.

Chen, Q., M.R. Larochelle, D.T. Weaver, et al. 2019. Prevention of prescription opioid misuse and projected overdose deaths in the United States. *JAMA Network Open* 2 (2): e187621–e187621.

Ciccarone, D. 2019. The triple wave epidemic: Supply and demand drivers of the US opioid overdose crisis. *The International Journal on Drug Policy* 71: 183–188. https://doi.org/10.1016/j.drugpo.2019.01.010.

Clarke, K.C. 1995. *Analytical and computer cartography*. Vol. 1. Upper Saddle River: Prentice Hall Englewood Cliffs.

Conrad, C., H.M. Bradley, D. Broz, et al. 2015. Community outbreak of HIV infection linked to injection drug use of Oxymorphone – Indiana, 2015. *MMWR. Morbidity and Mortality Weekly Report* 64 (16): 443–444.

Cooper, H., B. Bossak, B. Tempalski, D. Des Jarlais, and S. Friedman. 2009a. Geographic approaches to quantifying the risk environment: Drug-related law enforcement and access to syringe exchange programmes. *The International Journal on Drug Policy* 20 (3): 217–226.

Cooper, H.L., B.H. Bossak, B. Tempalski, S.R. Friedman, and D.C. Des Jarlais. 2009b. Temporal trends in spatial access to pharmacies that sell over-the-counter syringes in New York City health districts: Relationship to local racial/ethnic composition and need. *Journal of Urban Health* 86 (6): 929–945.

Cornish, J.W., and C.P. O'Brien. 1996. Crack cocaine abuse: An epidemic with many public health consequences. *Annual Review of Public Health* 17: 259–273.

Cranston, K., C. Alpren, B. John, et al. 2019. Notes from the field: HIV diagnoses among persons who inject drugs - northeastern Massachusetts, 2015-2018. *MMWR. Morbidity and Mortality Weekly Report* 68 (10): 253–254.

Cubbin C, Egerter S, Braveman P, Pedregon V. Where we live matters for our health: Neighborhoods and health. 2008.

Cummins, S., S. Curtis, A.V. Diez-Roux, and S. Macintyre. 2007. Understanding and representing 'place' in health research: A relational approach. *Social Science & Medicine* 65 (9): 1825–1838.

da Costa, A.C.C., C.T. Codeço, E.T. Krainski, M.F.D.C. Gomes, and A.A. Nobre. 2018. Spatiotemporal diffusion of influenza a (H1N1): Starting point and risk factors. *PLoS One* 13 (9): e0202832.

Dasgupta, N., L. Beletsky, and D. Ciccarone. 2018. Opioid crisis: No easy fix to its social and economic determinants. *American Journal of Public Health* 108 (2): 182–186.

Davidson, P.J., S. Scholar, and M. Howe. 2011. A GIS-based methodology for improving needle exchange service delivery. *The International Journal on Drug Policy* 22 (2): 140–144.

Drucker, E., P. Lurie, A. Wodak, and P. Alcabes. 1998. Measuring harm reduction: The effects of needle and syringe exchange programs and methadone maintenance on the ecology of HIV. *AIDS* 12: S217–S230.

Dunn, J.R., and M.V. Hayes. 1999. Toward a lexicon of population health. *Canadian Journal of Public Health* 90 (1): S7–S10.

Dworkis, D.A., L.A. Taylor, D.A. Peak, and B. Bearnot. 2017. Geospatial analysis of emergency department visits for targeting community-based responses to the opioid epidemic. *PLoS One* 12 (3): e0175115.

Earnest, A., M.I. Chen, D. Ng, and L.Y. Sin. 2005. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research* 5 (1): 36.

Elliott, P., and D. Wartenberg. 2004. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* 112 (9): 998–1006.

Evans, R.G., M.L. Barer, and T.R. Marmor. 1994. *Why are some people healthy and others not?: The determinants of the health of populations*. Transaction Publishers.

Fedorova, E.V., R.V. Skochilov, R. Heimer, et al. 2013. Access to syringes for HIV prevention for injection drug users in St. Petersburg, Russia: Syringe purchase test study. *BMC Public Health* 13: 183.

Fletcher-Lartey, S.M., and G. Caprarelli. 2016. Application of GIS technology in public health: Successes and challenges. *Parasitology* 143 (4): 401–415.

Florence, C., F. Luo, L. Xu, and C. Zhou. 2016. The economic burden of prescription opioid overdose, abuse and dependence in the United States, 2013. *Medical Care* 54 (10): 901.

Getis, A., and J.K. Ord. 2010. The analysis of spatial association by use of distance statistics. In *Perspectives on spatial data analysis*, 127–145. Berlin, Heidelberg: Springer.

———. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Getis A. Cliff, A.D. and J.K. Ord. 1973. Spatial autocorrelation. *London: Pion. Progress in Human Geography*. 1995;19(2):245–249. https://doi.org/10.1177/030913259501900205.

Ghertner, R., and L. Groves. 2018. The opioid crisis and economic opportunity: Geographic and economic trends. *ASPE Research Brief*: 1–22.

Guagliardo, M.F. 2004. Spatial accessibility of primary care: Concepts, methods and challenges. *International Journal of Health Geographics* 3 (1): 3.

Haffajee, R.L., L.A. Lin, A.S.B. Bohnert, and J.E. Goldstick. 2019. Characteristics of US counties with high opioid overdose mortality and low capacity to deliver medications for opioid use disorder. *JAMA Network Open* 2 (6): e196373.

Harduar Morano, L., A.L. Steege, and S.E. Luckhaupt. 2018. Occupational patterns in unintentional and undetermined drug-involved and opioid-involved overdose deaths - United States, 2007-2012. *MMWR. Morbidity and Mortality Weekly Report* 67 (33): 925–930.

Hoffman, L.A., B. Lewis, and S.J. Nixon. 2017. Opioid misuse trends in treatment seeking populations: Revised prescription opioid policy and temporally corresponding changes. *Substance Use & Misuse* 52 (14): 1850–1858.

Hudson, T.L., B.G. Klekamp, and S.D. Matthews. 2017. Local public health surveillance of heroin-related morbidity and mortality, Orange County, Florida, 2010-2014. *Public Health Reports* 132 (1_suppl): 80S–87S.

Jago, R., T. Baranowski, and J.C. Baranowski. 2006. Observed, GIS, and self-reported environmental features and adolescent physical activity. *American Journal of Health Promotion* 20 (6): 422–428.

Kimerling, A.J. 2009. Dotting the dot map, revisited. *Cartography and Geographic Information Science* 36 (2): 165–182.

Kline, D., S. Hepler, A. Bonny, and E. McKnight. 2019. A joint spatial model of opioid-associated deaths and treatment admissions in Ohio. *Annals of Epidemiology* 33: 19–23. https://doi.org/10.1016/j.annepidem.2019.02.004.

Kohli, S., K. Sahlen, A. Sivertun, O. Lofman, E. Trell, and O. Wigertz. 1995. Distance from the primary health center: A GIS method to study geographical access to health care. *Journal of Medical Systems* 19 (6): 425–436.

Kohli, S., H. Noorlind Brage, and O. Löfman. 2000. Childhood leukaemia in areas with different radon levels: A spatial and temporal analysis using GIS. *Journal of Epidemiology and Community Health* 54 (11): 822–826.

Koo, D., and S.B. Thacker. 2010. In Snow's footsteps: Commentary on shoe-leather and applied epidemiology. *American Journal of Epidemiology* 172 (6): 737–739.

Kulldorff, M., E.J. Feuer, B.A. Miller, and L.S. Freedman. 1997. Breast cancer clusters in the Northeast United States: A geographic analysis. *American Journal of Epidemiology* 146 (2): 161–170.

Lantos P.M, J. Tsao, L.E. Nigrovic, et al. Geographic expansion of Lyme disease in Michigan, 2000–2014. 2017.

Li, Z.R., E. Xie, F.W. Crawford, et al. 2019. Suspected heroin-related overdoses incidents in Cincinnati, Ohio: A spatiotemporal analysis. *PLoS Medicine* 16 (11): e1002956.

Light, R.U. 1944. The progress of medical geography. *Geographical Review* 34 (4): 636–641.

Luo, W. 2004. Using a GIS-based floating catchment method to assess areas with shortage of physicians. *Health & Place* 10 (1): 1–11.

Luo, W., and Y. Qi. 2009. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & Place* 15 (4): 1100–1107.

Luo, W., and F. Wang. 2003. Measures of spatial accessibility to health care in a GIS environment: Synthesis and a Case study in the Chicago region. *Environment and Planning B: Planning and Design* 30 (6): 865–884.

MacDonald, M., M. Law, J. Kaldor, J. Hales, and G.J. Dore. 2003. Effectiveness of needle and syringe programmes for preventing HIV transmission. *International Journal of Drug Policy* 14 (5–6): 353–357.

Macintyre, S., A. Ellaway, and S. Cummins. 2002. Place effects on health: How can we conceptualise, operationalise and measure them? *Social Science & Medicine* 55 (1): 125–139.

Manchikanti, L., S. Helm, B. Fellows, et al. 2012. Opioid epidemic in the United States. *Pain Physician* 15 (3 Suppl): ES9–E38.

Martinez, A.N., L.R. Mobley, J. Lorvick, S.P. Novak, A. Lopez, and A.H. Kral. 2014. Spatial analysis of HIV positive injection drug users in San Francisco, 1987 to 2005. *International Journal of Environmental Research and Public Health* 11 (4): 3937–3955.

McLafferty, S.L. 2003. GIS and health care. *Annual Review of Public Health* 24: 25–42.

MDPH. An Assessment of Fatal and Non-Fatal Opioid Overdoses in Massachusetts (2011–2015). http://www.mass.gov/eohhs/docs/dph/stop-addiction/legislative-report-chapter-55-aug-2017.pdf. In:2017.

Meyers, D.J., M.E. Hood, and T.J. Stopka. 2014. HIV and hepatitis C mortality in Massachusetts, 2002–2011: Spatial cluster and trend analysis of HIV and HCV using multiple cause of death. *PLoS One* 9 (12): e114822.

Monnat, S.M. 2018. Factors associated with county-level differences in U.S. drug-related mortality rates. *American Journal of Preventive Medicine* 54 (5): 611–619.

Moore, D.A., and T.E. Carpenter. 1999. Spatial analytical methods and geographic information systems: Use in health research and epidemiology. *Epidemiologic Reviews* 21 (2): 143–161.

Newburger, H.B., E.L. Birch, and S.M. Wachter. 2011. *Neighborhood and life chances: How place matters in modern America*. Philadelphia: University of Pennsylvania Press.

Nunn, A., A. Yolken, B. Cutler, et al. 2014. Geography should not be destiny: Focusing HIV/AIDS implementation research and programs on microepidemics in US neighborhoods. *American Journal of Public Health* 104 (5): 775–780.

Oliver, M.N., K.A. Matthews, M. Siadaty, F.R. Hauck, and L.W. Pickle. 2005. Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics* 4: 29.

Oliver, L.N., N. Schuurman, and A.W. Hall. 2007. Comparing circular and network buffers to examine the influence of land use on walking

for leisure and errands. *International Journal of Health Geographics* 6: 41.

Openshaw, S. 1979. A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Science*: 127–144.

———. 1984. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*.

Openshaw, S., M. Charlton, and A. Craft. 1988. Searching for leukaemia clusters using a geographical analysis machine. *Papers in Regional Science* 64 (1): 95–106.

Ortiz, M.R., N.K. Le, V. Sharma, I. Hoare, E. Quizhpe, E. Teran, et al. 2017. Post-earthquake Zika virus surge: Disaster and public health threat amid climatic conduciveness. *Scientific Reports* 7 (1): 1–10.

Overdose Prevention and Intervention Taskforce. PreventOverdose, RI © 2019. https://preventoverdoseri.org/overdose-deaths/. Accessed.

Pacurucu-Castillo, S.F., J.M. Ordóñez-Mancheno, A. Hernández-Cruz, and R.D. Alarcón. 2019. World opioid and substance use epidemic: A Latin American perspective. *Psychiatric Research and Clinical Practice* 1 (1): 32–38.

Peacock J. *Minnesota stroke registry hospital service areas and population distribution,* 2012 Minnesota Department of Health.

Peters, P.J., P. Pontones, K.W. Hoover, et al. 2016. HIV infection linked to injection use of Oxymorphone in Indiana, 2014-2015. *The New England Journal of Medicine* 375 (3): 229–239.

Pollini, R.A., A.E. Rudolph, and P. Case. 2015. Nonprescription syringe sales: A missed opportunity for HIV prevention in California. *Journal of the American Pharmaceutical Association (2003)* 55 (1): 31–40.

Rossen, L.M., D. Khan, and M. Warner. 2014. Hot spots in mortality from drug poisoning in the United States, 2007–2009. *Health & Place* 26: 14–20.

Rowe, C., G.M. Santos, E. Vittinghoff, E. Wheeler, P. Davidson, and P.O. Coffin. 2016. Neighborhood-level and spatial characteristics associated with lay naloxone reversal events and opioid overdose deaths. *Journal of Urban Health* 93 (1): 117–130.

Rudd, R.A., N. Aleshire, J.E. Zibbell, and Gladden R. Matthew. 2016. Increases in drug and opioid overdose deaths—United States, 2000–2014. *American Journal of Transplantation* 16 (4): 1323–1327.

Sakai, T., H. Suzuki, A. Sasaki, R. Saito, N. Tanabe, and K. Taniguchi. 2004. Geographic and temporal trends in influenzalike illness, Japan, 1992–1999. *Emerging Infectious Diseases* 10 (10): 1822.

Schnurrer F. Charte uÈber die geographische Ausbreitung der Krankheiten. *MuÈnchen [Hand coloured map 495 Â 345 cm World scale 1: 85,000,000 See notice in Isis 1828 XXI: 5 & 6 p 520 for a very brief note of presentation on Sept 22, 1827].* 1827.

Schoenfeld, E.R., G.S. Leibowitz, Y. Wang, et al. 2019. Geographic, temporal, and sociodemographic differences in opioid poisoning. *American Journal of Preventive Medicine* 57 (2): 153–164.

Scholl, L., P. Seth, M. Kariisa, N. Wilson, and G. Baldwin. 2018. Drug and opioid-involved overdose deaths - United States, 2013-2017. *MMWR. Morbidity and Mortality Weekly Report* 67 (5152): 1419–1427.

———. 2019. Drug and opioid-involved overdose deaths—United States, 2013–2017. *Morbidity and Mortality Weekly Report* 67 (5152): 1419.

Sifuna, P., L. Otieno, B. Andagalu, et al. 2018. A spatiotemporal analysis of HIV-associated mortality in rural Western Kenya 2011–2015. *Journal of Acquired Immune Deficiency Syndromes (1999)* 78 (5): 483.

Smith C.M, S.C. Le Comber, H. Fry, M. Bull, S. Leach, and A.C. Hayward. 2015. Spatial methods for infectious disease outbreak investigations: Systematic literature review. Euro Surveillance 20(39).

Snow J. *On the mode of communication of cholera.* John Churchill. 1855.

Somerville, N.J., J. O'Donnell, R.M. Gladden, J.E. Zibbell, T.C. Green, M. Younkin, S. Ruiz, H. Babakhanlou-Chase, M. Chan, B.P. Callis, J. Kuramoto-Crawford, H.M. Nields, and A.Y. Walley. 2017. Characteristics of fentanyl overdose - Massachusetts, 2014-2016. *MMWR. Morbidity and Mortality Weekly Report* 66 (14): 382–386. https://doi.org/10.15585/mmwr.mm6614a2.

Spiller, H., D.J. Lorenz, E.J. Bailey, and R.C. Dart. 2009. Epidemiological trends in abuse and misuse of prescription opioids. *Journal of Addictive Diseases* 28 (2): 130–136.

Stahler, G.J., J. Mennis, and D.A. Baron. 2013. Geospatial technology and the "exposome": New perspectives on addiction. *American Journal of Public Health* 103 (8): 1354–1356.

Stopka, T.J., A. Lutnick, L.D. Wenger, K. Deriemer, E.M. Geraghty, and A.H. Kral. 2012. Demographic, risk, and spatial factors associated with over-the-counter syringe purchase among injection drug users. *American Journal of Epidemiology* 176 (1): 14–23.

Stopka, T.J., E.M. Geraghty, R. Azari, E.B. Gold, and K. Deriemer. 2013. Factors associated with presence of pharmacies and pharmacies that sell syringes over-the-counter in Los Angeles County. *Journal of Urban Health* 90 (6): 1079–1090.

Stopka, T.J., C. Krawczyk, P. Gradziel, and E.M. Geraghty. 2014a. Use of spatial epidemiology and hot spot analysis to target women eligible for prenatal women, infants, and children services. *American Journal of Public Health* 104 (S1): S183–S189.

Stopka, T.J., E.M. Geraghty, R. Azari, E.B. Gold, and K. DeRiemer. 2014b. Is crime associated with over-the-counter pharmacy syringe sales? Findings from Los Angeles, California. *The International Journal on Drug Policy* 25 (2): 244–250.

Stopka, T.J., A. Donahue, M. Hutcheson, and T.C. Green. 2017a. Nonprescription naloxone and syringe sales in the midst of opioid overdose and hepatitis C virus epidemics: Massachusetts, 2015. *Journal of the American Pharmaceutical Association (2003)* 57 (2S): S34–S44.

Stopka, T.J., M.A. Goulart, D.J. Meyers, et al. 2017b. Identifying and characterizing hepatitis C virus hotspots in Massachusetts: A spatial epidemiological approach. *BMC Infectious Diseases* 17 (1): 294.

Stopka, T.J., L. Brinkley-Rubinstein, K. Johnson, et al. 2018. HIV clustering in Mississippi: Spatial epidemiological study to inform implementation science in the deep south. *JMIR Public Health and Surveillance* 4 (2): e35.

Stopka, T.J., H. Amaravadi, A.R. Kaplan, et al. 2019a. Opioid overdose deaths and potentially inappropriate opioid prescribing practices (PIP): A spatial epidemiological study. *The International Journal on Drug Policy* 68: 37–45.

Stopka, T.J., E. Jacque, P. Kelso, et al. 2019b. The opioid epidemic in rural northern New England: An approach to epidemiologic, policy, and legal surveillance. *Preventive Medicine* 128: 105740.

Thomas, K.K. 2016. *Health and humanity: A history of the Johns Hopkins Bloomberg School of Public Health, 1935–1985*. JHU Press.

Thornton, L.E., J.R. Pearce, and A.M. Kavanagh. 2011. Using geographic information systems (GIS) to assess the role of the built environment in influencing obesity: A glossary. *International Journal of Behavioral Nutrition and Physical Activity* 8 (1): 71.

Trooskin, S.B., J. Hadler, T. St Louis, and V.J. Navarro. 2005. Geospatial analysis of hepatitis C in Connecticut: A novel application of a public health tool. *Public Health* 119 (11): 1042–1047.

Tsai, P.-J., M.-L. Lin, C.-M. Chu, and C.-H. Perng. 2009. Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006. *BMC Public Health* 9 (1): 464.

Vindenes, T., M.R. Jordan, A. Tibbs, T.J. Stopka, D. Johnson, and J. Cochran. 2018. A genotypic and spatial epidemiologic analysis of Massachusetts' Mycobacterium tuberculosis cases from 2012 to 2015. *Tuberculosis (Edinburgh, Scotland)* 112: 20–26.

Wang, F., and W. Luo. 2005. Assessing spatial and nonspatial factors for healthcare access: Towards an integrated approach to defining health professional shortage areas. *Health & Place* 11 (2): 131–146.

Wangia, V., and T.I. Shireman. 2013. A review of geographic variation and geographic information systems (GIS) applications in prescrip-

tion drug use research. *Research in Social & Administrative Pharmacy* 9 (6): 666–687.

Wennberg, J. 1973. Gittelsohn. Small area variations in health care delivery. *Science* 182 (4117): 1102–1108.

Wennberg, D.E. 1998. Variation in the delivery of health care: The stakes are high. *Annals of Internal Medicine* 128 (10): 866–868.

Wodak, A., and L. McLeod. 2008. The role of harm reduction in controlling HIV among injecting drug users. *AIDS (London, England)* 22 (Suppl 2): S81.

World Health Organization. Constitution of the world health organization. 1995.

Yoo, Y., and A.P. Wheeler. 2019. Using risk terrain modeling to predict homeless related crime in Los Angeles, California. *Applied Geography* 109: 102039.

Yu, H., W.J. Alonso, L. Feng, et al. 2013. Characterization of regional influenza seasonality patterns in China and implications for vaccination strategies: Spatio-temporal modeling of surveillance data. *PLoS Medicine* 10 (11): e1001552.

Understanding Health Data by Mobility Analytics

Qiang Qu, Susheng Zhang, Seyed Mojtaba Hosseini Bamakan, Christos Doulkeridis, and George Vouros

Introduction

Basic Concepts

Space and time are two dimensions of importance in modeling data for mobility analysis, and in many realms of studies, space and time are vital in recognizing various mobility patterns for analyzing massive data. Data featuring both the dimensions of space and time is called spatiotemporal data. Spatiotemporal data have been recorded in studies such as climate science, neuroscience, social sciences, mobile health, epidemiology, transportation, criminology, and earth science (Liu and Qu 2016; Atluri et al. 2017), to generate useful information. In fact, these fields experience a rapid transformation with the proliferation of vast amount of available spatiotemporal data.

Spatiotemporal data often indicate the temporal relationship between locations or regions in space, often in time series. This is particularly useful in data sensemaking, where data collected at different locations should be exploited to generate meaningful representation of how particular features vary throughout space and time (Cao et al. 2012; Nobari et al. 2017; Zhao et al. 2017). Research into spatiotemporal data for mobility analytics has grown rapidly in recent years.

Q. Qu (✉)
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
e-mail: qiang.qu@siat.ac.cn

S. Zhang
University of Cambridge, Cambridge, UK

S. M. Hosseini Bamakan
Department of Industrial Management, Data Science Research Center, Yazd University, Yazd, Iran

C. Doulkeridis · G. Vouros
Department of Digital Systems, University of Piraeus, Piraeus, Greece

Spatial/Nonspatial Attributes and Preferences

Spatial data identifies the geographic location of features and regional boundaries. They are often stored as coordinates and topology and can be mapped. Usually they are in the form of graphic primitives that are points, lines, polygons, or pixels. For instance, a circle is defined by its center (location), its shape, its diameter (size), etc. Quite often, spatial data is multidimensional and autocorrelated. Spatial data have wide applications in archeological analysis, marketing research, and urban planning. Spatial data consists of various features in database management. The Open Geospatial Consortium (OGC) standard enlists the following features (Wikipedia 2018):

- Spatial measurements: including area, length, distance, etc.
- Spatial functions: modification of existing features including buffer area, intersections, etc.
- Spatial predicates: true/false relationships between geometries
- Geometry constructors: creation of new geometries by specification of vertices
- Observer functions: queries returning a specific feature

Spatial relationships are defined for spatial data objects, often indicating the distance and angle between objects represented by spatial data (Liu and Qu 2015). Spatial relationships between spatial data objects can be classified as topological, directional, and distance relationships (Mamoulis 2011). Topological relationships include overlap, inside, contains, and disjoint, and they are defined on the geometric extents of objects. Directional relationships (such as north/south or left/right) compare relative locations of objects with respect to some coordinate system. Finally, distance relationships express distance information between objects (such as "nearby").

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_5

Nonspatial data is independent of locations and does not concern spatial arrangement, itself. Nonspatial data may comprise aggregated data (e.g., numerical), can be numerical, but may also be in the form of text, images, or any media and modality. For example, in the monitoring of disease spread, spatial data would be the location of residence of the patients, and nonspatial data would be other information related to the patients such as index, age, diagnosis, and supporting evidence and so forth (Qu et al. 2014b).

Spatial analysis is the quantitative study of phenomena that are located in space (Bailey and Gatrell 1995), analyzing topological, geometric, or geographic properties. One important area in healthcare is *network analysis*, computing traveling times and measuring spatial proximity of spatial processes, thus outlining the relationship between different locations and the possibility of co-occurrence of events. Another area where spatial attributes come up useful is in the analysis of observational data on the movement of people between locations, i.e., in the *mobility analytics*. Tracking how people move through space is important in many healthcare applications, such as monitoring the spread of disease. In other areas such as transportation and geography, movement of people is important in understanding processes such as migration, urbanization, and decentralization. Spatial analysis is widely adopted in many research areas, as meaningful results and insights are often generated when the spatial dimension is considered (Qu et al. 2014a, b; Liu et al. 2015; Liu and Qu 2016; Nikitopoulos et al. 2016).

Temporal Concepts

Temporal data delineates the dimension of time. They can be classified into past, present, and future time. The major attributes of temporal data include (1) valid time and (2) transaction time (Jensen et al. 1994). The valid time refers to the time period during which a fact is true in the modeled reality. Such a time is represented by a variable indicating the start and the end of the valid instance, i.e., the time interval during which a fact occurs. For example, if a person contracted a disease on March 22, 2018, we indicate a start on that date for the fact "contraction of disease." On April 22, 2018, he or she was cured, and we indicate an end on that date for the same fact.

Transaction time is the time period during which a fact stored in the database is considered to be correct, demonstrating the state of the database at a given time. It can only occur in the past, until the current time. It can be quite useful when original data recorded in the database is to be updated. By having an attribute of transaction time, the original data can be labeled as invalid, while new data can be inserted and labeled as valid. This allows a record keeping of the data during a specific time interval. For example, when originally it was indicated that a person recovered from a disease on 22nd of April and it was later realized that he or she recovered on 1st of April, then we indicate the original time period from 1st of April to 22nd of April as invalid and insert the new entry of a cure on 1st of April, marked as valid.

Bitemporal data combine the valid time and the transaction time. One popular application is in bitemporal modeling, which is designed to handle historical data along two different timelines. It is useful in keeping track of past records, making it possible to recreate a past event through the database. There are various features of the temporal database that can be highly informative in managing temporal data, and these features include a time period data type, system-maintained transaction time, temporal primary keys, temporal constraints, temporal queries, predicates for querying time periods, etc. Allen's interval algebra (Allen 1983), a calculus for temporal reasoning that defines relations between time intervals (before, meets, overlaps, starts, during, finishes, equals), is widely used to express temporal predicates.

Mobility Analytics

Mobility analytics received increased attention when vast amount of data tracking the location of people at specific times in the form of spatiotemporal data are recorded by low-cost sensors and are accumulated at a rapid rate. As a result, there is increasing awareness that utilizing such data will be highly beneficial to multiple areas of research as well as real-life applications (Tan et al. 2014; Qu et al. 2015; Zhao et al. 2017).

Mobility data exploration and analytics (Pelekis and Theodoridis 2014) entail a wide range of techniques from spatiotemporal data mining, including trajectory pattern mining Liu et al. (2017), clustering, classification, outlier detection, and prediction, which can reveal useful knowledge and hidden patterns from the underlying data. In the following, an overview of applying such mobility analytics techniques to the domain of health data is provided.

Mobility Analytics Approaches for Health Data Sensemaking

Health data is becoming increasingly complex nowadays, covering clinical, administrative, financial, behavioral, and social data (Qu et al. 2015). The complexity, high dimensionality, volume, pace of arrival, diversity/heterogeneity, and interdependencies of the data have exceeded the capacity of human brain to digest and draw conclusions, and the complexity is expected to grow exponentially (Burke 2013). The field of healthcare is thus becoming more data-driven requiring more advanced analytics to generate useful information for patients and physicians. Networked sensors enable the

gathering of rich amount of health-related data collected continuously over geographical boundaries. Analysis of such data, in conjunction to archival data, has the potential to transform the healthcare landscape.

The recent development in the technology of mobility analytics holds great potential to reduce costs, streamline inefficiencies, and improve quality for the healthcare industry. New equipment and technology, such as sensors, IoT devices, mobile computers, and software (e.g., for gathering and associating data), increases our abilities to track patient activities and improves the healthcare environment by supporting the constant provision of rich information concerning the changes in patients' health conditions. For instance, healthcare technology using a real-time location system (RTLS) can identify, track, locate, and monitor the condition of patients, assisting clinical decision-making. Such a system aids in the collection of medical data and records spatiotemporal data for further analysis. In the following sections, we discuss some approaches to mobility data analysis.

Visualization Techniques

Color plays a major role in common practices for the visualization of spatial data. Color, in conjunction to graphical patterns/visual motifs, plays an important role in visualization of mobility data as it is a key discriminative attribute of our visual perception (Hassanalieragh et al. 2015). Color distance/contrast and color category enables an easier comprehension and differentiation of analytical data.

One typical tool used in the visualization of mobility data is the density map. Density maps make use of color, spatially distributed, to represent the spatial component of data in correlation to a map location. It is often used in combination with GIS tools, e.g., with ArcGIS Online,[1] to provide valuable insight into natural and social phenomena (Qu et al. 2016). Through density mapping, points or lines concentrated in a given area on the map can be represented by the intensity of colors. Figure 1 shows population density in and around Wuhan, where epidemiologists traced the likely source of the COVID-19 coronavirus outbreak to the Huanan Seafood Wholesale Market in downtown Wuhan. In the figure, darker areas indicate high population, especially the downtown region. A specific type of density map is the kernel density, which smooths point estimates to create a surface of density estimates. An example of a density map is a kernel density map recording the habit of smoking based on Twitter data (Silva 2016). This assists traditional health data analysis to identify the hot spots in aggregated rates of infection events. Therefore, historical spatiotemporal data can be collected, enriched with other features, and be analyzed, enabling the plotting of dynamic density maps with short-term, long-term, or current trends in the occurrence of events or mobility patterns.

Dynamic density maps (i.e., those showing evolution in spatiotemporal dimensions) allow analysts to look for specific spatiotemporal characteristics in combination with other features (e.g., healthcare facilities), applying various filtering techniques either regarding space or time or other features. For instance, they can enable the mapping or analysis of differences across time to demonstrate regions of healthcare improvement/escalation or changes in health data. Figure 2 shows a live example of confirmed COVID-19 coronavirus cases, which changes in terms of time scaled by the population. Other research makes use of spatiotemporal data to assess the effects of mobility on space-time clusters, the changes in geographic features on healthcare provision, the uncertainty in locational and attribute histories on estimated data, and mobility histories on exposure (Meliker and Sloan 2011).

In spatiotemporal data analysis, there are two major types of visualization techniques adopted, the individual-based and aggregation-based visualization. While both methods are popular in the research community, the visualization of dynamic human activities and movement through space and time still remains as a challenge in computation and visualization (Cao et al. 2015). Individual-based movement representation adopts a space-time path (3D polyline) to connect time-related positions with spatial coordinates. This means that the first two dimensions delineate the location of a person or object in space and the third dimension is time. This helps to visualize continuous spatial and temporal movement patterns, such as examples in the study (Cao et al. 2015). Analytical techniques for such a visualization include space-time prism, composite path-prisms, stations, bundling, and intersections (Miller 1991, 2005).

The aggregation-based visualization methods solve the problem of multiple trajectories to generalize and aggregate massive movement data. Such methods include traffic-oriented view and trajectory- oriented view (Andrienko and Andrienko 2008; Liu et al. 2015). For deeper insight of visual analysis methods for mobility data, please refer to the study by Andrienko et al. (2013).

Statistical Techniques

The most basic statistical method is to fit a standard multiple regression model to spatial data. This entails finding the best linear combination of the covariates that explains variation in the dependent variable. The residuals (the difference between the regression-predicted value and the actual value) are mapped based on locations to observe for spatial pattern. Clustering of positive or negative residuals indicates some

[1] www.arcgis.com

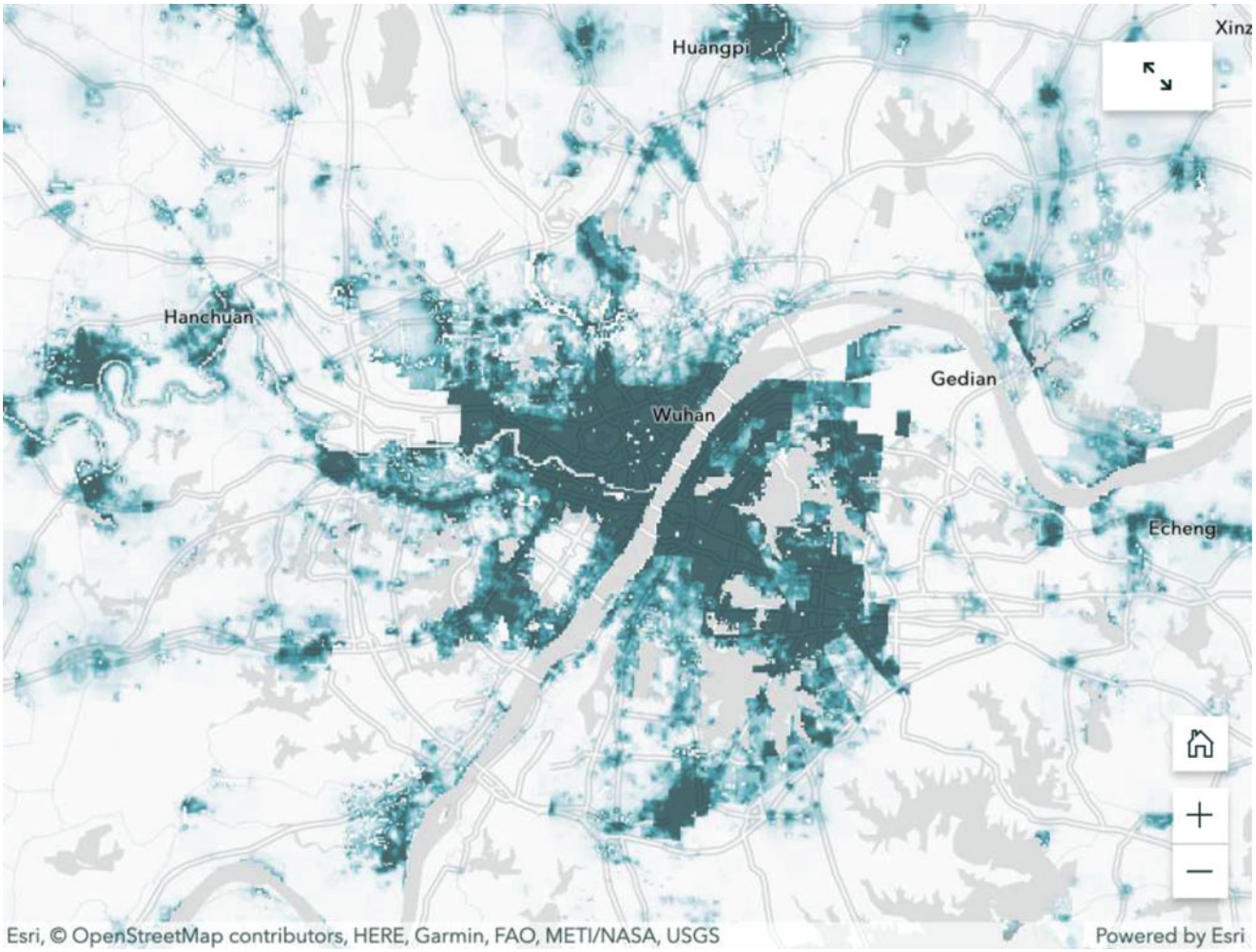


Fig. 1 A population density map of Wuhan

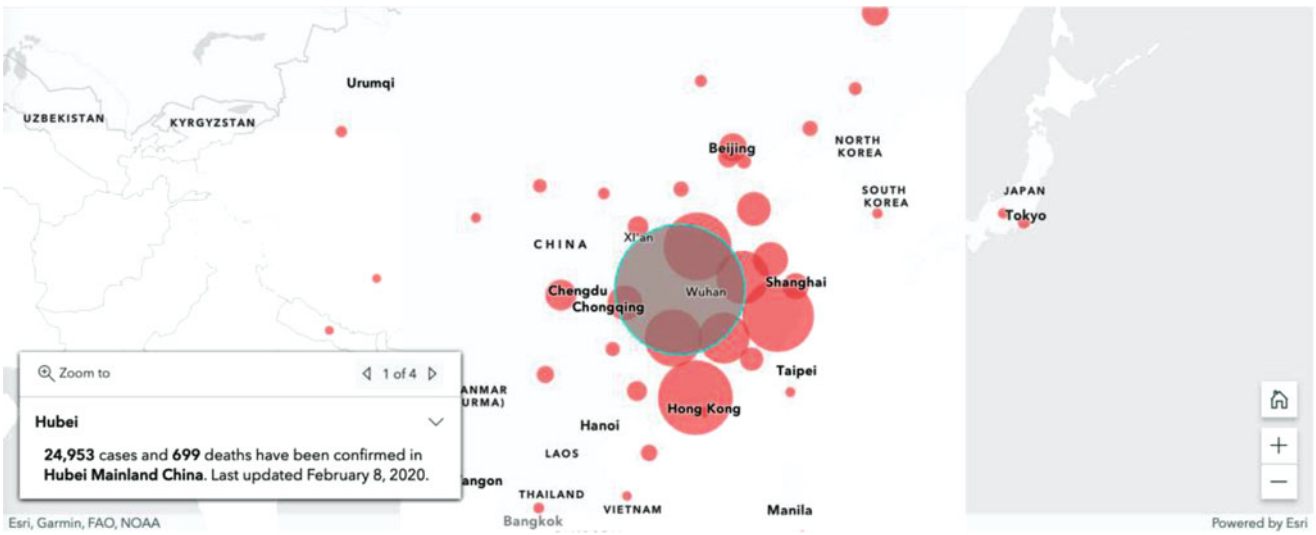


Fig. 2 A live map with updates of COVID19 coronavirus cases on February 8, 2020

locational importance in the prediction of values in that spatial region. Further modification of the model to incorporate temporal data can be performed. When there is no obvious spatial relationship in the final residuals, spatial correlation can be checked by discovering data similarities in regions.

Another prominent statistical technique for testing space-time interaction was the Knox method (Knox and Bartlett 1964). The method measures the closeness of a pair of events based on its spatial and temporal distance. Closeness between the pair of events indicates a space-time interaction. For example, food poisoning at a location at specific time may indicate a higher likelihood of food poisoning at a close-by location at the next point in time. The level of closeness can be defined in different ways depending on application, goals, and other data considered. For example, the authors define the closeness by fertility transition, i.e., transitions in a set of given figures as close in space and as close in time if they occurred in the same intercensal period (Schmertmann et al. 2010). The advantage of the method is that it is simple and straightforward to calculate. However, there are several biases related to this test. First, the population shift bias (Mantel 1967) indicates that when the rate of population growth is nonconstant for all geographic subareas, the variation in population distribution will generate space-time interactions that may not be created by the target phenomenon, such as spread of infection. Second, the choice of critical distance that defines closeness affects the accuracy of the cluster identification. Separate tests are required to justify critical distances (Gilman and Knox 1995).

Disease and Virus Propagation Analysis

In the book *On Airs, Waters and Places*, Hippocrates wrote: "Whoever wishes to investigate medicine properly, should proceed thus: in the first place to consider the seasons of the year, and what effects each of them produces for they are not at all alike, but differ much from themselves in regard to their changes. Then the winds, the hot and the cold, especially such as are common to all countries, and then such as are peculiar to each locality . . . " (Adams 1886). Thus, expanding from Hippocrates' statement, an understanding of person, place, and time has been an essential component to characterize disease events in the study of introductory epidemiology (Gerstman 2003). With the development in technology, changes in location can be tracked through time by means of trajectories and exploited as such. The study of the generated spatiotemporal data can be discussed in the context of epidemiology, and the analysis of such data, i.e., mobility analytics, has been an important constituent of modern epidemiology.

Spatial epidemiology combines the knowledge from statistics, geography, and epidemiology, focusing on the analysis of exposure, disease, and their relationships (Meliker and Sloan 2011). Early studies of spatial epidemiology involved associating a specific disease with a region. In John Snow's study, for example, a hand-drawn map of cholera infections associated the disease with the Broad Street drinking water pump (Snow 1856). Nowadays, with the use of GIS, mapping disease location has become more widely applied, and the temporal component can be included through the use of dynamic maps. This yields the development of spatiotemporal epidemiology, a branch of epidemiology that monitors disease propagation through the analytics of spatiotemporal data.

The analysis of spatiotemporal epidemiology often focuses on the following five domains (Meliker and Sloan 2011):

- Spatiotemporal epidemiologic theory
- Selection of scale of analysis comparing points with aggregated data in spatial or spatiotemporal dimension
- Pattern recognition techniques using spatial and spatiotemporal methods
- Exposure assessment in terms of spatial regions
- Assessment and consideration of spatial/temporal uncertainty

Theories regarding the tracking of infectious disease have a long tradition, focusing on key nodes in the spread of infection (longest time spent in a specific location) and the identification of episodes. From analysis of past patterns, future disease development can be predicted. Acute diseases are often tracked by short-term mobility patterns, and chronic diseases are often monitored by historical mobility, residence, and employment data of a longer time interval. In either case, an understanding of the underlying biological causes of the disease helps in developing useful theories to predict the disease spread. For example, in chronic diseases, latency usually appears as a confounding factor. Cancer patients are diagnosed not only after an accumulated exposure to carcinogens but also followed by a window of latency. Mapping of such diseases may involve patterns of both location of residence and duration spent in the region. Other diseases may have an exposure period in the utero but only develop later in life. Mapping the mobility data of these diseases may require tracing back the mobility data to the period of conception. Thus, studying etiology is important in epidemiology combined with mobility analytics (Elliott and Wartenberg 2004; Riley et al. 2003).

The scale of analysis affects the accuracy and precision of measurement by dividing a geographical zone into different-sized units. The modified areal unit problem (MAUP), for example, is a source of statistical bias that results when point-based measures of spatial phenomena are aggregated into districts. The resulting summary values are influenced

by both shape and scale of aggregation unit. Sometimes evidence of association occurs at one geographic scale but disappears or becomes inverse at other geographic scales (Dumbrell et al. 2008; Monmonier 2018). There may also be changes in the division of geographic regions due to the change in census tract boundaries. The use of small geographic unit can preserve the precision of geographic data but may become pointless for mapping of rare disease where only a few cases occur in a small area. Smoothing techniques can be used to introduce data from adjacent areas, producing smoothing disease maps (Best et al. 2005; Richardson et al. 2004; Richardson et al. 2004). However, the limitation would be the presence of artifacts and autocorrelation (Gelman and Price 1999).

The selection of methods for pattern recognition is critical and has the potential for the discovery of new etiologic factors. In identifying the patterns for spread of disease, there are usually multiple clustering statistics methods available, and research has demonstrated the effectiveness of SaTScan[2] in the analysis of aggregated data (Kulldorff et al. 2006; Meliker et al. 2009).

Assessment of exposure to environmental contaminants involves analysis of contaminants, activity, and mobility. This is discussed in Case Study 1 where contaminants in the water are modeled by concentration of contaminants, water consumption, and locations of exposure. It is a common practice to use GIS data where transport characteristics, spatial autocorrelation, and activity data can be analyzed. The recent trends of exposure assessment include the adoption of spatiotemporal dynamics in individual-level long-term exposure estimates, to retain intact spatiotemporal database and to incorporate extensive spatiotemporal data histories of mobility.

The big challenge in spatiotemporal applications is the uncertainty related to the data collected, especially those further back in time. Uncertainty can be incorporated into assessment as an attribute or location varying with time, propagated into epidemiologic analysis. However, more research into the assessment of the impact and propagation of uncertainty on epidemiologic analysis is required.

The above analysis of the five domains of spatiotemporal data applications shows that more research into the adoption of mobility data can be incorporated for epidemiologic studies. By conducting research with detailed investigations into these five domains (Meliker and Sloan 2011), the validity of the spatiotemporal models can be assessed, and simulations can be performed to evaluate whether a model performs as intended.

[2] A free software that analyzes spatial, temporal, and space-time data using the spatial, temporal, or space-time scan statistics, available at https://www.satscan.org/

In the following sections, we outline several case studies from research that involve spatiotemporal analysis of healthcare data. The case studies approach epidemiologic studies from different angles and are representative of research in the respective area.

Case Study 1: Water-Based Drug Loads Modeling

One potential area of epidemiology is the water-based epidemiology that quantitatively studies the use of illicit drugs in populations through extraction of samples from wastewater and analyzes the composition of wastewater to identify the presence and the amount of illicit drug. It involves the tracking of drug loads and exposure to environmental contaminants through connectivity to a wastewater treatment plant (Thomas et al. 2017). The endogenous and exogenous biomarkers of wastewater from humans are collected for analysis to study the presence of absorbed illicit drugs, and the average drug dose per person is calculated by estimating population size from the mobility data. The population-normalized drug loads (PNDL) are calculated as follows:

$$\text{PNDL} = \frac{C \times F}{P}$$

where C is the measured drug-biomarker concentration in wastewater, F is the total wastewater flow, and P is the mobility-analytics-derived time-weighted average population figure.

To calculate the value of P, mobile device-based data that record human mobility are used. For example, in the reference, the authors extract population patterns from the signaling data generated by handsets interacting with the mobile phone network provider within the catchment area. The study of mobile device-based population activity patterns is made possible and convenient by linking signals generated from headsets with mobile phone network provider. There is great potential to cooperate with mobile network providers in the generation of such data because the data can also be used for infrastructure deployment and maintenance. The key considerations of adopting such a system include the following:

- *Reliability:* Signaling data from mobile phones are highly reliable for analysis of population dynamics. It allows real-time tracking and represents not only the registered residents in the area but also visitors to the area. This reduces the uncertainty in population prediction and generates reliably monthly, weekly, and even daily variability in population size.
- *Dynamics:* The nature of mobility data is dynamic. A static population estimate can be altered due to commuting and holidays. This may distort the results and fail to normalize drug loads to a correct level. The use of mobile phone activity data allows dynamic monitoring of the

population. The selected period in this case study is from June to July, the summer months where people tend to commute to recreational destinations such as the seaside. Thus, the dynamic monitoring of population is important in generating accurate results.

- *Privacy:* Data preprocessing is necessary to protect the privacy of the mobile device users. The identifiers from the dataset are usually removed to prevent backtracking through methods such as IMSI,[3] MSISDN,[4] etc. The results can be further aggregated to conceal base station.

By calculating the PNDL across time, the temporal pattern of the drug load can be studied and compared, for example, by plotting the diagrams as in the study (Thomas et al. 2017). From the results of the study (Thomas et al. 2017), several conclusions can be drawn, and the patterns can be observed from the analysis of drug load data. Indicative results that can be revealed from such an analysis are as follows:

- While the absolute drug loads seem fluctuating around a constant value from June to July, the population size has dropped by 30.5% over these 2 months. Thus, the per capita consumption of drugs increased significantly from June to July.
- The ratio between weekend and weekday drug-biomarker loads represents recreational or "party" drug use. For example, the ratio for MDMA,[5] a psychoactive drug, is 2.4:1 and for cocaine is 1.6:1, indicating a high level of usage of these two drugs over the weekend for parties and recreational activities. On the other hand, amphetamine and methamphetamine have ratios of 1.2:1 and 1.1:1, indicating the little usage of the drugs for entertainments and the likelihood that these drugs are prescribed and used based on doctor's advice.

Case Study 2: Dengue Fever Patterns in French Guiana

Dengue fever is an arboviral disease spread by the mosquito of the *Aedes* genus (Monath 1989). The lack of effective disease treatment methods makes it critical to control the spread of disease vectors through mapping of risk areas and periods. In 2001, a dengue fever outbreak occurred in Iracoubo, French Guiana. This rural municipality is highly prone to disease spread as it is surrounded by rain forest, mangrove forest, and coastal wetlands, perfect environment for the breeding of mosquitoes. Dengue fever had been recognized as an endemic in the place and had been occurring repeatedly since 1965 (Tran et al. 2004). This case study experiments on the enhancement of a prevention strategy by analyzing the spatiotemporal clustering pattern of the infected patients through the GIS. The GIS was developed to record the location of patient homes and when symptoms were first observed.

Patients that visited the local healthcare center demonstrated symptoms of dengue fever, including arthralgia, headache, and myalgia, and were suspected and sent for further serological tests to obtain confirmed cases. The locations of residence of all suspected patients were recorded on a georeferenced aerial photograph, and geographic coordinates were integrated into a GIS. Nonspatial attributes, such as identification number, date of onset of symptoms, age, sex, and diagnosis, were also incorporated with the spatial GIS information.

The spatiotemporal analysis of dengue fever spread adopts a Knox method (cf. section "Statistical Techniques"), a classic space-time analysis technique to detect spatiotemporal clustering. This method is a test used to determine whether a pair of events are closer in spatial and temporal dimensions than that by chance assuming random distribution. This effectively predicts the likelihood of occurrence of infection at a specific time and space. When performed on data of suspected and confirmed cases, the location of those susceptible areas can be mapped using a relative risk map.

Initial mapping demonstrated that one case of infection can rapidly spread to all areas of the small municipality. Clusters of infection were also observed in nearby neighborhoods. Mapping of the relative risk within a space-time window was also plotted (Tran et al. 2004), where the global representation of the relative risk calculated from the confirmed cases and the three-dimensional representation of the main risk area are shown. By analyzing the plots shown in the reference, a spatiotemporal risk region of (400 m, 40 days) was identified with high-risk regions at (15 m, 6 days). This defines the area of a cluster where spread of infection is highly likely.

The study (Tran et al. 2004) uses the spatial breaks, i.e., the spatial distance between houses affecting the spread of disease, for the analysis of the risk of infection. The authors plot the relevance between spatial break and risk of the infection, indicating the relevance of the relative risk map obtained.

The risk space and time developed from spatiotemporal data analysis can be effectively used for the development of disease prevention and control strategies. During dengue epidemics, the risk area and period can be controlled to limit further spread by reducing breeding sites.

Case Study 3: Spatial Clustering of Severe Acute Respiratory Syndrome (SARS) in Hong Kong

From late 2002 to early 2003, the lethal virus of severe acute respiratory syndrome (SARS) has become widespread

[3] https://en.wikipedia.org/wiki/International_mobile_subscriber_identity

[4] https://en.wikipedia.org/wiki/MSISDN

[5] https://en.wikipedia.org/wiki/MDMA

in Hong Kong. While research has progressed much into understanding the mechanisms of spread, exploration is still needed for the "super-spreading events" (SSEs). The study of epidemiology is important in this case to understand the spread patterns for further control and prevention. Therefore, mobility analytics adopting GIS technology is applied to map and visualize SARS outbreak in Hong Kong.

Elementary analysis plotted the instances of infection at residential addresses of patients obtained through GIS. An example map of SARS-infected cases in Hong Kong from February to June 2003 can be found in the study (Lai et al. 2004), where cumulative counts of cases of infection were mapped with proportionally sized circles. Thus, the influence of SARS in Hong Kong is explicitly shown from the map.

Case Study 4: Epidemiology Study Through Social Media

The most important issue in epidemiology is the monitoring of real-time disease spread (Riley et al. 2003). However, this is often impeded by the lack of publicly available health data, and there is usually a lag between the development of symptoms and a visit to the hospital. An alternative solution to the problem is to make use of self-reported health conditions through analysis of posts on social media. This allows instant compilation and monitoring of disease distribution and enables more prompt reaction from health authorities and healthcare providers. In this case study, Twitter data is used to give a real-time modeling of infectious disease propagation. The technology combines the cutting-edge techniques of natural language processing and supervised machine learning to analyze the spread of disease, and the developed system facilitates faster response to unknown infectious diseases.

The greatest challenge in analyzing social media data is the presence of noise, subjecting the analysis of tweets to news and media hype regarding rare diseases, such as Ebola. Therefore, an important task in processing the social media data is to de-noise data. This is achieved through a staged process, aiming at noise invariance. The stages are as follows:

- Categorizing tweets into categories of self-reported, non-self-reported, and spam. Only self- reported tweets (those from infected individual or someone associated with the infected individual) are used in the subsequent analysis.
- Identifying hashtags linked to a specific disease and assigning a popularity term to rank the relevance of each hashtag.
- Identifying relevant keywords beyond disease names by obtaining the unigrams and bigrams that appeared frequently among the chosen hashtags. Then, finding disease-related tweets based on the identified keywords.

Fig. 3 Directed graph demonstrating disease relationship, with nodes indicating locations and the thickness of the edges indicating stronger connections

- Using TF-IDF (term frequency-inverse document frequency) feature vectors (Ramos 2003) to eliminate irrelevant tweets identified in the previous step.
- Clustering the tweets by means of the cosine similarity measure to group tweets with similar themes.
- Isolating salient tweet clusters by applying linguistic attribute-based random forest classifiers to randomly selected subsets of each cluster and rejecting clusters with a larger proportion of non-self-reporting tweets.
- Testing the effectiveness of the identified tweets in predicting the frequency of disease occurrence by comparing instances of infection with official data.

After processing of data, models of spread of disease using Twitter user relationships are built by means of network analysis. The locations of Twitter users were identified through Microsoft Bing Maps' reverse-geocoding API, thus obtaining a random sample of potentially infected Twitter followers. A directed graph representing locations as nodes and connections between individuals as edges was plotted, displaying the most prevalent connections between locations with thick lines as shown in Fig. 3.

The directed graph reflected the mobility patterns of people in a specific location, which produce underlying knowledge from massive dataset. For example, Mexico is well-connected with India in terms of disease spread. The two seemingly unrelated countries were well-connected due to the large number of Mexican tourists choosing India as a traveling destination. Thus, from analysis of Twitter data, the infectious disease often travels from Mexico to India,

indicating the need to construct quarantine zones in India if Mexicans were infected heavily.

Spatiotemporal Pattern Mining and Mobility Feature Learning to Health Data

Spatiotemporal Pattern Mining and Analysis

In this section, we list out some interesting applications generated from the mining of spatiotemporal data.

Case Study 1: Accessibility to Healthcare Services

Spatiotemporal pattern mining involves the use of geographic information systems (GIS) that record patients' movement across time. It has multiple applications such as the one reported in Hasan et al. (2018) and Muzammal et al. (2018).

For instance, it can be adopted to evaluate spatial accessibility to primary healthcare services. The adopted model is the floating catchment area model (Jamtsho et al. 2015). The model involves three parameters: an attractiveness component of the service center, travel time or distance between the locations of the service center and the population, and population demand for healthcare services (Jamtsho et al. 2015). In particular, nearest-neighbor modified two-step floating catchment area (NN-M2SFCA) model is an effective model often adopted by researchers to analyze mobility data. The model is defined by the following equation:

$$A_i = \sum_{j=1}^{n} \frac{S_j W_{ij} W_{ij}}{\sum_{k=1}^{m} P_k W_{kj}}$$

where A_i is the spatial accessibility index of population cluster i, n is the total number of healthcare service provider locations associated with population cluster i, S_{js} is the number of healthcare providers available at location j, W_{ij} and W_{kj} are distance weights computed using a distance decay function (e.g., step (E2SFCA) and continuous (KD2SFCA) decay functions as in the reference), m is the total number of population clusters associated with the health facility, and P_k is the population at the population cluster location k (Jamtsho et al. 2015).

The model can be understood as such: each health center has a population catchment area and each population cluster has a service catchment area. The areas are finite and overlapping. The spatial accessibility index represents the accessibility provided to a particular cluster of population to nearby healthcare centers. The index of spatial accessibility can be mapped to visualize the distribution of healthcare services. An example on accessibility index of doctor's services mapped for the country of Bhutan is shown in the study (Jamtsho et al. 2015).

When combined with temporal data, the accessibility can be evaluated across time to observe the changes in amount of healthcare providers. By plotting A_i against different clusters over years, trends of change can be observed: For example, the study reported in Jamtsho et al. (2015) shows population clusters' accessibility indices of Thimphu District, Bhutan, from 2010 to 2013.

Case Study 2: Mapping of Urban Violent Injuries

Another interesting example of application of spatiotemporal pattern mining stems from the study of epidemiology and determines the location of urban violent injury (Cusimano et al. 2010). While it is highly useful in allocating the providers of healthcare services, this can be extended to other applications such as car accident injuries. Thus, it is also of interest as a legal issue. Temporal analysis focuses on the hourly distribution of assault injuries, and spatial analysis maps out the concentration (density map) of cases of assault injuries aggregated over a 24-hour period. An example of the density map is shown in the study (Cusimano et al. 2010) where assault injury densities in Toronto are given from the EMS (i.e., Emergency Medical Services of Toronto) dataset over a 24-hour period and from the NACRS (the National Ambulatory Care Reporting System) dataset over a 24-hour period.

Spatiotemporal data is adopted to monitor the movement of urban assault through time. A dynamic analysis of the data patterns can generate useful information with regard to the spread of urban violence. For example, two distinct high-risk locations over time may be related to moving populations. Mobility data are thus useful in determining the movement of people through time, indicating the sources and locations of potential crime-related injuries and the major perpetrators. Observing the dynamic changes of the density maps, one may reveal that hot spots of urban assault shift from areas of relative social deprivation to higher income, lower residential density, and higher densities of drinking establishment throughout the day (Cusimano et al. 2010).

Predict Specific Diseases/Curing Methods

The availability of mobility data at hitherto unimagined scales and temporal longitudes can be effectively mined to transform the current post facto diagnose-and-treat reactive paradigm to a proactive framework for prognosis of disease at an incipient stage (Hassanalieragh et al. 2015). It also has the potential to allow more precise and personalized medical treatment through more targeted solutions to specific circumstances of patients, as demonstrated.

Disease Mapping

Infodemiology is the science of collecting and processing real-time data to map location of users and their input on

search engines to investigate locational diseases. In recent years with the development of mobile technology, mobile phone applications designed to assist tracking the spread of flu have been implemented under different circumstances. *Flu Near You*, for example, is an application jointly created by Skoll Global Threats Fund and the American Public Health Association. Its interactive interface obtains input from users about their self-reported symptoms before the sickness develops and gathers data regarding flu activity in a region for patients to prevent exposure. *Germ Tracker* is a similar application that obtains information regarding sickness from social media: an effective platform to identify cases unreported to doctors. Similarly, the application designed by the pharmaceutical company Help Remedies, *Help, I Have the Flu*, generates data from search engines regarding keywords such as "flu" and "cough." From the results of individuals' searches, it maps out the likelihood of a spread of disease (Nambiar et al. 2013).

Big Mobility Data Aspects

With the advent of big data, massive amounts of mobility data are made available, often requiring online processing, thereby challenging the processing capabilities of modern data management systems. Consequently, research prototypes and parallel data processing frameworks have been developed lately for spatiotemporal data, which are briefly reviewed in the following. It should be noted that even though several prototypes for big spatial data have been proposed, only few systems exist for big spatiotemporal data.

ST-Hadoop (Alarabi et al. 2017) is an open-source MapReduce extension of Hadoop tailored for spatiotemporal data processing, developed by the University of Minnesota. Support for spatiotemporal indexing is a core feature of ST-Hadoop. It is achieved by means of a multilevel temporal hierarchy of spatial indexes. Each level corresponds to a specific time resolution (e.g., day, month, etc.). Also, the entire dataset is replicated and spatiotemporally partitioned at each level based on the temporal resolution of that particular level. ST-Hadoop supports spatiotemporal range queries, aggregations, and spatiotemporal joins.

STARK (Hagedorn and Räth 2017) is another solution targeting big spatiotemporal data. STARK addresses query processing of spatiotemporal data in Spark, whereas other approaches only consider the spatial dimensions. STARK supports spatiotemporal partitioning and indexing using R-trees. Thus, it supports spatiotemporal filtering and join operations. However, the temporal dimension is not treated equally to the spatial dimensions. For example, partitioning is performed solely based on spatial criteria, and the temporal part of a query is used to filter out records that do not satisfy the temporal constraint.

Most recently, a couple of research prototypes have appeared for big trajectory data management, most notably UlTraMan (Ding et al. 2018) and DITA (Shang et al. 2018). UlTraMan (Ding et al. 2018) proposes a unified platform for the complete management cycle of big trajectory data. It provides both storage and processing layer for trajectory data. Interestingly, this is one of the few approaches that target the entire life cycle of big trajectory data, from data loading and indexing to processing and analytics. Supported query operators include range queries, KNN queries, and aggregation queries. In addition, co-movement pattern mining on trajectory data is also supported, demonstrating the trajectory analytics capabilities of UlTraMan. DITA (Shang et al. 2018) extends Apache Spark to offer in-memory trajectory analytics. It offers an extended Spark SQL language that facilitates the declarative specification of queries but also index construction.

Challenges and Concluding Remarks

Potential Challenges

Although mobility analytics in conjunction to health data has demonstrated potential to become a useful technology transforming the healthcare industry, there are several challenges related to the implementation and its widespread application.

First, conventional medical instrument evolves at a low speed because the implementation requires regulatory approval and training of medical personnel, despite the fact that electronic devices upgrade rapidly. This means that although mobility data can be collected with fair accuracy and speed and although there are means to their analysis, there is a gap to their interpretation and operational use toward improving healthcare services provided to patients and to the medical community.

Second, there lies the challenge of potential for large-scale theft or breach of sensitive data: This in conjunction to integrating sensitive data from disparate data silos that cannot be easily shared or moved from its original store presents the challenge of performing analytics using integrated views of data from disparate sources of sensitive data. Specifically, the protection of individuals' privacy by exploiting mobility data revealing real-world physical movement of individuals is essential. The use of big data analytics may require long-term availability of sensitive data, which, if not done properly, can be a potential threat to personal information security.

Fourth, there is the challenge of leveraging the patient data correlations in longitudinal records. The large volumes of patient data consist of various data types, domains, and uncertainties with underlying knowledge of both individuals and groups, which requires deep correlation analysis in space

and time for adequate understanding and usages of the medical big data.

Fifth, understanding the implicit spatiotemporal information from unstructured data (e.g., clinical notes) is a challenge as this data poses great difficulties in data manipulation (Priyanka and Kulennavar 2014), information extraction, interpretation, and exploitation in analytics.

Future Work

The most essential step for mobility analytics to become widespread in the medical world is to gain clinical support and approval. There are several future directions that this technology can be applied to enhance the level of medical care. First, it can be used in clinical decision support. Second, it can promote personalized care. Third, mobility analytics can be used to monitor public and population health. Big data from web-based and social media can be used to predict epidemics (e.g., due to flu) based on consumer's search, social content, and query activity (e.g., as reported in Priyanka and Kulennavar (2014)).

It is important to note that although mobility analytics have the advantage of generating insights to support disease control and medical policy making, it is essential to have cooperation from all stakeholders in the medical community: physicians, hospitals, pathology laboratories, and vector control agencies. Without cooperation, the results of mobility analytics will be data only, and the positive social impact it produces will be vastly limited. On the other hand, mobility analytics must promote and facilitate such a cooperation among stakeholders via the provision of analysis results that would support effective decision-making and action.

Another issue with mobility analytics is the protection of privacy in conjunction with the preprocessing of data. Medical privacy is a long-lasted issue. Future health reforms concerning the use of mobility data should aim to improve data quality and safety while preserving clinical security and transparency, e.g., considering big data and Blockchain technology (Qu et al. 2019). The ultimate aim of utilizing mobility data can serve to engage patients, improve care coordination, enhance clinical outcomes, empower individuals, and further encourage research into medical big data.

Acknowledgments This work was partially supported by the CAS Pioneer Hundred Talents Program, China, with grant number 2017063, and the National Natural Science Foundation of China with grant number 61902385, and the work of Christos Doulkeridis was supported by EU project Track&Know: Big Data for Mobility Tracking Knowledge Extraction in Urban Areas under Grant 780754.

References

Adams, F. 1886. *The genuine works of Hippocrates*. Vol. 1. New York: W. Wood.

Alarabi, L., M.F. Mokbel, and M. Musleh. 2017. ST-Hadoop: A MapReduce framework for spatio- temporal data. *SSTD* 2017: 84–104.

Allen, J.F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26 (11): 832–843.

Andrienko, G., and N. Andrienko. 2008. Spatio-temporal aggregation for visual analysis of movements. In *Proceedings of IEEE Symposium on the Visual Analytics Science and Technology*.

Andrienko, G., N. Andrienko, P. Bak, D.A. Keim, and S. Wrobel. 2013. *Visual analytics of movement*, 1–387. Berlin, Heidelberg: Springer. ISBN 978-3-642-37582-8, pp. I-XVIII.

Atluri, G., A. Karpatne, and V. Kumar. 2017. Spatio-temporal data mining: A survey of problems and methods. *arXiv preprint arXiv:1711.04710*.

Bailey, T.C., and A.C. Gatrell. 1995. *Interactive spatial data analysis*. Vol. 413. Essex: Longman Scientific & Technical.

Best, N., S. Richardson, and A. Thomson. 2005. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14 (1): 35–59.

Burke, J. 2013. *Health analytics: Gaining the insights to transform health care*. Vol. 71. Hoboken: John Wiley & Sons.

Cao, X., L. Chen, G. Cong, C.S. Jensen, Q. Qu, A. Skovsgaard, et al. 2012. Spatial keyword querying. In *Proceedings of International Conference on Conceptual Modeling*.

Cao, G., S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, and K. Soltani. 2015. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems* 51: 70–82.

Cusimano, M., S. Marshall, C. Rinner, D. Jiang, and M. Chipman. 2010. Patterns of urban violent injury: A spatio-temporal analysis. *PLoS One* 5 (1): e8669.

Ding, X., L. Chen, Y. Gao, C.S. Jensen, and H. Bao. 2018. UlTraMan: A unified platform for big trajectory data management and analytics. *PVLDB* 11 (7): 787–799.

Dumbrell, A.J., E.J. Clark, G.A. Frost, T.E. Randell, J.W. Pitchford, and J.K. Hill. 2008. Changes in species diversity following habitat disturbance are dependent on spatial scale: Theoretical and empirical evidence. *Journal of Applied Ecology* 45 (5): 1531–1539.

Elliott, P., and D. Wartenberg. 2004. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* 112 (9): 998–1006.

Gelman, A., and P.N. Price. 1999. All maps of parameter estimates are misleading. *Statistics in Medicine* 18 (23): 3221–3234.

Gerstman, B. 2003. *Epidemiology kept simple*. Hoboken: Wiley-Liss.

Gilman, E., and E. Knox. 1995. Childhood cancers: Space-time distribution in Britain. *Journal of Epidemiology & Community Health* 49 (2): 158–163.

Hagedorn, S., and T. Räth. 2017. Efficient spatio-temporal event processing with STARK. *EDBT* 2017: 570–573.

Hasan, A., Q. Qu, C. Li, L. Chen, and Q. Jiang. 2018. An effective privacy architecture to preserve user trajectories in reward-based LBS applications. *ISPRS International Journal of Geo-Information* 7 (2): 53.

Hassanalieragh, M., A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, et al. 2015. Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges. In *Proceedings of 2015 IEEE International conference on the Services Computing (SCC)*.

Jamtsho, S., R. Corner, and A. Dewan. 2015. Spatio-temporal analysis of spatial accessibility to primary health care in Bhutan. *ISPRS International Journal of Geo-Information* 4 (3): 1584–1604.

Jensen, C.S., J. Clifford, R. Elmasri, S.K. Gadia, P.J. Hayes, and S. Jajodia. 1994. A consensus glossary of temporal database concepts. *SIGMOD Record* 23 (1): 52–64.

Knox, E., and M. Bartlett. 1964. The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 13 (1): 25–30.

Kulldorff, M., C. Song, D. Gregorio, H. Samociuk, and L. DeChello. 2006. Cancer map patterns: Are they random or not? *American Journal of Preventive Medicine* 30 (2): S37–S49.

Lai, P., C. Wong, A. Hedley, S. Lo, P. Leung, J. Kong, and G. Leung. 2004. Understanding the spatial clustering of severe acute respiratory syndrome (SARS) in Hong Kong. *Environmental Health Perspectives* 112 (15): 1550.

Liu, C., and Q. Qu. 2015. Trip fare estimation study from taxi routing behaviors and localizing traces. In *Proceedings of 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

Liu, S., and Q. Qu. 2016. Dynamic collective routing using crowdsourcing data. *Transportation Research Part B Methodology* 93: 450–469.

Liu, S., Q. Qu, and S. Wang. 2015. Rationality analytics from trajectories. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10 (1): 10.

Liu, S., S. Wang, and Q. Qu. 2017. Trajectory mining. In *Encyclopedia of GIS*, 2310–2313. Cham: Springer.

Mamoulis, N. 2011. *Spatial data management*. Synthesis lectures on data management. San Rafael: Morgan & Claypool Publishers.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27 (2 Part 1): 209–220.

Meliker, J.R., and C.D. Sloan. 2011. Spatio-temporal epidemiology: Principles and opportunities. *Spatial and Spatio-Temporal Epidemiology* 2 (1): 1–9.

Meliker, J.R., G.M. Jacquez, P. Goovaerts, G. Copeland, and M. Yassine. 2009. Spatial cluster analysis of early stage breast cancer: A method for public health practice using cancer registry data. *Cancer Causes & Control* 20 (7): 1061–1069.

Miller, H.J. 1991. Modeling accessibility using space-time prism concepts within geographic information systems. *Geographical Information Systems* 5 (3): 287–301.

———. 2005. A measurement theory for time geography. *Geographical Analysis* 37: 17–45.

Monath, T.P. 1989. *The arboviruses: Epidemiology and ecology*. Vol. V. Boca Raton: CRC Press, Inc.

Monmonier, M. 2018. *How to lie with maps*. Chicago: University of Chicago Press.

Muzammal, M., M. Gohar, A.U. Rahman, Q. Qu, A. Ahmad, and G. Jeon. 2018. Trajectory mining using uncertain sensor data. *IEEE Access* 6: 4895–4903.

Nambiar, R., R. Bhardwaj, A. Sethi, and R. Vargheese. 2013. A look at challenges and opportunities of big data analytics in healthcare. In *Proceedings of 2013 IEEE International Conference on Big Data*.

Nikitopoulos, P., A.-I. Paraskevopoulos, C. Doulkeridis, N. Pelekis, and Y. Theodoridis. 2016. BigCAB: Distributed hot-spot analysis over big spatio-temporal data using apache spark (GIS Cup).In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographical Information Systems*.

Nobari, S., Q. Qu, and C.S. Jensen. 2017. In-memory spatial join: The data matters! In *Proceedings of EDBT*.

Pelekis, N., and Y. Theodoridis. 2014. *Mobility data management and exploration*, 1–298. New York: Springer. ISBN 978-1-4939-0391-7.

Priyanka, K., and N. Kulennavar. 2014. A survey on big data analytics in health care. *International Journal of Computer Science and Information Technologies* 5 (4): 5865–5868.

Qu, Q., S. Liu, B. Yang, and C.S. Jensen. 2014a. Efficient top-k spatial locality search for co-located spatial web objects. In *Proceedings of 2014 IEEE 15th International Conference on Mobile Data Management*.

———. 2014b. Integrating non-spatial preferences into spatial location queries. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*.

Qu, Q., C. Chen, C.S. Jensen, and A. Skovsgaard. 2015. Space-time aware behavioral topic modeling for microblog posts. *IEEE Database Engineering Bulletin* 38 (2): 58–67.

Qu, Q., S. Liu, F. Zhu, and C.J. Jensen. 2016. Efficient online summarization of large-scale dynamic networks. *IEEE Transactions on Knowledge and Data Engineering* 28 (12): 3231–3245.

Ramos, J. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Richardson, S., A. Thomson, N. Best, and P. Elliott. 2004. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives* 112 (9): 1016.

Riley, S., C. Fraser, C.A. Donnelly, A.C. Ghani, L.J. Abu-Raddad, A.J. Hedley, et al. 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* 300 (5627): 1961–1966.

Shang, Z., G. Li, and Z. Bao. 2018. DITA: Distributed in-memory trajectory analytics. *SIGMOD Conference* 2018: 725–740.

Silva, J.P. 2016. Mapping unhealthy behavior among economically active men using GIS in suburban and rural areas of Sri Lanka. *Asia-Pacific Journal of Public Health 28* (1_suppl): 10S–16S.

Snow, J. 1856. On the mode of communication of cholera. *Edinburgh Medical Journal* 1 (7): 668.

Tan, B., F. Zhu, Q. Qu, and S. Liu. 2014. Online community transition detection. In *Proceedings of International Conference on Web-Age Information Management*.

Thomas, K.V., A. Amador, J.A. Baz-Lomba, and M. Reid. 2017. Use of mobile device data to better estimate dynamic population size for wastewater-based epidemiology. *Environmental Science & Technology* 51 (19): 11363–11370.

Tran, A., X. Deparis, P. Dussart, J. Morvan, P. Rabarison, F. Remy, et al. 2004. Dengue spatial and temporal patterns, French Guiana, 2001. *Emerging Infectious Diseases* 10 (4): 615.

Wikipedia. 2018. Spatial database. *The Free Encyclopedia.*

Zhao, J., Q. Qu, F. Zhang, C. Xu, and S. Liu. 2017. Spatio-temporal analysis of passenger travel patterns in massive smart card data. *IEEE Transactions on Intelligent Transportation Systems* 18 (11): 3135–3146.

Health Line *Saúde24*: An Econometric Spatial Analysis of Its Use

Paula Simões, Isabel Natário, M. Lucília Carvalho, Sandra Aleixo, and Sérgio Gomes

Introduction

The attention given to spatial data in Statistics, more specifically to spatial patterns, goes back to the pioneering results of Whittle (1954), followed by other classic articles such as Besag (1974), Besag and Moran (1975), Ord (1975), and Ripley's book in (1981). The recognition of a spatial structure in the data led to the formulation of models and theories that could describe, model, and predict the phenomena that depend on their location in space. These data can be classified into three main categories, namely, geostatistical data, areal data, and point patterns data (Anselin 1988; Cressie 1993). In recent years, interest in spatial analysis in general and spatial econometrics in particular has been growing (Anselin et al. 2004), much in the social sciences (Goodchild et al. 2000). This increase is undoubtedly due to the high availability of growing volumes of geo-referenced data as well as to the development of easily manipulated technology to handle this type of geographical information (Fischer 2006; Goodchild et al. 1992). These factors potentiated new theoretical analysis perspectives of the geographical phenomena (Bivand 2008; Manski 2000).

Application of Bayesian methods in the adjustment of spatial models has spread widely, essentially due to the flexibility that this approach enables. Bayesian estimation has seen a boom in application with the development of computational methods and algorithms that use approximate methods or, more often, iterative simulation methods. In this context, Monte Carlo simulation methods via Markov chains (MCMC) have prevailed (Gamerman and Lopes 2006). This estimation approach also provides formal solutions to a wide range of spatial econometric estimation problems. Lesage has greatly contributed to the diffusion of Bayesian techniques in spatial econometrics (LeSage 1999; LeSage and Pace 2009; LeSage 2014, 2015).

Very recently, another methodological approach for Bayesian estimation has been developed using some well-known approximation methods (the Laplace approximation), known as Integrated Nested Laplace Approximation (INLA) (Rue et al. 2009). This is based on the idea of using deterministic approximations for the posterior marginal distributions of the parameters, in a computationally efficient way. Bivand, Goméz-Rubio, and Rue have explored the use of INLA for Bayesian inference in some widely used models in spatial econometrics (Bivand et al. 2014).

Despite all the several studies and developments in spatial models, spatial econometric models, and Bayesian inference, there is still a gap for handling count data within econometric models, explicitly assuming a Poisson distribution for counts. Certain types of spatial econometrics models for discrete data, such as the case of binary responses, have received more attention leading to the estimation, among others, of a spatial probit model (LeSage 1999; Bivand et al. 2014; Gomez-

P. Simões (✉)
CMA - Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, Portugal
e-mail: paula.simoes@isel.pt

I. Natário
CMA; Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, Portugal
e-mail: icn@fct.unl.pt

M. L. Carvalho
CEAUL; Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal
e-mail: mlucilia.carvalho@gmail.com

S. Aleixo
CEAUL; ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Lisboa, Portugal
e-mail: sandra.aleixo@adm.isel.pt

S. Gomes
Direção Geral de Saúde, Lisboa, Portugal
e-mail: sergiogomes@dgs.pt

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_6

Rubio et al. 2016). One of the main objectives of this work is to improve the understanding of the fundamental process behind spatial data correlation, in order to better describe the dynamics that result from this in the econometric models for count data, as well as to contribute methodologically for this.

Spatial econometrics traditionally rely on autoregressive models, such as the spatial lag model (SLM) or the spatial error model (SEM), that assume the Gaussianity of the response variable, which does not hold for counts. Consequently, their usage for count data demands data transformation to meet the assumptions of the models. New possible modelling strategies using autoregressive models for count data are investigated here in order to avoid that (Simões and Natário 2016).

However, in spatial statistics, the hierarchical modelling approach is the natural way to handle areal count data better considered under the Bayesian paradigm (Banerjee et al. 2004). This approach allows data to have any distribution, being for count data typically chosen the Poisson distribution, resulting on Bayesian Poisson hierarchical models. The spatial structure assumed for the risk is included in the first level of the hierarchy, through a prior distribution of spatially structured random effects. In addition, non-spatially structured random effects to account for risk variation can be considered.

For the hierarchical approach, spatial autocorrelation is accounted for in the disturbances and not in the observed responses, but a spatial autoregressive approach might also be considered. The fact that it is plausible to think that the risk of what is being counted in one area is related to the risk in the areas of its neighborhood, driven by effects of important covariables that may certainly impact the risk in a neighboring area, justifying the use of the autoregressive approach (LeSage and Pace 2009). In this case, the response variable in a given area is most certainly a good predictor of the response variable in its neighboring areas, meeting the different modelling strategies of the autoregressive models.

Both hierarchical and autoregressive approaches are not yet much explored for spatial econometric models for non-Gaussian data, but their development is surely more adequate for these, avoiding data transformation and corresponding to more realistic models.

The application and implementation of the methodologies studied can be considered in several areas of activity with scientific and technological interest. One main important social application is explored in this chapter, on hospital management context, more specifically the calls for the Portuguese national health line Saúde 24 (S24).

Urgency admissions is one of the most important factors regarding hospital costs, which can possibly be mitigated by the use of national health lines such as the Portuguese Saúde24 line (S24) (Portal of the National Portuguese Health Service 2015). For future development of decision support indicators in a hospital savings context, based on the economic impact of the use of S24 rather than hospital urgency services, the considered application investigates spatial dependencies in the number of calls to S24 in each Portuguese municipality for the year 2014, considering different spatial perspectives. Resorting to INLA methodology, the spatial structure is modelled through a set of autocorrelated random effects both in terms of Poisson Bayesian hierarchical models and within a spatial lag Poisson Bayesian model (Simões et al. 2017).

Spatial Modelling in Econometrics

The inclusion of spatial effects in econometric modelling evolved to form "one of the branches of econometrics" (Anselin 2010). The definition and scope of spatial econometrics has expanded substantially over the last three decades, moving from the "margins of urban and regional modeling" to the mainstream of econometric methodology (Anselin 2010).

When sample data have a location component, fundamental assumptions of traditional statistical methods are no longer guaranteed. Traditional econometrics has largely ignore this violation of the Gauss-Markov assumptions used in regression modeling (Anselin 1990). There are alternative estimation approaches that can be used when dealing with spatial data samples (LeSage 1999). An adequate alternative is to implement spatial econometric models that allow to assess the magnitude of the space influence by considering a specific weighting scheme in which relationships among spatial areas are specified (Anselin 1988). The topology or spatial pattern of data is taken care by the choice of a spatial weights or contiguity matrix, commonly denoted by the letter W, and represents our comprehension of the spatial association among data in different spatial units (Fischer 2006). Spatial econometrics is an appropriate area when dealing with data reflecting geographical events, which can accommodate spatial influences maintaining other factors or variables considered important to explain the phenomenon of interest (Anselin 2010).

Spatial Data

The availability of increasing volumes of geo-referenced data and a user-friendly technology to manipulate these in geographic information systems has been stimulating an increasing interest in spatial analysis (Anselin 2010; Fischer 2006). Data for which location attributes are taken into account cry for a spatial modelling approach. The recognition and incorporation of the spatial dimension can give more relevant results than an analysis that ignores it (Cressie 1993). Observations for which the absolute location or relative position are explicitly taken into account are defined as spatial data. Such data are the subject of many research fields, such

as epidemiology, econometrics, climatology, ecology, and sociology, among others.

Analyses of spatial data focus on detecting patterns and exploring and modelling relationships between data that form such patterns, in order to understand the processes responsible for them. Taking spatial patterns into account enables statistical analyses, for example, to emphasize the role of space as a potentially important explanatory variable of socioeconomic systems. Three main classes of spatial data can be distinguished: geostatistical or spatially continuous data, referring to observations associated with a continuous variation measure over space, taken at fixed sampling points; areal or lattice data, related to some measured attribute in partitions of the region of interest; and spatial point patterns, for which objects are the point locations where the events of interest have occurred (Cressie 1993).

Spatial Econometrics

Over the last three decades, the interpretation and range of spatial econometrics developed gradually in the literature (Anselin 2010). The definition provided by Anselin (1988) states that spatial econometrics is "the collection of spatial techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models" (Anselin 1988). At that time, when comparing spatial econometrics to standard econometrics, a definition was given for spatial econometrics: "the specific spatial aspects of data and models in regional science that precludes a straightforward application of standard econometric methods" (Anselin 1988). The referred spatial aspects may be classified into spatial dependence or spatial heterogeneity (Anselin 2010; Cressie 1993).

Twenty years later, this definition, whose subject and range were restricted to urban and regional modelling, has changed. The importance and application of spatial techniques registered an enormous growth in economics as well as in other mainstream sciences. According to Anselin (2006), spatial econometrics is now defined as a "subset of econometric methods that is concerned with spatial aspects present in cross-sectional and space time observations" (Anselin 2006, 2010).

When sample data have a location component associated, two settings can be considered: spatial autocorrelation between observations or spatial heterogeneity in relations. Under these, many fundamental assumptions of the classical statistical methods, namely, that data values are derived from independent observations or that exists a single relationship with constant variance across the sample data, are no longer guaranteed (LeSage 1999). Spatial econometrics constitutes an adequate alternative that can be used when dealing with observations linked to geographic economic phenomena or events (Fischer 2006). Variables related to location, distance, and patterns are considered in model specification, estimation, diagnostic checking, and prediction (LeSage 1999).

Similarly to what happens in any statistical modelling, four important steps that define the modern spatial econometric methodology must be followed: model specification (which deals with the formal mathematical expression for spatial dependence and spatial heterogeneity in econometric models), estimation methods, testing, and spatial prediction (Anselin 2010).

Geostatistical data, also termed field data, play an important role in environmental sciences (see, e.g., Cressie (1993) and references there in for more details) but are less important in spatial econometrics (Fischer 2006). Areal data and spatial point processes are more used in spatial econometric analyses and their applications. In this work, the focus is on areal data.

Spatial Dependence

Spatial association, also referred to as spatial autocorrelation, is present in situations where observations or spatial units are non-independent over space, that is, when nearby spatial units are associated in some way (Cressie 1993). Such association can be identified in a number of ways, for example, using a scatterplot where each observed value is plotted against the mean of observations in neighboring areas—the Moran's scatterplot—or using a spatial autocorrelation statistic such as Moran's I or Geary's C. Moran's I is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation (Carvalho and Natário 2008).

Both these statistics require the choice of a spatial weights matrix, usually symmetric and denoted by the letter W (with elements w_{ij}, $i, j = 1, \ldots, n$, where n is the number of spatial units), which represents the topology or spatial arrangement of the data and our understanding of spatial association among all areas units (Fischer 2006). Usually $w_{ii} = 0$, $i = 1, \ldots, n$, but for $i \neq j$, the association measure between area i and area j, w_{ij} can be defined in many different ways, being the most usual the contiguity criterion between areas for which $w_{ij} = 1$ only if areas i and j share a common border and $w_{ij} = 0$ elsewhere (Carvalho and Natário 2008).

Moran's scatterplot is a graph that allows to visually explore spatial autocorrelation. Considering a spatial weights matrix W, this graph has on the x axis the values of the variable of interest and on the y axis the weighted mean (by w_{ij}) of the variable values measured for the remaining spatial units. In the case that W is a contiguity matrix, the y axis corresponds to the average of the variable values of the neighbors of each spatial unit (Carvalho and Natário 2008).

In terms of interpretation, a Moran's scatterplot depicting points essentially in the odd quadrants suggests the presence of a positive (direct) spatial correlation, with high or low values of the variable of interest tending to cluster in space; if the points are shown in the even quadrants, that suggests the

presence of a negative (inverse) spatial correlation, because locations tend to be surrounded by neighbors with very dissimilar values for the same variable; points around the origin indicate no spatial correlation.

Moran's I statistics is one of the most used statistics to measure spatial association. This statistics can be used directly with the dependent variable of interest or with the residuals of a fitted model, and it is formally given by:

$$I = \frac{n \sum_i \sum_j w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{(\sum_i \sum_j w_{ij}) \sum_i (y_i - \overline{y})^2} \quad (1)$$

representing y the quantity of interest.

Using I statistics, tests for the null hypotheses of spatial independence can be built under two different situations: using a randomized distribution of the statistics or Normal approximation. A significantly positive value of I indicates the presence of direct spatial correlation, a significantly negative value, an inverse spatial correlation, and when I is close to zero the absence of spatial correlation. Note that for relatively small values of n, the I distribution may be far apart from the normal distribution approximation, and the randomized test is preferred.

Another statistic used to measure the spatial association is the Geary's C statistic, given by:

$$c = \frac{(n-1) \sum_i \sum_j w_{ij}(y_i - y_j)^2}{2\left(\sum_i \sum_j w_{ij}\right) \sum_i (y_i - \overline{y})^2} \quad (2)$$

This statistics is always positive having expected value one. Values of c less than the expected value indicate the presence of direct spatial association, while otherwise the presence of inverse spatial association is signaled (Carvalho and Natário 2008).

When spatial autocorrelation is identified, a specialized set of methods is needed (Arbia 2006; LeSage 1999). In order to capture dependencies across spatial units, spatially correlated variables can be introduced in the model specification (Anselin 2010).

The definition of the spatial weights matrix W, where the spatial topology of the spatial units is specified, is very important since estimation results may critically depend on the choice of this matrix. There are several approaches to define it, but they can essentially be classified into two main groups: spatial contiguity and distance-based approaches. Typical types of neighboring matrices for spatial contiguity approach are the linear, the rook, the bishop, and the queen contiguity matrices, described below. For the distance approach, there are, for example, the k-nearest neighbors or the critical cutoff neighborhood matrices (LeSage 1999).

- **Contiguity matrix:** Represents a $n \times n$ symmetric matrix W, with elements $w_{ij} = 1$ when i and j are neighbors and 0 when they are not. By convention, the diagonal elements are set to zero. W is usually standardized so that all rows sum to one, $\widetilde{W} = (\widetilde{w}_{ij})_{n \times n}$, with $\widetilde{w}_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$ (LeSage 1999).
  - **Rook contiguity:** Two regions are considered neighbors if they share a common border, and for these $w_{ij} = 1$.
  - **Queen contiguity:** Regions that share a common border or a vertex are considered neighbors, and for these $w_{ij} = 1$.
- **Distance approach:** Makes direct use of the latitude-longitude coordinates associated with spatial data observations for defining W (Arbia 2006).

The SAR and CAR Models

In order to capture dependencies across spatial units, through spatially correlated variables, the two most common approaches are the simultaneous autoregressive (SAR) specification and the conditional autoregressive (CAR) specification. These autoregressive specifications are frequently used to model spatial structure underlying areal data and are known as areal or lattice models. The SAR models were first presented by Whittle (1954) and the CAR models by Besag (1974), being among the most commonly used spatial statistical models. Both correspond to special cases of a general spatial process $\{y_i : i \in S\}$ for which a neighboring structure is defined based on the shape of the area, formed by a countable set of locations, the indexing set S.

Choosing a matrix W for the neighborhood structure, both models CAR and SAR incorporate spatial dependence into the model covariance structure as a function of W and a fixed unknown spatial autoregressive parameter (Wall 2004).

In what follows, consider $\{y_i : i \in S\}$ a Gaussian random process where the regions $\{S_1, \ldots, S_n\}$ constitute a partition of S, that is, $S_1 \cup \ldots \cup S_n = S$ and $S_i \cap S_j = \emptyset, \forall i \neq j$; $i, j = 1, \ldots, n$.

This process $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ can be modeled using a simultaneous autoregressive (SAR) model by:

$$y_i = \sum_j b_{ij} y_j + \epsilon_i, \; i = 1, \ldots, n$$

with $E(y_i) = 0$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim N\left(0, \sigma^2 I_n\right)$, I_n the n dimensional identity matrix, and $B = (b_{ij})_{n \times n}$ a matrix containing constants b_{ij}. B allows $\boldsymbol{y}$ to relate to itself and is called the spatial dependence matrix. This model can be written as:

$$\boldsymbol{y} = B\boldsymbol{y} + \boldsymbol{\epsilon}$$

and rewritten so that $\boldsymbol{y}$ only appears on the left side as

$$\boldsymbol{y} = (I_n - B)^{-1}\boldsymbol{\epsilon}.$$

Note that spatial units cannot depend on themselves, so matrix B must have zeros on the diagonal. Additionally, that $(I_n - B)^{-1}$ must exist.

Matrix B is responsible for the spatial dependence in the SAR model; because the error terms ϵ_i are correlated with $\{y_j : j \neq i\}$, the model is called "simultaneous," leading to the simultaneous autoregression of $\boldsymbol{y}$ on its neighborhoods. The joint distribution of $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is then given by $\boldsymbol{y} \sim N(0, \Sigma_S)$, where:

$$\Sigma_S = \sigma^2(I_n - B)^{-1}((I_n - B)^{-1})^T. \tag{3}$$

The covariance matrix must be positive definite, which is ensured by the fact $(I_n - B)^{-1}$ exists.

For SAR models, the covariance matrix, Σ_S, must therefore comply with the following conditions: $(I_n - B)$ is nonsingular, and $b_{ii} = 0, \forall i$ (Hoef et al. 2017).

The conditional autoregressive (CAR) model is another possibility to model $\{y_i : i \in S\}$, in which each element of the random process is taken conditionally on the values of the neighboring units, defined by:

$$y_i|\boldsymbol{y}_{(-i)} \sim N(\sum_j c_{ij}y_j, \tau_i^2),\ i = 1, \ldots, n,$$

where $\boldsymbol{y}_{(-i)} = \{y_j : j \neq i\}$, $E(y_i) = 0$ and τ_i^2 is the conditional variance. $C = (c_{ij})_{n\times n}$ is the spatial dependence matrix; let D be the diagonal matrix with $d_{ii} = \tau_i^2$. The conditional variance often varies with unit i.

When $(I_n - C)^{-1}D$ is positive definite, the joint distribution of $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is a multivariate normal distribution, $\boldsymbol{y} \sim N(0, \Sigma_C)$, with zero mean and variance-covariance matrix:

$$\Sigma_C = (I_n - C)^{-1}D. \tag{4}$$

Σ_C must be symmetric requiring

$$\frac{c_{ij}}{d_{ii}} = \frac{c_{ji}}{d_{jj}}, \forall i, j.$$

For CAR models, the covariance matrix Σ_C must therefore comply with the following conditions: $(I_n - C)$ has positive eigenvalues, and $\frac{c_{ij}}{d_{ii}} = \frac{c_{ji}}{d_{jj}}, \forall i, j$ (Hoef et al. 2017).

Usually B and C are constructed with a single parameter that scales a defined neighborhood matrix W, that is, $B = \rho W$ or $C = \rho W$, with W as described in section "Spatial Dependence" and predefined by the user. ρ is referred to as the spatial correlation parameter or spatial autoregressive parameter. In order to satisfy the referred conditions on $(I_n - \rho W)$, for both CAR and SAR models, the restriction $\frac{1}{\lambda_1} < \rho < \frac{1}{\lambda_N}$, with λ_1 the smallest eigenvalue and λ_N the largest eigenvalue, must hold (Hoef et al. 2017).

If W is row standardized (recommended for internal consistent (Clayton 1992)), $B = \rho\widetilde{W}$ and $C = \rho\widetilde{W}$, the expected conditional means form an average rather than a sum. In this case, the restriction for ρ becomes $\frac{1}{\lambda_1} < \rho < 1$. Usually $\frac{1}{\lambda_1} < -1$, due to irregularities for negative values near the lower bound, and $-1 < \rho < 1$ is then considered (Hoef et al. 2017; Wall 2004).

These definitions for SAR and CAR models are widely used for modelling irregular lattices on different research areas (Wall 2004), as in econometrics (Anselin and Florax 1995) or in disease mapping (Stern and Cressie 2000).

Spatial Econometric Classical Models for Continuous Data

This section summarizes some of the available spatial autoregressive econometric models that are used to model Gaussian spatial data and the corresponding classical inference. Spatial econometric models commonly employ SAR models, and inference is typically carried out with the classical maximum likelihood method. For an exhaustive review on this topic see, for example, Anselin (2010) or LeSage (1999).

Spatial Autoregressive Model

Consider a vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ of observations on n spatial units and W an $n \times n$ spatial contiguity matrix. A first-order spatial autoregressive model on the response, a SAR model, is given by:

$$\begin{aligned} \boldsymbol{y} &= \rho W\boldsymbol{y} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{5}$$

Here variation on the response y is explained as a linear combination of the response variable in neighboring units and no other explanatory variables. Parameter ρ is the autoregressive parameter. This model is frequently used for checking the existence of spatial correlation of residuals. The error term $\boldsymbol{\epsilon}$ is supposed to follow a normal distribution with zero mean and variance-covariance matrix $\sigma^2 I_n$. σ^2 is a global variance parameter.

The ordinary least squares estimation is not appropriate here. It would result on a biased estimator $\widehat{\rho}$ of the spatial autoregressive parameter ρ, leading to inconsistent estimates. With:

$$\widehat{\rho} = (\boldsymbol{y}^T W^T W\boldsymbol{y})^{-1}\boldsymbol{y}^T W^T\boldsymbol{y}$$

one has:

$$\begin{aligned} E(\widehat{\rho}) &= E[(\boldsymbol{y}^T W^T W\boldsymbol{y})^{-1}\boldsymbol{y}^T W^T(\rho W\boldsymbol{y} + \boldsymbol{\epsilon})] = \\ &= \rho + E[(\boldsymbol{y}^T W^T W\boldsymbol{y})^{-1}\boldsymbol{y}^T W^T\boldsymbol{\epsilon}] \neq \rho. \end{aligned}$$

The possible spatial dependence between the observations in the vector $\boldsymbol{y}$ prevents the consistency of the least squares estimate of ρ according to Anselin (1988).

Consequently, for estimating ρ in this model, the maximum likelihood estimatorobtained numerically from a "simplex univariate optimization routine" is commonly used. The correspondent likelihood function is (LeSage 1999; LeSage and Pace 2009):

$$L(\rho, \sigma^2|\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{(n/2)}}|I_n - \rho W| \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \rho W\boldsymbol{y})^T(\boldsymbol{y} - \rho W\boldsymbol{y})\}.$$

To simplify the maximization problem, a concentrated log likelihood function is constructed eliminating the disturbance variance parameter (LeSage 1999), by considering $\widehat{\sigma}^2$ obtained first conditionally on ρ by maximizating the conditioned log-likelihood:

$$\widehat{\sigma}^2 = \frac{1}{n}[(\boldsymbol{y} - \rho\, W\boldsymbol{y})^T(\boldsymbol{y} - \rho\, W\boldsymbol{y})]$$

in the previous likelihood function, which yields conditioned

$$\log(L(\rho, \sigma^2|\boldsymbol{y})) = -\frac{n}{2}\left(\log(\pi\frac{2}{n}) + 1\right) - \frac{n}{2}\log\left((\boldsymbol{y} - \rho W\boldsymbol{y})^T(\boldsymbol{y} - \rho W\boldsymbol{y})\right) + \log(|I_n - \rho W|).$$

Using $\widehat{\rho}$, the maximum of the previous expression, to estimate ρ, an estimate for the parameter σ^2 is provided by:

$$\widehat{\sigma}^2 = \frac{1}{n}[(\boldsymbol{y} - \widehat{\rho}\, W\boldsymbol{y})^T(\boldsymbol{y} - \widehat{\rho}\, W\boldsymbol{y})].$$

Remember that within a SAR model, a constrain is imposed on the parameter ρ. This parameter can assume viable values in the range $\frac{1}{\lambda_1} < \rho < \frac{1}{\lambda_N}$, with λ_1 the smallest eigenvalue and λ_N the largest eigenvalue of matrix W, restraining optimization search to values of ρ within this range (Anselin and Florax 1995).

Spatial Lag Model

An extension of the spatial autoregressive model is known as the spatial lag model (SLM), defined as:

$$\begin{aligned} \boldsymbol{y} &= \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{6}$$

where X is a $n \times k$ matrix of explanatory variables and the vector of parameters $\boldsymbol{\beta}$ reflects the influence of these covariates on the $\boldsymbol{y}$ variation. This model is also named in the literature as "mixed regressive-autoregressive model" (LeSage 1999), because it combines the standard regression model with a spatially dependent variable model. It can be rewritten so that the response only appears on the left-hand side as:

$$\begin{aligned} \boldsymbol{y} &= (I_n - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \Leftrightarrow \\ \boldsymbol{y} &= (I_n - \rho W)^{-1}(X\boldsymbol{\beta}) + \boldsymbol{\epsilon}', \\ \boldsymbol{\epsilon}' &\sim N(0, \Sigma), \end{aligned}$$

with $\Sigma = \sigma^2(I_n - \rho W)^{-1}((I_n - \rho W)^{-1})^T$ being the variance-covariance matrix a simultaneous autoregressive (SAR) specification (Wall 2004). As in the previous model, a maximum likelihood iterative estimation procedure is carried out in order to estimate/obtain the autoregressive parameter ρ that maximizes the likelihood function, consequently allowing the estimation of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ (LeSage 1999; LeSage and Pace 2009).

Spatial Error Model

The spatial error model (SEM), a regression model with spatial autocorrelation in the residuals, corresponding to a (SAR) model in this error terms, is defined by:

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{u} &= \lambda W\boldsymbol{u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{7}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observations on the dependent variable and X is a $n \times k$ matrix of explanatory variables, for each observation with parameters' vector $\boldsymbol{\beta}$ reflecting the influence of these variables on the variation of $\boldsymbol{y}$. W is a known $n \times n$ spatial contiguity matrix, and λ is a spatial autocorrelation parameter of the error term $\boldsymbol{u}$. The error term $\boldsymbol{\epsilon}$ is assumed to follow a normal distribution with zero mean and variance-covariance matrix $\sigma^2 I_n$.

This model can also be rewritten as:

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + (I_n - \lambda W)^{-1}\boldsymbol{\epsilon}, \Leftrightarrow \\ \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{\epsilon}', \end{aligned}$$

$$\boldsymbol{\epsilon}' \sim N(0, \Sigma),$$

with $\Sigma = \sigma^2(I_n - \lambda W)^{-1}(I_n - \lambda W^T)^{-1}$ a non-diagonal variance-covariance matrix for the error term.

As in the previous models, a maximum likelihood iterative estimation procedure is carried out that allows to estimate conditionally the value of λ. The values of the other parameters $\boldsymbol{\beta}$ and σ^2 are estimated as a function of the conditional maximum likelihood estimator of λ and of the observed data $\boldsymbol{y}$ and X (LeSage 1999; LeSage and Pace 2009).

General Spatial Model

The most general form of a spatial autoregressive model, which includes both the spatial lag term and a spatially correlated error term, is:

$$\begin{aligned} \boldsymbol{y} &= \rho W_1 \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{u} &= \lambda W_2 \boldsymbol{u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{8}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observations on the dependent variable and X is an $n \times k$ matrix of explanatory variables. W_1 e W_2 are known $n \times n$ spatial weight matrices that define spatial relations between spatial units, using the contiguity or the distance-based approach. ρ, $\boldsymbol{\beta}$, λ, $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ are defined as in the previous models.

The log likelihood function for this model is given by:

$$\ln(L(\boldsymbol{\beta}, \lambda, \rho, \sigma^2|\boldsymbol{y})) = C - \tfrac{n}{2}\ln(\sigma^2) + \ln(|A|) + \ln(|B|) - \tfrac{1}{2\sigma^2}(\boldsymbol{a}^T B^T B\boldsymbol{a}) \tag{9}$$

where C denotes an inessential constant, $\boldsymbol{a} = (A\boldsymbol{y} - X\boldsymbol{\beta})$, $A = (I_n - \rho W_1)$, $B = (I_n - \lambda W_2)$.

The log likelihood function for this model can be maximized through an optimization algorithm that allows to estimate conditionally on $\boldsymbol{\beta}$ and σ^2 the values of ρ and λ. Then values of the parameters $\boldsymbol{\beta}$ and σ^2 are estimated as a function of the maximum likelihood values of ρ, λ and the observed data $\boldsymbol{y}$ and X (LeSage 1999).

Bayesian Inference

Bayesian methods have had a huge development in the last decades and are now present in several research areas, in general, and in spatial econometric analyses, in particular. With the development of computational methods and computational algorithms which use approximate methods or, more often, iterative simulation methods, Bayesian inference has become a reality. It stands out the Monte Carlo simulation methods via Markov chains (MCMC) (Doucet et al. 2001), as well as another very recently approach developed using some well-known approximation methods (the Laplace approximation) to do Bayesian inference known as INLA—Integrated Nested Laplace Approximation (Rue et al. 2009). The fundamentals of Bayesian inference and of the enounced methods can be found, for example, in the references (Bernardo and Smith 1994; Paulino et al. 2003).

The Bayesian paradigm is based on the subjectivist interpretation of probability: the probability of a certain event measures the degree of credibility assigned to it by a certain person, in possession of evidence.

Using subjective knowledge, we establish a probability distribution for the unknown model parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ (where k can be one or more than one), $\boldsymbol{\theta} \in \Theta$, which contains or formalizes our initial beliefs or what is known about this parameter, previous to data. It is called the a priori distribution, which can be expressed in terms of the distribution probability function, represented by $h(\boldsymbol{\theta})$.

Then the data is collected. The observed data are then used to update the initial information about the parameter through its probability distribution. This results on the posterior information of $\boldsymbol{\theta}$, $h(\boldsymbol{\theta}|\boldsymbol{y})$, described as the posterior probability distribution of $\boldsymbol{\theta}$, distribution of $\boldsymbol{\theta}$ knowing or given $\boldsymbol{y} = (y_1, \ldots, y_n)$.

Observing $(y_1, y_2, \ldots, y_n)$, one has:

$$h(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d(\boldsymbol{\theta})}, \ \boldsymbol{\theta} \in \Theta,$$

where $f(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal distribution. As the left-hand side is a density for $\boldsymbol{\theta}$ and $f(\boldsymbol{y})$ is a constant, $h(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})h(\boldsymbol{\theta})$, with $L(\boldsymbol{\theta}|\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta})$ the likelihood function of $\boldsymbol{\theta}$.

Usually $f(\boldsymbol{y})$ is not possible to be obtained analytically, and numerical methods must be used.

The use of prior information in Bayesian inference requires the specification of a prior distribution for the vector of interest $\boldsymbol{\theta}$. This distribution must represent, probabilistically, the existing knowledge about $\boldsymbol{\theta}$ before performing gathering evidence. Different forms of specifying the prior distribution can be used, namely, through subjective prior distributions, through conjugated prior distributions and through non-informative prior distributions. For more details of the enounced methods see, for example, reference (Paulino et al. 2003). Within a hierarchical formulation, Simpson et al. (2017) have recently proposed a new specification for the prior distributions for the hyperparameters of the random effects, the penalised complexity (PC) priors. This formulation handles the random effects scaling and provides a way to define priors by taking the model structure into account,

needed when different types of random effects are considered in the modelling (Riebler et al. 2016).

The posterior distribution $h(\boldsymbol{\theta}|\boldsymbol{y})$ describes completely the current knowledge about $\boldsymbol{\theta}$, obtained from combining the prior information in $h(\boldsymbol{\theta})$ and the sampling information in data in $f(\boldsymbol{y}|\boldsymbol{\theta})$. It is of interest to resume this actualized information, since it is often necessary to address specific inferential questions on the parameter. A summarized description of $h(\boldsymbol{\theta}|\boldsymbol{y})$ should contain a graphical representation and quantitative summaries of the location, dispersion, and shape of the distribution. The inferences about the non-observable parameter $\boldsymbol{\theta}$ should be based on the posterior probabilities associated with different values of $\boldsymbol{\theta}$, conditioned by the particular observed value of $\boldsymbol{y}$.

In conclusion, we can say that Bayesian inference is mainly done through the evaluation and description of the posterior distribution of the parameters of interest, using several ways to summarize the available information.

The posterior distribution can be summarized in terms of the expected value of some parameter function, in case of θ be a scalar,

$$E[g(\theta)|\boldsymbol{y}] = \int g(\theta)h(\theta|\boldsymbol{y})d\theta,$$

or by marginal posterior distributions, in case $\boldsymbol{\theta}$ is multidimensional, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Thus it is frequently necessary to calculate such integrals according to the posterior distribution.

Therefore the integration of functions, often complex and multidimensional, is extremely important in Bayesian inference. Exact inference will only be possible if these integrals can be calculated analytically; otherwise, approximations must be used. When one arrives to a posterior distribution $h(\boldsymbol{\theta}|\boldsymbol{y}) \propto h(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$, often there is not an easy way of finding the integrating constant, and one must resort to some numerical techniques, such as numerical or simulation methods (Lee 2012).

In this context, the simulation methods of Monte Carlo are an appropriate alternative, whose functionality is achieved by the application of the Law of Large Numbers (Doucet et al. 2001; Gamerman and Lopes 2006).

Bayesian Inference with Markov Chain Monte Carlo (MCMC)

A MCMC method is based on the Monte Carlo integration using Markov chains. Monte Carlo integration's main idea is based on expressing an integral that we want to calculate as an expected value, being the problem of calculating an integral transformed into another problem of calculating an expected value. Monte Carlo integration draw samples from the distribution of interest (the posterior distribution) and then takes sample averages to approximate expectations. When the distribution of interest is not available, other proposed distribution is used, and the corresponding sample values are corrected to be accepted as values of the distribution of interest. This is the basis of non-iterative methods (Albert 2009; Robert and Casella 2010).

MCMC is an alternative to these methods for which sampled values are generated independently. The MCMC approach keeps the idea of obtaining a sample from the posterior distribution and calculating samples averages. However, by using an iterative simulation technique based on Markov chains, it obtains generated values that are not independent. MCMC draws these samples by running a cleverly constructed Markov chain for a long time (Hoff 2009; Robert and Casella 2010).

The initial values influence the chain behavior, but it gradually forgets these initial values, and during the process, they should be discarded—burn-in period.

The most used algorithms for implementing this method are the Metropolis-Hastings algorithm and the Gibbs Sampler, producing the desired Markov chains.

The objective of Metropolis-Hastings algorithm is to simulate from a particular distribution. This algorithm starts with an initial value $\boldsymbol{\theta}$ and specifies a rule for simulating the next value in the sequence $\boldsymbol{\theta}'$ given the previous. This rule is based on a proposal density $q(.|\boldsymbol{\theta})$ from which is simulated a candidate value. The proposal distribution may depend on the chain current value. For example, it could be a normal distribution centered in $\boldsymbol{\theta}$. Then an acceptance probability is computed, indicating the probability of that candidate value to be accepted as the next value in the sequence. This correction mechanism is responsible for the chain convergency to the equilibrium distribution (Doucet et al. 2001; Gamerman and Lopes 2006).

The Gibbs sampler considers a joint distribution $\pi(\boldsymbol{\theta})$ from which the main focus is to sample from, for example, a posterior distribution, $\pi(\boldsymbol{\theta}|y)$, where $\boldsymbol{\theta}$ is a vector of parameters. The full conditional distribution is the distribution of the ith component of $\boldsymbol{\theta}$ conditioned on all the remaining components. It is derived from the joint distribution and consists in:

$$\pi(\theta_i|\boldsymbol{\theta}_{-i}) = \frac{\pi(\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})d(\theta_i)},$$

where $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$.

If the full conditional distribution for all parameters are known, the Gibbs sampler could be used to sample from the joint distribution, where the state transitions are made according to these distributions (Hoff 2009). An advantage of Gibbs sampler method is that the chain always moves to a new value; there is not rejection. One disadvantage is that we have to know all the full conditional distributions. If the full conditional distributions are known and can be sampled from, then the Gibbs sampling proceeds (Gamerman and Lopes 2006).

Monte Carlo and MCMC sampling methods cannot be seen as models, and they do not bring anymore information that already is in data $\boldsymbol{y}$ and in $h(\boldsymbol{\theta})$. They constitute a "simple way" of looking to the posterior distribution of $h(\boldsymbol{\theta}|\boldsymbol{y})$ and help to describe it, in order to make inference about the parameter $\boldsymbol{\theta}$ (Gamerman and Lopes 2006; Hoff 2009; Robert and Casella 2010).

In terms of the algorithm convergence, one must be aware that the sample size should be large enough. Markov chain convergency for the target distribution should be motorized and evaluated. Some important issues related with the Markov chain convergency must be addressed. It is necessary to ensure that the equilibrium condition for the generated chain was reached. Through graphical or numerical diagnostics, it is possible to decide if the chain has sufficiently explored the entire posterior distribution, assessing convergence. The Geweke method (Geweke 1991) or the approach suggested by Gelman and Rubin (1992), for example, allows to carry out this analysis.

The Gelman and Rubin approach suggests comparing the behavior of several generated chains in terms of the variance of some summary statistics as a way of monitoring convergence of a MCMC chain, as in the one-way analysis of variance (ANOVA): the between-sample and within-sample variances.

Let υ be a scalar summary statistics that estimates some parameter of the distribution of interest. Consider the generated values of the k chains $\{C_{ij} : 1 \leq i \leq k, 1 \leq j \leq m\}$ of length m. Compute $\{\upsilon_{im} = \upsilon(C_{i1}, \ldots, C_{im})\}$ for each chain. If the chains are converging to the distribution of interest, as $m \to \infty$, the sampling distributions of the considered statistics should converge to the same distribution (Rizzo 2007).

To estimate an upper bound and a lower bound for the variance of the summary statistic υ, $Var(\upsilon)$, this approach uses the between-sequence variance and the within-sequence variance of υ, which converges to the variance υ from above and below, respectively, as the chain converges to the distribution of interest.

The between-sequence variance is given by:

$$B = \frac{m}{k-1} \sum_{i=1}^{k} (\overline{\upsilon}_{i.} - \overline{\upsilon}_{..})^2 . \tag{10}$$

where $\overline{\upsilon}_{i.} = \frac{1}{m} \sum_{i=1}^{m} (\upsilon_{ij})$ and $\overline{\upsilon}_{..} = (\frac{1}{mk}) \sum_{i=1}^{k} \sum_{j=1}^{m} \upsilon_{ij.}$.

The sample variance, within the ith sequence, is:

$$s_i^2 = \frac{1}{m} \sum_{j=1}^{m} (\upsilon_{ij} - \overline{\upsilon}_{i.})^2.$$

The estimate of the within sample variance is:

$$W = \frac{1}{k} \sum_{i=1}^{k} s_i^2. \tag{11}$$

The between-sequence and within-sample estimates of the variance are combined to estimate an upper bound for $Var(\upsilon)$, given by:

$$\widehat{Var}(\upsilon) = \frac{m-1}{m} W + \frac{1}{m} B. \tag{12}$$

which is an unbiased estimator of $Var(\upsilon)$ if the chains can be considered as random samples from the distribution of interest (Rizzo 2007).

The Gelman-Rubin statistic is the estimated potential scale reduction:

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{Var}(\upsilon)}{W}} \tag{13}$$

$\sqrt{\widehat{R}}$ should be closer to one if the chains have approximately converged to the distribution of interest. It is suggested that $\widehat{R}$ should be less than 1.1 or 1.2 (Gelman and Rubin 1992).

In terms of the Geweke method, this is based on the application of the usual techniques of time series to monitor convergence of simulated sequences by MCMC. Considering the MCMC simulated values $\{\boldsymbol{\theta}^{(t)}, t = 0, 1, \ldots\}$ and being υ a function of $\boldsymbol{\theta}$, $\upsilon(\boldsymbol{\theta}^{(t)})$ defines a temporal series. Having a sufficiently large number of iterations M, the mean of the first m_f iterations, as well as the mean of the m_l last iterations, $\upsilon_f = \frac{1}{m_f} \sum \upsilon(\boldsymbol{\theta}^{(t)})$ and $\upsilon_l = \frac{1}{m_l} \sum \upsilon(\boldsymbol{\theta}^{(t)})$ are calculated.

If the chain converges then the referred means should be similar. Considering $\frac{m_f}{M}$ and $\frac{m_l}{M}$ fixed:

$$\frac{(\upsilon_f - \upsilon_l)}{\sqrt{\frac{s_f^2}{m_f} + \frac{s_l^2}{m_l}}}$$

is asymptotic normal (0,1) distributed ($M \to \infty$), with s_f^2 and s_l^2 independent estimates of the asymptotic variances of υ_f and υ_l, respectively.

So, in terms of the convergence diagnostic for the samples proposed by Geweke method, values within the interval $(-1.96, 1.96)$ for this statistic are indicative of convergence. For more details of the enounced methods, see, for example, references (Turkman and Paulino 2015; Rizzo 2007).

Bayesian Inference with Integrated Nested Laplace Approximation

Recently Rue et al. (2009) have developed an approximate method, known as the Integrated Nested Laplace Approximation (INLA), which allows to estimate the marginal posterior distribution of the parameters of interest in a Bayesian model, being particularly efficient in the estimation of latent Gaussian models and capable of providing accurate and fast results (Blangiardo et al. 2015; Rue et al. 2009). It is quite general in the type of model that it can fit, allowing for great automation of the inferential process. Nowadays, this

method is implemented in the package R-INLA of the R software. Next, it is described the INLA methodology and the corresponding context in which it should be applied, according to Blangiardo et al. (2015) and Natário (2013).

Consider n observed values of the response variable, $\boldsymbol{y} = (y_1, \ldots, y_n)$, which are assumed to be distributed according to one of the distributions in the exponential family, with mean parameter μ_i, related to a linear predictor through a link function $g(.)$:

$$g(\mu_i) = \eta_i.$$

This linear predictor η_i is defined as:

$$\eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \sum_{j=1}^{J} f_j(z_{ji}) \qquad (14)$$

where β_0 is a scalar that represents the intercept, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$ the linear effects of chosen covariates $\boldsymbol{X} = (X_1, \ldots, X_K)$ on the response, and $\boldsymbol{f} = \{f_1(.), \ldots, f_J(.)\}$ one non-linear effects, functions of variables $\boldsymbol{z} = (z_1, \ldots, z_J)$. The vector oflatent effects, $\boldsymbol{u} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{f})$, forms a Gaussian Markov Random Field (GMRF) with precision matrix $Q(\boldsymbol{\theta}_2)$, $\pi(\boldsymbol{u}|\boldsymbol{\theta}_2) \equiv N(0, Q^{-1}(\boldsymbol{\theta}_2))$ where $\boldsymbol{\theta}_2$ is a vector of hyperparameters. The distribution of $\boldsymbol{y}$ will depend on a number of parameters $\boldsymbol{\theta}_1$.

This class of models is very flexible, and the terms $f_j(.)$ can assume many different forms as nonlinear effects of covariates, seasonal effects, or temporal or spatial random effects, covering generalized linear models, hierarchical models, and spatial and spatiotemporal models.

Let $\pi(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}_1)$ be the conditional density function of $\boldsymbol{y}$. Assuming conditional independence given $\boldsymbol{u}$ and $\boldsymbol{\theta}_1$, the distribution of the n observations is given by:

$$\pi(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}_1) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{u}, \boldsymbol{\theta}_1). \qquad (15)$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be a single vector of parameters with prior density function $\pi(\boldsymbol{\theta})$. The posterior distribution of the latent effects $\boldsymbol{u}$ and of the parameters $\boldsymbol{\theta}$, with precision matrix $Q(\boldsymbol{\theta})$, is given by:

$$\begin{aligned}
\pi(\boldsymbol{u}, \boldsymbol{\theta}|y) &\propto \pi(\boldsymbol{\theta}) \times \pi(\boldsymbol{u}|\boldsymbol{\theta}) \times \pi(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta}) \times \pi(\boldsymbol{u}|\boldsymbol{\theta}) \times \textstyle\prod_{i=1}^{n} \pi(y_i|u_i, \boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta}) \times |Q(\boldsymbol{\theta})|^{\frac{1}{2}} \exp(-\tfrac{1}{2}\boldsymbol{u}^T Q(\boldsymbol{\theta})\boldsymbol{u}) \times \textstyle\prod_{i=1}^{n} \exp(\log(\pi(y_i|u_i, \boldsymbol{\theta}))) \\
&\propto \pi(\boldsymbol{\theta}) \times |Q(\boldsymbol{\theta})|^{\frac{1}{2}} \exp[(-\tfrac{1}{2}\boldsymbol{u}^T Q(\boldsymbol{\theta})\boldsymbol{u}) + \textstyle\sum_{i=1}^{n} \log(\pi(y_i|u_i, \boldsymbol{\theta}))],
\end{aligned} \qquad (16)$$

corresponding to the product of the likelihood (15), the GMRF prior density function for $\boldsymbol{u}$, and the parameter prior distribution $\pi(\boldsymbol{\theta})$.

INLA approach does not estimate the posterior marginal distributions of the latent effects $\pi(u_i|\boldsymbol{y})$ and the hyperparameters $\pi(\theta_k|\boldsymbol{y})$, given by:

$$\pi(u_i|\boldsymbol{y}) = \int \pi(u_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

$$\pi(\theta_k|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k},$$

but rather the whole posterior distribution, by constructing "nested approximations," numerical approximations based on the Laplace approximation method. This method allows one to approximate density functions by the first terms of Taylor series expansion of the log of the densities:

$$\widetilde{\pi}(u_i|\boldsymbol{y}) = \int \widetilde{\pi}(u_i|\boldsymbol{\theta}, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

$$\widetilde{\pi}(\theta_k|\boldsymbol{y}) = \int \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k},$$

where $\widetilde{\pi}$ corresponds to the approximate density function. The proposed Laplace approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is then given by:

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta})\pi(\boldsymbol{u}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{u}=\boldsymbol{u}^*(\boldsymbol{\theta})}, \qquad (17)$$

where $\widetilde{\pi}(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation, given by the Laplace approximation method, of the complete conditional distribution of $\boldsymbol{u}$, $\pi(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{y})$, and $\boldsymbol{u}^*(\boldsymbol{\theta})$ is the mode for a given $\boldsymbol{\theta}$.

The INLA approximation of $\pi(u_i|\boldsymbol{y})$ follows three main steps:

1. Computation of an approximation to the posterior distribution of the hyperparameters, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, as in (17);
2. New use of the Laplace approximation to obtain $\pi(u_i|\boldsymbol{\theta}, \boldsymbol{y})$. For example, rewriting the vector of parameters as $u = (u_i, \boldsymbol{u}_{-i})$:

$$\widetilde{\pi}(u_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \frac{\pi(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})}{\widetilde{\pi}(\boldsymbol{u}_{-i}|u_i, \boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{u}_{-i}=\boldsymbol{u}^*_{-i}(u_i,\boldsymbol{\theta})}$$

   where $\widetilde{\pi}(u_{-i}|u_i, \boldsymbol{\theta}, \boldsymbol{y})$ is the Laplace Gaussian approximation to $\pi(u_{-i}|u_i, \boldsymbol{\theta}, \boldsymbol{y})$ and $\boldsymbol{u}^*_{-i}(u_i, \boldsymbol{\theta})$ is its mode;
3. Using the previous steps and a numerical integration:

$$\widetilde{\pi}(u_i|\boldsymbol{y}) = \int \widetilde{\pi}(u_i|\boldsymbol{\theta}, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

can be solved through a finite weighted sum:

$$\widetilde{\pi}(u_i|\boldsymbol{y}) \approx \sum_m \widetilde{\pi}(u_i|\boldsymbol{\theta}^m, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}^m|\boldsymbol{y})\Delta_m$$

considering that $\boldsymbol{\theta}$ has m relevant elements $\{\boldsymbol{\theta}^m\}$, with a corresponding set of weights Δ_m, where m small.

Software

Statistical software plays a fundamental role in empirical studies. This section reviews some of the available software for spatial econometric analysis, in an areal data context.

In the past, difficulties associated with computer power, for the necessary routines, led to absence of dedicated software and consequently to slow diffusion of empirical studies in spatial econometric analysis. In recent years, this scenario has changed, and several options for applying spatial econometrics methodologies for real cases are available to researchers. Currently, for spatial data analysis and modelling, the SpaceStat software (Jacquez et al. 2014), the GeoDa software (Anselin et al. 2006), and some packages of the R software such as the spdep package for spatial regression analyses stand out (Anselin 2007; Bivand et al. 2014).

Bayesian inference has become a reality with the development of computational methods and computational algorithms. It has been largely used in spatial statistics in recent years, mainly due to the availability of these computational methods for fitting spatial models.

The R software has been a huge contribution for this development. It made available to the scientific community, as a free software, which allows the implementation of this methodology. Also noteworthy are another five specific softwares for fitting Bayesian models, for general purpose, using Markov chain Monte Carlo methods, namely, Jags (Plummer 2003), OpenBugs (Spiegelhalter et al. 2003), BayesX (Belitz et al. 2015), Stan (Stan 2014), and Nimble (Valpine et al. 2017) softwares. These may also be used in connection with the R software, in terms of monitoring chains convergence, which can be done using the R-packages CODA (Plummer et al. 2006) and BOA (Smith 2004; Turkman and Paulino 2015).

There are also specialized spatial modelling packages developed under software R that implement MCMC for more complex Bayesian models (which will in the meantime be presented) such as CARBayes (Lee 2013).

The R-INLA package offers an interface to INLA methodology being adequate for estimating a large number of the most common models. Simultaneously with the INLA methodology development, their authors have been developing a set of R-functions (R-INLA) to implement the method. Initially, in simple modelling settings, they have greatly developed since then, such that nowadays a huge number of models are already covered and readily available to be fitted in R-INLA. However, it is also possible for the user to develop new models (Rue et al. 2012).

Spatial Count Data Modelling in Econometrics

Understanding geographical variation in discrete data, in terms of highlighting some kind of spatial association, is less developed when compared to existing standard methods for continuous data. Two different modelling approaches are available to study spatial patterns in count data.

One, which is more used in econometrics, is the employment of classical autoregressive econometric models, as the ones presented in section "Spatial Econometric Classical Models for Continuous Data". These models were originally designed for continuous data, thus demanding count data transformation to meet the model's assumptions (Arbia 2006). However, there are some important exceptions, essentially for certain types of discrete spatial data, such as the case of a binary outcome with Bernoulli distribution. For this, the spatial autoregressive lag specification has been extended through the spatial probit model or spatial tobit models (LeSage 1999; LeSage and Pace 2009).

Another approach, more used in statistics, is the use of hierarchical Bayesian models, where data can be modeled as having any distribution, being Poisson the natural choice for count data. In spatial statistics, hierarchical modelling is the natural way to handle areal count data to account for data overdispersion as it is well established in the literature (Banerjee et al. 2004; Cressie 1993). Poisson log-linear models are typically used for the analyses where the linear predictor includes important factors for the phenomenon explanation. For that, non-observable random effects can be added to the effects of existing covariates in modelling extra variation that might exist in counts. This section provides an overview of the considered approaches to model Poisson count data.

Spatial patterns can be modelled differently through autoregressive models, very common in spatial econometrics literature, in which spatial dependence is included in a way such that the value of one observation is dependent on the value of its neighboring observations (Bivand et al. 2014). The autoregressive approach is also valid for count data when it is plausible to think that the space relation between these counts is driven by the effects of covariates, whose values in one area may impact the counts in that area is neighborhood, even if those variables are not considered in the model (LeSage and Pace 2009).

The majority of the classical spatial autoregressive econometric models assume a continuous response variable. How-

ever, there are alternatives for modelling counts that are explored here.

Hierarchical Bayesian Spatial Models for Count Data

In order to use the traditional spatial econometric models for continuous data to model count data, defined into spatial units of a lattice, it is necessary to transform the discrete dependent variable to meet the required assumptions (LeSage 1999). However, there are some alternative models which can be applied directly to count data wherein the spatial dependency structure is defined conditionally.

Part of the spatial autocorrelation can be accommodated by including known covariate risk factors in a generalized linear regression model.

The generalized linear models (GLM) introduced by Nelder and Wedderburn (1972) have been playing an increasingly important role in statistical analysis, due to the large number of models that they encompass and facility of analysis associated with the rapid computer development, in responding to situations which are not properly explained by the normal linear model (Turkman 2000).

In a GLM, the outcome $\boldsymbol{y}$ is assumed to be distributed as a member of the exponential family of distributions, with mean parameter μ, such as the Poisson distribution.

A family of distributions is said to belong to the exponential family if its probability or density function $f(\boldsymbol{y}|\boldsymbol{\theta})$ can be expressed as:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = h(\boldsymbol{y})c(\boldsymbol{\theta})exp(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(\boldsymbol{y})),$$

Here, $h(\boldsymbol{y}) \geq 0, t_1(\boldsymbol{y}), \ldots, t_k(\boldsymbol{y})$ are real-valued functions of $\boldsymbol{y}$, not dependent on $\boldsymbol{\theta}$, $c(\boldsymbol{\theta}) \geq 0$, and $w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})$ are real-valued functions of the possibly vector-value parameter $\boldsymbol{\theta}$, not dependent on $\boldsymbol{y}$.

Consider a vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ of observations and a vector of covariates $\boldsymbol{X}^T = (X_1, .., X_k)$ with parameters $\beta_1, \ldots, \beta_k$. The relationship between the mean of the ith observation on the dependent variable, μ_i, and a linear predictor on the vector of covariates of the ith observation, $\boldsymbol{X}_i$, $i = 1, \ldots, n$ defines the systematic component of the GLM model and is established through a link function $g(.)$:

$$g(\mu_i) = \eta_i = \boldsymbol{X}_i^T \beta.$$

Suppose that y_i is Poisson (μ_i) distribuited, $i = 1, \ldots, n$. A possible and more common link function is the logarithmic function, resulting into the general log-Poisson regression model defined as:

$$y_i|\boldsymbol{\beta} \backsim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \eta_i = \boldsymbol{X}_i^T \beta\,.$$

It is common that some spatial structure remains in the residuals of this fitted model, even after accounting for these covariate effects. In modelling the residual autocorrelation, the most common approach is to expand the linear predictor with a set of spatially correlated random effects, in terms of a generalized linear model with random effects (Banerjee et al. 2004). The generalized linear models with random effects are a different way of modelling the outcome $\boldsymbol{y}$ considering covariates and random effects, either spatially structured or not, to account for spatial autocorrelation in the analysis of spatial data (McCullagh and Nelder 1989).

The referred spatial random effects are usually modeled by a conditional autoregressive (CAR) model (Besag et al. 1991), which induces a priori spatial autocorrelation through the contiguity structure of the spatial units. Different CAR prior distributions commonly used for modelling spatial autocorrelation have been established in the literature: from the Besag, York, and Mollié (BYM) proposal (Besag et al. 1991) to the alternatives developed by Leroux et al. (1999) and Stern and Cressie (1999), where each model is a special case of a Gaussian Markov Random Field (GMRF).

The general model is a generalized linear mixed model for spatial areal unit data, a hierarchical model where the responses $\boldsymbol{y}$ are assumed to be Poisson distributed, better handled under the Bayesian paradigm.

The next subsection describes and explains different Bayesian hierarchical models for Poisson count data.

Hierarchical Log-Poisson Regression Models

Considering a spatial domain divided into n spatial units (or areas), let $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{e} = (e_1, \ldots, e_n)$ represent, respectively, the number of observed and expected cases of the phenomena that is being counted in each spatial unit, the latter obtained by some standardization procedure. The counts y_i are assumed to be Poisson distributed with expected value $E(y_i) = \mu_i = e_i\,\theta_i$, where θ_i is the relative risk in area i. Let $\boldsymbol{X}_i^T = (X_{i1}, \ldots, X_{ik})$ denote a set of k covariates measured in spatial unit i, for $i = 1, \ldots, n$, the first of which, X_{i1} corresponds to an intercept term and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ the corresponding regression coefficients.

The general hierarchical log-Poisson regression model is defined as:

$$y_i|\eta_i \sim \text{Poisson}(e_i\theta_i), \tag{18}$$

where $\eta_i = \log(\theta_i)$, $i = 1, \ldots, n$, are the log relative risks (Dass et al. 2010; Lee 2013). Note that $\log(e_i)$ enter as known offsets in the model. In (18) the log relative risks are decomposed into the effects of covariates plus some random

effects that are able to account for possible over-dispersion:

$$\eta_i = \log(\theta_i) = \boldsymbol{X}_i^T \beta + u_i. \tag{19}$$

Under formulation (19), it is possible to have a spatial hierarchical log-Poisson model if u_i includes spatially autocorrelated random effects ε_i, modelled by a conditional autoregressive (CAR) prior distribution.

Therefore, when spatial autocorrelation is detected in data, the spatial structure can be considered through a global CAR prior. From the existing possibilities, the ones to be used in this thesis are the Besag-York-Mollié and the Leroux models. The CAR specification defines prior conditional distributions of the spatial random effects ε, where the distribution of ε_i conditioned on $\varepsilon_{-i} = (\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n)$ is dependent only on the ε_j that are neighbours, according to the chosen spatial structure. CAR prior is then specified as a set of n univariate full conditional distributions, $f(\varepsilon_i|\varepsilon_{-i})$, for $i = 1, \ldots, n$, rather than via the multivariate specification (Besag 1974).

It is necessary to establish a neighboring criterion, by considering a symmetric non-negative weight matrix (an adjacency matrix in this work) W with elements w_{ij}, $i, j = 1, \ldots, n$, where n is the number of spatial units. Here it is considered the contiguity criterion between areas for which $w_{ij} = 1$ only if areas i and j share a common border and $w_{ij} = 0$ elsewhere (Carvalho and Natário 2008).

For the regression coefficients β_j, $j = 1, \ldots, k$ normal $(\mu_\beta, \sigma_\beta^2)$ prior distributions are considered.

Besag-York-Mollié (BYM) Model

The BYM model (Besag et al. 1991) comprises two sets of random effects, spatially correlated ε_i and unstructured γ_i random effects, that is, $u_i = \varepsilon_i + \gamma_i$ in (19). The unstructured random effects partially account for possible effects of over-dispersion and are implemented with the exchangeable prior:

$$\gamma_i \sim \mathrm{N}(0, \sigma^2), \tag{20}$$

with an an inverse-gamma prior distribution assigned to the variance parameter:

$$\sigma^2 \sim \text{Inverse-Gamma}(a, b). \tag{21}$$

For the spatial random effects, a CAR prior is proposed, where the conditional expectation of each effect is given as the average of the random effects in its neighboring areas, while the conditional variance is inversely proportional to the number of neighbors. Thus, the more areas that are close to area i and have similar values to ε_i results in reducing uncertainty. The prior distribution of the random effects is given by:

$$\varepsilon_i|\varepsilon_{-i} \sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{ij}\varepsilon_j}{\sum_{j=1}^n w_{ij}}, \frac{\sigma_B^2}{\sum_{j=1}^n w_{ij}}\right), \tag{22}$$

with an inverse-gamma prior distribution assigned to σ_B^2:

$$\sigma_B^2 \sim \text{Inverse-Gamma}(a, b). \tag{23}$$

This model accommodates both weak and strong spatial autocorrelation. The spatial structure is split into strongly spatial correlated variation and independent spatial variation.

Leroux, Lei, and Breslow Model

The previous model requires two random effects to be estimated for each data point, whereas only their sum is identifiable from data. To get through this, Leroux et al. (1999) proposed an alternative CAR prior distribution for modelling spatial autocorrelation, using a single set of random effects for modelling varying strengths of spatial autocorrelation, that is, $u_i = \varepsilon_i$ in (19). The prior distribution of the random effects is given by:

$$\begin{aligned} \varepsilon_i|\varepsilon_{-i} &\sim \mathrm{N}\left(\frac{\rho\sum_{j=1}^n w_{ij}\varepsilon_j}{\rho\sum_{j=1}^n w_{ij} + 1 - \rho}, \frac{\sigma_L^2}{\rho\sum_{j=1}^n w_{ij} + 1 - \rho}\right) \\ \sigma_L^2 &\sim \text{Inverse-Gamma}(a, b), \\ \rho &\sim \text{Uniform}(0, 1), \end{aligned} \tag{24}$$

where ρ is a spatial dependence parameter, with value zero in case of independence and values near one for strong spatial autocorrelation. A uniform prior distribution on the unit interval is specified for this parameter, ρ, and an inverse-gamma prior distribution is adopted for the variance of the random effects, σ_L^2. This model formulation makes a compromise between unstructured and structured variation using ρ as a mixing parameter.

The CAR prior distributions defined for these models enforce a single global level of spatial smoothing for the set of random effects, which for the Leroux model is controlled by ρ.

The inference for these methods is based on MCMC methods or based on approximation methods as the INLA.

Autoregressive Bayesian Spatial Models for Count Data

Traditional spatial econometric models, such as the spatial autoregressive model (SLM) and the spatial error model (SEM), rely on the Gaussian assumption of the distribution of the response variable (LeSage and Pace 2009), which does not hold for count data. Consequently their usage for this type of data demands data transformation to meet the assumptions of the models. In order to avoid that, this section presents

new possible spatial modelling strategies for handling count data, assuming a Poisson distribution for those. Within these spatial autoregressive models, there are alternatives for modelling counts that have been explored (Simões et al. 2017), more specifically, considering a standard spatial lag model, recently developed within a new class of latent models defined in Integrated Nested Laplace Approximations (INLA) (Rue et al. 2009), by Goméz-Rubio et al. (2015); a spatial autoregressive lag model of counts, developed by Lambert et al. (2010), under a classical perspective; and a spatial lag autoregressive component incorporated in the model for counts, under a Bayesian paradigm and using INLA methodology.

Bayesian Standard Spatial Lag Model

Consider the first-order spatial autoregressive model on the response with covariates, also known as SLM, presented in section "Spatial Lag Model", given by:

$$\begin{aligned} \boldsymbol{y} &= \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(0, \sigma^2 I_n). \end{aligned} \tag{25}$$

As referred there, spatial contiguity matrix W is usually row standardized. This model explains the variation on the response $\boldsymbol{y}$ as a linear combination of the response in neighboring units and some explanatory variables. Parameter ρ is the autoregressive parameter, and parameters in vector $\boldsymbol{\beta}$ reflect the influence of the covariate values X on the $\boldsymbol{y}$ variation over the spatial domain. The error term $\boldsymbol{\epsilon}$ is assumed to follow a normal distribution with zero mean and variance-covariance matrix $\sigma^2 I_n$, where σ^2 is a global variance parameter (Anselin 2010; LeSage 1999).

The methodology INLA (Rue et al. 2009) provides an alternative to the simulation methods for doing Bayesian inference, being based on numerical approximation techniques. It is quite broad in application, just requiring the models to be written in a special but quite general framework, as described in section "Bayesian Inference with Integrated Nested Laplace Approximation", where a function g of the expected value of the response variable $\boldsymbol{\mu} = E[\boldsymbol{y}]$ is decomposed into:

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \sum_{j=1}^{J} f_j \tag{26}$$

where f_j are random effects. The vector of latent effects $(\beta_0, \boldsymbol{\beta}, \boldsymbol{f})$ forms a Gaussian Markov random field (GMRF).

In practice, there are still models that are not implemented in INLA and in the software, which led Goméz-Rubio et al. (2015) to have recently implemented in R-INLA a new class of models which includes the standard spatial lag model (25).

For this particular case of Gaussian models, the spatial lag model (25) can be rewritten as:

$$\boldsymbol{y} = (I_n - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}).$$

The authors implement the expression:

$$\boldsymbol{u} = (I_n - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

as a random effect that includes, besides the nonlinear effect, the intercept and the linear effects of the chosen covariates:

$$g(\mu_i) = u_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \sum_{j=1}^{J} f_{ji},\ i = 1, \ldots, n,$$

where $\boldsymbol{\epsilon}$, related to f_j's, is assumed normal distributed:

$$\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 I_n),$$

and where ρ, β, and X, W are defined as in (25).

For this model, the prior distributions considered for the vector of parameters $\boldsymbol{\beta}$, to the spatial autoregressive parameter ρ and the precision error term $\tau = \frac{1}{\sigma^2}$, are:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}(0, Q), \\ \mathrm{logit}(\rho) &\sim \mathrm{N}(a, b), \\ \tau &\sim \mathrm{Gamma}(c, d), \end{aligned} \tag{27}$$

with Q a precision matrix (that has to be specified).

In the development of this Bayesian spatial lag model, a Gaussian distribution was considered for the response variable $\boldsymbol{y}$, but it is possible to extend this to other distributions due to the broad INLA methodology model's formulation (14). The case of a binary response, leading to the estimation of a spatial probit model is proposed in Gomez-Rubio et al. (2016) and exemplified in Bivand et al. (2014). The case of a Poisson response variable, suitable for counts, is proposed and developed in section "A Bayesian Poisson Spatial Lag Model".

A Classical Poisson Spatial Lag Model

In this subsection, a spatial autoregressive lag model of counts developed by Lambert et al. (2010) under a classical inference framework is described. The spatial autoregressive count model suggested by these authors was motivated by their previous work on estimating temporally lagged count processes. These processes are time series y_t, $t = 0, 1, \ldots$, with:

$$E(y_t) = \mu_t = \exp(\beta X_t)\, y_{t-1}^{\rho}. \tag{28}$$

It specifies a multiplicative relation between a predetermined count and future outcomes (Lambert et al. 2010). Their autoregressive model for spatial lagged means, for count responses, specifies a multiplicative relationship between the mean μ_i of the Poisson response y_i in each area and all the means μ_j of the response in its neighbors, similarly to the multiplicative time series model for count data (Lambert et al. 2010).

Consider the non-spatial log-Poisson regression model for a vector of counts $\boldsymbol{y} = (y_1, \ldots, y_n)$ assumed to be Poisson distributed with expected value $E(y_i) = \mu_i$:

$$f(y_i) = \frac{\mu^{y_i} \exp(-\mu_i)}{y_i!}. \tag{29}$$

Being $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})$ a set of covariates with associated parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ the response expected values is decomposed as:

$$E(y_i) = \mu_i = \sum_{k=1}^{K} \exp(\beta_k X_{ik}) = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}),\ i = 1, \ldots, n. \tag{30}$$

Inspired by the model (28), Lambert, Brown, and Florax developed a spatial autoregressive count process that lays in the specification of the expected mean of counts at location i as a function of its j neighbors, given by:

$$\mu_i = E(y_i) = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}). \prod_{j \neq i}^{n} E(y_j)^{\rho w_{ij}}, \tag{31}$$

where w_{ij} are the elements of a weight matrix W and ρ is a spatial autocorrelation parameter. This specification has a multiplicative autoregressive component $\prod_{j \neq i}^{n} (E(y_j))^{\rho w_{ij}}$, added to the non-spatial log-Poisson regression model (30). Including that in the exponential part leads to the structural model, written in terms of the predictor $\eta_i = \log(\mu_i)$, as follows:

$$\begin{aligned} &\mu_i = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta} + \log(\textstyle\prod_{j=1}^{N} (\mu_j)^{\rho w_{ij}})) \Leftrightarrow \\ &\mu_i = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta} + \rho \textstyle\sum_{j \neq i} w_{ij} \log(\mu_j)) \Leftrightarrow \\ &\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \rho \textstyle\sum_{j \neq i} w_{ij} \eta_j. \end{aligned} \tag{32}$$

Expressing (32) in matrix notation, including all spatial units, leads to the reduced form of the conditional log-mean function:

$$\eta = (I_n - \rho W)^{-1} (X\boldsymbol{\beta}), \tag{33}$$

where $(I_n - \rho W)^{-1}$ is called the spatial multiplier term. Inference is done by usual maximum likelihood (Lambert et al. 2010).

A Bayesian Poisson Spatial Lag Model

Consider a vector of counts $\boldsymbol{y} = (y_1, \ldots, y_n)$ assumed to be Poisson distributed with expected value $E(y_i) = \mu_i$, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})$ a set of covariates with parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$.

The Lambert, Brown, and Florax previous model can be seen as a GLM model with spatially structured random effects (with log as link function), defined through the relationship:

$$\eta_i = \log(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{34}$$

With:

$$\varepsilon_i = \rho \sum_{j \neq i} w_{ij} \eta_j. \tag{35}$$

These random effects ε_i include a spatial lag term on the log mean, resulting in a Poisson spatial autoregressive lag model. ρ represents the spatial autoregressive parameter, for a considered weight or adjacency matrix W.

It is proposed now that the spatial lag autoregressive component (35) is incorporated in a model for counts, as described in the classical Poisson spatial lag model, in section "A Classical Poisson Spatial Lag Model", being afterward the estimation done under the Bayesian paradigm. For this the Bayesian standard spatial model in section "Bayesian Standard Spatial Lag Model" is adapted, being INLA methodology used for doing inference, under formulation (26):

$$\begin{aligned} \boldsymbol{y} &\sim \text{Poisson} \\ \boldsymbol{\mu} &= E[\boldsymbol{y}] \\ \log(\boldsymbol{\mu}) = \boldsymbol{\eta} &= (I_n - \rho W)^{-1} (X\boldsymbol{\beta}). \end{aligned}$$

This construction allows a Bayesian spatial lag model for a Poisson response, considering η as a random effect in the linear predictor, borrowed from the classical spatial lag Poisson model from section "Bayesian Standard Spatial Lag Model", having:

$$\boldsymbol{u} = \boldsymbol{\eta} = (I_n - \rho W)^{-1} (X\boldsymbol{\beta} + \boldsymbol{\epsilon}),$$

where $\boldsymbol{\epsilon}$ is assumed normal distributed, $\boldsymbol{\epsilon} \sim \text{N}(0, \sigma^2 I_n)$.

This results in a Bayesian Poisson spatial lag model, an alternative to do Bayesian inference for spatial autoregressive econometric models for count data.

For this model, the prior distributions assigned to the spatial autoregressive parameter ρ and to the precision error term τ are chosen as:

$$\begin{aligned} \text{logit}(\rho) &\sim \text{N}(a, b), \\ \tau &\sim \text{Gamma}(c; d). \end{aligned} \tag{36}$$

A normal prior is assigned for the vector of parameters $\boldsymbol{\beta}$ with precision matrix Q:

$$\boldsymbol{\beta} \sim \mathrm{N}(0, Q). \tag{37}$$

Different prior distribution can be specified as well as other hyperparameters.

Note that an offset can be used as a correction factor in the model specification, considering $E(y_i) = \mu_i = e_i\theta_i$, where e_i represents the number of expected cases of what is being measured in each spatial unit $i = 1, \ldots, n$ and θ_i is the relative risk in area i. With this Bayesian Poisson spatial lag model, the risk of what is being counted in one area is related to the risk of what is being counted in the areas of its neighborhood, driven by effects of important covariables on explaining the phenomenon in one area. The use of this modelling strategy of the autoregressive models allows to evaluate if the risk of the phenomena that is being counted in a given location may be simultaneously determined by the risk in neighboring locations. This way of modelling spatial structure for areal data does not ignore the discrete nature of data whenever it applies, incorporating it in the model. In this case, the response variable in a given area is a good predictor of the response variable in its neighborhood areas, addressing a *global* spatial autocorrelation arising from dependence between counts.

To implement this Bayesian Poisson spatial lag model, in R-INLA, it was used the "slpm" function (Goméz-Rubio et al. 2015; Simões et al. 2017).

Model Selection

Bayesian models can be evaluated and compared by measuring their performance through their predictive accuracy. This can be estimated using cross-validation which requires training sets to re-fit the models, which is less convenient, or information criteria, which use functions of the deviance. Given the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $L(\boldsymbol{\theta}|\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta})$ as likelihood function, the deviance of the model is defined as

$$D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{y}|\boldsymbol{\theta})),$$

where $\boldsymbol{\theta}$ corresponds to the parameters of the likelihood. The deviance of a model measures the variability linked to the likelihood (Blangiardo et al. 2015). For the information criteria approach, measures like the Akaike information criterion (AIC) (Akaike 1998), the deviance information criterion (DIC) (Spiegelhalter et al. 2002; Blangiardo et al. 2015), and the Watanabe-Akaike information criterion (WAIC) (Watanabe 2010; Gelman et al. 1992) are the most used. DIC is a generalization of AIC, developed especially for Bayesian model comparison (Spiegelhalter et al. 2002), and WAIC can be seen as an improvement over the DIC (Watanabe 2010). The preferred model will be the one with lower values for the considered criteria. These criteria are defined below.

Akaike Information Criterion (AIC)

The AIC measure of predictive accuracy is composed by two components, one associated with the quality of the adjustment of the model and another associated with its complexity, quantified by two times the parametric dimension, given by:

$$AIC = -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}) + 2p,$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and p the number of parameters.

Deviance Information Criterion (DIC)

In Bayesian formulation, the deviance is a random variable using the posterior mean of the deviance $\overline{D} = E_{\boldsymbol{\theta}|\boldsymbol{y}}(D(\boldsymbol{\theta}))$ as a measure of fit. Replacing in AIC measure the maximum likelihood estimate of $\boldsymbol{\theta}$ by its posterior mean and replacing p with a data based bias correction, another measure of the predictive accuracy is:

$$DIC = \overline{D} + p_D,$$

where p_D, the effective numbers of parameters, is given by:

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})] - D(E_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]) = \overline{D} - D(\overline{\theta}),$$

where $D(\overline{\theta})$ is the deviance computed on the posterior mean of the parameters.

The DIC measure is also composed by two components, one for quantifying the model fit (measured through the posterior expectation of the deviance) and the other for evaluating the model complexity (measured through the effective number of parameters).

Note that DIC depends only on a data-dependent function that can be omitted when the models to compare are based on the same sampling model, although the model for $\boldsymbol{y}$ may differ on the parametric structured (Turkman and Paulino 2015).

Watanabe-Akaike Information Criterion (WAIC)

This measure of predictive accuracy introduced by Watanabe (2010) is given by:

$$WAIC = -2\sum_{i=1}^{n} \log E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[f(y_i|\boldsymbol{\theta})\right] + 2p_W,$$

with two different proposals for the effective number of parameters, p_W. One possibility uses the variance of individual terms in the log predictive density summed over the n data

points, calculated through the posterior variance of the log predictive density for each data point y_i:

$$p_{W_1} = \sum_{i=1}^{n} Var_{\boldsymbol{\theta}|\boldsymbol{y}}(\log(f(y_i|\boldsymbol{\theta})).$$

Another possibility for p_W is the one similar to that used to construct p_D:

$$p_{W_2} = -2\sum_{i=1}^{n}\left(\log(E_{\boldsymbol{\theta}|\boldsymbol{y}}[f(y_i|\boldsymbol{\theta})]) - E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\log(f(y_i|\boldsymbol{\theta}))\right]\right).$$

Either p_{W_1} or p_{W_2} can be used as a bias correction in WAIC.

There is still some disagreement on which one of the criteria should be used. For example, AIC does not perform well on settings with strong prior information; DIC can produce negative estimates of the effective number of parameters, and it is based on a point estimate, when the posterior distribution is not well summarized by its mean and provides nonsensical results; WAIC uses the posterior distribution rather than a point estimate, and it is invariant to re-parametrization, being referred to as "fully Bayesian." However, WAIC depends on data partition that might raise difficulties for structured models (Gelman et al. 1992). Nevertheless, according to recent studies, "WAIC has various advantages over simpler estimates of predictive error such as AIC and DIC" but because it requires an additional computational effort, it is less used in practice (Vehtari et al. 1999). Given the above, in this thesis, we focus on WAIC and DIC measures to compare models. Actually, DIC is the predictive measure most used in Bayesian applications, and WAIC has been shown to be more stable and particularly helpful with hierarchical and mixture structures, in which the number of parameters increases with sample size although when working with point estimates, it is not the most appropriate approach (Gelman et al. 1992).

Application

One of the most relevant factors regarding hospital costs in the Portuguese health-care system are urgency admissions, consuming large financial and human resources. It is possible that a considerable part of the admissions corresponds to non-urgent cases that could be handled by primary health-care services, namely, the family doctor, or in a self-care basis eventually assisted by a remote nursing service. This helps to understand why the Portuguese hospital urgency service became one of the most important worries of the Portuguese Health Ministry over the last years. The Portuguese national health line Saúde24 (S24) service directs users to the most appropriate institutions of the public health service or offers counsels on self-care measures. It is hoped that its use mitigates the unnecessary urgent care in hospitals and that the reached savings can be channeled toward other needy areas. This study aims to describe and evaluate the use of S24 by analyzing the number of calls received, at a municipal level, under two different spatial econometric approaches. This analysis is important for future development of decision support indicators in a hospital context, based on the economic impact of the use of this health line rather than on the criterion of hospital urgency.

Saúde24 Data Analysis

An initiative to improve accessibility to health care and to rationalize the use of existing resources was carried out by the Portuguese Health Ministry through the creation of a national health line, S24, in April 2007 (Portal of the National Portuguese Health Service 2015). These objectives are accomplished by the S24 service which directs users to the most appropriate institutions of the national public health service or by counseling self-care home measures.

The location attribute for S24 data is an important source of information to describe its use, which leads to analyze the number of calls to S24 at a municipal level. As space is an important feature of these data, and ignoring it results in a poorer analysis (Anselin et al. 1996; Cressie 1993).

To model the number of calls to S24, in each municipality, with spatial models, given the discrete nature of data (counts), an alternative is to use a hierarchical Bayesian model with covariates (Banerjee et al. 2004). For the hierarchical approach, spatial autocorrelation is accounted for in the disturbances and not in the observed responses, as happens with spatial autoregressive approaches. The latter is a different modelling strategy common in spatial econometrics literature that may also be considered for these data. It is plausible to think that the number of calls to S24 in one municipality is related to the number of calls in the municipalities of its neighborhood, driven by effects of covariates such as the number of hospitals in one municipality, which may certainly have an impact on the number of calls to S24 in a neighboring municipality, or others not considered in the modelling (LeSage and Pace 2009). Hierarchical and autoregressive modelling perspectives have already been used to model the same data sets (Bivand et al. 2014; Gomez-Rubio et al. 2016; Quddus 2008).

This analysis begins with the use of standard spatial econometric techniques to look for spatial dependence in the number of calls to S24 in each municipality, considering a neighborhood contiguity structure, as well as in the residuals of a baseline log-Poisson regression model with covariates. The number of calls is further analyzed, on one hand, through different hierarchical log-Poisson models and, on the other hand, through a Poisson spatial lag model, implementing

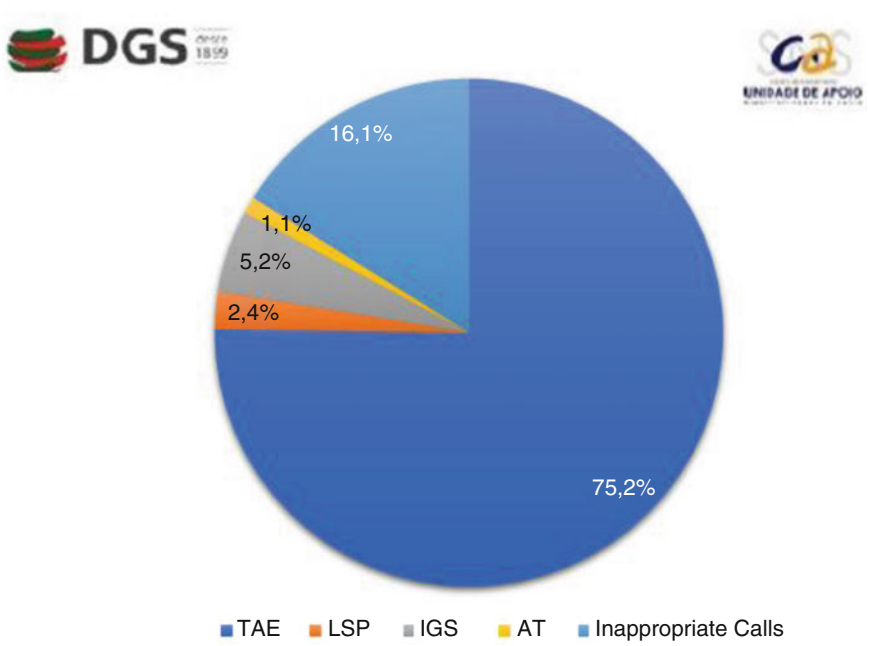


Fig. 1 Calls to S24 by service in 2014—Graph provided by DGS

different econometric approaches to model spatial structure in data. The results of this study are intended to be used in the near future in cooperation with the Portuguese Directorate-General of Health to analyze, test, implement, and predict consequences of different government management policies at the hospital level under distinct scenarios. The savings from the correct use the S24 will avoid unnecessary urgent care in hospitals that can then be channeled toward other needy areas.

The S24 Data

The data considered in this study were provided by the Support Unit of the Call Center of the National Health Service of the Portuguese Directorate-General of Health (DGS). It is a comprehensive data set of the calls recorded by the S24 health line in the year 2014 and includes information such as user's gender, residence, age, and call's day of the week, together with the health problem specification.

The S24 has two call centers and offers various services such as triage, counseling, and routing (TAE); therapeutic counseling (AT) to clarify issues relating to medication; assistance in public health (LSP) in specific topics such as flu, heat, poisoning, etc.; and general health information (IGS), such as the location of public health units and pharmacies, among others. The S24 service is provided by qualified nurses, trained to give the best advice or, when appropriate, to assist citizens in solving the situation by themselves. The service is available to the beneficiaries of all different kinds of health sub-systems. The S24 incorporates approximately 300 nurses and 16 clinical supervisors.

Most of the calls answered by S24, are catalogued as TAE, approximately 92% (calculated after removing inappropriate calls)—see Fig. 1. For those, the description of the health problem and the original intention of the user about how to solve it (e.g., go to an urgency room) are recorded, and then a decision algorithm follows. The final disposition given by this algorithm, jointly with the evaluation of the nurse, falls in one of two possibilities: emergent or non-emergent situation. The non-emergent situation calls are the ones analyzed in this study—see Fig. 2.

This study focuses on the number of TAE calls to S24 in 2014 at a municipality level, in Continental Portugal. For this year, 50% of the users were aged between 4 and 46, with a median of 26 years and a range of 111 years. Elderly users are less than 13%. The distribution of the number of TAE calls to S24, by municipality in 2014, is mapped in Fig. 3. The average raw call rate by municipality is 32 per 1000 inhabitants.

Non-spatial Modelling: The Log-Poisson Regression Model

The number of TAE calls to S24 in each of the 278 municipalities of Continental Portugal was first modelled via a log-Poisson regression model before considering the need of a spatial analysis.

An indirect standardization of these numbers has been carried out, applied to the resident population of each municipality in terms of age groups, namely, 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and +80. This method considers standard age rates

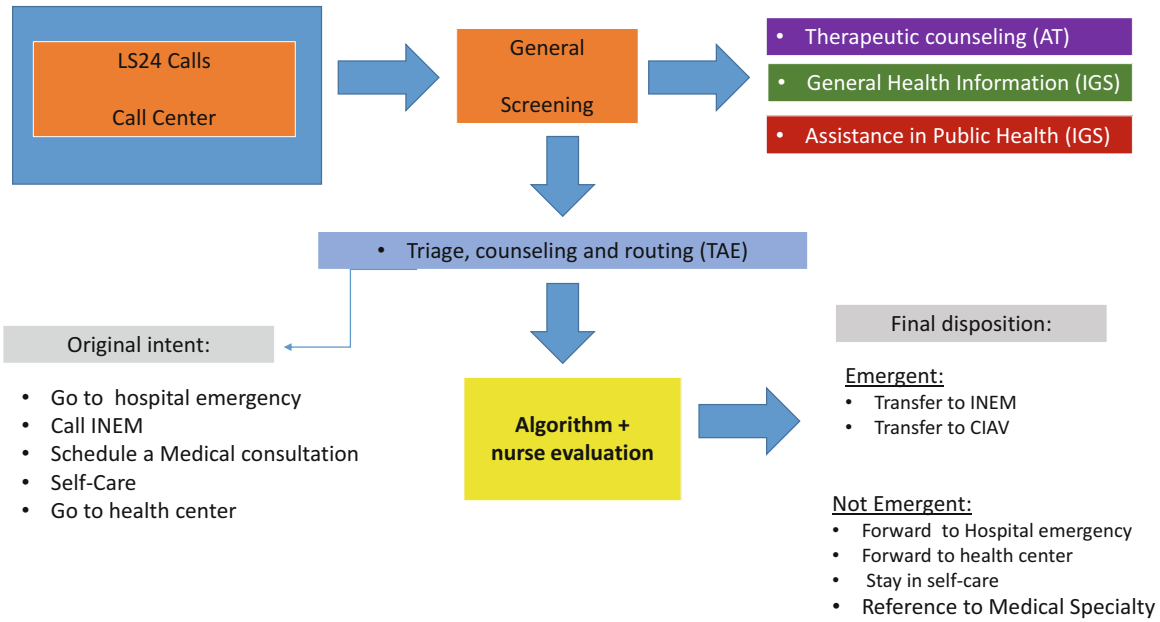


Fig. 2 The collection of information in S24

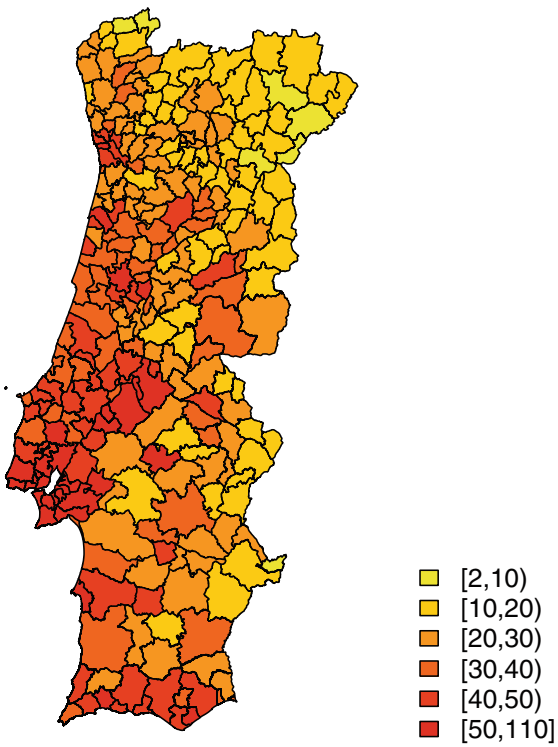


Fig. 3 Number of TAE calls to S24 per 1000 inhabitants, in 2014

$$\varphi_j = \frac{\sum_i y_{ij}}{\sum_i n_{ij}}, \; j = 1, \ldots, 9,$$

with y_{ij} the number of cases (calls) and n_{ij} the at-risk population (resident population), in municipality i and age group j, with $i = 1, \ldots, 278$, with $j = 1, \ldots, 9$, in order to obtain $e_i = \sum_i n_{ij}\varphi_j$, $i = 1, \ldots, 278$, the expected number of calls in each municipality, that is included in the model as an offset. So, in fact, what is modelled is the relative call risk, which can be roughly estimated by the Standard Call Rate (SCR), mapped in Fig. 4. This ratio is calculated between the observed number of cases and the expected number of cases, allowing comparisons across different populations,

$$SCR_i = \frac{y_i}{e_i}, i = 1, \ldots, 278.$$

The resident population of each municipality, in terms of age groups, was obtained from Census 2011 data and adjusted for subsequent years (Database of contemporary Portugal 2016).

Demographic and socioeconomic information, development indicators, as well as characteristics of the Portuguese health system at the municipal level were investigated as possible covariates for modelling the TAE call counts, in order to understand if the inclusion of certain covariates obviated the need for a spatial model. Using the stepwise methodology (Rawlings et al. 1998) for selecting covariates, under different scenarios, the two best sets of the most significant explanatory variables are:

Case 1: The *average number of years of schooling*, the *proportion of elderly residents*, the *unemployment rate*, the *rurality index*, the *number of hospitals and health centers* per 1000 inhabitants, and the *proportion of women*, in each municipality (AIC: 29530);

Case 2: The *monthly average income*, the *proportion of children*, the *unemployment rate*, the *rurality index*, the *number of hospital and health centers* (both per 1000 inhabitants), and the *proportion of women*, in each municipality (AIC: 36980).

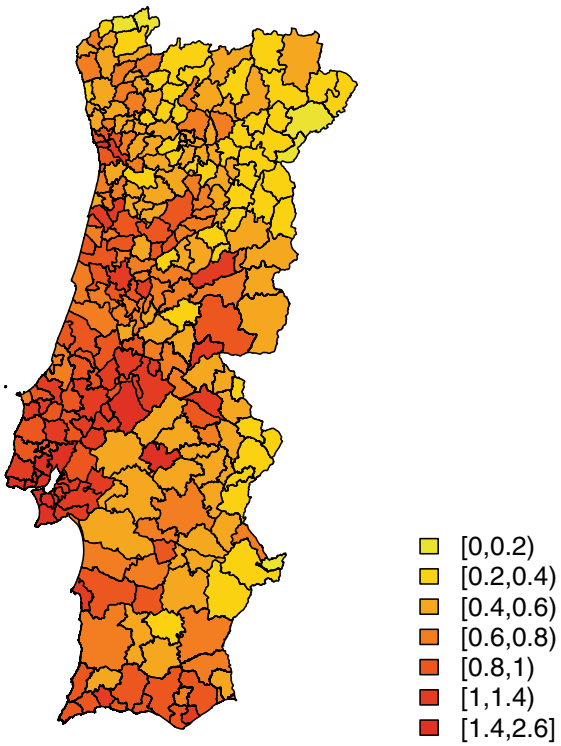


Fig. 4 Standard call rate to S24, in 2014

From these variables, the average number of years of schooling and the monthly average income are the ones that show a stronger positive correlation with the response variable (0.67 and 0.61, respectively), followed by the proportion of children (0.49). The rurality index and the proportion of elderly residents are negatively correlated with the response (-0.45 and -0.35, respectively). The study analysis is developed considering this two cases.

Over-dispersion in these Poisson data is expected, since space is suspected to be an important feature for their modelling. If this over-dispersion is ignored, the standard errors of the covariate effects are underestimated, resulting in an incorrect assessment of the significance of individual regression parameters. So, instead, it has been opted to fit a quasi-Poisson model to account for the over-dispersion, realizing that the significant covariates under this approach were in fact different from the ones of the Poisson model (although the estimated effects are, of course, the same).

Tables 1 and 2 depict the estimated coefficients of the considered quasi-Poisson log-regression models for these analyses, with:

$$\log(\theta_i) = \beta_0 + \sum_{j=1}^{7} \beta_j X_{ij}, \ i = 1, \ldots, 278,$$

where θ_i is the relative risk in the ith municipality. For case 1, the unemployment rate turned out to be not significant after all, and for case 2, the same happened with the rurality index, the number of hospital and health centers.

Package stats of R-project software was used to obtain the results presented in this section (R Core Team and contributors worldwide 2013).

Table 1 Covariates and their estimated coefficients for the quasi-Poisson log-regression model, case 1, for the S24 2014 data

Variable	Id	Coefficients	*p*-values
Average number of years of **schooling**	x_1	0.322	<2e−16
Proportion of **elderly** residents	x_2	4.456	9.52e−13
Unemployment rate	x_3	−0.743	0.3156
Rurality index	x_4	−0.741	4.10e−06
Number of **hospitals**	x_5	−3.822	6.68e−06
Number of **health centers**	x_6	−1.289	0.0437
Proportion of **women**	x_7	−5.509	0.0661
Intercept		−0.288	0.8390

Table 2 Covariates and their estimated coefficients for the quasi-Poisson log-regression model, case 2, for the S24 2014 data

Variable	Id	Coefficients	*p*-values
The monthly average **income**	x_1	0.001	5.97e−16
Proportion of **children**	x_2	5.727	0.0004
Unemployment rate	x_3	−1.679	0.0391
Rurality index	x_4	−0.212	0.1884
Number of **hospitals**	x_5	−0.398	0.6504
Number of **health centers**	x_6	−0.902	0.1702
Proportion of **women**	x_7	11.810	0.0003
Intercept		−7.930	7.64e−06

Spatial Correlation

In this subsection, standard spatial techniques are used to look for spatial dependence in the number of TAE calls, considering a contiguity neighborhood structure, and also in the residuals of the log-Poisson regression models fitted before.

For the considered contiguity neighborhood structure, in the first-order queen neighborhood, there are 1.9% non-zero weights, and the average number of neighbors is 5.3. Taking the corresponding queen neighborhood matrix, and using Moran's I statistics (1), both under normality ($I = 0.6182$, $p \leq 2.2e{-}16$), or considering a randomized distribution of the statistics ($I = 0.6182$, $p \leq 2.2e{-}16$), resulted in a clear rejection of the spatial independence hypothesis of the number of TAE calls, suggesting that there is a positive spatial correlation among these.

The spatial autocorrelation in the residuals of the log-Poisson regression models fitted in section "Non-spatial Modelling: The Log-Poisson Regression Model" was further investigated, using a randomized distribution of the statistic and a two-sided test, having $I = 0.1513$ ($p = 1.102e{-}05$) for case 1 and $I = 0.2702$ ($p = 2.276e{-}14$) for case 2. The results suggest a high positive spatial autocorrelation in the residuals. With spatially correlated residuals, the fitted models may be providing biased estimates of the parameters, leading to incorrect interpretations and misleading conclusions (LeSage 1999). It is then clear that space is an important feature of these data, and that must be considered in the modelling.

Package spdep (Bivand et al. 2014) of R-project software was used to obtain the results presented in this section according to Anselin (2007).

Spatial Bayesian Econometric Modelling

Spatial Hierarchical Log-Poisson Regression Model

In order to capture and model data spatial variability, the number of TAE calls in each municipality is now analyzed through different spatial hierarchical log-Poisson regression models. The residual autocorrelation of the log-Poisson regression model considered before can be explained, in a Bayesian setting, adding to the model's predictor a set of spatially structured $\boldsymbol{\varepsilon}$ random effects, considering the contiguity neighborhood structure mentioned before. Additional unstructured random effects $\boldsymbol{\gamma}$ can be considered, if needed. The prior distributions of the random effects define their structure, as described in section "Hierarchical Bayesian Spatial Models for Count Data". Two models were considered differing on the way the random effects are included, the BYM and the Leroux models.

The estimates were obtained via Markov chain Monte Carlo (MCMC) method, implemented in R-package CARBayes (Lee 2013). A few MCMC run of 1,000,000 iterations were made, discarding 50,000 burn-in iterations and thinning by 100, in order to reduce autocorrelation, resulting in 9500 sample points.

In general for most parameters, acceptance rates of the Metropolis-Hastings algorithm were about 40%. MCMC output convergence was assessed through visual inspection of the samples traces, autocorrelation function plots, and the application of the Geweke method (Geweke 1991) available in R-package CARBayes (Lee 2013).

BYM Model

The Besag-York-Mollié (BYM) model, as described in section "Hierarchical Bayesian Spatial Models for Count Data", for both cases 1 and 2, as in section "Non-spatial Modelling: The Log-Poisson Regression Model", is a log-Poisson regression model with the covariates considered before plus unstructured ($\boldsymbol{\gamma}$) and spatially structured random effects ($\boldsymbol{\varepsilon}$), for which a CAR prior is chosen. The main parameter estimates are summarized in Tables 3 and 4.

For case 1, only one of the covariates, the average number of years of schooling, showed to be significant, whereas in case 2, it was the monthly average income. The estimated random effects, given by $\exp(u_i) = \exp(\varepsilon_i + \gamma_i)$, still display some patterns for both cases—left panels of Figs. 5 and 6.

Table 3 Parameter estimates (median, 2.5% and 97.5% quantiles) for the BYM hierarchical log-Poisson model, case 1, for the S24 2014 data

Variable	Id	Median	2.5%	97.5%
Average number of years of **schooling**	x_1	0.1931	0.0062	3.5386
Proportion of **elderly** residents	x_2	0.4840	−1.6883	2.6921
Unemployment rate	x_3	1.8680	−1.5338	4.7276
Rurality index	x_4	−0.0930	−0.4980	0.3126
Number of **hospitals**	x_5	−0.2549	−1.9709	1.4916
Number of **health centers**	x_6	0.3777	−1.1932	1.8507
Proportion of **women**	x_7	−1.2479	−10.1867	7.7633
Intercept		−1.3506	−6.4541	3.5386
σ_B^2		0.2268	0.1440	0.3494
σ^2		0.0326	0.0140	0.0592

Table 4 Parameter estimates (median, 2.5% and 97.5% quantiles) for the BYM hierarchical log-Poisson model, case 2, for the S24 2014 data

Variable	Id	Median	2.5%	97.5%
The monthly average **income**	x_1	0.001	0.0	0.0021
Proportion of **children**	x_2	1.7348	−2.1687	6.3659
Unemployment rate	x_3	1.9042	−1.5689	5.4343
Rurality index	x_4	−0.1838	−0.5460	0.2105
Number of **hospitals**	x_5	−0.2282	−2.0447	1.3542
Number of **health centers**	x_6	−0.0613	−1.3820	1.3585
Proportion of **women**	x_7	0.6238	−8.9309	9.4824
Intercept		−1.9574	−7.2782	3.4761
σ_B^2		0.2443	0.1594	0.3690
σ^2		0.0313	0.0149	0.0540

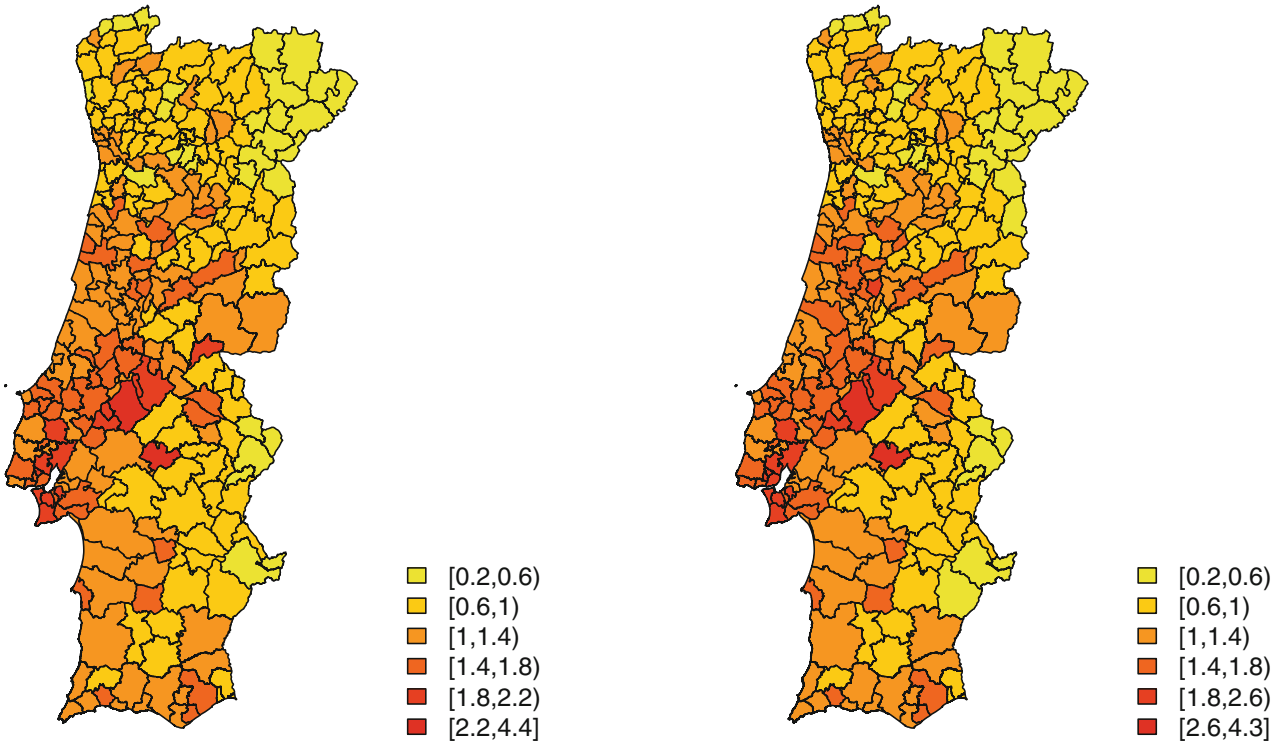


Fig. 5 Estimated random effects for BYM model(left) and for Leroux model (right), case 1

Leroux Model

The Leroux model, as described in section "Hierarchical Bayesian Spatial Models for Count Data", is a log-Poisson regression model with the covariates previously considered and the random effects for which the Leroux CAR prior is chosen. The main parameter estimates are displayed in Table 5 for case 1 and in Table 6 for case 2.

Here, for the first case, only one of the initial covariates showed to be significant, the average number of years of schooling. The estimates of the random effects, given by $\exp(\varepsilon_i)$, seem to indicate that there still is spatial variability in these data—right panel of Fig. 5, which is strongly confirmed by an estimated value of ρ of 0.90.

Considering this model for the second case, also only one of the initial covariates was significant, the monthly average income. This model has an estimated value of ρ of 0.89, and the estimates of the random effects seem to indicate that there still is spatial variability—right panel of Fig. 6.

The Bayesian Poisson Spatial Lag Model

A modelling alternative is to account for spatial autocorrelation in the observed responses instead of the disturbances, as before, using an autoregressive perspective. This approach may also be considered for these data.

Here, the TAE number of calls in each municipality is then analyzed through the Bayesian Poisson spatial lag model where a spatial autocorrelation lag is incorporated in the econometric model of counts. The estimates were obtained via INLA methodology in R-package R-INLA, according to the R-code available in Simões et al. (2017). The prior

Table 5 Parameter estimates (median, 2.5% and 97.5% quantiles) for the Leroux hierarchical log-Poisson model, case 1, for the S24 2014 data

Variable	Id	Median	2.5%	97.5%
Average number of years of **schooling**	x_1	0.1897	0.0141	0.3187
Proportion of **elderly** residents	x_2	0.8586	−1.3888	2.8322
Unemployment rate	x_3	2.1413	−0.7070	4.8630
Rurality index	x_4	−0.1129	−0.4562	0.2337
Number of **hospitals**	x_5	−0.2606	−1.7421	1.1745
Number of **health centers**	x_6	0.3458	−0.8881	1.6717
Proportion of **women**	x_7	−1.9482	−9.8121	6.2431
Intercept		−1.1342	−5.1422	3.1621
σ_L^2		0.3492	0.2829	0.4494
ρ		0.9059	0.7008	0.9888

Table 6 Parameter estimates (median, 2.5% and 97.5% quantiles) for the Leroux hierarchical log-Poisson model, case 2, for the S24 2014 data

Variable	Id	Median	2.5%	97.5%
The monthly average **income**	x_1	0.001	0.0001	0.0019
Proportion of **children**	x_2	1.8345	−1.9861	5.7947
Unemployment rate	x_3	1.6751	−1.3984	4.7253
Rurality index	x_4	−0.2145	−0.5562	0.0639
Number of **hospitals**	x_5	−0.3568	−1.8623	1.1085
Number of **health centers**	x_6	−0.0069	−1.2701	1.2484
Proportion of **women**	x_7	0.5789	−7.0199	8.6312
Intercept		−1.9231	−6.5246	2.4100
σ_L^2		0.3581	0.2911	0.4508
ρ		0.8936	0.6879	0.9855

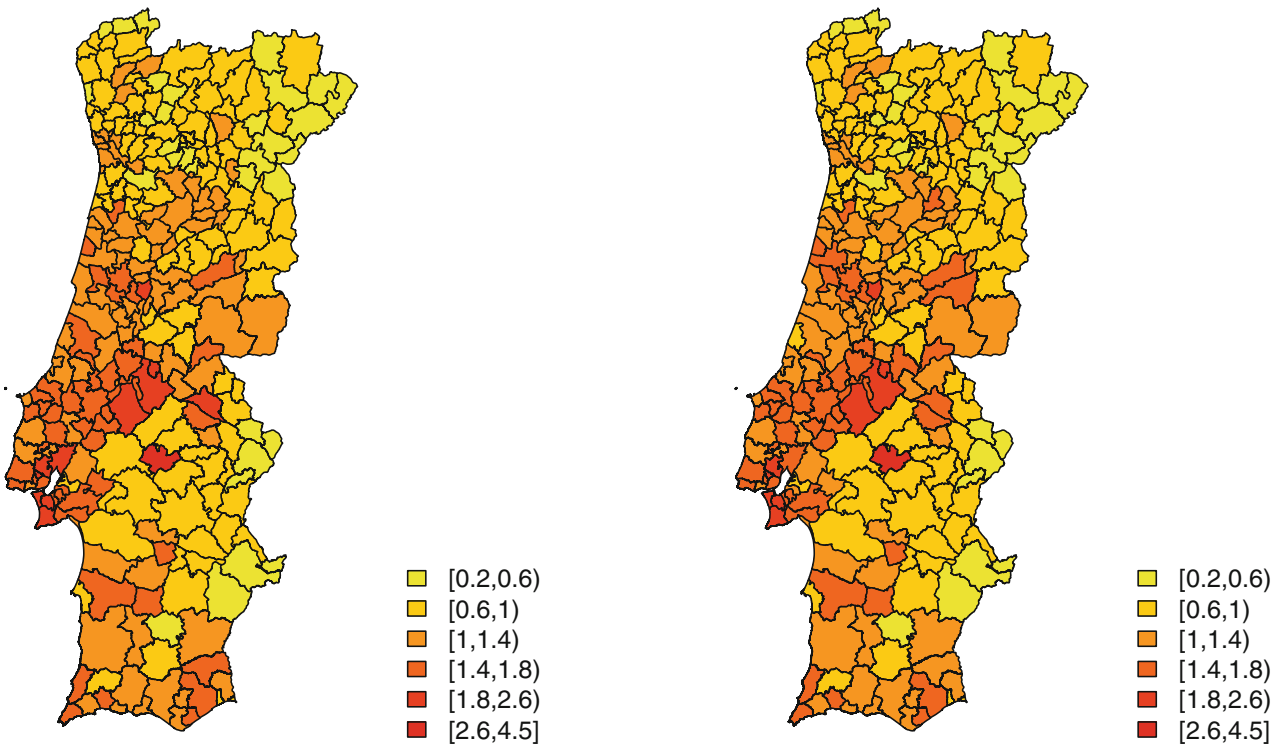


Fig. 6 Estimated random effects for BYM model (left) and for Leroux model (right), case 2

Table 7 Parameter estimates (mean, 2.5% and 97.5% quantiles) for the spatial lag Poisson model, case 1, for the S24 2014 data

Variable	Id	Mean	2.5%	97.5%
Average number of years of **schooling**	x_1	0.179	0.121	0.237
Proportion of **elderly** residents	x_2	0.591	−0.288	1.473
Unemployment rate	x_3	0.605	−0.851	2.048
Rurality index	x_4	−0.005	−0.209	0.197
Number of **hospitals**	x_5	−0.179	−1.300	0.941
Number of **health centers**	x_6	0.173	−0.316	0.660
Proportion of **women**	x_7	−0.919	−4.702	2.848
Intercept		−1.071	−3.012	0.868
σ^2		0.070	0.568	0.085
ρ		0.852	0.806	0.893

Table 8 Parameter estimates (mean, 2.5% and 97.5% quantiles) for the spatial lag Poisson model, case 2, for the S24 2014 data

Variable	Id	Mean	2.5%	97.5%
The monthly average **income**	x_1	0.001	0.000	0.001
Proportion of **children**	x_2	0.972	−1.232	3.190
Unemployment rate	x_3	0.065	−1.408	1.520
Rurality index	x_4	−0.098	−0.291	0.094
Number of **hospitals**	x_5	0.151	−0.989	1.290
Number of **health centers**	x_6	−0.095	−0.560	0.369
Proportion of **women**	x_7	0.307	−3.570	4.182
Intercept		−1.015	−3.180	1.141
σ^2		0.074	0.061	0.089
ρ		0.859	0.813	0.90

distributions assigned to the spatial autoregressive parameter ρ and to the precision error term τ are, by default, $\text{logit}(\rho) \sim \text{N}(0, 10)$ and $\tau \sim \text{Gamma}(1; 5 \times 10^{-5})$; however, other values can be chosen by the user.

Poisson Spatial Lag Model

This is the Bayesian Poisson spatial lag autoregressive model with the covariates initially considered significant. Tables 7 and 8 summarize the main parameter estimates for case 1 and case 2, respectively.

For case 1, only one of the previous covariates revealed to be significant, the average number of years of schooling. This model has an estimated value of ρ of 0.852. As for the second case, only the monthly average income is significant. This second model has an estimated value of ρ of 0.859. The estimated random effects, given by $\exp(\varepsilon_i)$, still display some patterns for both cases—Fig. 7 for case 1 and Fig. 8 for case 2.

Results Comparison

In the various spatial fits, the covariates considered important for explaining the number of calls and the corresponding effects were the same. These fits were further compared by means of their predictive accuracy, using the *Deviance Information Criterion* (DIC) measure and the *Watanabe-Akaike Information Criterion* (WAIC) measure. See Table 9 for case 1 and Table 10 for case 2. The *Relative Root Mean Square Error* (RRMSE) was also considered to measure goodness of fit:

$$RRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{y}_i)^2}{\widehat{y}_i^2}}.$$

Results are displayed in Tables 11 and 12.

In terms of spatial hierarchical log-Poisson regression models, the model with smaller DIC (preferred model) is the one including the covariates and the spatially structured random effects through Leroux CAR prior. This was confirmed by the RRMSE values. For the sake of comparison, the fit measures for the baseline log-Poisson regression model without random effects, fitted by MCMC, are further displayed in the first line of the tables. The log-Poisson regression model was also fitted, including only covariates and unstructured random effects (results not shown here), which performed worse, indicating that spatial random effects are indeed necessary in the models. This might indicate that there are possibly some relevant covariates that are not yet being included in the model. There is a spatial asymmetry that is not explained by the variables. Similar conclusions were reached when the autoregressive perspective was considered in terms of the Bayesian Poisson spatial lag model.

In order to compare both hierarchical and autoregressive model fits, WAIC measure was used, as it is more appropriate for comparing different model structures. The autoregressive model reveals better performance, according to this measure.

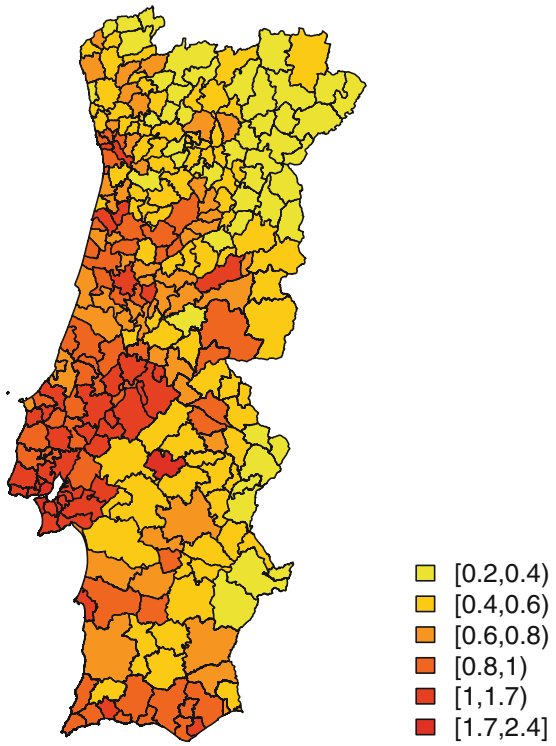


Fig. 7 Estimated random effects for the Poisson spatial lag model, case 1

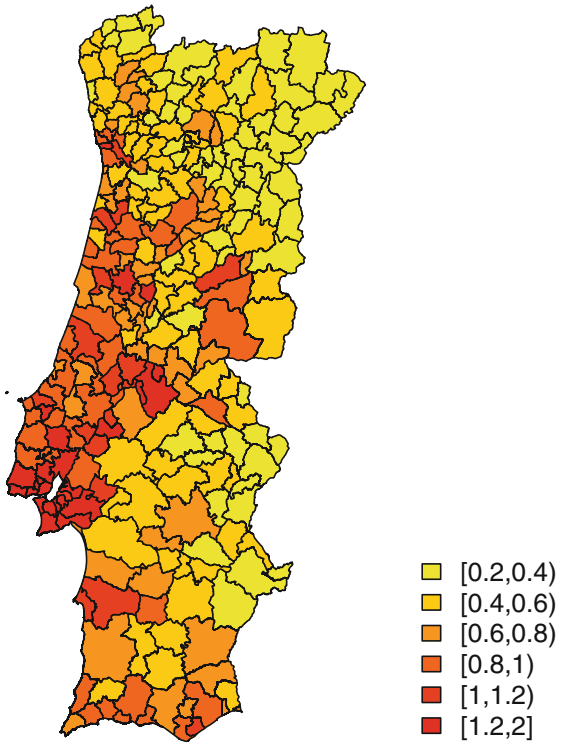


Fig. 8 Estimated random effects for the Poisson spatial lag model, case 2

Table 9 DIC and WAIC measured for the three models fitted for case 1

Model	DIC	pD	WAIC	pW
Baseline model MCMC	2816.9	287.2	2830.8	198.5
BYM model	2801.1	275.2	2773.8	179.9
Leroux model	2788.6	267.16	2744.5	157.2
Poisson spatial lag model	2778.63	261.96	2717.6	144.9

Table 10 DIC and WAIC measured for the three models fitted for case 2

Model	DIC	pD	WAIC	pW
Baseline model MCMC	2816.9	287.2	2830.8	198.5
BYM model	2800.0	275.6	2770.7	169.5
Leroux model	2793.2	272.4	2743.0	156.7
Poisson spatial lag model	2777.3	262.6	2714.1	143.8

As for the RRMSE values, they are very similar although they are somewhat smaller for the hierarchical models.

Final Remarks

This application study combines insights from classical spatial econometrics and the analysis of spatial data in order to handle spatial count data, both in a hierarchical and in an autoregressive perspective. The approaches applied here circumvent the limitations of the classical econometrics methods.

Table 11 RRMSE measured for the three models fitted for case 1

Model	RRMSE
Baseline model	0.588
BYM model	0.029
Leroux model	0.028
Poisson spatial lag model	0.037

Table 12 RRMSE measured for the three models fitted for case 2

Model	RRMSE
Baseline model	0.469
BYM model	0.025
Leroux model	0.026
Poisson spatial lag model	0.034

Within the scope of the spatial econometric methods and also resorting to Bayesian hierarchical and autoregressive methodology, their application to the study of the number of TAE calls to the national health line S24 revealed spatial correlation, and the addition of spatial structure in the models improved estimation.

The count data were first analyzed with a log-Poisson regression model, and then the inclusion of spatial random effects in a hierarchical Bayesian setting proved to be relevant, as expected, being the preferred model the one including the

covariates and the spatially structured random effects through Leroux CAR prior distribution. However, the modelling may possibly be improved by considering some other more adequate covariates. Additionally, a Bayesian Poisson spatial lag model was developed and implemented, an alternative to do Bayesian inference for spatial econometric models for count data. Similar conclusions were drawn when both the hierarchical and the autoregressive perspectives were considered.

The average number of years of schooling for case 1 of the analysis and the average monthly income for case 2 stand out as being important in explaining the use of S24. The spatial component for both cases was quite relevant, which was confirmed by the high values of the estimates of the spatial autocorrelation parameter.

It is intended to proceed with this application study of S24 data set in order to be able to describe and evaluate in which municipalities the use of S24 should be encouraged, as well as detecting those regions that most contribute to the economic success of the good use of the line for future assessment of hospital savings (Hughes and McGuire 2003).

Additionally, this analysis will be extended to include data available for the years between 2010 and 2016, fitting some spatiotemporal models (Cressie 1993) under an econometric approach and developing and implementing the temporal effects on Bayesian hierarchical models (Blangiardo et al. 2015; Lee et al. 2013), or on Bayesian autoregressive models (Blangiardo et al. 2015) for count data.

Acknowledgments This work is financed by national funds through FCT — Foundation for Science and Technology under the projects UID/MAT/00297/2019 and UID/MAT/00006/2013.

References

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, 199–213. New York: Springer.

Albert, J. 2009. *Bayesian Computation with R*. 2nd ed. New York: Springer.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. 1990. Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* 30: 185–207.

Anselin, L. 2006. Econometric theory. In *Palgrave Handbook of Econometrics*, ed. T. Mills and K. Patterson. Vol. 1, 901–969. Basingstoke: Springer.

Anselin, L. 2007. *Spatial regression analysis in R - A Workbook*. Center for Spatially Integrated Social Sciences, University of Illinois, Urbana-Champaign.

Anselin, L. 2010. Thirty years of spatial econometrics. *Papers in Regional Science* 89: 3–25.

Anselin, L., and R. Florax. 1995. Small sample properties of tests for spatial dependence in regression models: Some further results. In *New Directions in Spatial Econometrics*, 21–74. Berlin: Springer.

Anselin, L., A. Bera, R. Florax, and M. Yoon. 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26: 77–104.

Anselin, L., R. Florax, and S. Rey. 2004. Econometrics for spatial models: recent advances. In *Advances in Spatial Econometrics*, 1–25. Berlin: Springer.

Anselin, L., I. Syabri, and Y. Kho. 2006. GeoDa: an introduction to spatial data analysis. *Geographical Analysis* 38: 5–22.

Arbia, G. 2006. *Spatial Econometrics, Statistical Foundations and Applications to Regional Convergence*. Heidelberg: Springer.

Banerjee, S., B. Carlin, and A. Gelfand. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall/CRC.

Belitz, C., A. Brezger, T. Kneib, S. Lang, and N. Umlauf. 2015. *BayesX: Software for Bayesian inference in structured additive regression models*. Version 2.1.

Bernardo, J., and A. Smith. 1994. *Bayesian theory*. New York: Wiley.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* 36, N.2: 192–236.

Besag, J., and P. Moran. 1975. On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* 62, N.3: 555–562.

Besag, J., J. York, and A. Mollié. 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43: 1–59.

Bivand, R. 2008. Implementing representations of space in economic geography. *Journal of Regional Science* 48, N.1: 1–27.

Bivand, R., V. Goméz-Rubio, and H. Rue. 2014. Approximate Bayesian inference for spatial econometric models. *Spatial Statistics* 9: 146–165.

Bivand, R., L. Anselin, O. Berke, R. Bivand, M. Altman, L. Anselin, R. Assuncao, G.A. Bernat, W. Müller. 2014. spdep: Spatial dependence: weighting schemes, statistics and models. Available via http://cran.r-project.org/web/packages/spdep/index.html. Accessed Feb 2014.

Blangiardo, M., and M. Cameletti. 2015. *Spatial and Spatial-temporal Bayesian Models with R-INLA*. New York: Wiley.

Carvalho, M. L., and I. Natário. 2008. *Análise de Dados Espaciais*. Sociedade Portuguesa de Estatística.

Clayton, D. 1992. Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-area Studies*, 205–220.

Cressie, N. 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

Dass, S., C. Lim, and T. Maiti. 2010. Experiences with approximate Bayes inference for the Poisson-CAR model. Technical Report RM679, Department of Statistics and Probability, Michigan State University.

Database of contemporary Portugal, organized by Fundação Francisco Manuel dos Santos. Available via https://www.pordata.pt/. Accessed 3 April 2016.

Doucet, A., N. De Freitas, and N. Gordon. 2001. *Sequential Monte Carlo methods in practice*. New York: Springer.

Fischer, M. 2006. *Spatial analysis and geocomputation*. Vienna: Springer.

Gamerman, D., and H. Lopes. 2006. *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*. Milton Park: Chapman and Hall, CRC.

Gelman, A., and D. Rubin. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7(4): 457–511.

Gelman, A., J. Hwang, and A. Vehtari. 1992. Understanding predictive information criteria for Bayesian models. *Journal of Statistics and Computing* 24: 997–1016 (1992)

Geweke, J. 1991. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN USA.

Goméz-Rubio, V., R. Bivand, and H. Rue. 2015. A new latent class to fit spatial econometrics models with integrated nested Laplace approximations. *Spatial Statistics: Emerging Patterns-Part2* 27: 116–118. Details of the implementation in http://www.math.ntnu.no/inla/r-inla.org/doc/latent/slm.pdf. Accessed Sep 2016.

Goméz-Rubio, V., R. Bivand, and H. Rue. 2015. *Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximations*. Technical Report-Preprint to Elsevier. Available via DIALOG. http://previa.uclm.es/profesorado/vgomez/SSTMR/papers/INLA-slm.pdf. Accessed Jun 2016.

Goodchild, M., R. Haining, and S. Wise. 1992. Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems* 6(5): 407–423.

Goodchild, M., L. Anselin, R. Appelbaum, and B. Harthorn. 2000. Toward spatially integrated social science. *International Regional Science Review* 23(2): 139–159.

Hoef, J. M., E.M. Hanks, and M. B. Hooten. 2017. On the Relationship between Conditional (CAR) and Simultaneous (SAR) Autoregressive Models. Preprint. arXiv:1710.07000

Hoff, P. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.

Hughes, D., and A. McGuire. 2003. Stochastic demand, production responses and hospital costs. *Journal of Health Economics* 22: 999–1010.

Jacquez, G.M., P. Goovaerts, A. Kaufmann, and R. Rommel. 2014. *SpaceStat 4.0 User Manual: Software for the Space-time Analysis of Dynamic Complex Systems*. Ann Arbor: BioMedware.

Lambert, D., J. Brown, and R. Florax. 2010. A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application. *Regional Science and Urban Economics* 40: 241–252.

Lee, P.M. 2012. *Bayesian Statistics, An Introduction*. Chichester: Wiley.

Lee, D. 2013. CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software* 55(13): 1–24.

Lee, D., A. Rushworth, and G. Napier. 2013. *CARBayesST: An R Package for Spatio-temporal Areal Unit Modelling with Conditional Autoregressive Priors. R package version 2.2*. Available via DIALOG. http://CRAN.R-project.org/web/packages/CARBayesST/. Accessed Feb 2016

Leroux, B., X. Lei, and N. Breslow. 1999. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. M.E. Halloran, D. Berry, 135–178. New York: Springer-Verlag.

LeSage, J. 1999. *The Theory and Practice of Spatial Econometrics*. Toledo: University of Toledo.

LeSage, J. 2014. Spatial econometric panel data model specification: A Bayesian approach. *Spatial Statistics* 9: 122–145.

LeSage, J. 2015. Software for Bayesian cross section and panel spatial model comparison. *Journal of Geographical Systems* 17(4): 297–310.

LeSage, J., and R. Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

Manski, C.F. 2000. *Economic Analysis of Social Interactions*. Cambridge: National Bureau of Economic Research.

McCullagh, P., and J. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Boca Raton: Chapman and Hall, CRC Press.

Natário, I. 2013. *Métodos Computacionais: INLA, Integrated Nested Laplace Approximation*. Boletim da Sociedade Portuguesa de Estatística Outono de 2013, 52–56.

Nelder, J. A., and Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370–384.

Ord, K. 1975. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70(349): 120–126.

Paulino, C., M.A. Turkman, and B. Murteira. 2003. *Estatística Bayesiana*. Lisbon: Caloust Gulbenkian Foundation.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124, 125.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* 6(1): 7–11.

Portal of the National Portuguese Health Service. Available via DIALOG. https://www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/saude-24.aspx. Accessed 3 Sep 2015.

Quddus, M. 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity - An analysis of London crash data. *Accident Analysis and Prevention* 40: 1486–1497.

R Core Team and contributors worldwide. 2013. *The R Stats Package*. R package version 3.5.0. Available via http://CRAN.R-project.org/

Rawlings, J., S. Pantula, and D. Dickey. 1998. *Applied Regression Analysis: A Research Tool*, 2nd ed. New York: Springer.

Riebler, A., S. Sørbye, D. Simpson, and H. Rue. 2016. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research* 25(4): 1145–1165.

Ripley, B. 1981. *Spatial Statistics*. Cambridge: Cambridge University Press.

Rizzo, M.L. 2007. *Statistical Computing with R*. Boca Raton: CRC Press.

Robert, C., and G. Casella. 2010. *Introduction Monte Carlo Methods with R*. Berlin: Springer.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian Inference for latent gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* 71: 319–392.

Rue, H., S. Martino, and F. Lindgren. 2012. *The R-INLA project*. R-INLA. Available via http://www.r-inla.org

Simões, P., and I. Natário. 2016. Spatial econometric approaches for count data: an overview and new directions. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* 10(1): 348–356.

Simões, P., M.L. Carvalho, S. Aleixo, S. Gomes, and I. Natário. 2017. A Spatial Econometric Analysis of the Calls to The Portuguese National Health Line. *Econometrics, MDPI Journals* 5: 24.

Simpson, D., H. Rue, A. Riebler, T.G. Martins, S.H. Sørbye. 2017. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* 32(1): 1–28.

Smith, B. 2004. *Bayesian output analysis program (BOA) for MCMC*. R package version 1(5).

Spiegelhalter, D., N. Best, B. Carlin, and A. Van der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B* 64(4): 583–639.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2003. *WinBUGS version 1.4 user manual*. Cambridge: MRC Biostatistics Unit.

Stan, Development and Team. 2014. *Stan: A C++ library for probability and sampling*. Available via DIALOG. http://mc-stan.org

Stern, H., and N. Cressie. 1999. Inference for extremes in disease mapping. In *Disease Mapping and Risk Assessment for Public Health*, ed. A.B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.F. Viel, R. Bertollini, 63–84. Chichester: John Wiley & Sons.

Stern, H., and N. Cressie. 2000. Posterior predictive model checks for disease mapping models. *Statistics in Medicine* 19(17–18): 2377–2397.

Team, R. 2013. R development core team. *R: A Language and Environment for Statistical Computing* 55: 275–286.

Turkman, M.A., and C. Paulino. 2015. *Estatística Bayesiana Computacional*. Sociedade Portuguesa de Estatística.

Turkman, M.A., and G. Silva. 2000. *Modelos Lineares Generalizados-da teoria á prática*. Sociedade Portuguesa de Estatística.

Valpine, P., D. Turek, C. Paciorek, C. Anderson-Bergman, D. Lang, and R. Bodik. 2017. Programming with models: writ-

ing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* 26(2): 403–413.

Vehtari, A., A. Gelman, and J. Gabry. 1999. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Journal of Statistics and Computing*. https://doi.org/10.1007/s11222-016-9696-4

Wall, M. 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121(2): 311–324.

Watanabe, S. 2010. Asymptotic equivalence of Bayes cross-validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11: 3571–3594.

Whittle, P. 1954. On stationary process in the plane. Biometrika 41: 434–449.

Modeling and Predicting Influenza Circulations Using Earth Observing Data

Radina P. Soebiyanto and Richard K. Kiang

Introduction

The Burden of Influenza

Influenza is a very common infectious disease. Almost everyone has been infected with influenza before and often more than once. Each year 35% of the world populations (Huang et al. 2019) may get infected with 290,000–650,000 deaths (World Health Organization 2019). The symptoms – fevers, chills, sore throats, cough headache, fatigue, etc. – may not be severe. Without vaccination, fever and other symptoms may last about a week, and cough and weakness may remain for 1 to 2 weeks longer. But for the young, the old, the pregnant, the immunologically impaired, and those with chronic medical conditions, influenza may lead to other infections and become fatal. Because the infected may miss work or school, need medical attentions or hospital cares, or simply perform less efficiently, the economic burdens due to lost productivity and lives may reach $90 billion in the USA alone in more severe influenza epidemics (Molinari et al. 2007).

Types of Influenza Viruses

Influenza viruses (Centers for Disease Control and Prevention 2019) are RNA viruses in the family *Orthomyxoviridae* and classified into four antigenic types – A, B, C, and D – based on their nucleoproteins. Types A and B have eight RNA segments that encode ten or more proteins, while Types C and D have seven RNA segments that encode nine proteins. Type A's natural reservoir is wild aquatic birds. Besides humans and birds, Type A also infects pigs, horses, dogs, and seals. Types B, C, and D do not have natural reservoirs. Types B and C infect only humans. The recently discovered Type D infects cattle and is not known to infect humans. Each antigenic type affects human populations differently: Type A is the most severe and may cause serious epidemics and pandemics; Type B is less so but can still result in outbreaks; and Type C only causes minor symptoms.

Based on the hemagglutinin (H) and the neuraminidase (N) protein on the surface of the Type A virus, Type A influenza viruses can further be divided into subtypes. Hemagglutinin is a glycoprotein used for binding to the host cell, and neuraminidase is an enzyme the virus uses to split the host's mucoprotein, in order to release the progeny of offspring viruses from the host cell. There are 18 types of hemagglutinin (H1–H18) and 11 types of neuraminidase (N1–N11). For example, the two main circulating subtypes in the 2018–2019 influenza season were A(H1N1)pdm09 and A(H3N2).

Antigenic Drift and Antigenic Shift

Influenza viruses are single-stranded RNA viruses and have a very high mutation rate compared with DNA viruses. The viruses mutate frequently through antigenic drifts with minor point mutations which allow the viruses to evade immune recognition. The World Health Organization (WHO) makes semi-annual recommendations for influenza vaccine composition. Vaccine production is always a race against time. Even in a normal influenza season, vaccination is often prioritized. Since it is difficult to make effective and long-

R. P. Soebiyanto (✉)
Universities Space Research Association, Goddard Earth Sciences Technology and Research, Columbia, MD, USA
e-mail: rsoebiyanto@gmail.com

NASA Goddard Space Flight Center, Greenbelt, MD, USA

R. K. Kiang
NASA Goddard Space Flight Center, Greenbelt, MD, USA

Current affiliation: Ecoepidemia, New York, NY, USA
e-mail: rkiang08@gmail.com

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_7

lasting vaccines, annual influenza epidemics continue to take place.

Occasionally the viruses undergo antigenic shifts by reassorting the genetic materials from different Type A's subtypes. Since the populations rarely have immunity again such reassorted strains, antigenic shifts may lead to pandemics with massive illnesses and deaths. Of the three types of influenza viruses that infect humans, only the Type A viruses can undergo antigenic shift and cause pandemics.

Pandemics

The deadliest influenza pandemic was the 1918–1920 Spanish flu (Palese 2004), in which up to 100 million people worldwide might have perished. The subsequent pandemics include the 1957–1958 Asian flu, the 1968–1969 Hong Kong flu, the 1977–1978 Russian flu, and the 2009–2010 A(H1N1)pdm09 flu. The last four pandemics, with 0.3–1.5 million deaths, were much less deadly than the Spanish flu.

The world has experienced five influenza pandemics over the past hundred years. No one can predict how and when the next pandemic will appear. However, a pandemic does not just appear overnight. It starts with Type A subtypes circulating in the human populations and other species' populations. It is possible to estimate how likely reassortments among the subtypes may take place, pathogenicity of the reassorted strains, and the probable severity if they bring on a pandemic. Good surveillance of the circulating subtypes in human and other species' populations is obviously important. Good influenza models may also help detecting a pandemic in formation.

The Roles of Climate and Weather in Influenza Transmission

In the Northern Hemisphere's temperate zone, influenza typically occurs seasonally. Every year influenza starts in the fall, picks up strength as it becomes colder in the winter, and then gradually recedes when spring comes. It is also known that the influenza trend in the Southern Hemisphere's temperate zone has a phase difference of 6 months. In general, November to March in the Northern Hemisphere and May to September in the Southern Hemisphere are considered the influenza seasons. This suggests that temperature and humidity, both are low in the winter, play a role in influenza transmission. However, in the tropics where temperature and humidity are continuously high, influenza circulates year-round.

Because influenza transmits through contacts with infected people's respiratory droplets drifting either in the air or on contaminated objects, any factors that promote such contacts enhance influenza transmission. Since precipitation encourages indoor crowding, it has been hypothesized that rains lead to crowding and promote influenza transmission. Furthermore, sunlight exposure and level of ultraviolet radiation, which stimulate vitamin D production and antagonize virus survival, have also been proposed as the environmental factors affecting influenza transmission.

Laboratory studies showed that viral stability has a nonlinear relationship with relative humidity (Schaffer et al. 1976) – maximum at low humidity (20–40%), minimum at intermediate humidity (50%), and turns high again at higher humidity (60–80%). In addition, low humidity (Tellier 2006) promotes the formation of respiratory droplets from bioaerosols. But at high humidity, droplets absorb water, become larger, and fall out of the air such that the opportunity of transmitting influenza is reduced. Several modeling studies have indicated the role of specific humidity in influenza transmission (Shaman et al. 2010; Tamerius et al. 2013; Soebiyanto et al. 2014, 2015a, b).

It has also been shown (Lowen et al. 2007) that cold air increases the viscosity of mucous layer, reduces mucociliary clearance, and makes it easier for viruses to travel along the respiratory tract. As discussed earlier, raining encourages crowding, in either warmer or cold weather, and increases the direct and indirect contacts between viruses and hosts. Therefore, precipitation is also an important predictor for influenza circulation independent of humidity and temperature.

As temperature, humidity, and precipitation are most directly related to viral survivorship, host susceptibility, transmission efficiency, and crowding, these three parameters are most frequently used in modeling and predicting influenza circulation.

Earth Observation Data and Models that Provide Meteorological Information

Table 1 shows satellite-derived data from NASA and NOAA that can be used to model influenza. Some of these datasets have been used in previous influenza studies, and specific humidity especially has been widely used as an indicator for forecasting influenza (Shaman et al. 2010; Tamerius et al. 2013; Soebiyanto et al. 2014, 2015a, b). Specific humidity measures the mass of water vapor in a unit mass of air (expressed in g/kg). It is different from relative humidity as it does not depend on temperature and is conceptually similar to absolute humidity, which measures the mass of water vapor in a unit volume of air (expressed in g/m^3). In modeling influenza, we have previously used near-surface specific humidity variable obtained from the Global Land Data Assimilation System (GLDAS) dataset (Rodell et al. 2004). It is a NASA-NOAA system that assimilates ground and satellite measurements to model global terrestrial geo-

Table 1 Remote sensing and assimilated data products and their geophysical parameter that can be used to model influenza

Dataset/sensor/platform	Geophysical parameter	Resolution	
		Spatial	Temporal
MODIS (moderate resolution imaging Spectroradiometer)	Land surface temperature	0.05° (5 km)	Daily
Global Precipitation Measurement Mission (GPM) – 3IMERGHH	Rainfall	0.1° (11 km)	30 minutes
NOAA climate prediction Center unified (CPC-UNI)	Rainfall	0.5° (50 km)	Daily
Global land data assimilation system	Air temperature	0.25° (25 km)	3-hourly
	Specific humidity		
	Rainfall		
	Solar radiation		
CRU CL (climate research unit)/Oxford university	Temperature	10-minute	Monthly
	Relative humidity		
	Precipitation		

physical parameters with contiguous spatial and temporal coverage. The dataset has global coverage with 0.25° resolution and is available at 3-hourly or monthly time step since 2000. Global near-surface air temperature (2 meter) and precipitation can also be obtained from the GLDAS dataset.

There are other precipitation measures (Table 1) including satellite-derived Global Precipitation Measurement (GPM) and gauge-based NOAA Climate Prediction Center (CPC) Unified (CPC-UNI). GPM is a constellation of satellites initiated by NASA and Japan Aerospace Exploration Agency (JAXA) that comprises of consortium of international space agencies including the Centre National d'Études Spatiales (CNES), the Indian Space Research Organization (ISRO), the National Oceanic and Atmospheric Administration (NOAA), and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), among others. It is a successor of NASA-JAXA Tropical Rainfall Measuring Mission (TRMM) that was decommissioned in 2014. The GPM Core Observatory that was launched in 2014 carries the first space-borne dual-frequency precipitation radar (DPR) and a multi-channel GPM Microwave Imager (GMI). These sensors allow for better detection of rainfall and snowfall with a global coverage as compared to sensors aboard TRMM. GPM dataset is available at 0.1° resolution every 30 minute since 2014. The CPC UNI dataset, on the other hand, is a reanalysis dataset based on daily gauge measurements worldwide. It is derived from quality-controlled daily reports from more than 30,000 stations worldwide, and they are interpolated using optimal interpolation (OI) method with orographic consideration (Chen et al. 2008). The data is produced every day at real time with 0.5° resolution. Although this dataset has lower resolution, it is produced at near real time with worldwide coverage and temporal coverage since 1979, which makes it suitable to detect any anomalous conditions (i.e., above normal rain) and operational purpose (i.e., monitoring).

We have previously used land surface temperature (LST) measure in modeling influenza in Hong Kong (Soebiyanto et al. 2010). LST can be obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument which has 36 bands spanning from the visible to the long-wave infrared spectra. Both Terra and Aqua missions of NASA's Earth Observing System carry this instrument. Although there are differences between LST and air temperature, changes in LST may induce convection at the boundary layer and influence air temperature, winds, cloudiness, and precipitation – all of which affect the influenza transmission. LST daily dataset is available globally since 2000 at 0.05° resolution.

Common Methodologies for Modeling Influenza Circulations

Any mathematical techniques that map one set of parameters to another set can be used to model the association of influenza circulations with meteorological variables (Thompson et al. 2006; Lofgren et al. 2007). Several methods which we have used – ranging from time series analysis to machine intelligence and spectral analysis – are described briefly below.

Logistic regression is a common method for modeling disease occurrence (Hosmer and Lemeshow 2000). As a person can be tested either positive or negative with influenza infection, logistic regressions are suitable for estimating the infected population. It can model strictly bounded response variable that is suitable for data on proportions. The predictors, or the independent variables, may include binary and continuous variables. For example, holiday or non-holiday is a binary variable, and all meteorological parameters are continuous variables. We previously used logistic regression to model the weekly proportion of influenza samples that are tested positive, referred as influenza-positive proportion (Soebiyanto et al. 2014, 2015a). Given the meteorological condition in week t, the odds for proportion (or probability) of samples that were tested positive for influenza would be higher if the meteorological condition was suitable for influenza transmission because more people would likely be infected.

For each week t, if Y_t denotes the number of samples tested positive for influenza out of total samples examined at that

week (N_t), then Y_t is a binomial random variable. That is, Y_t Bin (N_t, p_t) where p_t is the proportion of influenza samples that are tested positive in week t. If we denote the logit of the influenza-positive proportion as

$$z_t = \ln\left(\frac{p_t}{1-p_t}\right)$$

the logistic regression is then:

$$z_t = \alpha + \sum_j \beta_j x_{jt} + \sum_l \gamma_l v_{lt} + \sum_m \lambda_m z_{(t-m)} + \sum_n \theta_n w_t^n$$

where

x_{jt} Meteorological variable j in at week t; $j \in$ {temperature, specific humidity, rainfall}

v_{lkt} Proportion of samples that are positive for virus l at week t; $l \in$ {respiratory syncytial virus (RSV), adenovirus, parainfluenza virus}

w_t^n Week number (1–52)

α Intercept

$\beta, \gamma, \lambda, \theta$ Regression coefficients

In the equation above, the explanatory variables included were the meteorological variables (temperature, specific humidity, and rainfall), the proportion of samples that are tested positive for other respiratory viruses that co-circulated with influenza (such as RSV, parainfluenza viruses, and adenoviruses), lagged of the influenzas positive proportion (up to week 4 lags), and a polynomial function of the week number. We included the co-circulating viruses to adjust for any potential confounding associations between influenza-positive proportion and the meteorological variables. The lagged dependent variable was included because influenza activity in a particular week depends on previous week activity (i.e., how many people were infected previously which can potentially infect the susceptible population). The inclusion of the lagged dependent variable also accounts for autocorrelation. Lastly, the week number was included to represent influenza seasonality and other nonlinear relationships that were not represented by the meteorological variables in the model.

Variations of logistic regression may be used to accommodate the characteristics of the disease data. When assumptions can be made that the influenza data – for example, the proportion of the samples tested positive – follows binomial or Poisson distribution, the generalized linear model (GLM) (Dobson and Barnett 2008) which allows for non-normal error distributions for the influenza data may be used. A GLM uses a link function to relate the influenza data to linear models. A logit link is normally used for binomial regression and a log link for Poisson regression.

One of regression models that can account for nonlinear relationship is generalized additive model (GAM) through the use of smoothing spline. This type of model has been used to assess the relationship between meteorological factors and influenza-associated mortality (Barreca and Shimshack 2012). In addition, we previously employed this method to successfully modeled influenza and meteorological condition relationships for cities in Europe (Berlin, Germany; Slovenia, Ljubljana; Castile and León in Spain) and the districts of Israel (Soebiyanto et al. 2015b). In this study, influenza activity was represented by weekly new cases of influenza among patients with influenza-like illnesses (ILI). The GAM model for influenza can be written as:

$$\ln(y_t) = \alpha + \sum_j s\left(x_{jt}\right) + s\left(\ln\left(y_{t-1}\right)\right)$$

where y_t is the influenza activity in week t, x_{jt} is meteorological variable j at time t, y_{t-1} is influenza activity during the previous week, and $s(.)$ indicates the smooth spline function. There are different choices of smoothing functions, and in the study, we opted for the penalized cubic regression smoothing splines:

$$s(x) = \sum_k b_k(x)\beta_k$$

β_k are the parameters to be estimated and $b_k(x)$ are the basis functions for the cubic splines.

Neural network, a machine intelligence technique, may also be used for modeling (Haykin 1999). A common type of neural network is multilayer perceptron, which consists of an input layer and an output layer and optionally one or more hidden layers. More complex network gains training accuracy but sacrifices generalization. As a universal mapping tool, neural network is easy to set up. But it can be challenging to interpret the modeling rationale and obtain statistical significance.

Because a disease time series can be noisy, time-frequency decomposition with wavelet transform (Mallat and Peyré 2009) and empirical mode decomposition using Hilbert-Huang transform (Huang et al. 1998, 2019) may reveal better association with meteorological and social events, such as holidays, school closing, etc. The selected decomposed components are further analyzed with other modeling methods.

Case Study: Influenza in Austria

To illustrate the use of satellite-derived meteorological data to model influenza, we will assess influenza activity in Austria. We extracted influenza data from WHO FluNet system where we obtained the weekly number of influenza specimens tested and the number of specimens tested positive for

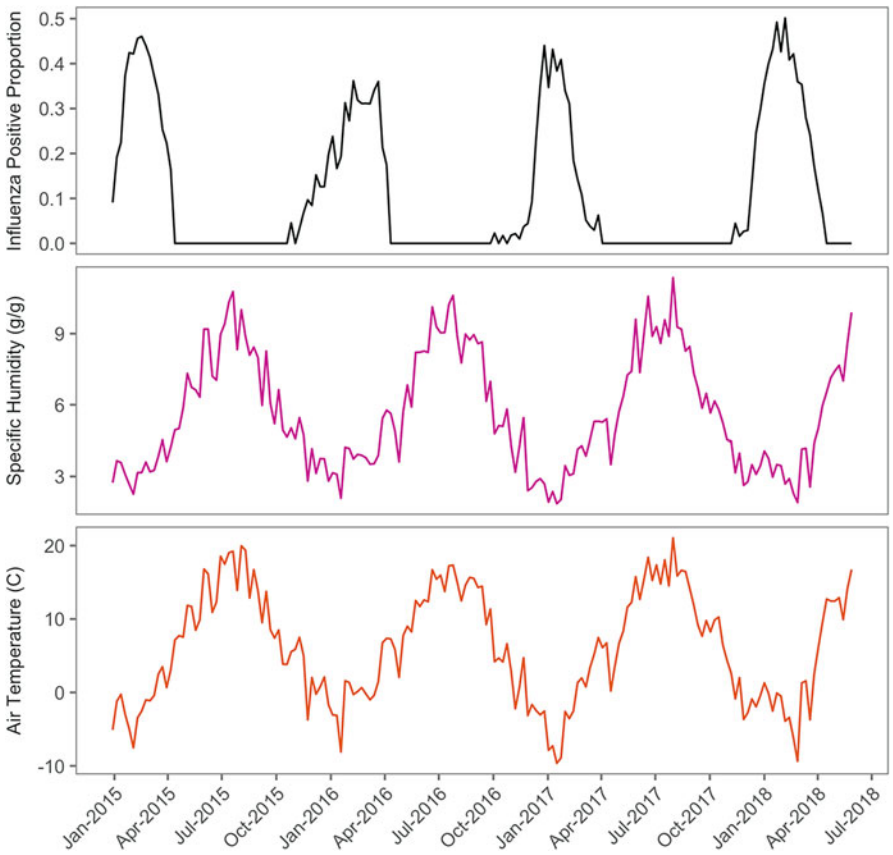


Fig. 1 Weekly influenza-positive proportion in Austria (**a**), weekly specific humidity (**b**), and air temperature (**c**)

influenza (all types and subtypes) between January 2015 and June 2018. Meteorological data was obtained from the Global Land Data Assimilation System (GLDAS) for near-surface temperature and specific humidity. Our previous study had indicated that rainfall was not a significant determinant for influenza in the temperate region (Soebiyanto et al. 2015b) and therefore we exclude it in this analysis. GLDAS is a NASA-NOAA system that utilizes ground and satellite measurements to model global terrestrial geophysical parameters with contiguous spatial and temporal coverage. In order to obtain weekly time series that matches influenza data, we first averaged the pixels that lie within Austria, followed by averaging the 3-hourly data into weekly data. For each of the meteorological variable, we created 1-week lag composite. We used univariate generalized linear model regression to model the weekly influenza-positive proportion, which is the number of influenza specimens tested positive for influenza divided by the total specimens tested. Since specific humidity and temperature were highly correlated, we tested each variable in univariate regression so as to avoid collinearity.

The weekly influenza-positive proportion in Austria showed a strong seasonality that typically starts in December and peaks in February (Fig. 1a). During this time, specific humidity and air temperature are typically at their lows (Fig. 1b, c) indicating dry and cold conditions. The long-term mean of specific humidity during winter months (January to March) when influenza is typically at its peak is remarkably low (Fig. 2a) when compared to the summer months (June to August) (Fig. 2b). We observed similar pattern in air temperature long-term mean (Fig. 2c, d). Scatterplot of influenza-positive proportion with specific humidity and air temperature showed inverse relationship (Fig. 3).

We assessed the relationship between influenza-positive proportion and the lagged (1 week) specific humidity and air temperature, separately, using univariate regression. Our results (Table 2) indicated that influenza-positive proportion in Austria was inversely associated with specific humidity and air temperature (p-value <0.05). These findings are consistent with ours and other studies of influenza in temperate regions (Shaman et al. 2010; Soebiyanto et al. 2010, 2015b). The modeled influenza from these regression models showed agreeable pattern with the observed data (Fig. 4).

Although there are variations in magnitudes between the modeled and observed influenza activity (Fig. 4), the modeled influenza activity timing is relatively in agreement with the observed data. We observed similar results in our studies for influenza in other regions (Soebiyanto et al. 2010, 2014, 2015a, b). Using such models, combined with influenza surveillance data and seasonal meteorological forecasts, one can estimate the timing of influenza onset and/or peak a

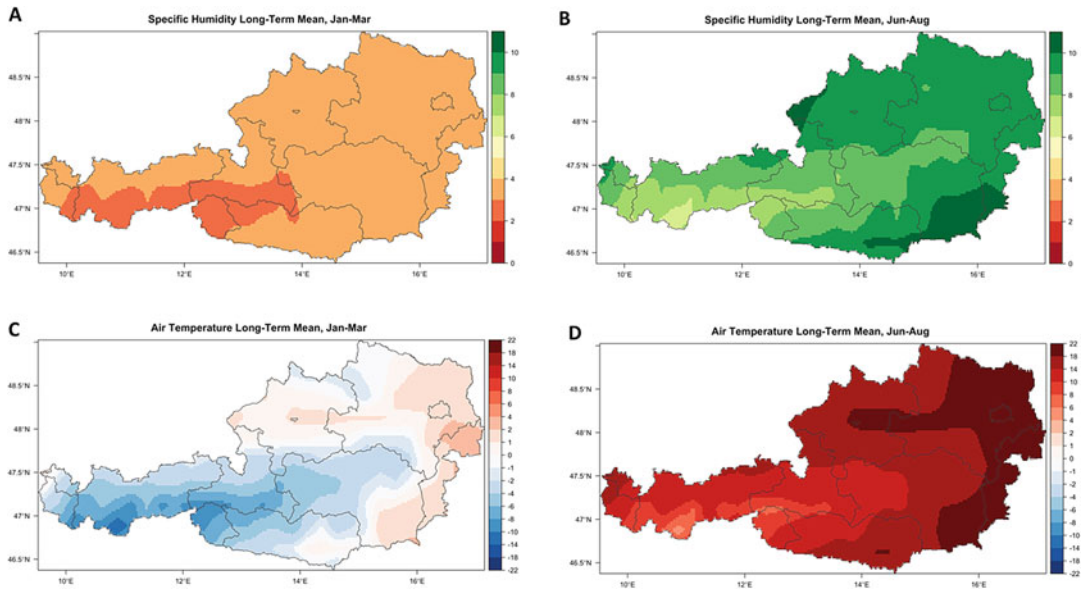


Fig. 2 Climatology during winter (January to March) and summer (June to August) months in Austria: specific humidity long-term mean (**a** and **b**) and air temperature (**c** and **d**)

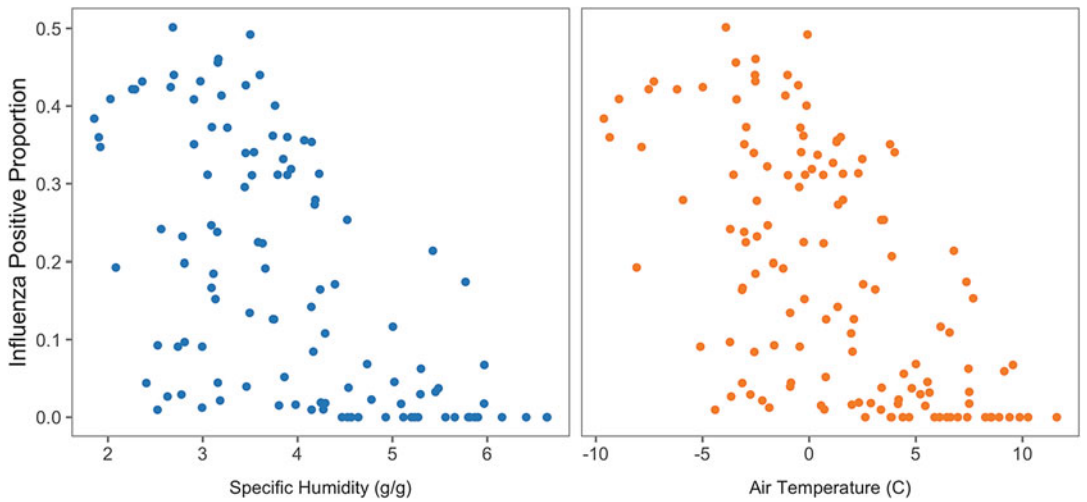


Fig. 3 Scatterplot of influenza-positive proportion and specific humidity (left) and air temperature (right)

few months ahead. Such information can provide a guide for public health agencies to plan for prevention and control efforts with sufficient window.

Discussions

Although it is the most common infectious disease, each year influenza incurs huge economic loss as well as significant morbidity and mortality. If an influenza epidemic becomes a serious pandemic, the mortality rate may reach 20% (as in the 1918 pandemic), and the total loss to all the affected countries would be immeasurable.

Vaccination of the general population, taking precautions in public places, and timely and effective treatment of those infected are practical ways to respond to influenza epidemics. For infected individuals, the effective window of treatment to reduce symptoms and the likelihood of infecting others is narrow. Reliable modeling and predictive capabilities for influenza circulation will help the public health organizations to more effectively respond to the epidemic.

Table 2 Univariate regression for influenza-positive proportion in Austria

Meteorological determinant	Coefficient	*p*-value	Adjusted R^2
Specific humidity (1 week lag)	−0.044	<0.05	0.47
Temperature (1 week lag)	−0.016	<0.05	0.56

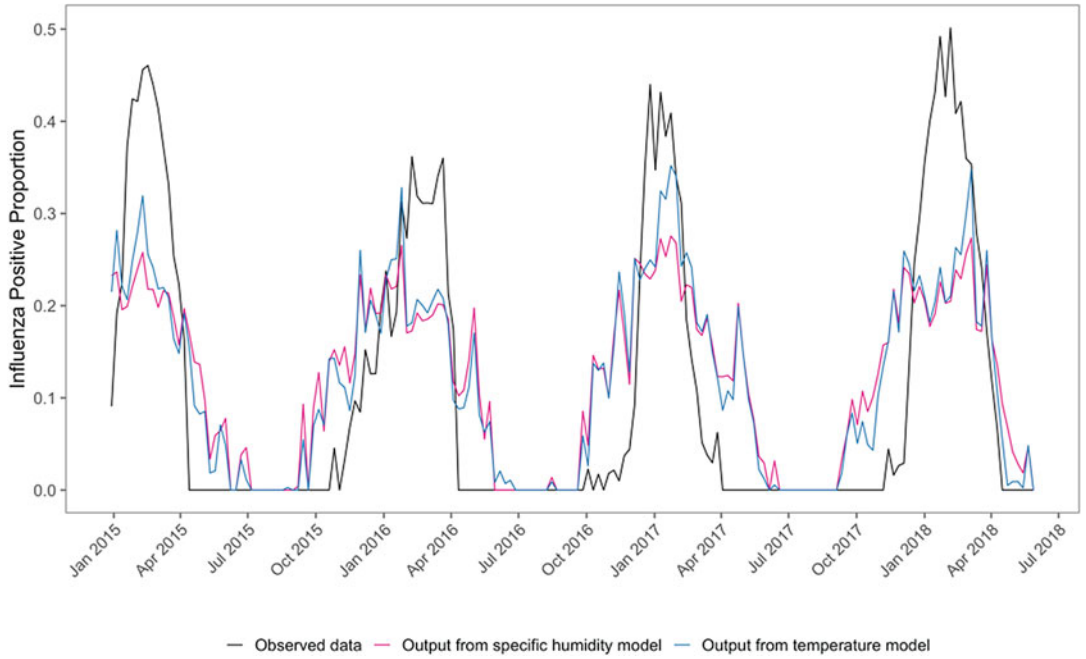


Fig. 4 Observed (black line) and modeled influenza-positive proportion using specific humidity (pink line) and temperature (blue line)

In this modern, interconnected world, epidemic-prone respiratory infections, including influenza, spread rapidly. Infections can be introduced into the home country almost immediately through airlines passengers regardless of whether it is in or out of the influenza season. Therefore in addition to paying attention to influenza surveillance in the home country, it is also essential to monitor the influenza circulations in other countries.

We have shown that reasonable accuracies can be obtained when using Earth observing data to model and predict influenza circulations (Shaman et al. 2010; Soebiyanto et al. 2010, 2014, 2015a, b). Beyond meteorological conditions, however, there are other important factors that determine influenza's epidemic potential. For example, vaccine selection and manufacturing, timely availability of the vaccine to the general population, similarity of the circulating strains to those in previous seasons, large social gatherings, and population movement such as pilgrimage or refugees from military conflicts all contribute to influenza circulation. Furthermore, the meteorological variables where influenza transmission takes place, often indoor or in public transportation systems, correlate with but differ from those obtained from Earth observing data. It is conceivable that the more urbanized the region is, the more the true meteorological variables differ from those derived from Earth observing data. All the factors described here may limit the modeling and prediction accuracies.

References

Barreca, A.I., and J.P. Shimshack. 2012. Absolute humidity, temperature, and influenza mortality; 30 years of county-level evidence from the United States. *American Journal of Epidemiology* 176: S114–S122.

Centers for Disease Control and Prevention. 2019. *Understanding Influenza Viruses*. Available at: https://www.cdc.gov/flu/about/viruses/ . Last Accessed 4 Nov 2019.

Chen, M., W. Shi, P. Xie, V.B.S. Silva, V.E. Kouscy, R.W. Higgins, et al. 2008. Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research* 113: D04110.

Dobson, A.J., and A.G. Barnett. 2008. *An introduction to generalized linear models*. CRC Press, Boca Raton, Florida.

Haykin, S.S. 1999. *Neural networks : A comprehensive foundation*. Prentice Hall, Hoboken, New Jersey.

Hosmer, D.W., and S. Lemeshow. 2000. *Applied logistic regression*. Hoboken: Wiley.

Huang, N., Z. Shen, S.R. Long, M. Wu, H. Shih, and Z. Quanan. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceeding of the Royal Society of London* 454: 903–995.

Huang, Q.S., D. Bandaranayake, T. Wood, E.C. Newbern, R. Seeds, J. Ralston, et al. 2019. Risk factors and attack rates of seasonal influenza infection: Results of the southern hemisphere influenza and vaccine effectiveness research and surveillance (SHIVERS) Seroepidemiologic cohort study. *The Journal of Infectious Diseases* 219: 347–357.

Lofgren, E., N.H. Fefferman, Y.N. Naumov, J. Gorski, and E.N. Naumova. 2007. Influenza seasonality: Underlying causes and modeling theories. *Journal of Virology* 81: 5429–5436.

Lowen, A.C., S. Mubareka, J. Steel, and P. Palese. 2007. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathogens* 3: 1470–1476.

Mallat, S.G., G. Stéphane, and G. Peyré. 2009. *A wavelet tour of signal processing : The sparse way*. Elsevier/Academic Press, Amsterdam.

Molinari, N.-A.M., I.R. Ortega-Sanchez, M.L. Messonnier, W.W. Thompson, P.M. Wortley, E. Weintraub, et al. 2007. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine* 25: 5086–5096.

Palese, P. 2004. The great influenza the epic story of the deadliest plague in history. *The Journal of Clinical Investigation* 114: 146–146.

Rodell, M., P.R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, et al. 2004. The global land data assimilation system. *Bulletin of the American Meteorological Society* 85: 381–394.

Schaffer F. L., Soergel M. E., Straube D. C. 1976. Survival of airborne influenza virus: effects of propagating host, relative humidity, and composition of spray fluids. *Arch. Virol.* 51, 263–273.

Shaman, J., V.E. Pitzer, C. Viboud, B.T. Grenfell, and M. Lipsitch. 2010. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology* 8: e1000316.

Soebiyanto, R.P., F. Adimi, and R.K. Kiang. 2010. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PLoS One* 5: e9450.

Soebiyanto, R.P., W. Clara, L. Castillo, O. Sorto, S. Marinero, M. De Antinori, et al. 2014. The role of temperature and humidity on seasonal influenza in tropical areas: Guatemala, El Salvador and Panama, 2008-2013. *PLoS One* 9: e100659.

Soebiyanto, R.P., W.A. Clara, J. Jara, A. Balmaseda, J. Lara, M. Lopez Moya, et al. 2015a. Associations between seasonal influenza and meteorological parameters in Costa Rica, Honduras and Nicaragua. *Geospatial Health* 10.

Soebiyanto, R.P., D. Gross, P. Jorgensen, S. Buda, M. Bromberg, Z. Kaufman, et al. 2015b. Associations between meteorological parameters and influenza activity in Berlin (Germany), Ljubljana (Slovenia), castile and León (Spain) and Israeli districts. *PLoS One* 10.

Tamerius, J.D., J. Shaman, W.J. Alonso, K. Bloom-Feshbach, C.K. Uejio, A. Comrie, et al. 2013. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathogens* 9: e1003194.

Tellier R. 2006. Review of aerosol transmission of Influenza A virus. *Emerging Infectious Diseases* 12, 1657–1662.

Thompson, W.W., L. Comanor, and D.K. Shay. 2006. Epidemiology of seasonal influenza: Use of surveillance data and statistical models to estimate the burden of disease. *The Journal of Infectious Diseases* 194 (Suppl): S82–S91.

World Health Organization. 2019. *Influenza (Seasonal)*. Available at: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Last Accessed 4 Nov 2019.

Using the NASA Giovanni System to Assess and Evaluate Remotely-Sensed and Model Data Variables Relevant to Public Health Issues

James G. Acker

Introduction

At the advent of the twenty-first century, the field of Earth science remote sensing was in the midst of profound evolutionary changes with regard to both the marked increase in data volume being acquired by recently launched observational missions and the computational resources that were being developed and expanded to archive, distribute, and analyze these data. The (somewhat unexpected) swift expansion of the World Wide Web, and related network capabilities, forced an examination of the current state of the data distribution process from NASA missions and how it would be altered by the increasing capability of individual researchers to access and use these data.

The NASA Earth Observing System (EOS), conceived in the late 1980s and early 1990s (McElroy and Williamson 2004), anticipated its data archive and distribution system based on technology that was current at that time. Thus, the Earth Observing System Data and Information System (EOSDIS) was a centralized, hardware-intensive system. Remotely-sensed land, atmospheric, and oceanic data from instruments on the EOS satellites, which evolved to become the Terra (launched in 1999), Aqua (launched in 2002), and Aura (launched in 2004) satellites, were distributed to Distributed Active Archive Centers (DAACs), which were responsible for archiving data from specific instrument missions and distributing it to the research community. In the mid- to late-1990s, the expectation was that the distribution of the data would primarily be on physical media, primarily magnetic tapes (McElroy and Williamson 2004).

J. G. Acker (✉)
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), Adnet Systems, Inc., Greenbelt, MD, USA
e-mail: james.g.acker@nasa.gov

The remarkable increase in data network capabilities (embodied in the public mind as the Internet and World Wide Web) that occurred essentially simultaneously with the launch of the EOS satellites required NASA to reevaluate how its data distribution function for satellite data from EOS and other Earth observing missions would be carried out. It was realized across the enterprise that distribution of data via the network, that is, electronically, would come to dominate over the "old way" that the data distribution function had been performed and eventually lead to its phase-out. This remarkable shift required EOSDIS and the NASA DAACs to dramatically increase their networking capabilities. So the first major shift in DAAC data distribution procedure was to transmit the data to research users electronically, over the computing network, rather than by mailing data tapes to them (Acker 2015).

The World Wide Web markedly increased the amount of information that could accompany the data and the information that the NASA DAACs could provide to users online, such as digital documents, Web pages, and data archives. Data placed in online archives could be accessed and downloaded by users using simple protocols. "Anonymous FTP" sites proliferated, making it quite easy to acquire large volumes of data from NASA observational missions.

However, there was still one particular aspect of the data acquisition and analysis process which still adhered to the older model of the data center. Even though data could now be acquired quickly via the network, it was still downloaded to a user system essentially in raw form, just as it had been previously read from tapes. To process and analyze the data, additional software was required to read, translate, and transform the data into a format that could be used by a user's personal and potentially idiosyncratic, hardware and software system. Furthermore, doing so frequently required individuals with computer programming skills, which resulted in

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_8

separating scientists who were not skilled programmers from the data until it had been translated for their use.

The situation described above summarizes the "state of the art" in the early years of the twenty-first century for the NASA DAAC system (which now encompasses 12 data centers). At Goddard Space Flight Center, atmospheric scientist Yoram Kaufman expressed to Goddard Earth Sciences DAAC Science Data Manager Gregory Leptoukh that he would like to have a way to examine and analyze the data he used as a researcher on atmospheric aerosols without needing a team of programmers to translate it for him first. Based on these conversations, and also due to other nascent online data services existing at the time, Leptoukh envisioned a system where the analysis of the data could take place at the data center, rather than on the user's own system. Such a system would obviate the need for knowledge of data formats and software required to specifically unpack scientific data files and even reduce the need for software to analyze the data. Leptoukh conceived of the system as a way to explore the data, identifying features of interest and compiling preliminary guiding analyses, before using data at higher spatial and temporal resolution—which at that time could involve lengthy processing times and considerable data storage resources. (Note: the Goddard Earth Sciences DAAC is now named the Goddard Earth Sciences Data and Information Services Center, GES DISC.)

The initial system, which utilized the Grid Analysis and Display System (GrADS) as its analytical core (Doty and Kinter III 1995), was named either the *MO*derate-Resolution Imaging Spectroradiometer (MODIS) *V*isualization and *A*nalysis *S*ystem (MOVAS) or the *T*ropical Rainfall Measurement Mission (TRMM) *V*isualization and *A*nalysis *S*ystem (TOVAS). MOVAS was set up for MODIS atmospheric data variables, including atmospheric aerosols, and TOVAS was set up for precipitation variables. When additional missions were added, a broader acronym was required, which resulted in the system being named the *G*oddard Earth Sciences *I*nteractive *O*nline *V*isualization *AN*d a*N*alysis *I*nfrastructure—GIOVANNI. After a few years, it became simpler to refer to the system as *Giovanni* without reference to its acronym definition, though the "G" now officially represents the word "Geospatial."

Since this beginning, Giovanni has gone through several evolutionary stages, and it now exists in its fourth version. New releases of the system are designated numerically, so the current release is Giovanni 4.35. The system now offers 22 different analysis options. The most popular analysis options are maps, which can be averaged over specified time periods; area-averaged time series, which can be constructed for a user-specified region; animations, showing maps of successive time steps for specific data variables; Hovmöller diagrams, for which both latitude-time and longitude-time options are available; and atmospheric profiles for sounding instruments which collect data at several atmospheric altitudes. As an example of the expansion of Giovanni's analytical capabilities, time series can now be created for specific months or seasons over multi-year periods. Table 1 lists the current analytical options in Giovanni.

While the analytical options in Giovanni are important, of greater importance to scientists and application users (including researchers concerned with public health issues) are the data holdings in Giovanni. Data available in Giovanni now include precipitation data from multiple missions, as well as hydrological models; additional hydrological data variables from assimilation models; detailed meteorological data variables from a reanalysis model; atmospheric chemistry variables; atmospheric aerosol variables; selected ocean color data variables; vegetation indices; and also tailored data products for specific socioeconomic issues.

In the following text, we will first provide several examples of the use of Giovanni for public health issues to demonstrate the system's broad applicability in the public health regime. The subsequent section will highlight specific data variables and data sets in relation to the areas of health that they can be applied to. The final section provides a case history of the use of Giovanni for a public health issue that has wide concern—discomfort index and heat stress—which will demonstrate Giovanni's ease of use for such investigations.

Table 1 Analysis options currently available in Giovanni

Maps	Comparisons	Vertical choices	Time-series	Miscellaneous
Time-averaged variable	Correlation map (two variables)	Cross-section, latitude vs. pressure	Hovmöller, longitude-averaged	Zonal mean
Animations	Area-averaged scatter plot (static)	Cross-section, longitude vs.pressure	Hovmöller, latitude-averaged	Histogram
Difference of time-averaged variables	Area-averaged scatter plot (interactive)	Cross-section, time vs. pressure	Area-averaged differences	
Accumulated variable	Scatter plot (static)	Vertical profile	Area-averaged	
Time-averaged overlay	Time-averaged scatter plot (interactive)			
Monthly and seasonal averages				

Categorization of Current Data Holdings in Giovanni

The data variables in Giovanni can be categorized with respect to their applicability to public health issues. In the following, three tiers of applicability will be presented: Tier 1, data that have a strong relationship to public health, and which are thus directly applicable in public health research; Tier 2, data that have indirect yet established relationships with an area of public health concern; and Tier 3, data that are related to weather or climate with an effect on public health and well-being.

Tier 1 Data Variables

Tier 1 data types include:

- Precipitation
- Temperature
- Aerosol optical depth (AOD)
- Nitrogen dioxide (NO_2)
- Carbon monoxide (CO)
- Ozone (O_3) erythemal ultraviolet (UV) daily dose
- Relative humidity

The GES DISC is NASA's designated archive for remotely-sensed and related precipitation variables, and thus, Giovanni has many different types, and these range in temporal resolution from monthly to hourly. Remotely-sensed quantities from TRMM and the Global Precipitation Measurement (GPM) mission are available, alongside modeled precipitation variables from the Land Data Assimilation System (LDAS) data sets.

Temperature data are available as a remotely-sensed variable (from both MODIS and the Atmospheric Infrared Sounder, AIRS) and as an assimilated model variable as well. Relative humidity is provided as an AIRS data variable.

Widely used AOD variables, directly relevant to air quality issues, are provided by the Ozone Measuring Instrument (OMI) and MODIS. OMI also provides NO_2, SO_2, ozone concentration (in Dobson units), and erythemal UV daily dose.

Tier 2 Data Variables

Tier 2 health-related variables in Giovanni are:

- Chlorophyll concentration (phytoplankton)
- Euphotic depth
- Sea surface temperature
- Normalized difference and enhanced vegetation indices (NDVI/EVI)
- Soil moisture

As the discussion of research papers noted, chl *a* indicates phytoplankton activity, and increased phytoplankton concentrations are connected to disease (cholera), harmful algal blooms (HABs), and eutrophication causing hypoxia or anoxia in the water column. Chl *a* is also a foundational variable for fisheries research. Euphotic depth, as an indicator of turbidity, is also related to water quality. SST data can also be related to eutrophication potential, fisheries (as an indicator of currents and upwelling), and factors contributing to tropical storm system intensities.

Vegetation indices are related to droughts and agricultural impact and can also indicate where insect-borne diseases may be an increased risk, which can happen for both drier-than-normal and wetter-than-normal conditions. Soil moisture offers similar research applications, as well as indicating flood potential and the after-effects of severe storms, and is also a critical variable for agriculture.

Tier 3 Data Variables

Tier 3 data types may be related to weather and climate, with effects on public health and well-being. Many of these data types measure quantities that are important to water resources:

- Snow depth
- Snow mass
- Snowfall rate
- Snowmelt
- Fractional snow cover
- Cloud cover
- Snow/ice frequency
- Wind speed
- Runoff

The snow variables listed here have the potential to be used for water resource studies, as snowpack is an important water resource for many communities in mountainous regions and contributes to reservoir levels. Trends in snow cover and related variables are also climate change indicators.

Cloud cover and wind speed both represent important weather factors and can be used in examination of weather system and interannual variability. Wind speed is also useful for severe storm research, and cloud cover can be related to solar radiation exposure, drought, and even HAB occurrence.

Examples of Public Health Research Using Giovanni

One of the best ways that the use of Giovanni has been continuously assessed is via the compilation of peer-reviewed journal research papers that cite the use of the system. Over 2,000 papers citing Giovanni have been published since the first paper which used the system appeared in 2004. Many of these papers have dealt with public health topics. We have characterized the use of the system for public health by the following categories:

- Air Quality
- Water Quality
- Epidemiology
- Erythemal Radiation Exposure
- Disaster Assessment
- Agriculture and Nutrition

The following provides examples of published papers for each category.

Air Quality

Due to the availability of atmospheric aerosol data from multiple instruments, including MODIS and OMI, and the multi-year length of the data sets from these instruments, a large number of journal papers have evaluated public health in relation to atmospheric aerosols. OMI also provides nitrogen dioxide (NO_2) data that are indicative of combustion processes, so it has been used for both urban air quality investigations and wildfire smoke research.

The Atmospheric Infrared Sounder (AIRS) data variables include carbon monoxide (CO) and methane (CH_4). Many CO data variables acquired by the Measurement of Pollution in the Troposphere (MOPITT) instrument were recently added to Giovanni.

Several data variables in Giovanni's reanalysis data sets also allow insight into air quality. The Modern Era Retrospective-analysis for Research and Applications – 2 (MERRA-2) data set provides such variables as Black Carbon, CO, dust (dry deposition, wet deposition, mass density, etc.), sulfur dioxide (SO_2), and atmospheric aerosols.

Ozone data are provided by both instrument observations and modeled data sets. OMI, AIRS, and MERRA-2 all include ozone data variables. In addition, Giovanni provides access to ozone data from the Total Ozone Mapping Spectrometer (TOMS) missions. Ozone data extend back to November 1978 for data from the TOMS instrument on the Nimbus 7 satellite.

The following discussion briefly describes a variety of papers which utilized Giovanni for research into air quality topics.

A 2010 paper by Lu et al. (2010) described the creation of an inventory of SO_2 and carbonaceous aerosol emissions for China and India from 1996 to 2010, noting the variability of emission trends and its relationship to economic growth. The researchers compared their model-based inventories and emission trend estimates to satellite data and reported good agreement.

The atmosphere over India was also the subject of Kishcha et al. (2011), which examined the trends in aerosol optical thickness (AOT; also referred to as Aerosol Optical Depth, AOD) and the relationship of these trends to population growth. Kishcha et al. used data from MODIS and the Multi-Angle Imaging Spectroradiometer (MISR) for the period March 2000–February 2008. They found that AOT was highest near the largest population centers, and the regions with the most rapid population growth also had the fastest-increasing trends in AOT. They inferred from the AOT trends that high-population centers were currently experiencing deterioration in air quality, and these trends could worsen, causing an increase in health-related problems.

Urban patterns of temperature, precipitation, and atmospheric aerosols vary systemically on a daily and weekly basis, according to the clear results of Sitnov (2011) for the city of Moscow. The patterns are most pronounced during the warm months of summer. Temperature showed a maxima for Monday–Thursday and a minima on Saturday and Sunday. This pattern was similar for precipitation when it exceeded 10 mm of accumulation or more. Atmospheric aerosols had an out-of-phase relationship with temperature and precipitation, with the minima occurring Tuesday–Friday and the significantly higher AOD values occurring on Saturday and Sunday. Sitnov states that the patterns indicated the likelihood of anthropogenic forcing on temperature and precipitation, related to the aerosol pollution weekly cycle, but that the actual mechanisms were difficult to determine and could result from the interplay of both anthropogenic and natural factors.

Simha et al. (2013) looked at a somewhat unique topic—the potential effect on regional climate caused by a famous celebration in India, the Holi festival. During the Holi festival, celebrants cover each other with colored paint and powder, and large fires are ignited during the evening, accompanied in many places by fireworks. The researchers concerned themselves with the effects of the Holi festival in the city of Mumbai. During the Holi festival, ground-based AOD data collected with a sun photometer showed a distinct increase. After correcting for the time of MODIS overpass, the ground-based data and MODIS data exhibited a very

good correlation. Higher water content in the atmosphere was likely a contributing factor to an increase in particle size during the festival as well. The increase in aerosols during the festival contributed to an increase in aerosol radiative forcing, decreasing the shortwave solar irradiance.

In their 2015 paper, Buchholz et al. (2015) examined air quality in Wollongong, located on the southeast coast of Australia. They used measurements of CO, CH_4, and carbon dioxide (CO_2). The main cause of CO variability was biomass burning in northern Australia, while CH_4 was influenced by nearby coal mining activities. MODIS fire pixel data were used to show a decrease in northeast Australian fires from 2011 to 2014, which helped to explain a decrease in the background concentrations of CO.

In the United States, summer fire seasons in the Pacific Northwest and California contribute to air quality hazards over much of the country, depending on the intensity and duration of the fires and the efficiency and distance of transport on continent-spanning winds. Creamean et al. (2016) analyzed the effect of smoke from Pacific Northwest fires on air quality in Colorado. Ground-based aerosol data were collected at a site in Boulder. The ground-based data for three different events showed increases in both metals and mineral content in the aerosols contributed by wildfire smoke and from mineral dust, which was believed to have been lifted into the air during the fires. The latter process had not been previously reported. MODIS AOD data in Giovanni were used to delineate the sources of the aerosols, combined with fire pixel and surface thermal anomaly data also acquired by MODIS.

Smoke from fires in Indonesia induced by a dry weather pattern due to an El Niño event was the subject of an investigation by Koplitz et al. (2016). The smoke from the widespread fires caused extreme haze conditions over much of equatorial Asia. The results indicate that the haze contributed to 100,300 excess deaths, which was more than twice as many caused by a similar, though less intense, 2006 event. Fires in South Sumatra Province were identified as an important factor contributing to the increase. AOD measurements from MODIS and OMI were compared to data from surface sensors in the AERONET network and showed approximately double the aerosol concentrations in 2015 compared to 2006. Koplitz et al. complete their paper by stating that their modeling approach could assist government agencies in prioritization of peatland and forest restoration areas to reduce the potential health impacts on downwind populations and also aid policy decisions and law enforcement directed at illegal forest burning.

Water Quality

Water quality, like air quality, is a significant public health factor that has been the subject of research and monitoring around the world, both for fresh (lakes, rivers) and saline (estuaries, seas, oceans) bodies of water. The data variables in Giovanni that are classified generally as "ocean color" are the primary variables utilized for water quality-related investigations. Of these, chlorophyll *a* concentration (chl a) and sea surface temperature (SST) dominate the usage patterns, but variables that measure water clarity and turbidity have also been utilized.

In addition, Giovanni also provides numerous hydrological variables that are strongly related to water quality. The NASA GES DISC is the primary NASA data archive for precipitation data, and it is obvious that precipitation affects both the volume and quality of freshwater resources. Also available are hydrological data from land data assimilation system (LDAS) models, both for North America (NLDAS) and global (GLDAS). The NLDAS and GLDAS variables are similar, but NLDAS data are at higher spatial and temporal resolution.

In the following brief summaries, a variety of water quality-related investigations that have accessed data in Giovanni are described.

In 2006, military hostilities between Lebanon and Israel caused significant damage to the Jiyyeh power station located on the Mediterranean Sea coast south of Beirut. An oil spill occurred in mid-July, and heavy fuel oil continued to flow into the coastal waters until early in August. The release of approximately 15,000 tons of oil in total extended over approximately 150 km of coastline. Pan et al. (2012) present the results of phytoplankton monitoring efforts conducted with satellite data following the oil spill. Chl *a* data acquired by both the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) and MODIS discovered a short-lived bloom occurring adjacent to the spill zone in May 2007, about 10 months after the oil spill. The data did not indicate a bloom occurring in this region at this time in any other year. Because anomalous blooms have been reported several months after other major oil spills, the authors concluded that this bloom was related to the power station oil spill in the previous year.

HABs appear to have been increasing in many freshwater and saltwater water bodies over the past decade. In a study that described changing conditions and which anticipated increasing difficulties with HABs in Lake Erie, Stumpf et al. (2012) used SST data from MODIS in conjunction with other ocean color data sources to examine the variability

of algal blooms in Lake Erie from 2002 to 2011. SST was important in identifying conditions conducive to the growth of *Microcystis*, a harmful algal species. Two years after the publication of this paper, a massive algal bloom in eastern Lake Erie, with high *Microcystis* activity, forced the city of Toledo to shut down their water supply for 2 days in August.

As noted above, the NASA GES DISC is the designated archive for precipitation data from NASA satellite observation missions, and data products from these missions are available in Giovanni. Tesi et al. (2013) created a map of rainfall accumulation with Tropical Rainfall Measuring Mission (TRMM) precipitation data for their study of the November 2011 flood of the Po River in Italy. The researchers studied the effects of this intermediate-sized flood (2.5-year return period) on total suspended sediments, nitrogen concentrations, nutrients, and particulate material in the river and exported by the flood flow into the northern Aegean Sea.

The alteration of the Arabian Sea ecosystem due to climate change, resulting in a shift from diatoms to *Noctiluca* (a dinoflagellate) as the primary phytoplankton producer, has been documented in several studies. This shift has had significant effects on both coastal water quality and Arabian Sea fisheries. A 2015 study by Dwivedi et al. (2015) described the detection of *Noctiluca scintillans* "hot spots" in the Arabian Sea.

Both tourism interest and coastal fisheries can be affected by adverse coastal water quality conditions and both are important along the coast of Maine. Tilburg et al. (2015) describe models that employed multiple variables to predict noxious or hazardous water quality conditions for this famous coastal region, particularly focused on fecal coliform bacteria concentrations. They found that accurate predictions required more than just river discharge data and precipitation data. The authors state that this effort was one of the first attempts to predict water quality in coastal ocean waters. The daily TRMM precipitation data for this modeling effort were acquired with Giovanni.

Another coastal region with water quality concerns is the long coastline of Italy. In their 2016 study, Corbari et al. (2016) investigated how intense short rainfall events influence coastal water quality in the vicinity of the discharge zone for four different rivers draining small watersheds. The researchers combined daily TRMM precipitation data with high-resolution data from MODIS, allowing the generation of turbidity and suspended sediment concentration estimates, to examine the occurrence of intense rainfall events and the effects on the discharge zone. The expected correlation of high-volume rainfall with increases in offshore turbidity parameters was observed, and the authors indicate that remote sensing could be used to detect changes in water quality due to weather events in a more rapid manner than ground-based water quality sampling.

Epidemiology

There are many different aspects to the use of remote sensing data in epidemiology research. One aspect concerns the effort to understand the causes of certain types of disease outbreaks with respect to factors in the natural environment. Another aspect is monitoring for conditions that might be conducive to the occurrence of a disease and potentially using the data to predict disease outbreaks beforehand. A third aspect is demonstrating linkage between actual causality (as opposed to correlation) and disease occurrence and in so doing perhaps identify methods to prevent disease pandemics and epidemics. Giovanni has been used in published research that is related to all three of these epidemiological aspects.

One of the strongest relationships between a disease and a remotely-sensed parameter is the relationship between cholera and chl *a*. The reason for this relationship is that *Vibrio cholerae* bacteria attach to zooplankton (copepods), which are primary consumers of phytoplankton. Thus, increases in phytoplankton concentration are correlated with increased copepod populations and the associated *V. cholerae* bacteria. Remotely-sensed chl *a* may thus indicate regions where cholera incidence is increased. Using SeaWiFS chl *a* data, Jutla et al. (2010) showed a clear seasonal relationship between river discharge, chl *a,* and cholera in the Bay of Bengal region over a 10-year period.

Diseases that spread with mosquitoes as the vector, the most notable being malaria, are related to the availability of standing water where the mosquitoes can breed. Precipitation is therefore an important environmental variable related to malaria occurrence patterns. In Midekisa et al. (2012), a malaria early-warning model for mountainous regions of Ethiopia was developed. The researchers were able to construct and test the model using TRMM precipitation data and both land surface temperature (LST) and vegetation index data from MODIS.

Another mosquito-borne disease, dengue fever, was the subject of Moreno-Madrinan et al. (2014). The large research group used MODIS LST and TRMM rainfall accumulation data to estimate the abundance of the dengue virus mosquito carrier *Aedes aegypti*. Estimates of abundance were compared to *A. aegypti* pupae counts acquired by field collection. The researchers concluded "Strong correlations were found between the abundance of the dengue virus mosquito vector, *Ae. aegypti*, and RS-derived nighttime LST, elevation or rainfall along a geographic climate/elevation gradient in Central Mexico."

Reducing the incidence of cholera in Haiti was the focus of the paper by Rebaudet et al. (2013). A cholera epidemic in 2010 accounted for over 8000 deaths and over 650,000 diagnosed cases. Rainfall amounts during Haiti's dry and wet seasons exhibit a very strong positive relationship with the

occurrence of cholera, as shown in their Fig. 1, constructed with TRMM rainfall data acquired through Giovanni. The reason that cholera in Haiti persisted in subsequent years was due to the survival of the *V. cholerae* bacteria in just a few locations. The researchers therefore concluded that proper water treatment measures taken during the dry season could significantly reduce or eliminate the occurrence of cholera in Haiti.

Influenza is one of the most widespread contagious diseases affecting the world population. Thus, studying factors that affect the spread of influenza is important to public health concerns around the world. Unlike temperate regions, where influenza has a primary season of occurrence (generally the colder winter months), the "flu season" is not as well defined in the tropics. Despite this difference, there is a detectable seasonal pattern, as described by Soebiyanto et al. (2014). They examined temperature, specific humidity, and rainfall in three Central American countries to determine correlations between these factors and influenza affliction frequency. Specific humidity was the strongest factor, as it showed a positive association with influenza in El Salvador and Panama and a negative association in Guatemala. Temperature and rainfall exhibited positive associations in sub-regions of each of the countries studied. Both environmental and social patterns may influence disease transmission, accounting for the different positive and negative associations observed.

According to Wu et al. (2016), shallow wells are an important source of drinking water in rural Bangladesh. Thus, extreme rainfall events and land use can both influence the presence of fecal coliform bacteria in these wells. This study found that the presence of fecal coliform bacteria was much more likely if heavy rain events occurred in the three-day period before water testing. Particular types of land use combined with heavy rainfall amplified the likelihood of fecal coliform contamination in the shallow wells.

Erythemal Radiation Exposure

The availability of a particular data product in Giovanni has enabled one area of public health research with a similar theme. The data product is Erythemal (UV) Radiation. "Erythemal" refers to reddening of the skin, which is caused by exposure to solar UV radiation. Such exposure is an immediate cause of sunburn and is a causal factor in skin cancer, cataracts, and immune system disorders. The Erythemal Radiation data product is available from both the TOMS and OMI sensors.

In 2014, a paper by Serrano et al. (2014) examined the various levels of UV exposure that occur for outdoor sports: tennis, hiking, and running. Hikers received the highest amount of daily UV exposure of these three groups. Because some of the hikes took place in locations that were not near a ground solar radiation instrument, daily OMI data were used for these locations to estimate the ambient erythemal UV radiation.

Workers on ships at sea are also exposed to sunlight for extended periods, and this can be particularly acute along tropical sea lanes. Feister et al. (2015) modeled the exposure of deck crew members using both ship-based sensors and satellite data. Makgabutlane and Wright (2015) performed a similar study for outdoor workers in Pretoria, South Africa.

Exposure to UV radiation from the Sun is not entirely a deleterious health factor, as the skin synthesizes Vitamin D when exposed to the Sun. Wainwright et al. (2016) described the development of a dosimeter capable of measuring both erythemal radiation and Vitamin D-effective radiation. To calibrate the dosimeter, the researchers applied both OMI AOT and ozone concentration data.

Humans are not the only living beings that are affected by solar UV radiation. Plants can also be affected, and damaged, by UV, and this potential also requires measurement. Parisi et al. (2017) describe their long-term dosimeter designed to measure the UV exposure of plants. This group obtained the average ozone concentration during the one-month exposure period with OMI data.

As noted earlier, UV exposure can affect the human body's immune system. An effect of this is to reduce the effectiveness of vaccinations given to prevent disease. This interaction of health concerns was examined by Wright et al. (2017) for children in rural areas. In these areas, children and their families may walk long distances to clinics where the vaccinations are obtained. The sun exposure during the walks can reduce vaccine effectiveness, so the authors studied practices including counseling parents on sun protection, and providing sun protection (clothing, umbrellas, and sunscreen) for the parents of children receiving measles vaccine. The annual erythemal radiation exposure cycle for the Limpopo province of South Africa, where the study took place, was generated with OMI data in Giovanni.

Natural Hazards Prediction and Assessment

Natural hazards are a public health factor for many different reasons. Injuries and death can result from the conditions occurring during earthquakes, hazardous weather (hurricanes, tornadoes, and severe storms), wildfires, landslides, and floods. Subsequent to the actual event, public health may be affected by disease outbreaks due to poor air and water quality. Effects of disastrous events can extend long distances from the immediately affected area as well. Data products acquired from and visualized with Giovanni have been used to assess the cause of natural hazard events, the effects of

such events, and even to provide predictions of certain types of events.

In 2004, a massive tsunami inundated large stretches of the Indian Ocean coast, including the coastlines of Thailand, Malaysia, Indonesia, India, and Sri Lanka, causing thousands of fatalities and extreme damage. After the catastrophic waves went inland, the outflow of these waves into the ocean resulted in pollution, turbidity, and hazardous water quality. Tan et al. (2007) analyzed the effects of the tsunami on the ocean color along the coasts of Sumatra and Thailand, finding elevated turbidity due to enhanced sediment concentrations in some coastal areas affected by the tsunami. While chl *a* did not appear to be influenced after the tsunami, as normal oceanic processes were observed affecting chl *a*, the sedimentation and land erosion caused by the waves were indicated as a potential factor affecting local marine resources.

In the Bihar state of India, a devastating flood of the Kosi River in August 2008 resulted in both heavy loss of life and widespread destruction of crops, according to Singh et al. (2011). During the flood, the river breached its normal channel and agricultural canals and flooded several villages that are located on its alluvial fan, along with extensive flooding of croplands. Consultation of TRMM rainfall data in Giovanni indicated that the region, particularly the upper catchment basin, received considerably higher-than-normal rainfall in June, July, and August 2008, creating an increased susceptibility to flooding due to soil saturation.

Both the prediction of wildfire danger in Greece, and the consequences of a destructive wildfire in August 2007, constituted the subject of Athanasopoulou et al. (2014). They found that the enhanced fire risk for this period was reproduced by a Fire Weather Index model, which could enable future predictive success. The atmospheric effects of the August 2007 fires were evaluated using MODIS AOD.

Nearly 5 years after the Kosi River flood disaster, another flood caused thousands of deaths in northern India. A glacial lake above the village of Kedarnath rapidly expanded due to anomalous snow melt, and the water volume in the lake was increased by heavy monsoon rains earlier than normal in the monsoon season. These factors combined to cause a lake outburst and debris flow on June 16 and 17, 2013. A ground weather station measured 325 mm of rainfall on June 15–16 adjacent to the glacier feeding the lake, and TRMM average daily precipitation data demonstrated that this was an unusually high amount for these dates, substantially in excess of even 90th-percentile values for the 1998–2012 period. TRMM rainfall accumulation maps also showed that the heaviest precipitation fell very near to the lake. These events were described in Allen et al. (2015).

Northern India and the Himalayan region may seem to be plagued with various kinds of natural disasters; the April 2015 earthquake in Nepal was also the subject of a study which used Giovanni (Ganguly 2016). In this remarkable paper, AOD was shown to increase near the earthquake epicenter 26 days prior to the earthquake. Post-earthquake AOD increases were attributed to natural dust and building damage and did not exceed the peak prior to the earthquake. Ozone concentrations were elevated approximately 20 days prior to the earthquake. The author discusses the pre-earthquake subsurface processes that could lead to the elevation of AOD and ozone, which have also been described for other earthquakes.

Prediction of natural disaster effects was the subject of both Wijesundera et al. (2016) and Yu et al. (2017). Flooding in eastern Australia due to tropical cyclones was the subject of the former paper, which used rainfall data in a case history of tropical cyclone Yasi as a test of their predictive model. The latter paper described the prediction of wildfire risk in Cambodia, using TRMM rainfall data from Giovanni, in conjunction with several MODIS data products (temperature, vegetation index, and thermal anomalies) obtained from the NASA Land Processes DAAC. One important element of the Yu et al. fire risk model was that it only used publically available remotely-sensed data products.

Agriculture, Fisheries, and Natural Resources

The last category of public health concerns that have been investigated with Giovanni's data holdings and analytical capabilities is that of agriculture, fisheries, and natural resources. Giovanni provides insight into both the cause of agricultural problems and resource depletion, as well as providing data to monitor conditions and provide information to policymakers and stakeholders. In addition to the research papers summarized here, Giovanni has found frequent use in government and consultant reports in this particular focus area.

Over the past decades, reduction of ice area and volume in Earth's cryosphere has been described in many different venues. In 2008, Kehrwald et al. (2008) used ice cores on Naimona'nyi Glacier in the Himalayan mountain range to study the mass loss of this high elevation glacier. Mass loss of these glaciers was deemed a concern because the glaciers feed the headwaters of major rivers on the Indian subcontinent. Precipitation data in Giovanni were used to characterize the precipitation delivered to the region, which decreases substantially from the southwest to the northeast.

Agricultural diseases are a major concern of farmers, and precipitation again is an important factor. Farrow et al. (2011) examined the relationship between bean root rot and precipitation in East Africa. Heavy rainfall events during the cropping season were the primary cause of root rot, but the data from rain gauges were too sparse to allow accurate ground-based assessment. While initial results using remotely-sensed data were inconclusive, the researchers noted that the rapid

availability of TRMM daily data in Giovanni could enable better estimates of the probability of root rot and also enable an "early-warning system" for the growing season.

Moving from the African continent to the offshore Atlantic Ocean, one of the most productive ocean regions in the world is the Benguela Upwelling Zone off of the coast of western South Africa and Namibia. Jury's 2012 paper described the physical oceanographic factors that affect the fish catch, which is normally reliable but which has experienced occasional "crashes" in fish populations. Jury found that a weakening of southeasterly winds and a half-degree Centigrade increase in SST were associated with a higher fish catch. Chl *a* and SST from Giovanni characterized the oceanic environment and also provided direct information on the annual seasonal cycle.

Both the short-term effects of air pollution and the long-term effects of climate change on agriculture in India were the subject of Burney and Ramanathan (2014). They estimated that climate change and "short-lived climate pollutants" (SLCPs) decreased wheat yields by 36% over India, with some areas having decreases up to 50%, compared to conditions with neither factor present. The influence of short-lived pollutants was significantly larger than the effects attributed to climate change, through the year 2010. AOD from MODIS and surface ozone from MERRA were used in this pioneering study.

Another agricultural concern for regions that are near major continental deserts is the process of desertification, which can reduce the available area for farming and livestock grazing. Lamchin et al. (2015) used rainfall and air temperature data acquired from Giovanni in their examination of desertification processes near the Hogno Khaan protected area in Mongolia. These parameters did not exhibit notable trends during the study period. The primary process identified for desertification in this area was increased livestock grazing pressure, reducing vegetation, which allows greater movement of sand dunes. The pace of desertification was related to the slope and elevation of the land surface.

The Atlantic Forest ecosystem of northern Argentina was the focus of the study performed by Zaninovich et al. (2016). Necromass, which consists of fallen vegetation and organic debris, was compared for the native forest environment and non-native pine tree plantations. The necromass for the native forest was found to be more diverse and had much greater moisture retention and longer carbon storage times than the pine forest plantations. The necromass in the pine forest plantations also increased fire risk, due to the higher amount of fine detritus and lower water retention characteristics. TRMM rainfall data provided context for the annual precipitation cycle in the region.

According to Adama and Mochiah (2017), the African armyworm is an "important migratory pest" in sub-Saharan Africa, capable of causing significant damage to crops. In Ghana, the African armyworm can damage both maize and rangeland vegetation. Grain loss during armyworm outbreaks can range from 60% to 100% of production. In an effort to determine the climatic factors related to such outbreaks, the authors used TRMM daily rainfall data from Giovanni in conjunction with air temperature data and NDVI data from other sources. Two outbreaks in 2006 and 2009 were studied. In 2006, the conditions, including moderate rainfall, favored the attraction of moths to existing vegetation where eggs could be deposited. However, the conditions prior to the 2009 outbreak differed, including heavier rainfall, with no clear reason indicating favorability for moth egg-laying activity. Given the seriousness of armyworm outbreaks, the authors advocated additional study with an examination of more climate-related variables.

Sub-Saharan African was also the region of study for McNally et al. (2017), who describe the creation of a land data assimilation system concerned with food and water resource security. The system, the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS), is a custom instance of the NASA Land Information System (LIS), which also creates the NLDAS and GLDAS data sets. The FLDAS data variables are all available for analysis in Giovanni. In their paper, McNally and her co-authors examine correlations with several FLDAS variables and data from other sources to validate the accuracy of these data variables.

Now that the range of public health topics that can be addressed with data in Giovanni has been presented, the next step is to demonstrate how these variables can be useful for public health applications research.

Case Study—Using Giovanni to Examine the Factors Contributing to Heat Stress Dangers

The following example case study was partly prepared by a high school student intern working in conjunction with our GES DISC staff. This work demonstrates how easily Giovanni can be applied by "novice" users to areas of public health concern. The papers previously discussed indicate that the analytical capabilities and data variable archive in Giovanni provide valuable data for sophisticated public health research. Giovanni can also be used to contribute to basic monitoring, trend analysis, and reporting and also provide data baselines for understanding the factors that affect many different public health topics.

The topic under consideration here is heat stress. Two different variables, temperature and humidity, are used to

compute the Discomfort Index (or "Feels Like" index) for high temperature and high humidity conditions. The National Weather Service issues Heat Advisories and Extreme Heat Warnings when there is an increased risk of heat exhaustion and heat stroke. Such advisories are important for many different outdoor activities, including construction work, sports activities, and child care.

For this study, we wished to investigate if the primary factor that contributes to a higher Discomfort Index (which will be referred to as DI subsequently) varied depending on the region under consideration. In areas near the ocean coast, humidity was anticipated to be a more important factor than for inland areas more distant from a source of humidity.

The regions selected were the US states California, Texas, Arizona, and Florida, and the countries of Sudan, Saudi Arabia, and Yemen. For each of these regions, the time period of interest was selected to be June 2018, generally a warm summer month for each of these Northern Hemisphere locations. Daily surface relative humidity data from AIRS and daily mean 2-meter air temperature data from MERRA-2 were selected. The DI formula employed was from the South African Weather Service Web page "What is the discomfort index?" (http://www.weathersa.co.za/learning/educational-questions/58-what-is-the-discomfort-index) and is expressed as the following:

$$\text{Discomfort Index} = (2 \times T) + (\text{RH}/100 \times T) + 24$$

where T is the dry bulb or air temperature in degrees Celsius and RH is the relative humidity in percent.

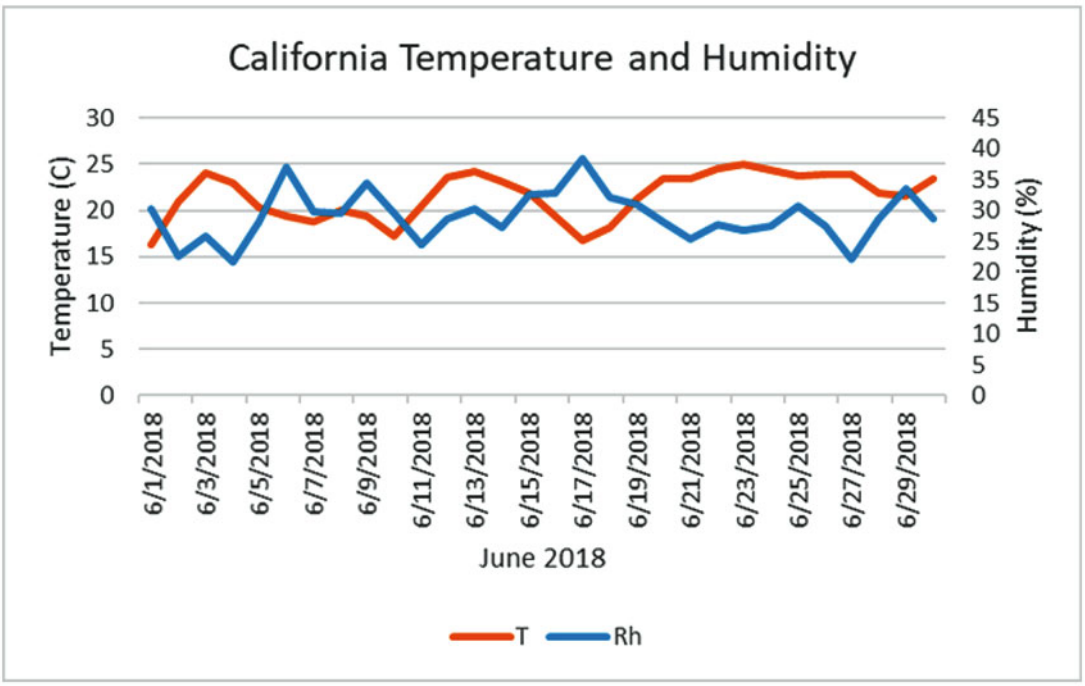
The time-series plots were created by generating daily time series of temperature and RH with Giovanni. A shapefile for either the state or country was used as the region of interest, and the data values were averaged over the entire region. Because both temperature and RH could vary considerably over a large state like California, a more in-depth study could focus on smaller coastal or inland regions rather than an entire state.

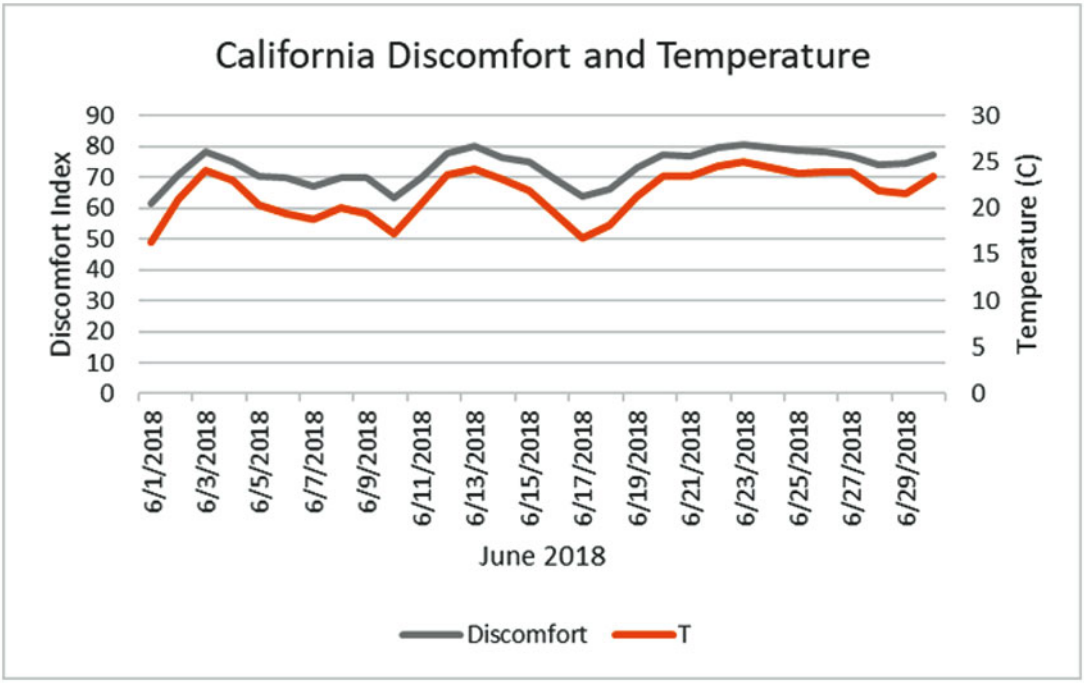
After each individual time series was created, the time-series data (date and data value) were downloaded from Giovanni in comma-separated variable (CSV) format and entered into an Excel spreadsheet. The DI formula given above was used to calculate DI for each date in the third column of the spreadsheet. After the DI values were calculated, time-series plots of temperature and RH vs. time, temperature and DI vs. time, and RH and DI vs. time were created for each region of interest.

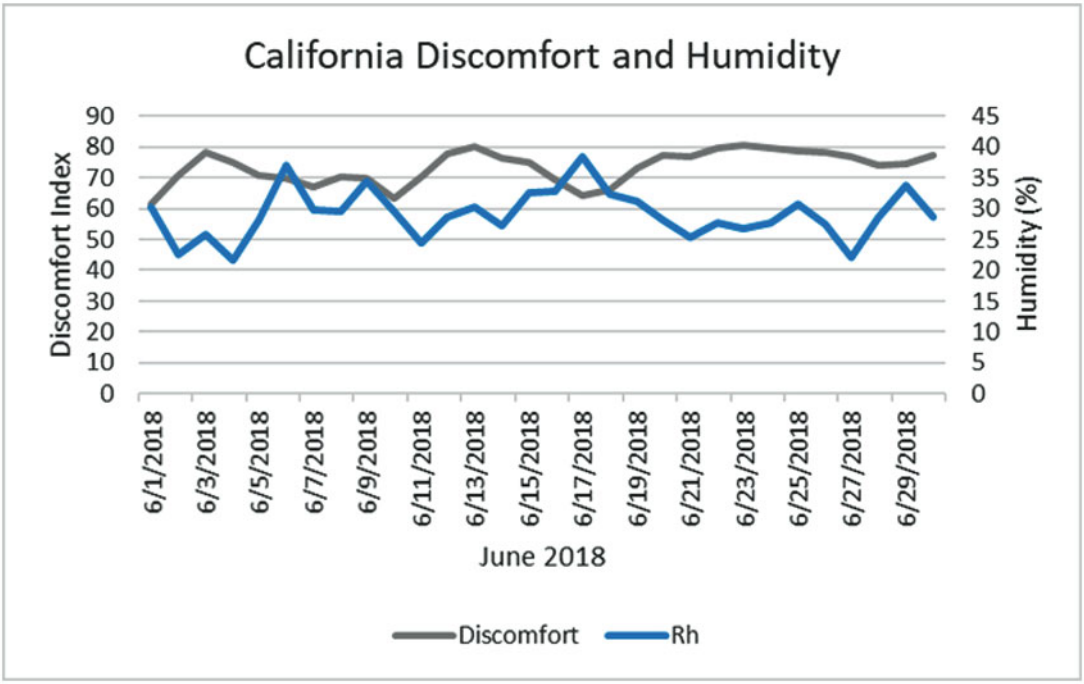
The results for each region are presented and briefly discussed below.

California

Temperature and RH both varied in a relatively narrow range in the state of California during June 2018. Temperature varied between 15 and 25 °C and was near 25 °C for most of the latter half of the month (note that these are daily



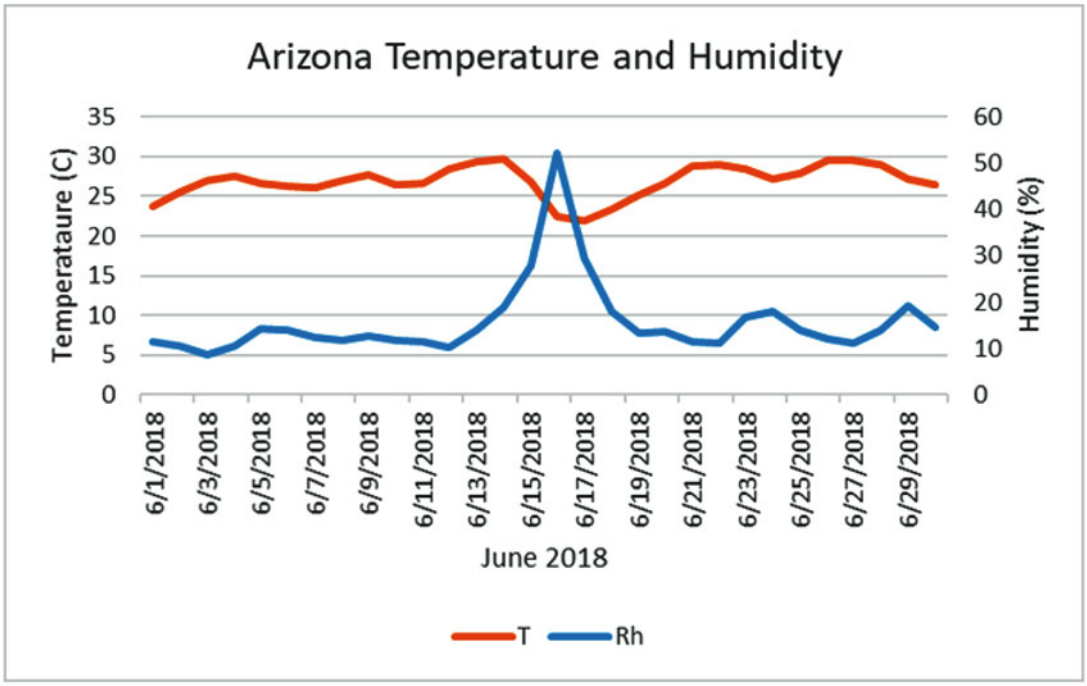


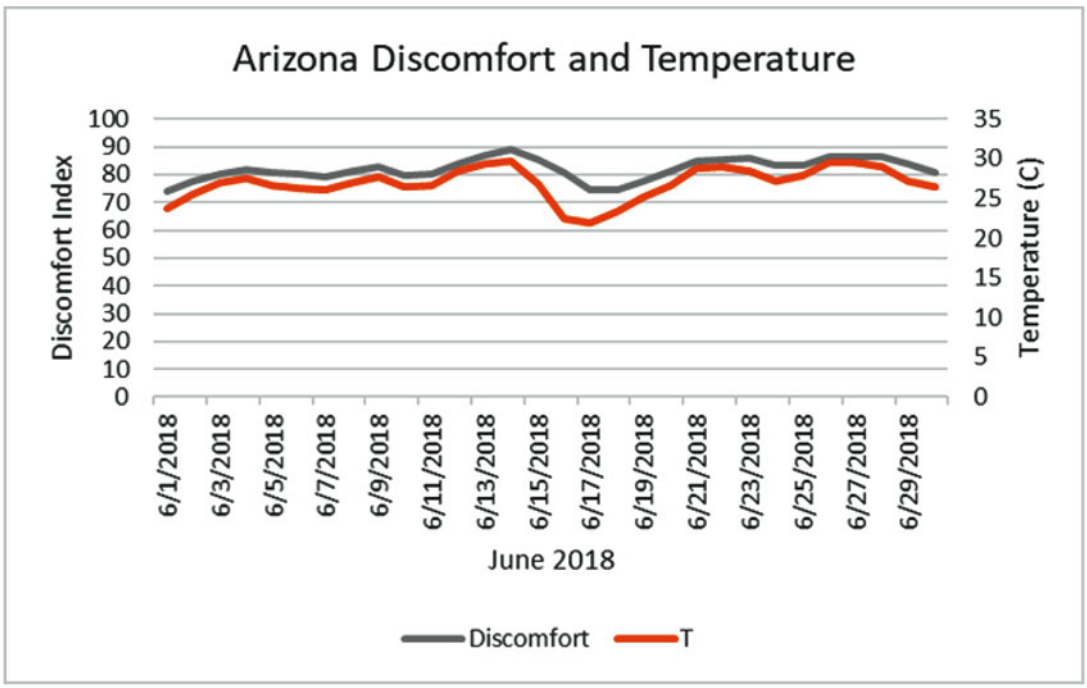


averages). RH varied between 25% and 35%, which likely indicates that this statewide average value was dominated by the drier interior of the state rather than the coastal region. The DI tracks closely with the daily temperature but does not appear to have a strong connection with RH at these low humidity values. The DI ranged from a low of about 60 to a high of 80 seen in the latter half of the month.

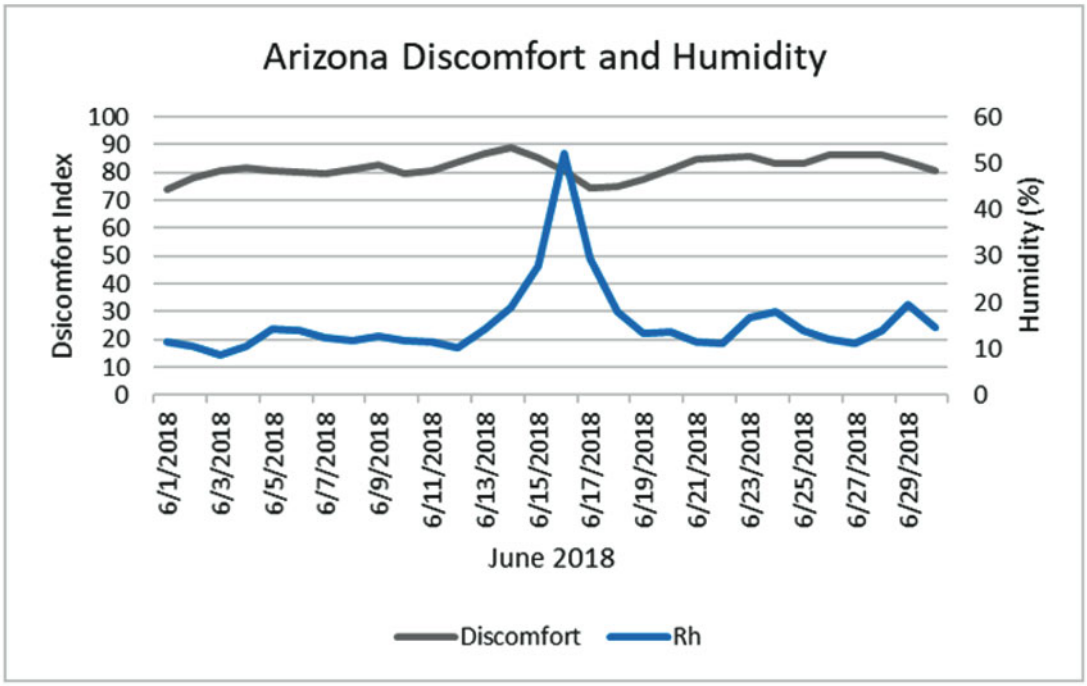
Arizona

The temperature and RH data for Arizona in June 2018 show an interesting pattern. While the average temperature was relatively constant between about 22 and 29 °C, the RH values increased quite markedly during the middle of the month to over 50%, while they were normally in the dry 10–20% envelope. This increase in RH was accompanied by a



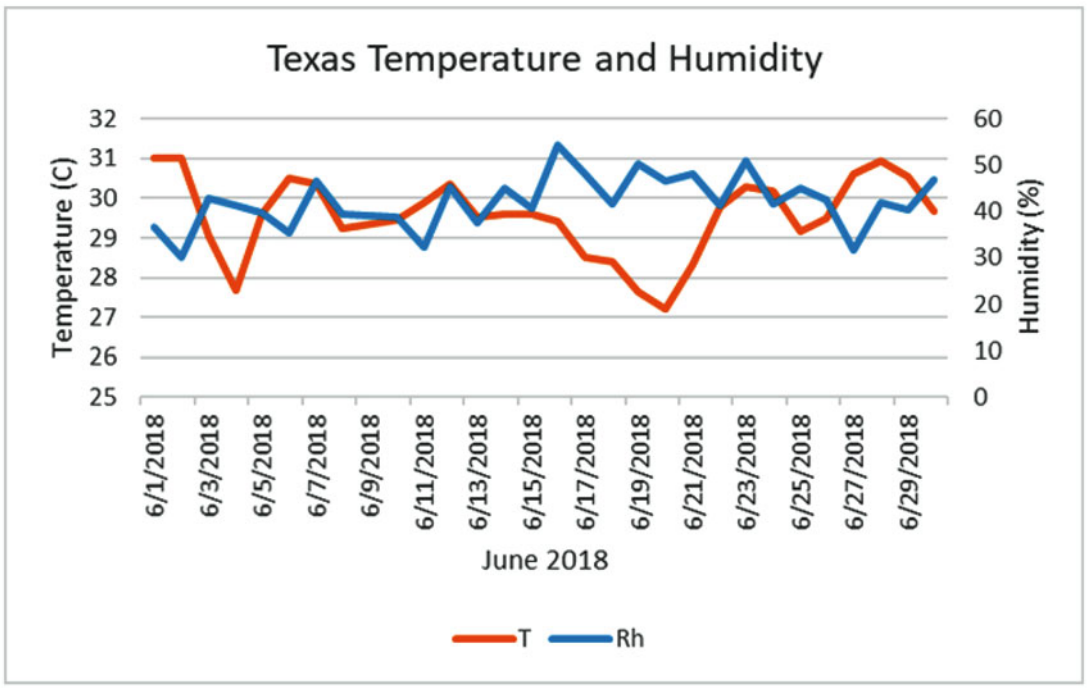


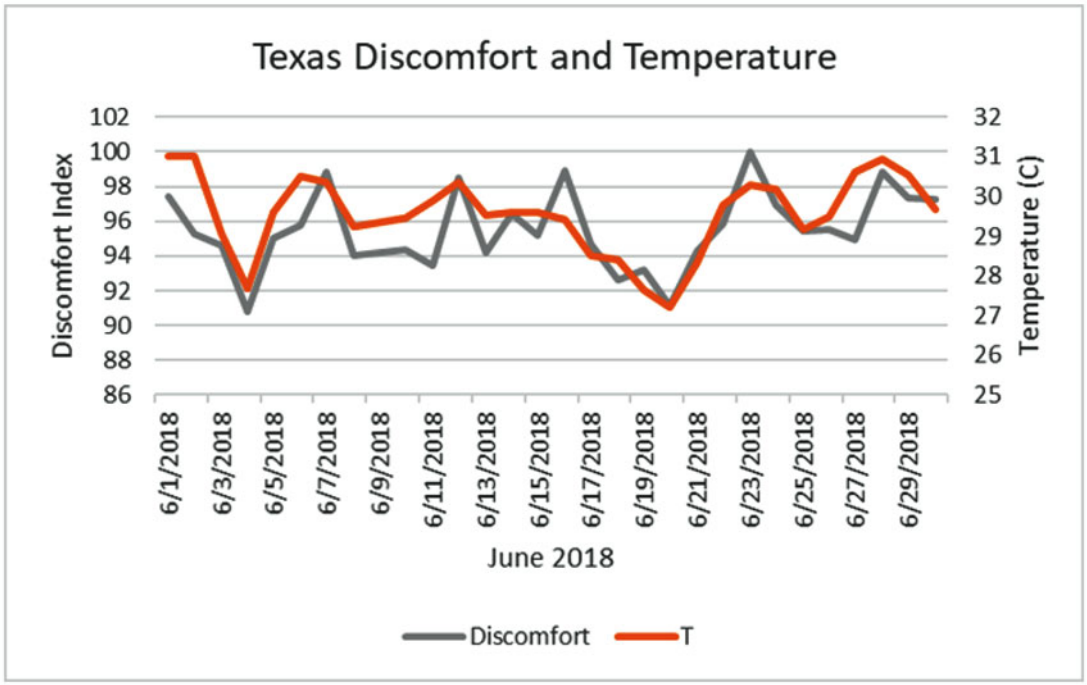
decrease in average temperature, indicating a rainfall event, which was confirmed by examination of a June 2018 time-series of the *I*ntegrated *M*ulti-satellitE Retrievals for GPM (IMERG) Final Run data product. So the combination of significantly increased RH with a lower temperature only caused a slight decrease in DI. The influence of the higher humidity can be seen in the DI vs. temperature plot, where the separation of the two time-series lines at mid-month is the only time that temperature and DI do not match closely. This congruence also indicates that for most of the dry desert summer, temperature changes are the dominant cause of DI variability. In fact, the DI exhibited only slight changes in the 70–90 range.

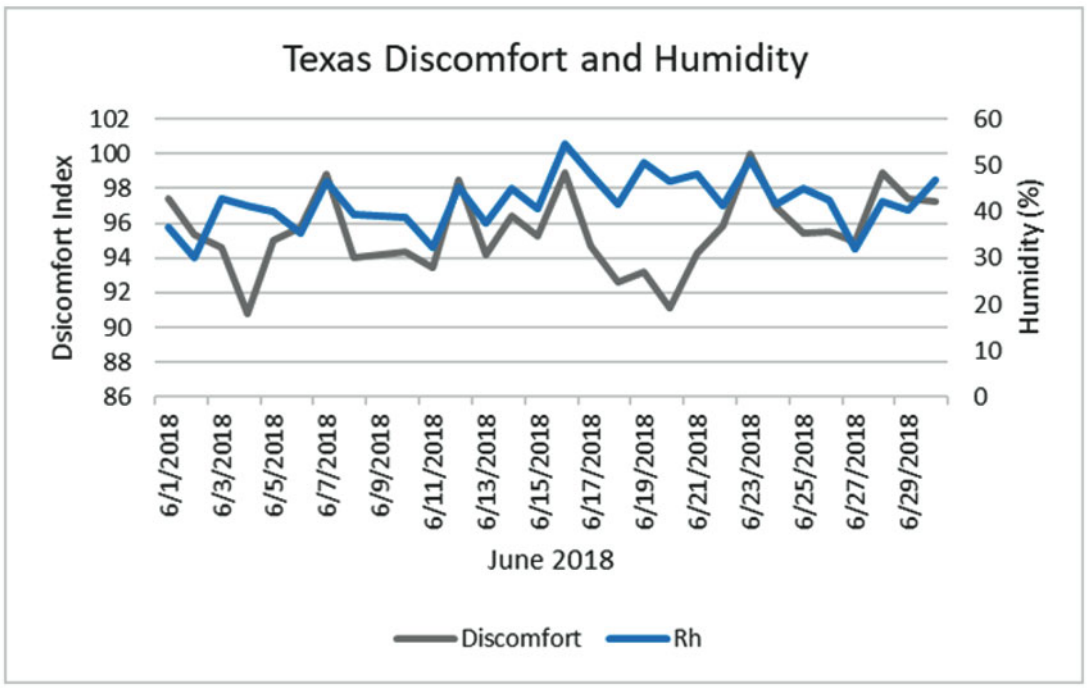


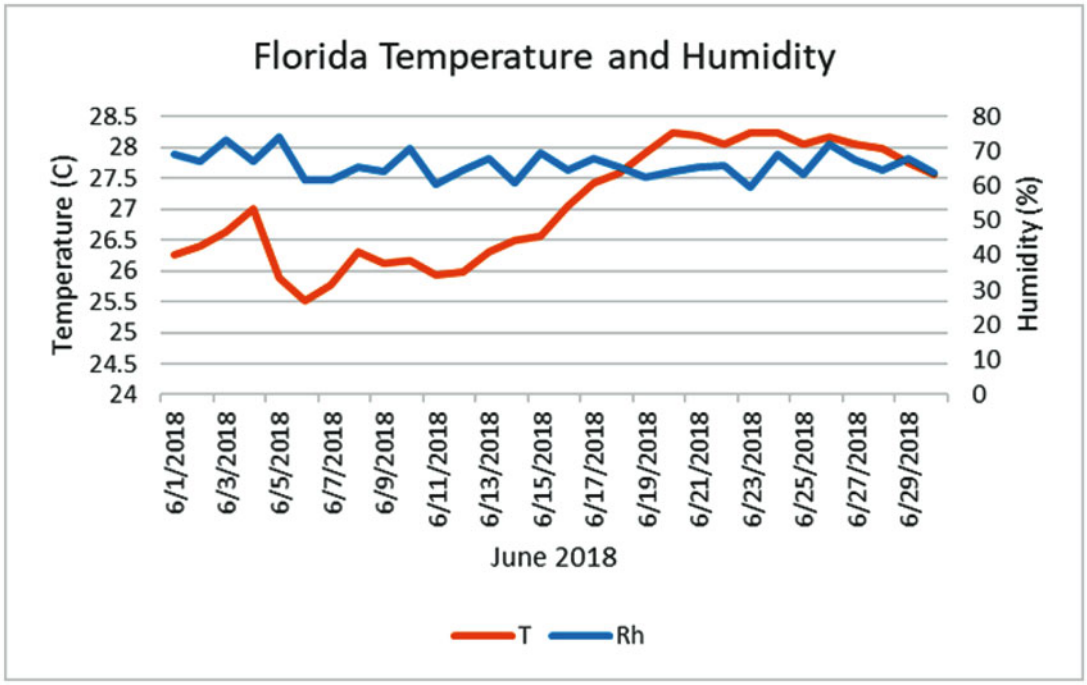
Texas

The state of Texas is large and includes a semi-arid western region and a southeastern region with a long Gulf Coast shoreline. Thus, examining the DI over the entire state will generate an average that is likely not fully indicative of regional conditions. For the month of June 2018, the average temperature ranged from a high of 31 °C to a low of 27 °C. The RH range was from approximately 30 to 50%, which likely combines the influence of low humidity conditions in the west and higher humidity in the east. The time-series show the DI matching the variability of temperature quite closely and also showing corresponding evolution with RH, with the exception of the minimum temperature days on June 4th and the cooler period from June 17 to 22. So, over the entire state, RH is a secondary influence on the DI compared




Texas Discomfort and Temperature
Discomfort Index
Temperature (C)
102
100
98
96
94
92
90
88
86
32
31
30
29
28
27
26
25
6/1/2018
6/3/2018
6/5/2018
6/7/2018
6/9/2018
6/11/2018
6/13/2018
6/15/2018
6/17/2018
6/19/2018
6/21/2018
6/23/2018
6/25/2018
6/27/2018
6/29/2018
June 2018
Discomfort
T

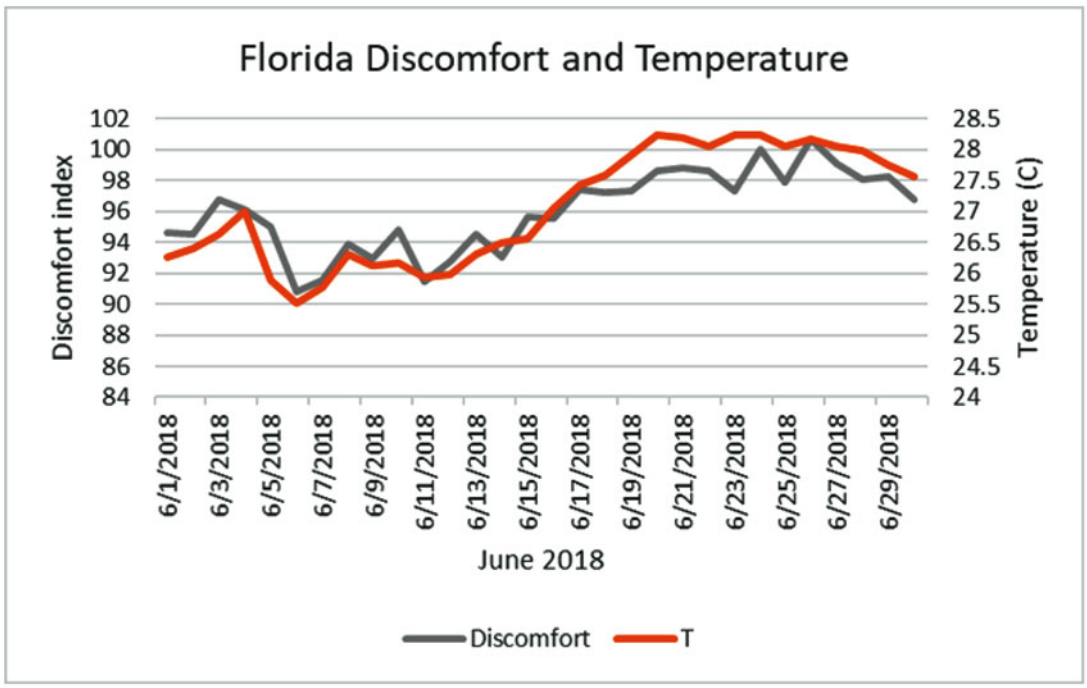

Texas Discomfort and Humidity
Dsicomfort Index
Humidity (%)
102
100
98
96
94
92
90
88
86
60
50
40
30
20
10
0
6/1/2018
6/3/2018
6/5/2018
6/7/2018
6/9/2018
6/11/2018
6/13/2018
6/15/2018
6/17/2018
6/19/2018
6/21/2018
6/23/2018
6/25/2018
6/27/2018
6/29/2018
June 2018
Discomfort
Rh

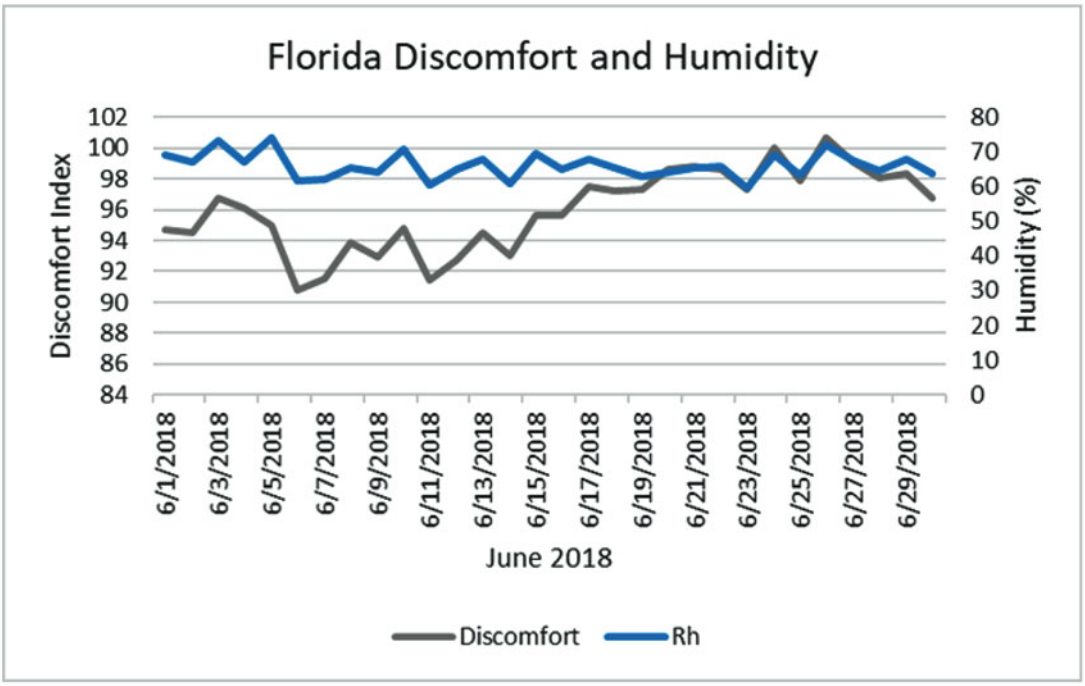


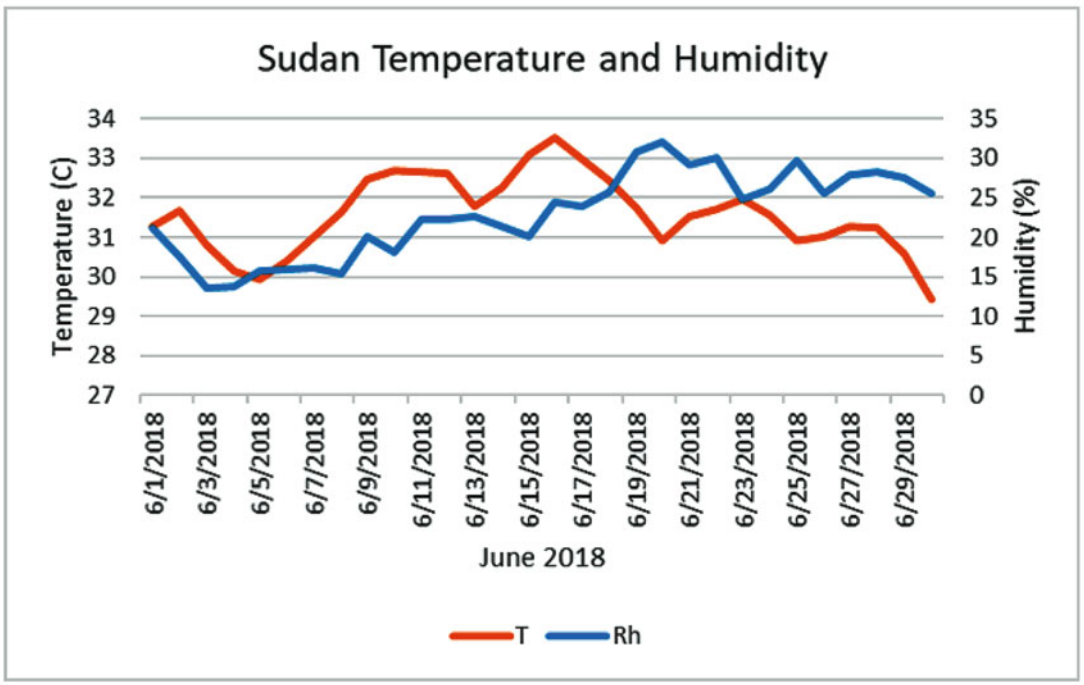
to the surface temperature. However, the maximum DI was just over 98 on the most humid day of the month and nearly reached 100 on the second most humid day, indicating that humidity is a factor during uncomfortable days in Texas. Considering how much higher the average humidity in Texas is compared to Arizona in June, the higher DI values are not surprising.

Florida

When the weather in Florida in summer is discussed, *hot* and *humid* are two words that can be generally used to describe it. For June 2018, the average RH just barely fell below 60% on one day and was usually between 60% and 70%. The temperature in June rose from a low of 25.5 °C early in the month to a high of above 28 °C for much of the second half





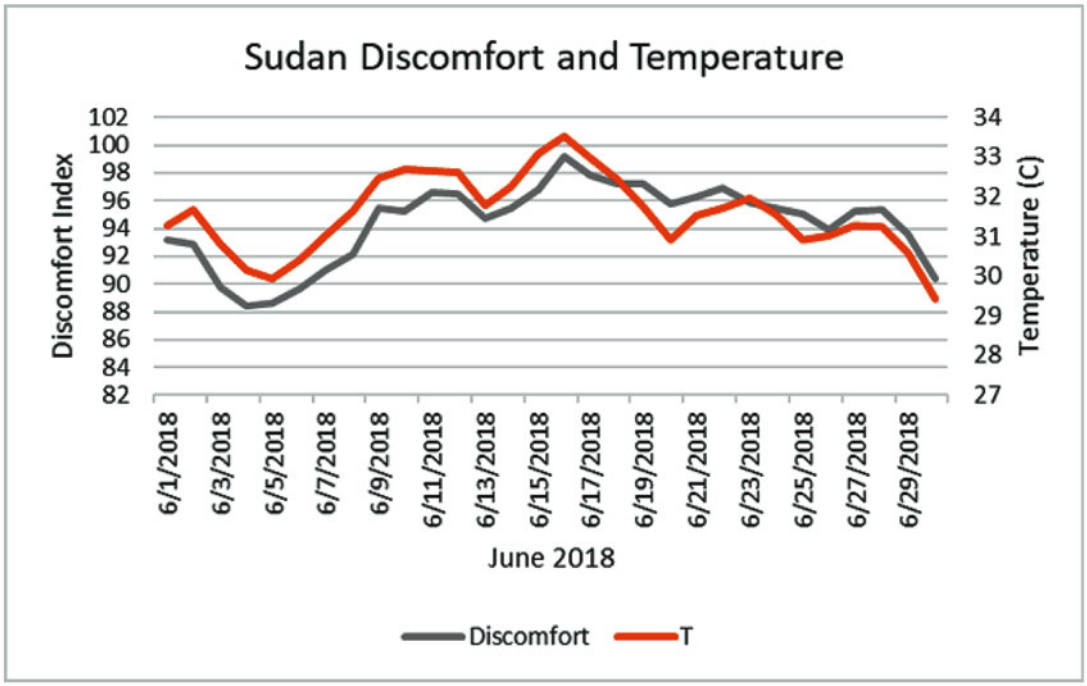


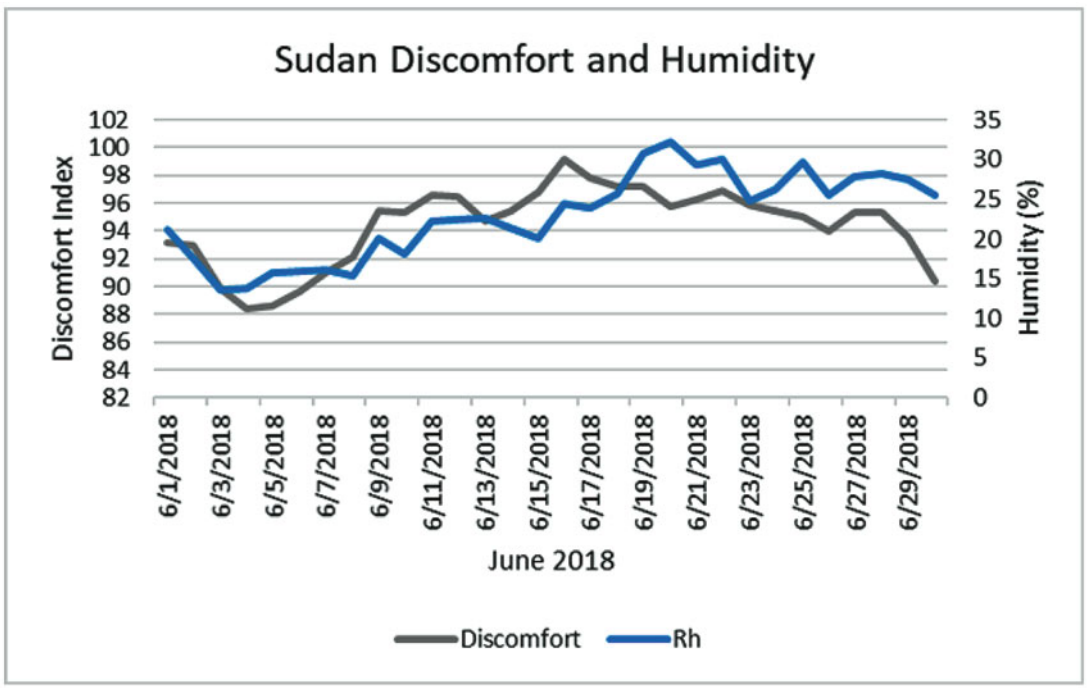
of the month. Because the RH was so constant, the variability of the DI was mostly determined by the temperature change. The DI rose from about 91 early in the month to 98–100 late in the month for several days. So, progressing from west to east, the DI ranges for each state rise with increasing humidity, even though the primary factor causing variability in DI is surface temperature.

Next we will consider three countries in Africa and the Middle East: Sudan, Saudi Arabia, and Yemen.

Sudan

Sudan is a huge country, located in the sub-Saharan climate region. The shapefile in Giovanni (acquired from the Humanitarian Information Unit of the US State Department) used for Sudan is the now-recognized boundary of the country that does not include South Sudan, which is adjacent to the northern tropical region of Africa. Hence, the climate conditions in Sudan are semi-arid to arid. This can be clearly

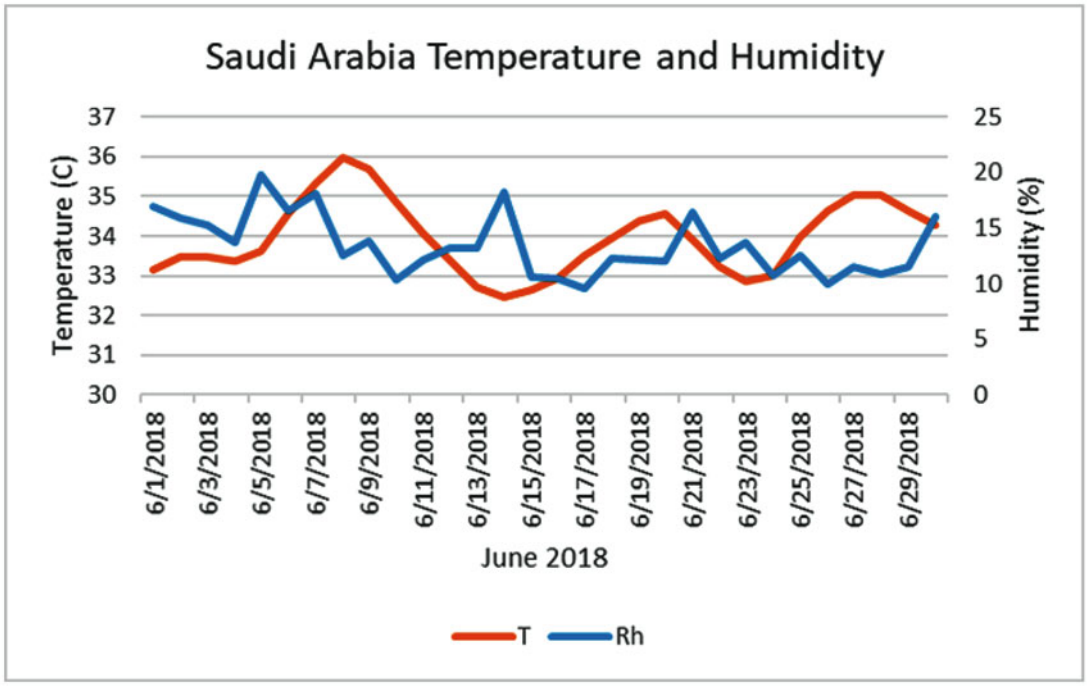


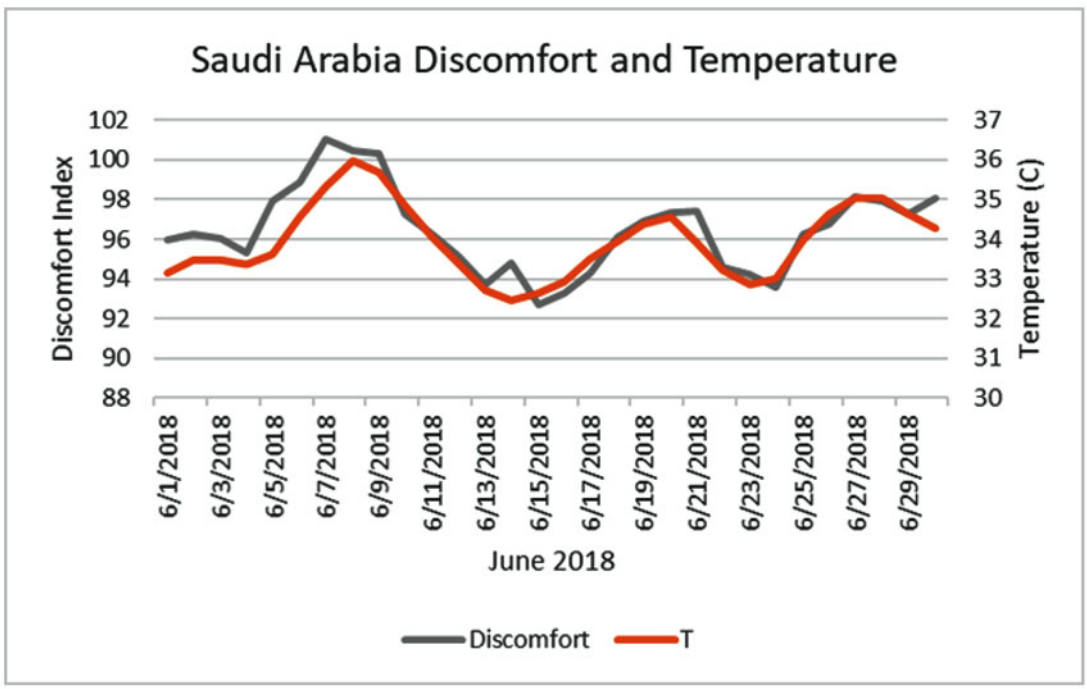


seen in the average temperature, which ranged from above 29 °C at the end of the month and a high over 33 °C at mid-month. The humidity was in the moderate range between 15% and 30%. As both temperature and humidity rose in the first half of the month, the DI rose from a low value of 88 to a high value of 97. During the final days of the month, as the humidity remained between 25% and 30%, slightly decreased temperatures led to a decrease in the DI.

Saudi Arabia

The country of Saudi Arabia is very arid, much like the climate of Arizona, though the presence of the warm waters of the Red Sea on the western coast of the state does contribute some humidity to the country's climate. Clearly, this is not a large influence on most of the country, as the RH only ranged from 10% to 17.5% in June. The surface

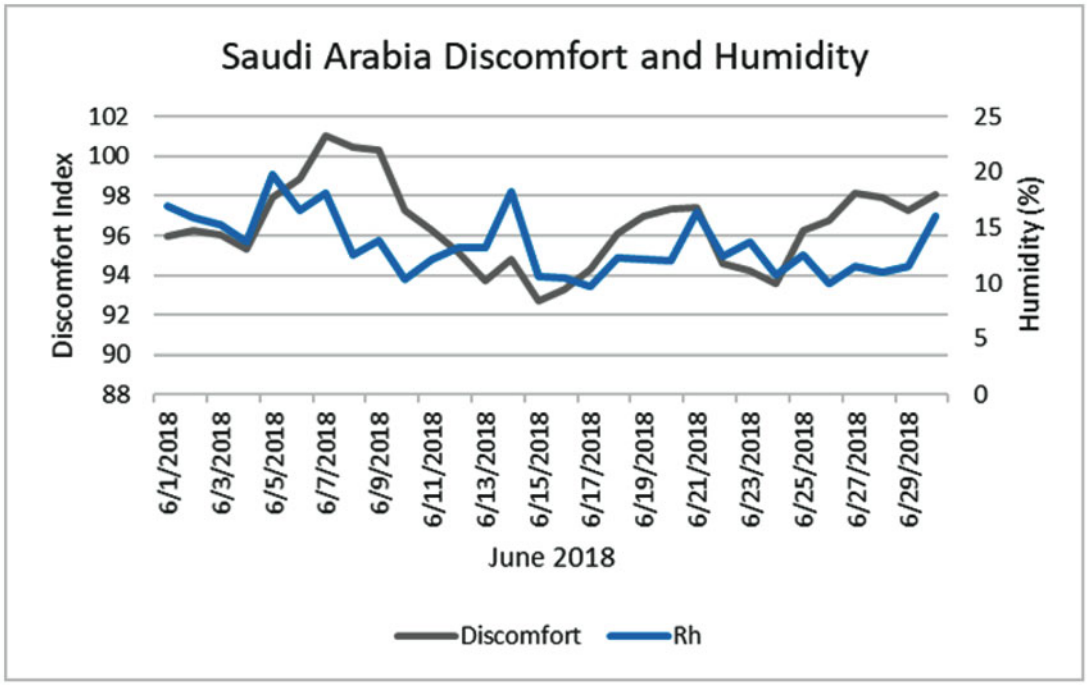


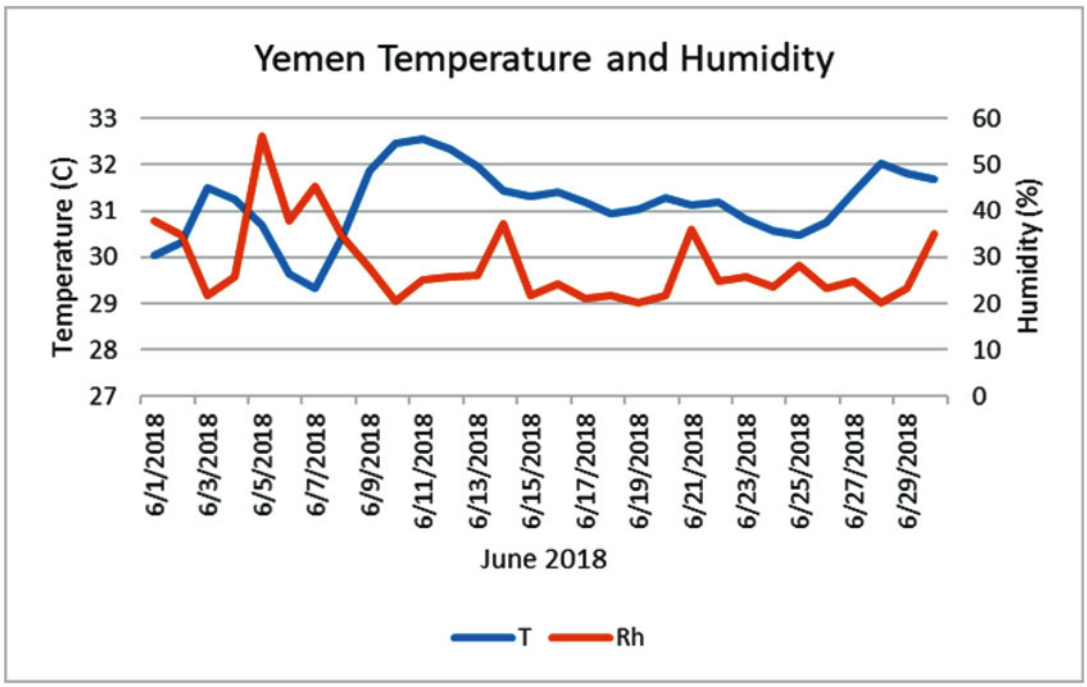


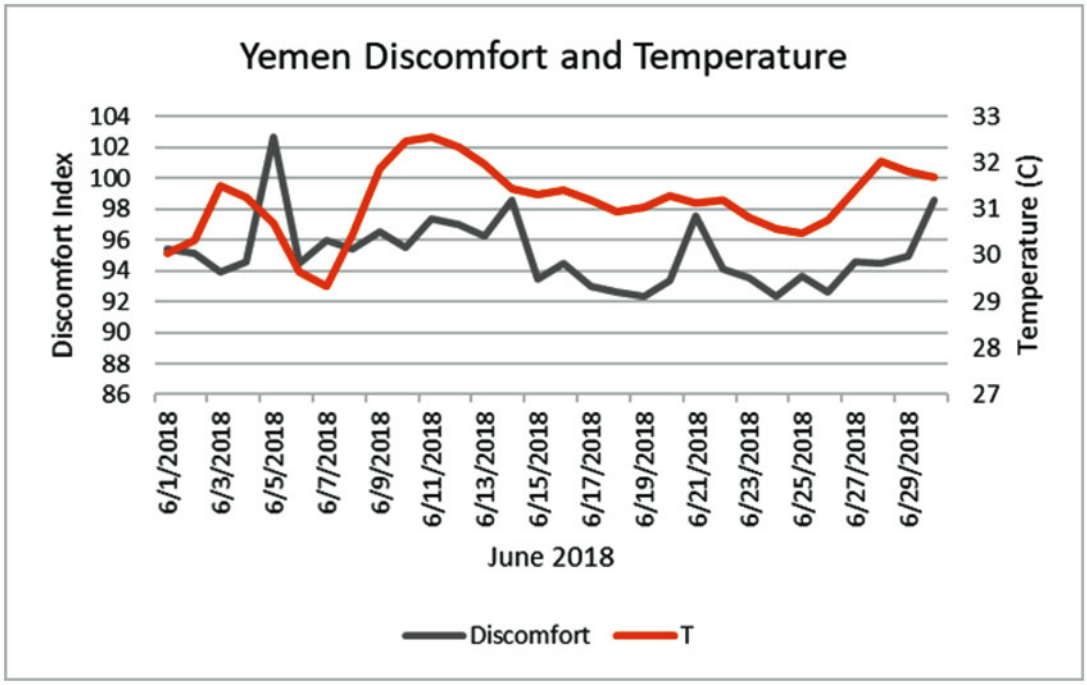
temperature, as would be expected, is fiercely hot, between about 33 and 36 °C. Thus, given the low RH values over most of the country and the high temperatures, the DI pattern in the time series closely matches the temperature pattern. It is interesting to note that even though the temperatures are considerably higher compared to Texas, the DI ranged between similar values, 91–100 for Texas and 96–101 for Saudi Arabia, though for the latter half of the month the DI in Saudi Arabia was in the 92–98 range.

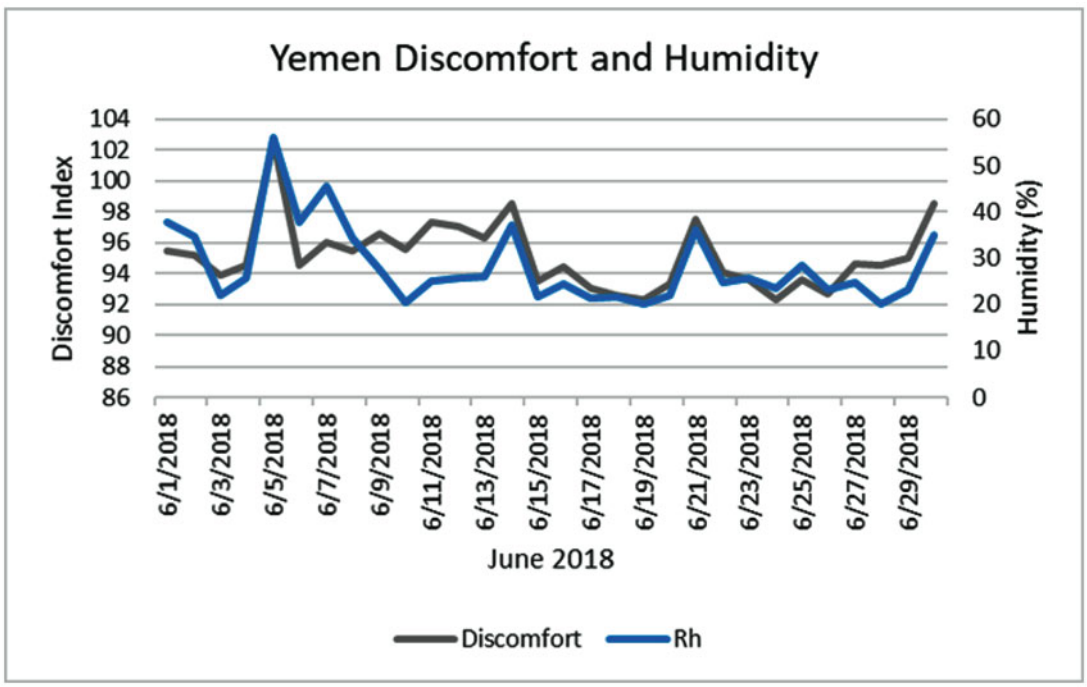
Yemen

The country of Yemen is located south of Saudi Arabia and has a coastline on both the Arabian Sea and the Gulf of Aden.


Saudi Arabia Discomfort and Humidity
Discomfort Index
Humidity (%)
102
100
98
96
94
92
90
88
25
20
15
10
5
0
6/1/2018
6/3/2018
6/5/2018
6/7/2018
6/9/2018
6/11/2018
6/13/2018
6/15/2018
6/17/2018
6/19/2018
6/21/2018
6/23/2018
6/25/2018
6/27/2018
6/29/2018
June 2018
Discomfort
Rh


Yemen Temperature and Humidity
Temperature (C)
Humidity (%)
33
32
31
30
29
28
27
60
50
40
30
20
10
0
6/1/2018
6/3/2018
6/5/2018
6/7/2018
6/9/2018
6/11/2018
6/13/2018
6/15/2018
6/17/2018
6/19/2018
6/21/2018
6/23/2018
6/25/2018
6/27/2018
6/29/2018
June 2018
T
Rh





Given this location, and the size of the country, increased humidity is a definite aspect of the country's climate. Still, its location on the Arabian Peninsula keeps its average humidity significantly lower than that of Florida. During June 2018, the RH saw a high of just over 50% early in the month and remained in the range 20–40% after June 9. The temperature was lowest (about 29.5 °C) when the humidity was highest and rose to between 30.5 and 32 °C after June 9. In our examination, Yemen is the only region of interest where there is a closer match between the time series of RH and DI than between temperature and DI. Unlike the other states and regions, higher temperatures did not contribute to the DI value as much as lower RH values. This is particularly evident in the latter half of the month. So, for Yemen, the discomfort of the populace during the warmth of summer is determined more by the humidity than the temperature,

and for this location, that is likely closely related to the direction that the wind blows, either from the dry interior of the Arabian Peninsula or the sultry waters of the Red Sea and Gulf of Arabia.

Summary and Conclusions

This examination of public health research topics in which Giovanni was employed, and a case study showing how Giovanni can be used to examine heat stress conditions, provides ample demonstration that the system allows investigative research and useful results for many different topics. In many of the research papers we have discussed, Giovanni contributed data or data visualizations as part of an entire analysis related to a particular public health issue. The nature of the contribution varies with the actual issue being investigated—data variables in Giovanni are more directly related to air quality, for example, compared to the factors contributing to the spread of influenza. This difference in applicability was the motivation for our classification of the data variables in Giovanni as Tier 1, Tier 2, or Tier 3 with respect to their connection to public health issues of import.

We also realize that evaluating research papers published in journals provides insight into only one aspect of Giovanni usage. Throughout our years of experience with the system, we have also seen data and data visualizations used in meeting presentations, government reports, consultant reports, monitoring compilations, and popular media articles, all having some relationship to public health. So we are confident that the system provides a valuable source of data for both researchers and professionals in the public health field.

Our case study provides an example of a topic for which we have not seen Giovanni used in a research journal, but one which is a significant public health concern, and one which could be the subject of more in-depth research. There have been recent events of higher mortality related to heat waves in Europe (2003), India (2015), and recently a summer 2018 heat wave in southern Quebec. The method of analysis shown in our case study could be applied to each of these events as part of a research effort examining the heat stress-related mortality and its relationship to climate change. Furthermore, for regions that are known to have potential for heat stress injuries and mortality, notably tropical or semi-tropical coastal areas, our simple analysis method could be applied to identify the combination and pattern of meteorological factors that lead to dangerous conditions in a retrospective analysis. This information would be useful to many different organizations concerned with the management of outdoor activities and with the care and protection of individuals at heightened risk.

We have thus shown that Giovanni has contributed to public health research, even in the system's earliest instantiation at the beginning of its scientific usage. Giovanni can and will continue to inform the arena of public health science with regard to many different environmental influences on local, regional, and global health issues.

Acknowledgments Melanie Ventura, summer intern at the GES DISC from Fairmont Heights High School, Landover, Maryland, performed the relative humidity-temperature-discomfort index analyses presented in the case study. The contributions of the data providers, mission specialists, and research scientists who provide the various datasets in Giovanni are recognized with gratitude. Finally, all of the research and applications performed with the system would not have been possible without the dedicated efforts of the Giovanni development team members, both past and present.

References

Acker, J.G. 2015. Chapter 5: High stakes and bright seas. In *The color of the atmosphere with the ocean below: A history of NASA's ocean color missions*. North Charleston: CreateSpace Independent Publishing Platform, ISBN 1507699220, 362 pages.

Adama, I., and M.B. Mochiah. 2017. Assessing the relationship between outbreaks of the African armyworm and climatic factors in the forest transition zone of Ghana. *British Journal of Environment & Climate Change* 7 (2): 69–82. https://doi.org/10.9734/BJECC/2017/30588.

Allen, S.K., P. Rastner, M. Arora, C. Huggel, and M. Stoffel. 2015. Lake outburst and debris flow disaster at Kedarnath, June 2013: Hydrometeorological triggering and topographic predisposition. *Landslides*. https://doi.org/10.1007/s10346-015-0584-3.

Athanasopoulou, E., D. Rieger, C. Walter, H. Vogel, A. Karali, M. Hatzaki, E. Gerasopoulos, B. Vogel, C. Giannakopoulos, M. Gratsea, and A. Roussos. 2014. Fire risk, atmospheric chemistry and radiative forcing assessment of wildfires in eastern Mediterranean. *Atmospheric Environment* 95: 113–125. https://doi.org/10.1016/j.atmosenv.2014.05.077.

Buchholz, R.R., C. Paton-Walsh, D.W.T. Griffith, D. Kubistin, C. Caldow, J.A. Fisher, N.M. Deutscher, G. Kettlewell, M. Riggenbach, R. Macatangay, P.B. Krummel, and R.L. Langenfelds. 2015. Source and meteorological influences on air quality (CO, CH4 & CO2) at a Southern Hemisphere urban site. *Atmospheric Environment*. https://doi.org/10.1016/j.atmosenv.2015.11.041.

Burney, J., and V. Ramanathan. 2014. Recent climate and air pollution impacts on Indian agriculture. *Proceedings of the National Academy of Sciences* 111 (46): 16319–16324. https://doi.org/10.1073/pnas.1317275111.

Corbari, C., F. Lassini, and M. Mancini. 2016. Effect of intense short rainfall events on coastal water quality parameters from remote sensing data. *Continental Shelf Research* 123: 18–28. https://doi.org/10.1016/j.csr.2016.04.009.

Creamean, J.M., P.J. Neiman, T. Coleman, C.J. Senff, G. Kirgis, R.J. Alvarez, and A. Yamamoto. 2016. Colorado air quality impacted by long-range-transported aerosol: A set of case studies during the 2015 Pacific Northwest fires. *Atmospheric Chemistry and Physics* 16 (18): 12329–12345. https://doi.org/10.5194/acp-16-12329-2016.

Doty, B.E., and J.L. Kinter III. 1995. Geophysical data analysis and visualization using GrADS. In *Visualization techniques in space and atmospheric sciences*, ed. E.P. Szuszczewicz and J.H. Bredekamp, 209–219. Washington, D.C.: NASA.

Dwivedi, R., P. Priyaja, M. Rafeeq, and M. Sudhakar. 2015. MODIS-Aqua detects *Noctiluca scintillans* and hotspots in the central Arabian Sea. *Environmental Monitoring and Assessment* 188 (50)., published online December 2015. https://doi.org/10.1007/s10661-015-5041-1.

Farrow, A., M. Didace, S. Cook, and R. Buruchara. 2011. Assessing the risk of root rots in common beans in East Africa using simulated,

estimated and observed daily rainfall data. *Experimental Agriculture* 47: 357. https://doi.org/10.1017/S0014479710000980.

Feister, U., G. Meyer, G. Laschewski, and C. Boettcher. 2015. Validation of modeled daily erythemal exposure along tropical and subtropical shipping routes by ship-based and satellite-based measurements. *Journal of Geophysical Research - Atmospheres* 120. https://doi.org/10.1002/2014JD023005.

Ganguly, N.D. 2016. Atmospheric changes observed during April 2015 Nepal earthquake. *Journal of Atmospheric and Solar-Terrestrial Physics* 140: 16–22. https://doi.org/10.1016/j.jastp.2016.01.017.

Jury, M.R. 2012. Physical oceanographic influences on central Benguela fish catch. *Earth Interactions* 16: 1–15. https://doi.org/10.1175/2012EI421.1.

Jutla, A.S., A.S. Akanda, and S. Islam. 2010. Tracking cholera in coastal regions using satellite observations. *Journal of the American Water Resources Association (JAWRA)*: 1–12. https://doi.org/10.1111/j.1752-1688.2010.00448.x.

Kehrwald, N.M., L.G. Thompson, Y. Tandong, E. Mosley-Thompson, U. Schotterer, V. Alfimov, J. Beer, J. Eikenberg, and M.E. Davis. 2008. Mass loss on Himalayan glacier endangers water resources. *Geophysical Research Letters* 35: L22503. https://doi.org/10.1029/2008GL035556.

Kishcha, P., B. Starobinets, O. Kalashnikova, and P. Alpert. 2011. Aerosol optical thickness trends and population growth in the Indian subcontinent. *International Journal of Remote Sensing*: 13 pages. https://doi.org/10.1080/01431161.2010.550333.

Koplitz, S.N., L.J. Mickley, M.E. Marlier, J.J. Buonocore, P.S. Kim, T. Liu, M.P. Sulprizio, R.S. DeFries, D.J. Jacob, J. Schwartz, M. Pongsiri, and S.S. Myers. 2016. Public health impacts of the severe haze in Equatorial Asia in September–October 2015: Demonstration of a new framework for informing fire management strategies to reduce downwind smoke exposure. *Environmental Research Letters* 11 (9): 11 pages. https://doi.org/10.1088/1748-9326/11/9/094023.

Lamchin, M., J.-Y. Lee, W.-Y. Lee, E.J. Lee, M. Kim, C.-H. Lim, H.-A. Choi, and S.-R. Kim. 2015. Assessment of land cover change and desertification using remote sensing technology in a local region of Mongolia. *Advances in Space Research*. https://doi.org/10.1016/j.asr.2015.10.006, available online October 14, 2015, 14 pages.

Lu, Z., D.G. Streets, Q. Zhang, S. Wang, G.R. Carmichael, Y.F. Cheng, C. Wei, M. Chin, T. Diehl, and Q. Tan. 2010. Sulfur dioxide emissions in China and sulfur trends in East Asia since 2000. *Atmospheric Chemistry and Physics* 10: 6311–6331, www.atmos-chem-phys.net/10/6311/2010/. https://doi.org/10.5194/acp-10-6311-2010.

Makgabutlane, M., and C.Y. Wright. 2015. Real-time measurement of outdoor worker's exposure to solar ultraviolet radiation in Pretoria, South Africa. *South African Journal of Science* 111 (5/6)., 7 pages. https://doi.org/10.17159/sajs.2015/20140133.

McElroy, J.H., and R.A. Williamson. 2004. The evolution of earth science research from space: NASA's earth observing system. In *Chapter 4 in Exploring the unknown, Volume VI: Space and earth science*, ed. J.M. Logsdon, S.J. Garber, R.D. Launius, and R.A. Williamson. Washington, D.C.: U.S. Government Printing Office, ISBN 0160731356.

McNally, A., K. Arsenault, S. Kumar, S. Shukla, P. Peterson, S. Wang, C. Funk, C. Peters-Lidard, and James P. Verdin. 2017. A land data assimilation system for sub-Saharan Africa food and water security applications. *Scientific Data* 4: 170012, 19 pages. https://doi.org/10.1038/sdata.2017.12.

Midekisa, A., G. Senay, G.M. Henebry, P. Semuniguse, and M.C. Wimberly. 2012. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. *Malaria Journal* 11: 165–181. https://doi.org/10.1186/1475-2875-11-165.

Moreno-Madrinan, M.J., W.L. Crosson, L. Eisen, S.M. Estes, M.G. Estes Jr., M. Hayden, S.N. Hemmings, D.E. Irwin, S. Lozano-Fuentes, A.J. Monaghan, D. Quattrochi, C.M. Welsh-Rodriguez, and E. Zielinski-Gutierrez. 2014. Correlating remote sensing data with the abundance of pupae of the dengue virus mosquito vector, *Aedes aegypti*, in Central Mexico. *ISPRS International Journal of Geo-Informatics* 3 (2): 732–749. https://doi.org/10.3390/ijgi3020732.

Pan, G., D. Tang, and Y. Zhang. 2012. Satellite monitoring of phytoplankton in the East Mediterranean Sea after the 2006 Lebanon oil spill. *International Journal of Remote Sensing* 33 (23): 7482–7490. https://doi.org/10.1080/01431161.2012.685982.

Parisi, A.V., A. Amar, and D.P. Igoe. 2017. Long-term UV dosimeter based on polyvinyl chloride for plant damage effective UV exposure measurements. *Agricultural and Forest Meteorology* 243: 68–73. https://doi.org/10.1016/j.agrformet.2017.05.012.

Rebaudet, S., P. Gazin, R. Barrais, S. Moore, E. Rossignol, N. Barthelemy, J. Gaudart, J. Boncy, R. Magloire, and R. Piarroux. 2013. The dry season in Haiti: a window of opportunity to eliminate cholera. *PLOS Currents: Outbreaks*. 2013 Jun 10 [last modified: 2013 Jul 24]. Edition 1. https://doi.org/10.1371/currents.outbreaks.2193a0ec4401d9526203af12e5024ddc.

Serrano, M., J. Cañada, J.C. Moreno, and G. Gurrea. 2014. Personal UV exposure for different outdoor sports. *Photochemical and Photobiological Sciences* 13 (4): 671–679. https://doi.org/10.1039/C3PP50348H.

Simha, C.P., P.C.S. Devara, and S.K. Saha. 2013. Aerosol pollution and its impact on regional climate during Holi festival inferred from ground-based and satellite remote sensing observations. *Natural Hazards*: 1–15. https://doi.org/10.1007/s11069-013-0743-6.

Singh, S.K., A.C. Pandey, and M.S. Nathawat. 2011. Rainfall variability and spatio-temporal dynamics of flood inundation during the 2008 Kosi flood in Bihar State, India. *Asian Journal of Earth Sciences* 4 (1): 9–19.

Sitnov, S. 2011. Weekly variations in temperature and precipitation in Moscow: Relation with weekly cycles of air pollutants and synoptic variability. *Izvestiya Atmospheric and Oceanic Physics* 47 (4): 445–456. https://doi.org/10.1134/S0001433811040098.

Soebiyanto, R.P., W. Clara, J. Jara, L. Castillo, O.R. Sorto, S. Marinero, M.E. Barnett de Antinori, J.P. McCracken, M.-A. Widdowson, E. Azziz-Baumgartner, and R.K. Kiang. 2014. The role of temperature and humidity on seasonal influenza in tropical areas: Guatemala, El Salvador and Panama, 2008–2013. *PLoS One* 9 (6)., 11 pages. https://doi.org/10.1371/journal.pone.0100659.

Stumpf, R.P., T.T. Wynne, D.B. Baker, and G.L. Fahnenstiel. 2012. Interannual variability of cyanobacterial blooms in Lake Erie. *PLoS One* 7 (8): e42444. https://doi.org/10.1371/journal.pone.0042444.

Tan, C.K., J. Ishizaka, A. Manda, E. Siswanto, and S.C. Tripathy. 2007. Assessing post-tsunami effects on ocean colour at eastern Indian Ocean using MODIS Aqua satellite. In: "Satellite Observations Related to Sumatra Tsunami and Earthquake of 26 December 2004". *International Journal of Remote Sensing* 13–14: 3055–3069.

Tesi, T., S. Miserocchi, F. Acri, L. Langone, A. Boldrin, J.A. Hatten, and S. Albertazzi. 2013. Flood-driven transport of sediment, particulate organic matter, and nutrients from the Po River watershed to the Mediterranean Sea. *Journal of Hydrology* 498: 144–152. https://doi.org/10.1016/j.jhydrol.2013.06.001.

Tilburg, C.E., L.M. Jordan, A.E. Carlson, S.I. Zeeman, and P.O. Yund. 2015. The effects of precipitation, river discharge, land use and coastal circulation on water quality in coastal Maine. *Royal Society Open Science* 2: 140429. https://doi.org/10.1098/rsos.140429.

Wainwright, L., A.V. Parisi, and N. Downs. 2016. Dual calibrated dosimeter for simultaneous measurements of erythemal and vitamin D effective solar ultraviolet radiation. *Journal of Photochemistry and Photobiology B: Biology* 157: 15–21. https://doi.org/10.1016/j.jphotobiol.2016.02.003.

Wijesundera, I., M.N. Halgamuge, T. Nanayakkara, and T. Nirmalathas. 2016. Chapter 2: Predicting cyclone induced flood: A comprehensive case study. In *Natural disasters, when will they reach me?* Springer Natural Hazards Series, ed. I. Wijesundera, M.N. Hal-

gamuge, T. Nanayakkara, and T. Nirmalathas, 29–66. Singapore: Springer. https://doi.org/10.1007/978-981-10-1113-9_3.

Wright, C.Y., P.N. Albers, A. Mathee, Z. Kunene, C. D'Este, A. Swaminathan, and R.M. Lucas. 2017. Sun protection to improve vaccine effectiveness in children in a high ambient ultraviolet radiation and rural environment: An intervention study. *BMC Public Health* 17 (37): 8 pages. https://doi.org/10.1186/s12889-016-3966-0.

Wu, J., M. Yunus, Md.S. Islam, and M. Emch. 2016. Influence of climate extremes and land use on fecal contamination of shallow tubewells in Bangladesh. *Environmental Science and Technology* 50 (5): 2669–2676. https://doi.org/10.1021/acs.est.5b05193.

Yu, B., F. Chen, B. Li, L. Wang, and M. Wu. 2017. Fire risk prediction using remote sensed products: A case of Cambodia. *Photogrammetric Engineering & Remote Sensing* 83 (1): 19–25. https://doi.org/10.14358/PERS.83.1.19.

Zaninovich, S.C., J.L. Fontana, and M.G. Gatti. 2016. Atlantic Forest replacement by non-native tree plantations: Comparing aboveground necromass between native forest and pine plantation ecosystems. *Forest Ecology and Management* 363: 39–46. https://doi.org/10.1016/j.foreco.2015.12.022.

Geospatial Analysis of the Urban Health Environment

Juliana Maantay, Angelika Winner, and Andrew Maroko

Introduction to the Spatial Analysis of Urban Health

More than half of the world's population currently lives within urbanized areas, and this percentage is expected to increase in future. Many of the global public health and environmental challenges of the twenty-first century occur in urban areas, and there are some issues that are unique to, or greatly exacerbated in, cities (Freudenberg et al. 2009). Some of the potential issues adversely impacting the health of urban residents are as follows:

- High density of urban areas
- Housing overcrowding
- Poor-quality and hazard-prone housing
- Informal residential slum areas without services
- Racial and ethnic segregation
- Pressures on infrastructure (water supply, waste water treatment, solid waste disposal, energy distribution, etc.)
- Inadequate transportation systems
- Concentration of pollution and noxious land uses
- Food insecurity
- Urban violence and crime
- Extreme differential access to services (e.g., health providers, emergency care, public transportation, adequate fire and police protection, social services, other governmental services) and beneficial environmental amenities (e.g., parks, recreational and cultural opportunities, healthy food options, safe and attractive environments) among population groups within the same municipality
- Disproportionate exposures to risk and hazards (e.g., fires, flooding, unsafe living environments, landslides, earthquakes), climate change, and extreme weather events

Geospatial analysis and Geographic Information Science (GISc) technology can help address these issues in a number of ways. For instance, calculating the geographical extent of an environmental burden or benefit in order to estimate and identify the population potentially exposed to or impacted by the event or condition, illuminating the spatial relationships between environmental conditions and health outcomes, and examining the spatial and temporal patterns of how diseases cluster are just a few of the types of analyses conducted to help guide policy-makers, health experts, and public officials in improving health outcomes, reducing inequity, and mitigating vulnerability. These techniques and methods will be reviewed in this chapter.

The goals and purpose of medical geography (now more typically termed "health geography" or "health geograph-

J. Maantay (✉)
City University of New York, Lehman College, Department of Earth, Environmental, and Geospatial Sciences, New York, NY, USA

CUNY Graduate Center, Earth and Environmental Sciences Doctoral Program, New York, NY, USA

CUNY Graduate School of Public Health and Health Policy, Doctoral Program, New York, NY, USA

CUNY, City College of New York, NOAA-CREST Research Scientist, New York, NY, USA
e-mail: juliana.maantay@lehman.cuny.edu

A. Winner
City University of New York, Lehman College, Department of Earth, Environmental, and Geospatial Sciences, New York, NY, USA

CUNY Graduate Center, Earth and Environmental Sciences Doctoral Program, New York, NY, USA
e-mail: angelikawinner@gmail.com

A. Maroko
City University of New York, CUNY Graduate School of Public Health and Health Policy, Department of Environmental, Occupational, and Geospatial Health Sciences, New York, NY, USA

CUNY Graduate Center, Earth and Environmental Sciences Doctoral Program, New York, NY, USA
e-mail: Andrew.Maroko@sph.cuny.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_9

ics") can be thought of as an attempt to figure out the spatial patterns of disease – for instance, where a certain disease is geographically concentrated; whether a disease hot spot or cluster is more pronounced than we would expect based on the underlying population numbers; how disease spreads or diffuses; and how its location is connected with other environmental and socio-demographic variables. The early instances of medical mapping generally focused on contagious diseases (Koch 2005; LSHTM 2013). In more recent times, medical geography has expanded to examine the spatiality of chronic health conditions, rare diseases, genetic disorders, HIV/AIDS, and other health concerns, such as drug use and misuse, accidents, suicides, and interpersonal violence. Investigating the spatial patterns in these health outcomes is a bit like detective work, and some of the avenues of research are red herrings or lead to dead ends. Often the analyses cannot pinpoint correlations or causalities.

Even with current, comprehensive, and relatively reliable contemporary health data and state-of-the-art statistical and analytical techniques, there are many pitfalls and deficiencies in using geospatial analysis and mapping to prove causation, or even to provide reasonable explanations of health issues (Maantay 2002, 2007). However, mapping can be a stepping stone to greater understanding of critical spatial relationships and of space-time trends, thus potentially pointing the way to answers or, at least, to better questions.

Somewhat more tenuous in terms of definitive conclusions is the mapping of historical medical data, due mainly to limitations of data availability at an optimal scale, and questions about data accuracy and completeness. Historical medical mapping is one of the most fascinating topics in health geography, and much of it pertains to urban areas and urban populations since that is typically where it was undertaken first. It is often thought that Dr. John Snow's seminal work in mapping the 1854 cholera epidemic in London was the first real instance of the geospatial analysis of a disease. Dr. Snow, through statistics and mapping, was able to make the connection between the residential locations of fatalities from the disease and the probable source of the disease, a public water pump. This was significant because up until this time, cholera was considered to be caused by unhealthy air, a "miasma," when in reality it is a water-borne disease. Once the cause was demonstrated fairly definitively, authorities and the population could go about developing the remedies (Johnson 2007).

While Dr. Snow's work is the best-known example of historical medical mapping, it was hardly the first. A full decade earlier, Dr. Robert Perry, a surgeon in Glasgow's Royal Infirmary, also mapped a fever epidemic and found substantial spatial correspondence between the locations of the fever victims and environmental and socio-economic characteristics of the neighborhoods in Glasgow, Scotland (Perry 1844) (See Fig. 1). Even earlier, disease mapping had been undertaken during various plagues and epidemics (such as during the 1690 Black Plague in Bari, Italy, and the 1798 Yellow Fever outbreak in New York City (NYC)) to aid in the quarantine efforts of cities as well as to enhance understanding of the disease's causation and prevention. Additionally, medical mapping was employed to monitor disease occurrences during the eighteenth- and nineteenth-

Fig. 1 *Left:* The map of Glasgow, where Dr. Perry numbered the districts 1–17 and used different colors to indicate the level of the epidemic in each area. The black dots represent fever cases (likely typhus). *Right*: Detail of the map showing three of the districts most gravely affected by the epidemic. Perry also visually and statistically correlated the fever epidemic with overcrowding, poor sanitation, and poverty, and estimated that in districts 3 and 4, over 20 per cent of the population had been affected. These were the same areas that tended to be the poorest and most overcrowded (Perry 1844). (*Figure Source*: Perry 1844http://special.lib.gla.ac.uk/exhibns/month/feb2006.html)

century European military forays and exploratory colonizing and mercantile expeditions in Asia and Africa (Koch 2005; LSHTM 2013).

In recent times, the geospatial analysis of urban health has focused more on environmentally related health concerns rather than contagious diseases, although in some parts of the world, contagious diseases are still prioritized due to their prevalence. The recent COVID-19 pandemic has demonstrated once again the importance of disease mapping, as many worried people around the world avidly viewed maps and graphs updated on an almost daily basis, showing case rates, death rates, hospitalization rates, hot spots, the spread of the disease's geographic extent, and the changing locations of its highest prevalence. Mapping COVID-19 also informed us about the connection between the illness and various socio-demographic, economic, housing, and environmental factors of the affected population, aiding in a better understanding of the disease's transmission and risk to certain subpopulations (Chakraborty 2021).

Many vector-borne diseases, like malaria and dengue fever, can also be considered environmentally related health outcomes, for which it is important to conduct optimal habitat (or suitability) analyses to determine where the disease vector, such as the mosquito, is most likely to proliferate (Kleinschmidt et al. 2000; Thomson et al. 1997). We can now use spatial analysis and Geographic Information Science to investigate and model a wide range of health concerns, many of which are especially acute in urban areas.

Measuring and Quantifying Urban Exposures

A geospatial analysis of environmental health in the urban context is at the nexus of the geography of environmental hazards and the geography of the exposed population. GISc has proven to be a valuable tool for bringing these two geographies together, allowing us to map and model environmental hazards and the exposed population simultaneously as well as linking these two data groups together in order to measure and quantify exposure to environmental hazards (Maheswaran and Craglia 2004). GISc-based exposure assessment is concerned with determining the areas affected by adverse effects of environmental hazards, estimating the population and characteristics of those living in the affected areas, and analyzing if certain population groups such as minorities are disproportionally affected (Maantay and McLafferty 2011).

Before one gets started with a geospatial analysis of urban exposure, it is imperative to explore the given data set spatially in order to gain a more holistic understanding of the phenomena or processes under study. Exploratory Spatial Data Analysis (ESDA) represents a powerful tool-set for data exploration which allows for the revealing and clarification of relationships, patterns, and correlations. ESDA is very helpful not only with the generation of research questions and hypotheses but also with the refinement of research design (Maantay 2013). Figure 2 provides an example where ESDA is used to explore the spatial relationship between areas of poor mental health, high economic and social deprivation, and proximity to derelict land.

Delineation of the Boundaries of Adverse Environmental Exposure

Several spatial analytical methodologies have been used in urban exposure assessment, which measure proximity to environmental hazards and estimate the boundaries of potentially affected areas. These assessment techniques are used not only to measure exposure to environmental hazards but also to estimate access to environmental benefits and public services. These different methodologies can be broadly classified into four groups, as described in this section: spatial coincidence analysis, distance-based proximity methods, pollutant fate and transport modeling, and spatial statistical methods (Chakraborty and Maantay 2011).

Spatial Coincidence Method

The spatial coincidence method represents the simplest exposure assessment method (Maheswaran and Craglia 2004). This method assumes that exposure to environmental hazards occurs within and is restricted to pre-defined geographic entities or administrative units such as ZIP codes, census tracts, or block groups containing such hazards (Chakraborty and Maantay 2011). In other words, if a spatial unit contains an environmental hazard, it is assumed that all people residing within this spatial unit are exposed to the adverse effects of the given hazard(s). The socio-economic and demographic characteristics of the exposed spatial units, also called host units, are then statistically compared to all other (non-host) units that do not contain any hazards to evaluate if certain population groups are disproportionally exposed to environmental hazards.

Deciding upon the spatial unit which represents the host area is not a trivial process since the size of the unit will affect the accuracy as well as the statistical significance of the results – generally it can be assumed that choosing a smaller spatial unit will yield more accurate results, whereas the use of a larger unit increases the strength and significance of statistical relationships between environmental hazards and socio-demographic variables (Maantay 2007). However, when using population data aggregated at higher levels, e.g., county or metropolitan area, it will be harder to detect if specific population groups are disproportionately affected by environmental hazards.

There are also other limitations associated with the hazard coincidence method. First, there is the problem of the so-

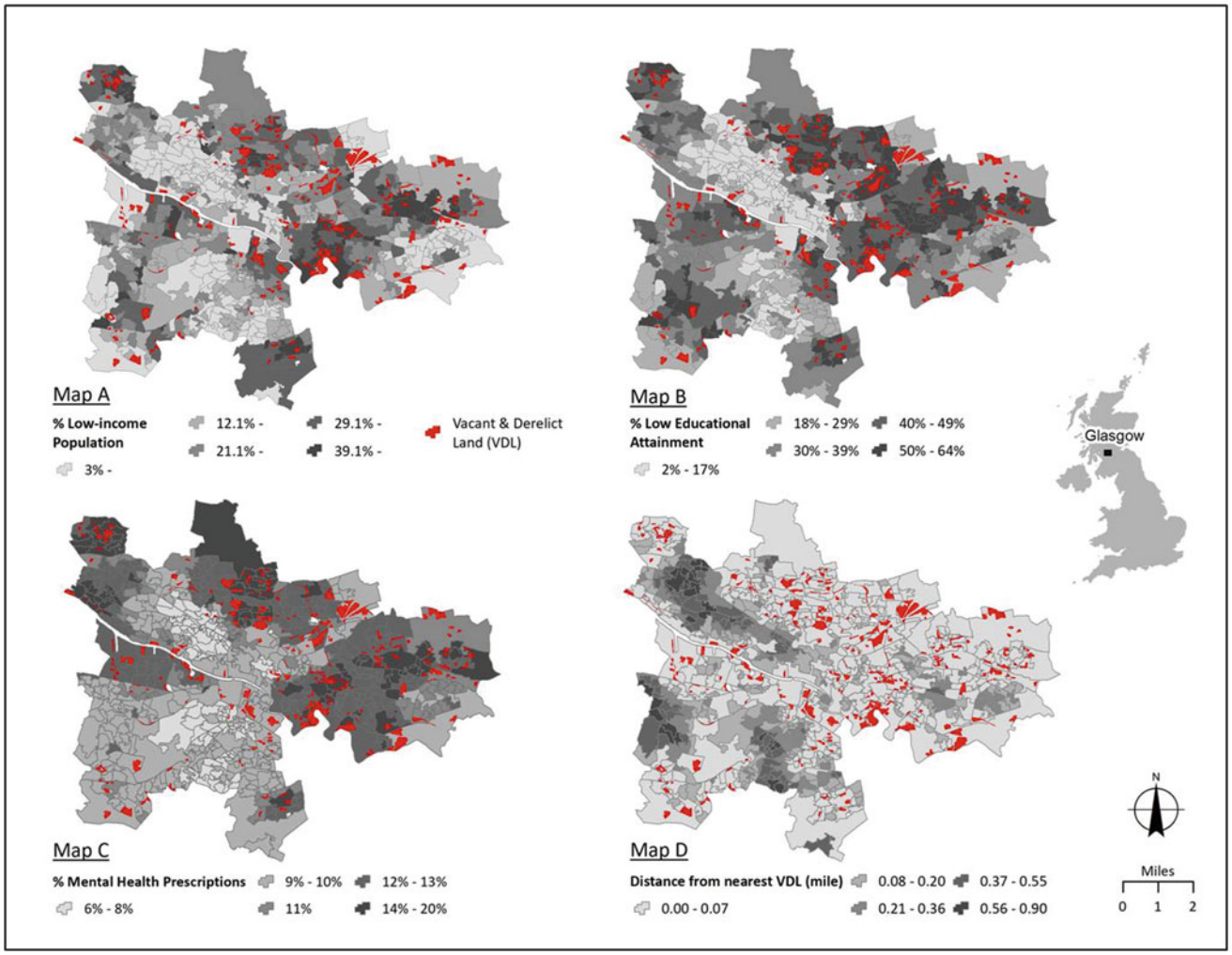


Fig. 2 GISc-based exploration of the spatial relationship between urban areas of high economic deprivation and low educational attainment (Maps A and B), mental health (Map C), and the proximity to vacant and derelict land (Map D) in Glasgow, Scotland. Data are aggregated to data zones (DZ) which are composed of about 750 people and resemble US census block groups. Distance from VDL represents a measure of exposure to various environmental stressors including the VDL itself, whereas DZ with high proportions of low educational attainment and low income represents a proxy for exposure to social stressors. The type of vacant and derelict land (VDL) included here is over 2 acres in size and has been vacant since 2004 or earlier. Mental health as shown in Map C is represented by the proxy variable of percentage of the population in each DZ who have been prescribed medication for anxiety, depression, or psychosis. Map D represents the distance in meters from each DZ centroid to the nearest VDL: the shorter the distance, the higher the likelihood of exposure to potential health hazards and environmental stressors related to VDL. Distance from the nearest VDL was calculated with the NEAR tool – a distance value of zero is assigned to DZs which either touch the boundary of or which contain a VDL. The maps highlight the fact that deprived areas not only contain most of the VDL in Glasgow but also host a higher percentage of the population prescribed medications for anxiety, depression, or psychosis. This simple example of ESDA may help with building a research design to further analyze the association between vacant and derelict land and mental health. For further detail, see Maantay and Maroko 2015; and Maantay 2013. (*Data sources*: Scottish Neighborhood Statistics 2011 and 2007; Office for National Statistics 2018; Scottish Government Vacant and Derelict Land Survey 2012. *Figure credit*: Angelika Winner)

called edge effect. The hazard coincidence method assumes that exposure is limited to the host unit only but ignores the fact that a hazard may be located very close to the edge of the unit, and, thus, a neighboring non-host unit could be equally exposed. Map (A) in Fig. 3 shows current and historic toxic release inventory (TRI) facilities in Northern Brooklyn, NY, with their host census tracts. The TRI program is run by the US EPA and it tracks the management of toxic chemicals posing a threat to human health and the environment. In map (A), many facilities are located close to the edge of their respective host unit. This is especially problematic in the case of large host units – see, for example, the large unit in the Northeast of the study area – as the hazard coincidence method could lead to the misidentification of large areas considered exposed or ignoring other areas that are potentially much more exposed that are right on the other side of the border of the host unit. Second, geographic or administrative

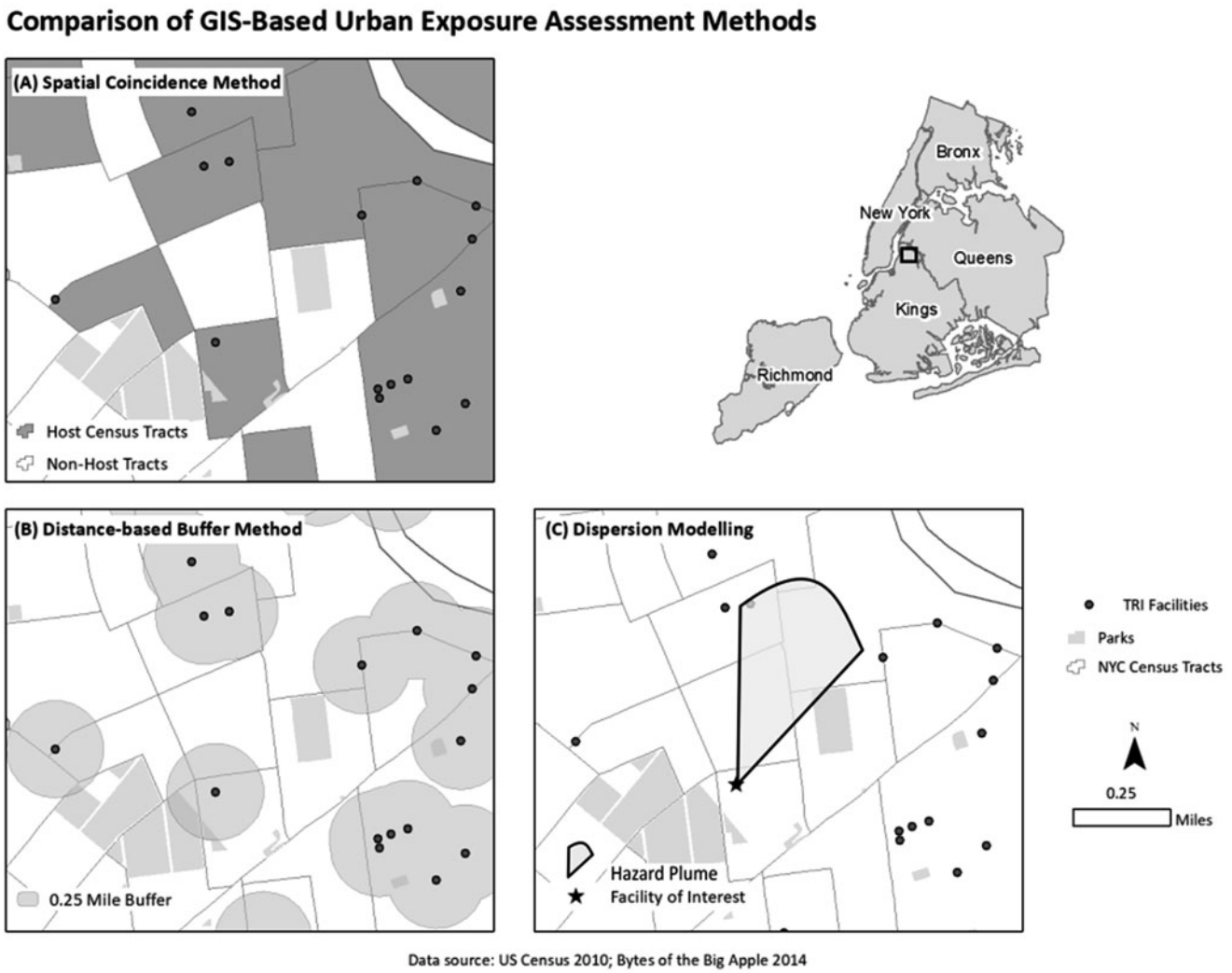


Fig. 3 This figure represents a comparison of three different GISc-based urban exposure assessment techniques discussed in the text. The maps show exposure to current and historic Toxic Release Inventory (TRI) facilities in a small section of North-Brooklyn (Kings County) with (**a**) the spatial coincidence method, (**b**) the distance-based buffer method, and (**c**) the dispersion modeling method. In the spatial coincidence method, any tract that hosts at least one facility is selected (its population is considered potentially impacted); (**b**) the distance-based buffer method is more exclusive since here only those areas within a quarter mile of the facilities are selected; (**c**) the dispersion technique only selects areas that are most likely to be exposed to environmental hazards released from the facility of interest based on facility-specific hazard release data and climatological data. For further details, see Chakraborty and Maantay 2011; and Maantay and Maroko 2017. (*Data sources*: US Census 2010; Bytes of the Big Apple 2014. *Figure credit*: Angelika Winner)

spatial units are typically not a good representation of the size or shape of the exposed areas. Third, it is assumed that everyone within the host unit is impacted equally; however, it is well documented that pollution does not disperse equally in all directions from a source (Maantay 2007). And fourth, most applications of the spatial coincidence method do not take environmental hazard density in consideration. In other words, they do not differentiate between spatial units hosting only one hazard and those hosting several. However, exposure works cumulatively – the more sources of pollution exist in a host unit, the higher will be the total exposure of its residents.

The last limitation can be mitigated by incorporating data on the quality and quantity of pollution emitted from each hazard source, which allows for the distinction between host spatial units on the basis of the magnitude of potential environmental risk (Chakraborty and Maantay 2011). However, even applications of the spatial coincidence that take into account the actual emissions and toxicity in a given host area cannot overcome the edge effect as well as the irregular spatial dispersion of pollutants.

Distance-Based Methods

Among the distance-based methods of exposure assessment, buffering is certainly one of the most widely used techniques,

and it rests on the basic principle that exposure declines with distance from the pollution source to a threshold beyond which the population is considered unexposed (Maheswaran and Craglia 2004). Buffer analysis is available for point, line, or polygon features depending on the geographic feature they represent – buffers around point features (e.g., toxic facilities) are generally circular, whereas buffers around lines (e.g., roads or powerlines) and polygons (e.g., noxious land uses and superfund sites) are irregularly shaped. Map (B) in Fig. 3 shows the same example of current and historic TRI facilities in North-Brooklyn, but here exposed areas are identified with 0.25-mile buffers around the toxic facilities. The toxic facilities are represented as point features and thus the buffers are circular. In Fig. 4, circular buffers around TRI facilities are combined with irregular line buffers around limited access highways to investigate the cumulative effect of both sources of air pollution on asthma rates in the Bronx.

In creating the buffers, one must determine the appropriate distance beyond which people are considered unexposed. The identification of an appropriate buffer distance, however, is most often guesswork and is typically not based on empirical data (Maantay 2007). Once the buffers have been computed, the user must select the areal units falling within the buffers to identify the units and the corresponding number of people that are potentially exposed to the environmental hazard. Then, the socio-demographic characteristics of exposed areas (units inside the buffers) may be statistically compared to the rest of the study area (outside the buffers) to determine disproportionate exposure to the given hazards (Chakraborty and Maantay 2011).

When focusing in on the same facility located in the center of the study area, it becomes clear how much of an improvement the buffer method is over the spatial coincidence method. Now, all surrounding spatial units within a quarter mile of the toxic facility intersect the buffer and thus must be considered as potential host units, and not just the spatial unit hosting the facility as with the spatial coincidence method in the previous example.

Even though buffer analysis represents an improvement over the spatial coincidence method, there are several limitations associated with its application. One limitation is that the buffer distance is usually chosen arbitrarily and that all buffers have the same radius. The determination of the buffer radii is typically based on assumptions and ignores the nature and quantity of pollutants released in the environment, as well as operational parameters and environmental conditions at the time of release (Chakraborty and Maantay 2011). Another problem is that like the spatial coincidence method the buffer method is based on the dichotomous assumption that the adverse effects of a hazard are restricted only to the buffer area, whereas areas outside the buffer are considered unaffected. Thus, the results of a buffer analysis are very sensitive to the choice of buffer distance. In addition, just as the simpler spatial coincidence method, the buffer areas typically do not accurately represent the dispersion of the environmental hazard(s) of concern. Pollutants are typically not dispersed evenly in all directions and their concentrations decrease more gradually. The last limitation can be overcome to a certain degree by using multiple ring buffers combined with estimated release volumes and emissions – however, the determination of buffer distances remains largely subjective, and multiple ring buffers do not necessarily improve exposure assessment.

A more promising solution is the use of continuous distances, based on the calculation of the exact distance between locations of the potentially exposed population and environmental hazards. Kernel density estimation (KDE, also often called "heat mapping") represents an alternative method of exposure assessment as it can give a more nuanced estimation with different levels such as no, low, medium, or high exposure, and not just a binary answer of "within the buffer" or "not within the buffer" as in the discrete buffer analysis (Maroko et al. 2009). The KDE approach does not just consider the location of environmental hazards but considers the density of the features in a search area defined by the user. The KDE approach calculates the density of hazards in a neighborhood and then creates a smoothly curved surface over each point of the area with the highest surface values at the location of the hazard and lower values with increasing distance from the hazard. One limitation with this method is that it only allows for circular search areas.

When one wants to measure access to environmental goods such as parks or social services such as health centers, a network buffer analysis may be a big improvement over the simple buffer method since this buffer technique takes into consideration the actual street network available for traveling by foot and/or car when calculating buffer distances, rather than just "as-the-crow-flies" distances (see Fig. 12 in the next section).

Pollutant Fate and Transport Modeling

One of the limitations that both distance-based methods and the spatial coincidence method of exposure assessment have in common is the lack of acknowledgment of the physical process of dispersion which is often fundamental to the spatial distribution of pollutants in the environment. The potential for environmental exposure depends not just on the distance to a pollution source but much more so on the effects of pollution dispersion in the environment (Maheswaran and Craglia 2004). Because dispersion is a physical process affected by environmental conditions such as flow speed (wind or water) and flow direction, the two previously discussed methods of exposure assessment can only ever provide approximate estimates of exposure extent. In order to provide a more accurate spatial representation of exposure extent, detailed information on toxic chemical emissions,

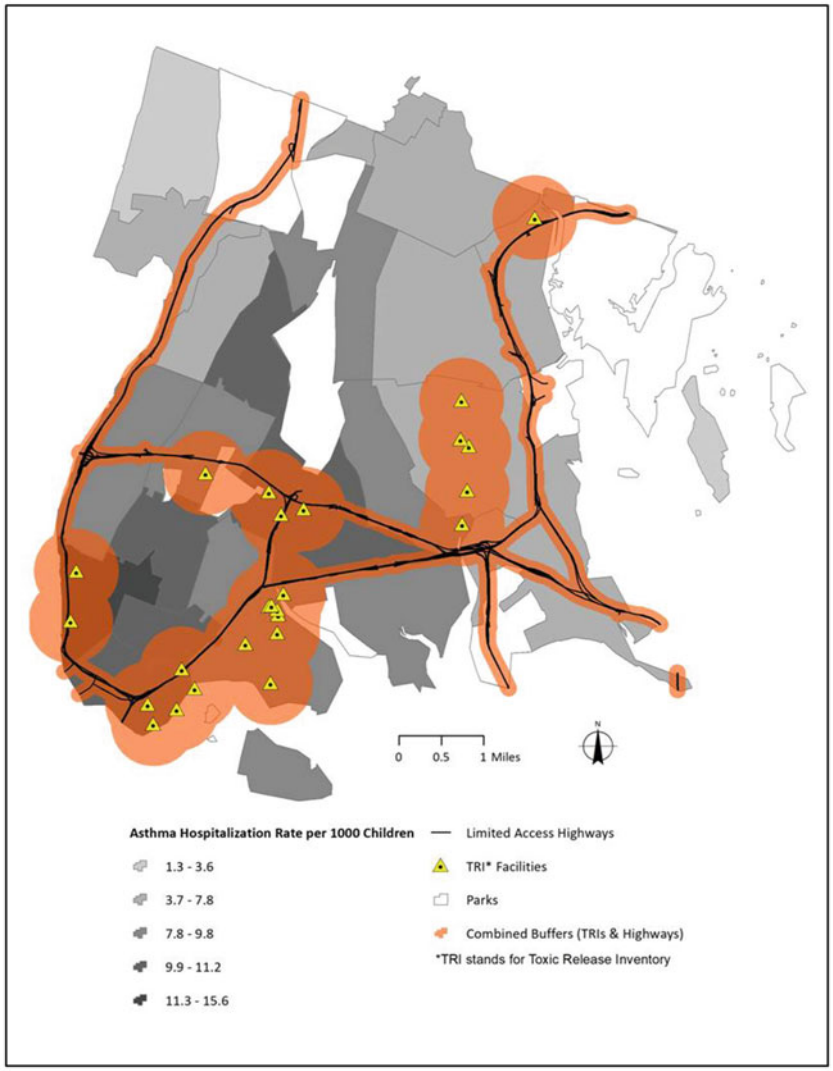


Fig. 4 Childhood asthma hospitalization rates, major air pollution sources, and likely areas of impact in the Bronx, NY (Maantay 2007). Here, areas exposed to current and historic TRI facilities and limited access highways, both sources of air pollutants linked to asthma, are estimated with the distance-based buffer method. The map shows the combined buffers of TRIs (stationary point pollutions sources) and limited access highways (linear mobile source pollution) as well as asthma hospitalization rates per 1000 children aggregated to ZIP code areas. The chosen TRI buffer distance was 0.5 miles, whereas for highway buffers, the distance was 500 feet – the decision was made based on previous research (Wilson et al. 2012; Maantay 2007; Mohai and Saha 2006; Chakraborty and Armstrong 1997). However, the decision is still based on best "guesstimates" based on limited empirical evidence for exposure which is one of the limitations of the buffer technique. Since TRI facilities are point features, the associated buffers are circular. Highways are represented as line features which results in more irregular polygon buffers. Even though many ZIP code areas containing TRIs and highways exhibit higher asthma hospitalization rates, there are ZIP code areas with high rates that are not within the combined buffer and ZIP code areas within the combined buffers with lower rates. There are two main reasons for this. First, the asthma data are aggregated to spatial units (ZIP code areas) too large to accurately represent the spatial relationship between pollution sources and health impacts – a very common problem associated with health data. Here, the spatial association between asthma and major air pollution sources is strong only in those areas where pollution sources cluster and thus the combined buffers cover one or more ZIP code areas. Second, there are many other sources of outdoor air pollution not included by simply looking at major point sources of air pollution and limited access highways, such as major truck routes, smaller polluting facilities, waste and recycling facilities, construction sites, and energy-producing facilities, even those in neighboring boroughs. Third, asthma is not only influenced by outdoor air quality but also by indoor air quality, and individual and family behaviors and health histories. In contrast to this example, Maantay (2007) had access to geolocated individual, patient-level asthma hospitalization cases in the Bronx. Her research found that the people living within the proximity buffer boundaries of the TRI facilities were up to 60% more likely to be hospitalized for asthma than those living outside the buffers, showing that the higher asthma hospitalization rates were associated with closer proximity to local air pollution sources. For further details, see Maantay 2007. (*Data Sources*: NYC Dept. of City Planning 2010; NYC Health Department 2000. *Figure credit*: Angelika Winner)

local weather conditions, and other physical landscape and built environmental characteristics are needed to model the environmental fate and dispersal of pollutants released from the hazard source (Chakraborty and Maantay 2011).

Dispersion modeling of air (and water) pollutants is a long-established field and a wide range of models have been developed ranging from "simple plume dispersion models used for a single point or line source, to numerical grid models that incorporate interactions among numerous pollutants and produce three-dimensional spatial estimates for relatively large regions" (Setton et al. 2011). Air dispersion models are mathematical models which require a large amount of data about emission quantities and other physical characteristics of the pollutants, meteorological, and topographical factors. Dispersion models describe chemical and physical processes within the plume over time and space, calculate pollutant concentrations over the study area dispersion, and provide a more accurate assessment of potential exposure without the need for extensive monitoring networks (Maantay et al. 2009). A full exposure assessment with the dispersion method requires an integration of a dispersion model and a GIS, which usually involves intensive expertise in computer programming, GISc, and meteorology (Chakraborty and Maantay 2011). Map (C) in Fig. 3 exemplifies the output of a dispersion transport model showing the hazard plume of a selected TRI facility. The plume's shape delineates the modeled dispersion of the hazard, showing an approximation of how it is affected not only by wind direction and speed but also by operational parameters at the point of release. This leads to a much more accurate exposure assessment than with the fixed-distance buffer method which would select all areas around the facility within the given buffer radius and result in a circular or linear buffer of constant dimension. Figure 5 provides a more detailed example of a dispersion model output showing a sample pollution source with its pollution plume of PM_{10} (particulate matter of 10 microns or smaller).

Land-Use Regression Analysis

The land-use regression (LUR) technique differs from the previously discussed methods of exposure assessment because it is designed to estimate total ambient pollution rather than the pollutant(s) emitted from any specific source and represents the best suited method to analyze the effect of accumulative effects of total pollution burden on health outcomes (Maroko 2010). This analysis uses multivariate regressions of monitored air pollution data and the physical environment surrounding the monitors to predict air pollution concentrations applicable to any location in the study area such as a census tract centroid or place of residence (Hennig et al. 2016). Common LUR variables are related to road types, traffic count, and land cover. Figure 6 shows the result of a LUR analysis predicting particulate matter concentration ($PM_{2.5}$) based on proximity to a major truck route and land use in New York State.

The advantage of the LUR method is its easy applicability, but the quality of the model depends largely on the initial choice of locations as well as the total number of the measurement sites. Additionally, LUR was developed for long-term predictions of air pollutant concentrations that are temporally relatively stable – this means that it is not the appropriate method to assess short-term exposures or exposure to pollution sources that change over time.

Estimating Population Characteristics in Proximate Areas

Once the areal extent of potential exposure has been delineated, there are a variety of methods available to estimate the number of people residing in these exposed areas, as well as their socio-demographic characteristics. In order to estimate the number of people exposed to a given hazard and to obtain information about their socio-demographic characteristics, a method of areal selection is necessary to transfer data from the census units to boundaries of the exposed areas since fixed-distance or plume-based buffers are unlikely to match the size and shape of the census units (Maantay et al. 2008). These methods may be broadly classified into point estimation and areal estimation methods depending on the level of spatial aggregation of the socio-demographic data (Chakraborty and Maantay 2011).

If the addresses of all individuals or households in the study area are known, they can be represented on a map with the help of street network reference data and geocoding tools available with GIS software. Once these locations are located on the map, the number of people exposed to a given hazard as well as their socio-demographic characteristics can be estimated with a point-in-polygon overlay by determining the address points located within distance-based or plume-based buffers.

Although the point-in-polygon overlay to estimate exposure is easy to execute, data on socio-demographics are not publicly available and can only be obtained with the help of extensive surveys of all individuals or households in the study area. Since that is typically not feasible due to time and money constrains, researchers have relied mainly on census data which are aggregated at the level of administrative or statistical spatial units.

Several areal selection methods are available to implement a polygon-on-polygon overlay needed to transfer data from the census units to the buffer units. These areal selection techniques are illustrated in Fig. 7 using circular buffers around current and historic TRI facilities in Brooklyn, NY.

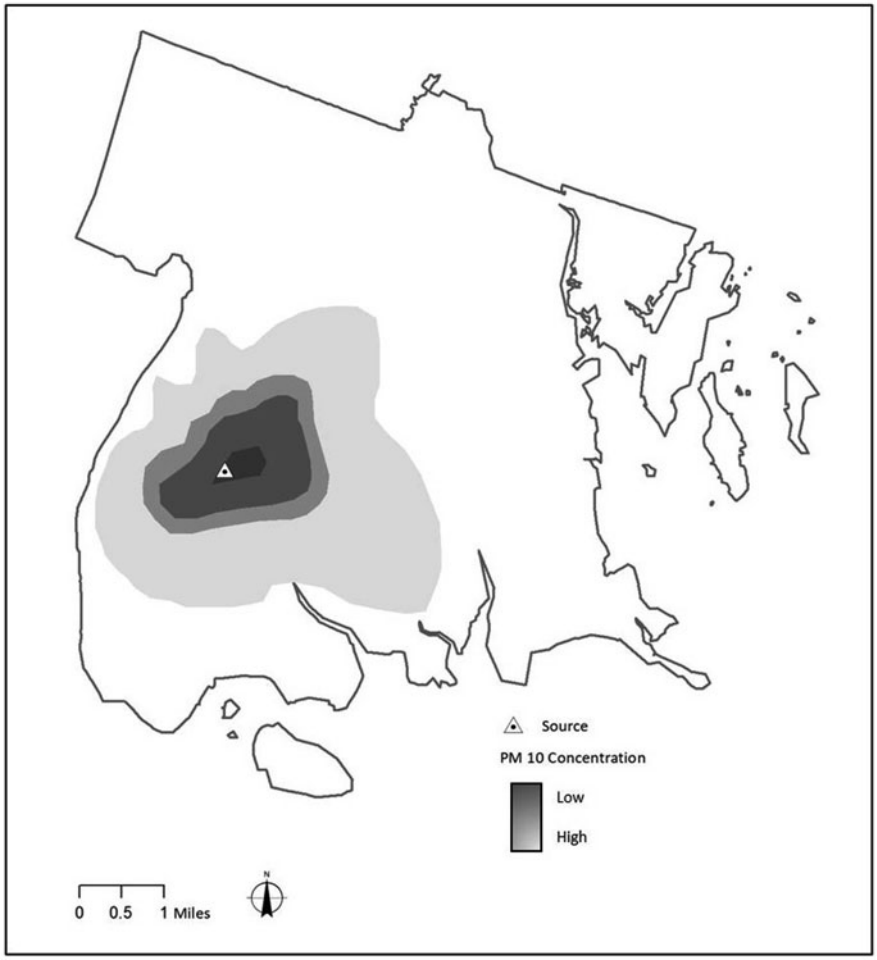


Fig. 5 PM_{10} air dispersion for sample pollution source (TRI facility) modeled with AERMOD in Bronx, NY (Maantay et al. 2009). In this study, AERMOD (American Meteorological Society/Environmental Protection Agency Regulatory Model) was integrated with a GIS to study the association between asthma hospitalizations and air pollution sources in the Bronx. AERMOD is an advanced steady-state plume model aiming to simulate the air dispersion from sources to a distance up to 50 km. AERMOD combines boundary layer theory with an understanding of turbulence and dispersion while also considering the influence of building wakes on plume rise and dispersion to create a continuous pollutant surface for the study area. Thus, this case study required a large amount of data including meteorological data, stationary source data (emissions and release parameters), buildings data (location and height), as well as asthma data and population data. The plume buffers were created based on the AERMOD continuous pollution surface in ArcGIS in order to define the geographic impact extent of the highest pollutant values from each individual major stationary point source. The highest PM_{10} for a source can be found at or close to the location of the source, and the PM_{10} concentration decreases as the air pollutant disperses out from the source. Then, a series of contours were generated based on the PM_{10} value at each point over the study area. The resulting contour buffers are not circular due to meteorological and building effects, which are the major differences from proximity buffers. The values of contours decrease as distance increases away from the source, meaning that for each contour, the impact of the source on ambient air was higher inside the contour than outside. The contours can then be used to identify the difference in asthma hospitalizations inside and outside the buffer. Maantay et al. (2009) showed that asthma hospitalization rates were higher inside of both plume than those outside and that plume buffers captured more people and asthma hospitalizations than the proximity buffers indicating that more people may be exposed than previously calculated based on proximity buffer analysis. For further detail, see Maantay et al. 2009. (*Data source*s: NYC Dept. of Urban Planning 2010; National Climatic Data Center 1999; US EPA 2002 National Emissions Inventory 2002; NYC DOITT 2005. *Figure credit*: Angelika Winner)

The simplest method of areal selection is polygon containment, or selection by polygon intersection. In this technique, all census units that either fall completely within or intersect with a distance or plume-based buffer are selected as shown in Fig. 7a. The population data are then simply aggregated for all the census units that have been selected. Since the polygon containment method does not differentiate between spatial units that are completely within the buffer area or those that are only partially inside the buffer, this method may lead to overestimation of exposed populations if most people live outside of the area intersecting with the buffer (Chakraborty and Maantay 2011). One way to improve upon the polygon containment method is to use a cutoff value to exclude census units with only a small area intersecting with the buffer. Most

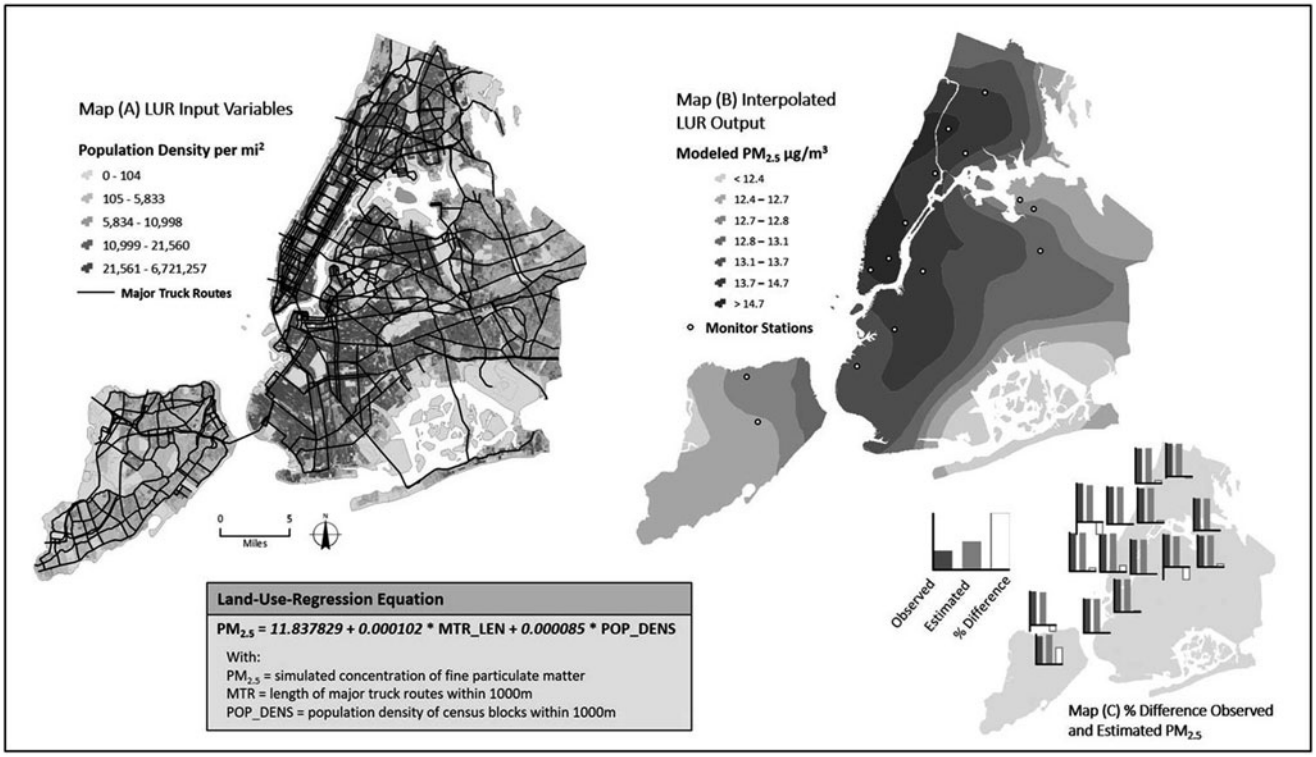


Fig. 6 Modeling of fine particulate matter in New York City with land use regression (LUR). The goal of this case study was to estimate the average annual concentration of fine particulate matter ($PM_{2.5}$) in New York City, based on the limited data available from a small number of air quality monitoring locations. Since the length of major truck routes (MTR) and population density within 1000 m from the monitor site are highly correlated to $PM_{2.5}$ concentrations, these variables were used to estimate $PM_{2.5}$ with a LUR (see Map A). A $PM_{2.5}$ continuous surface was interpolated utilizing Kriging which was the spatial interpolation method with the lowest mean-square error between observed $PM_{2.5}$ concentrations and the modeled one (see Map B). The LUR model was calibrated using EPA air quality data from 15 monitoring sites across New York City from the year 2002. Overall, modeled $PM_{2.5}$ concentrations were lower at a majority of the stations compared to observed concentrations (see Map C). The LUR model was overall significant with an adjusted R^2 of 0.87 indicating that truck routes and land use close to the monitoring stations were able to explain almost 90% of the variation in $PM_{2.5}$ concentrations. The clustering of monitoring stations in and close to Manhattan may lead to an underestimation of the exposure to $PM_{2.5}$ air pollution of outer-borough residents living in areas with high population density and close to truck routes. Additionally, LUR is sensitive to the edge effect. This means that in those areas along the edge of the study area with land-based borders such as Westchester County to the North and Nassau County to the East, the modeled emissions may be underestimated because data on emissions from these neighboring counties are not considered in the LUR model. (*Data sources*: EPA 2002; NYS Dept. of Transportation 2007. *Figure credit*: Angelika Winner)

commonly used is the 50% area containment method, where only census units with more than half of their area within the buffer zone are selected.

Another common method for estimating socio-economic characteristics of the exposed population is known as centroid-based selection. This technique is more exclusive than the polygon containment method since it selects only those census polygons that have their geographic centers or centroids located inside the buffer area, limiting the number of census units that are selected as shown in Fig. 7b. However, both selection methods discussed so far produce effective buffer zones that do not resemble the original distance or plume-based buffer areas. This is because the effective buffer zones are based on the boundaries of census units and not on the boundaries of the original buffer areas. In addition, the centroid containment method may not deliver accurate estimates of the exposed population if the actual place of residence of people inside the selected census units is not concentrated near the centroid (Chakraborty and Maantay 2011).

The most widely used areal estimation method is buffer selection – a method which selects all census units completely within the buffer as well as a fraction of the population from units intersected by the buffer (Chakraborty and Maantay 2011). The advantage with this method is that the effective buffer zone retains the shape and size of the original one as shown in Fig. 7 map (C). In order to determine the fraction of the population from units intersecting with the buffer areas, an areal weighting technique is utilized weighing the population of each census unit by the proportion of its area within the exposed areas as shown in the middle panel of Fig. 8 (Maantay and Maroko 2009). Despite its improvements compared to the polygon and the centroid-based estimation methods, the buffer technique assumes that the pop-

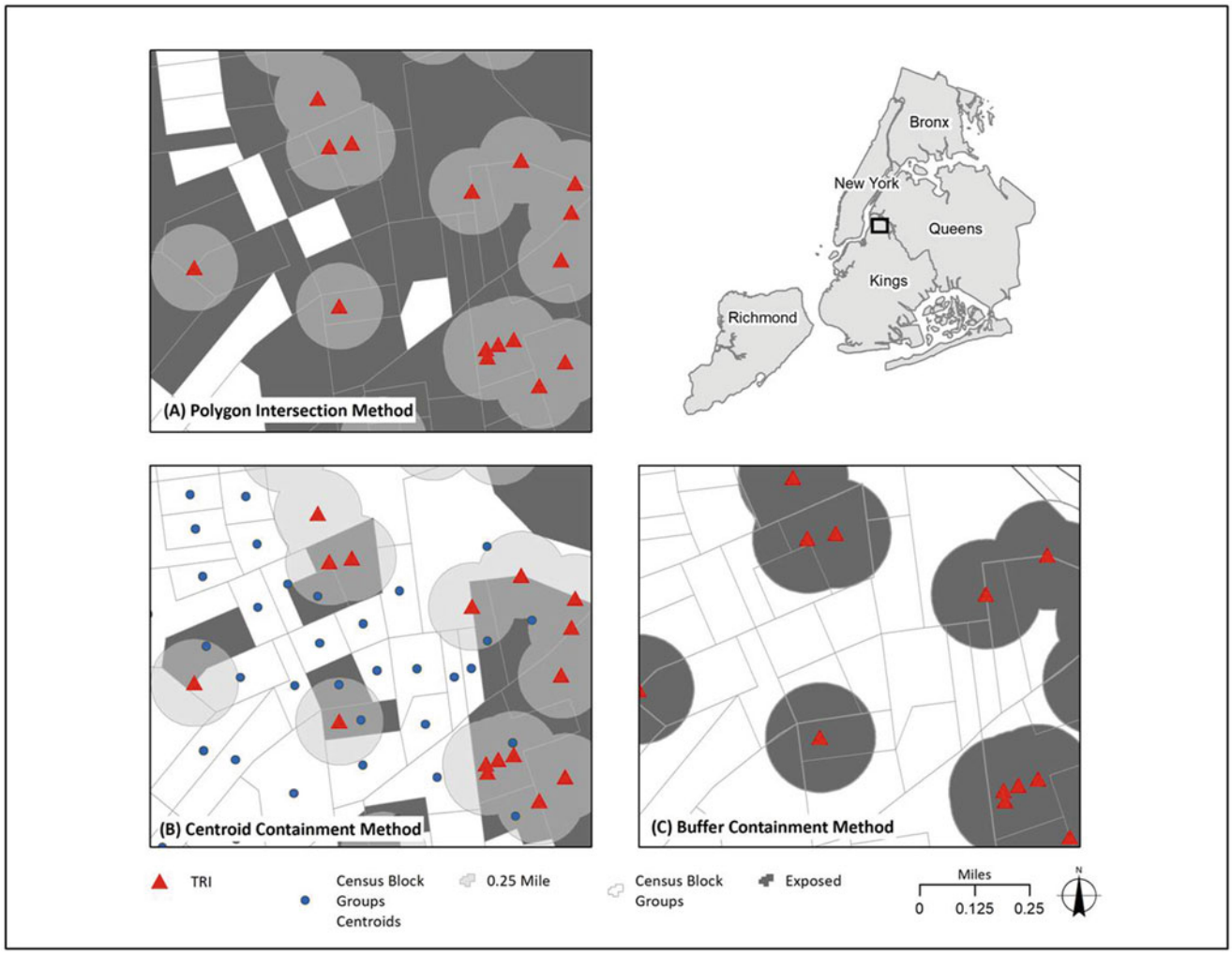


Fig. 7 Comparison of areal interpolation methods to estimate the population exposed to current and historic TRI facilities in North-Brooklyn. Map (**A**) shows the simple polygon intersection method: here all census block groups that intersect a TRI facility are selected regardless of how little of the block group is actually within the buffer, and the resident population of these units is assumed to be exposed. In this case, many of the selected census block groups only share a small fraction of their area with the quarter mile buffers around the TRI facilities leading to an overestimation of the exposed population. Map **B** shows the same data, but now the exposed population has been estimated with the centroid containment method. Here, only those block groups have been selected that have their centroid within the buffer areas. This method may lead to an underestimation of the exposed population if few centroids happen to fall within the buffers. Map **C** estimates the exposed population with the buffer containment method. Here, the exposed areas are limited to the buffer areas, but because the spatial units of the population data (census block groups) do not match the spatial units of the buffer areas, one still needs some form of areal interpolation in order to estimate who is affected. Examples for such methods are areal weighting and dasymetric mapping which are shown in Fig. 6. The buffer method delivers a more realistic estimate of the exposed population when compared to the previous two methods, but it still is only an estimate as it does not take into consideration how the pollutants are dispersed in the environment which depends on meteorological conditions and release parameters. For further details, see Chakraborty and Maantay 2011; and Maantay and Maroko 2017. (*Data sources*: US Census 2010; Bytes of the Big Apple 2014. *Figure credit*: Angelika Winner)

ulation distribution of a census unit and all its characteristics are homogeneous within its boundary which could lead to inaccurate estimates of the exposed population especially in very heterogeneous urban areas (Maantay et al. 2008).

A further refinement of areal weighting is dasymetric mapping which refers to the process of disaggregating spatial data to a finer unit of analysis, using ancillary data such as land cover or land use to help refine locations of population or other phenomena being mapped (Maantay and Maroko 2017). Filtered areal weighting represents a simple case of dasymetric mapping – here limited land use is used to exclude areas with no resident population such as parks, empty lots, water bodies, and open space, and residents are redistributed to all remaining areas (Chakraborty and Maantay 2011). Thus, the ancillary data set acts to mask the census data so that the uninhabited land is left without any population. By excluding sparsely inhabited industrial, commercial, or institutional areas using land cover data from satellite imagery

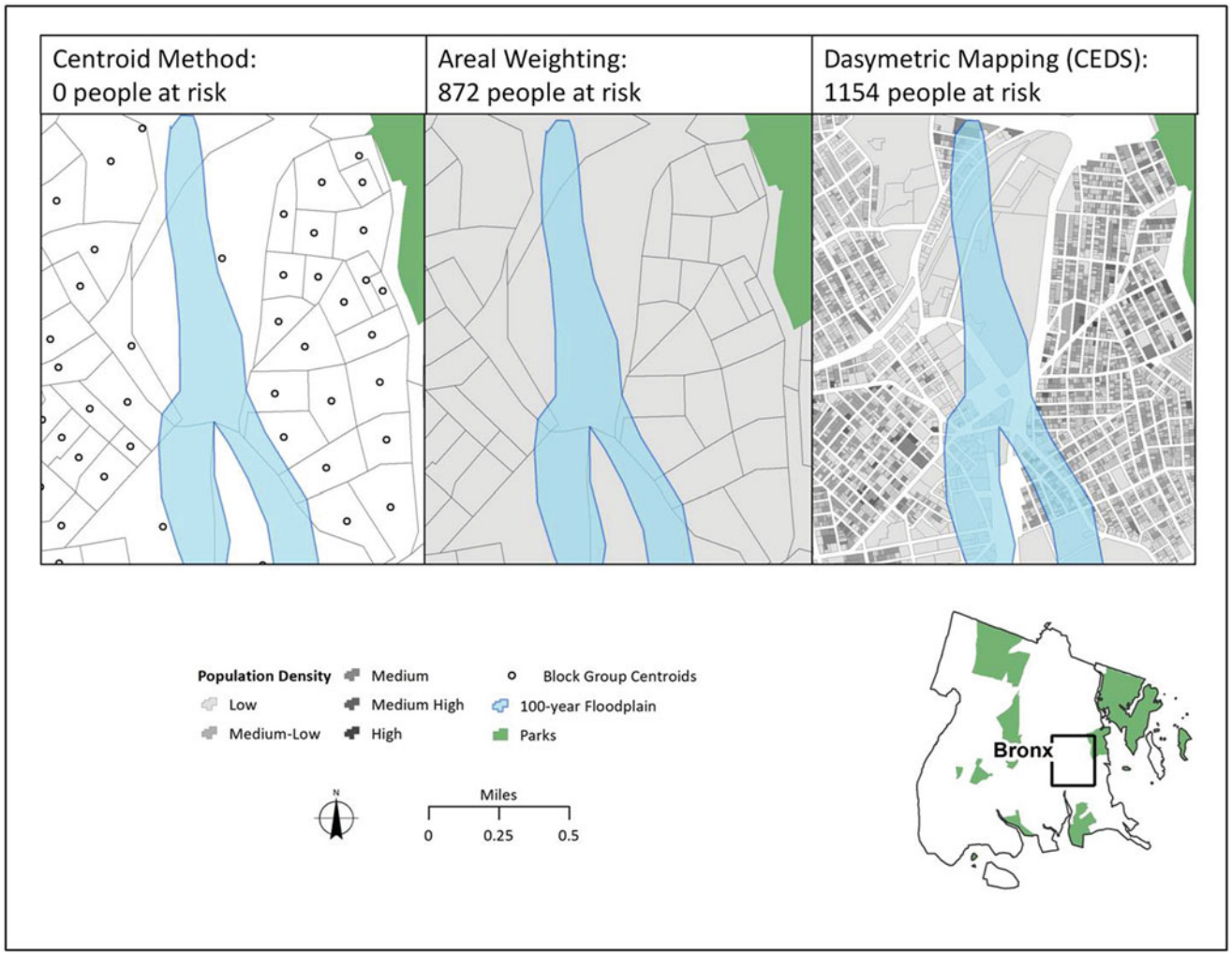


Fig. 8 A comparison of population estimation methods to approximate the number of people at risk of flooding in the Bronx, NY (Maantay and Maroko 2009). The left panel shows the results for the centroid method. Here, only census block groups with their centroids within the 100-year floodplain will be considered exposed. Since none of the centroids happen to fall within the flood plain, zero people will be considered at risk. The middle panel shows results for the areal weighting method. Here, the population per census block group is weighted based on the proportion of area shared with the floodplain – e.g., if 25% of the block group's area falls within the floodplain, 25% of its resident population will be considered at risk. The number of the people considered at risk increases dramatically with this method to 872 people. Finally, the right panel shows results for dasymetric mapping. Here land-use data at the tax lot level are used to get a more accurate estimate of exposed people. In this case, the number of at-risk people increased again to 1154 people highlighting the potential discrepancies between areal interpolation methods. It should be mentioned that the more refined methods do not always yield the largest population numbers. It is possible, for instance, for centroid containment to produce higher numbers that either of the other methods, or areal weighting may do so, depending upon the actual data and configuration of the spatial units. However, dasymetric method will almost always provide more accurate numbers, if not necessarily the largest ones, because the method is based on finer-grained data and more precise locational information. For further details, see Maantay and Maroko 2009; Maantay et al. 2010. (*Data sources*: US Census 2010; NYC Dept. of Urban Planning 2010; FEMA 2007. *Figure credit*: Angelika Winner)

or zoning/land-use maps, one can further refine the simple filtered areal weighting technique.

A special case of dasymetric mapping is the cadastral-based mapping which utilizes tax or property lot data to redistribute people based on the number of residential units or the total residential area per property lot (Maantay and Maroko 2017). Thus, the cadastral-based expert dasymetric system (CEDS) approach assumes that the population is not distributed homogeneously across census units. Property tax lot information on residential units and residential area are used as proxies for the population in each tax lot allowing for a more accurate representation of the population distribution and thus the exposed or at-risk population as shown in the right panel of Fig. 8. However, tax lot data does not contain information on how many people really live in each tax lot or actual square footage per resident. Thus, the residential units or the residential area per tax lot can only ever be a proxy

and we can only estimate the population by disaggregating census data (Maantay et al. 2008).

Spatial Statistics

Once the potentially exposed areas and their populations have been identified, it is often determined whether race/ethnicity or socio-economic status are indicators of disproportionate exposure to environmental hazards. However, spatial data and variables rarely meet the two assumptions of independence and homogeneity that standard statistical tests such as least squares regression or correlation are based upon. The independence assumption of linear regression is violated when it comes to spatial data since proximate observations will exhibit more similarity in value when mapping socio-demographic variables than what can be expected on a random basis. This was expressed best by Tobler (1970) who stated that even though everything is related to everything else, proximate observations are more related than distant ones, which has come to be known as Tobler's first law of geography.

Tobler's observation is more formally known as (positive) spatial autocorrelation and is a fundamental concept in geospatial analysis (Chakraborty and Maantay 2011). Positive spatial autocorrelation means that close-by values will be similar or in other words they will be correlated with each other. The most commonly used measure of spatial autocorrelation is Moran's I which represents a global measure of autocorrelation ranging from −1 to 1 (Chakraborty and Maantay 2011). A Moran's I value approaching zero means that there is no autocorrelation between proximate values, whereas a value approaching 1 or −1 means there is positive autocorrelation (clustering of similar values) or negative autocorrelation (dispersion of similar values). A standardized Z score is available with Moran's I analysis that can be used to measure statistical significance.

In order to account for spatial autocorrelation, spatial regression models have been developed where spatial autocorrelation is considered as an additional variable in the regression equation. The application of spatial regression models has become a standard statistical tool, thanks to the availability and user-friendliness of current GIS and spatial statistical software packages. An example of a spatial regression model is geographically weighted regression (GWR) analysis which accounts for local differences in statistical relationship between dependent and independent variables, also known as spatial non-stationarity (Chakraborty and Maantay 2011). GWR produces a separate regression equation for each spatial unit in the study area allowing for relationships among variables to vary locally (See Fig. 9).

Issues of Equity – Environmental Justice and Health Disparities

Issues of Equity

In this section, we explore several topics having substantial ramifications for urban health. All are aspects of the environmental justice (EJ) problematic: health inequities, relative inequality, segregation, disparities of vulnerability and risk, and differential impacts of social and environmental stressors. Why do health outcomes differ among various populations within a city? Is this due to the environment where people live or is it due primarily to their socio-demographic characteristics? This is a question of "context" versus "composition" and it is a controversial subject in public health today because it has considerable implications for policy, regulation, and the allocation of resources in the urban environment (Gatrell and Elliott 2015). These questions can be explored and addressed, to a large extent, through various spatial analytical methods.

Environmental Justice and Health Disparities

Environmental justice (EJ) and the related issue of health disparities are of particular importance to cities. EJ is broadly defined as the concept that less-affluent populations, communities of color, and other marginalized groups bear a disproportionate burden of environmental "bads" (pollution, urban blight, noxious land uses, traffic congestion, unsafe living conditions, poor housing, and urban incivilities), while conversely they have distinctly deficient access to environmental "goods" (healthy food options, quality health care, and health-promoting amenities such as parks and open spaces) compared with the rest of the population (Bryant 1995; Bullard 1994; Hofrichter 1993; Johnston 1994; Maantay 2019; Pulido 2000; United Church of Christ's Commission for Racial Justice 1987). This phenomenon has been demonstrated over the past several decades through an extensive number of research studies and has been borne out by considerable case study evidence (Boer et al. 1997; Chakraborty and Armstrong 1997; Maantay 2007; Maroko et al. 2009, 2011; Talen 1997; Wolch et al. 2005). Due in part to the disproportionate exposure to environmental burdens and lack of environmental benefits for some communities, there are often extreme health disparities between sectors of the population, and due to the prevalence in urban areas of residential segregation based on class or race/ethnicity, the burdens and benefits are unevenly distributed geographically.

Health outcomes and health conditions vary from place to place around the world by nation, region, state, and city. People in more affluent countries tend to live longer and remain in better health longer than those in less affluent

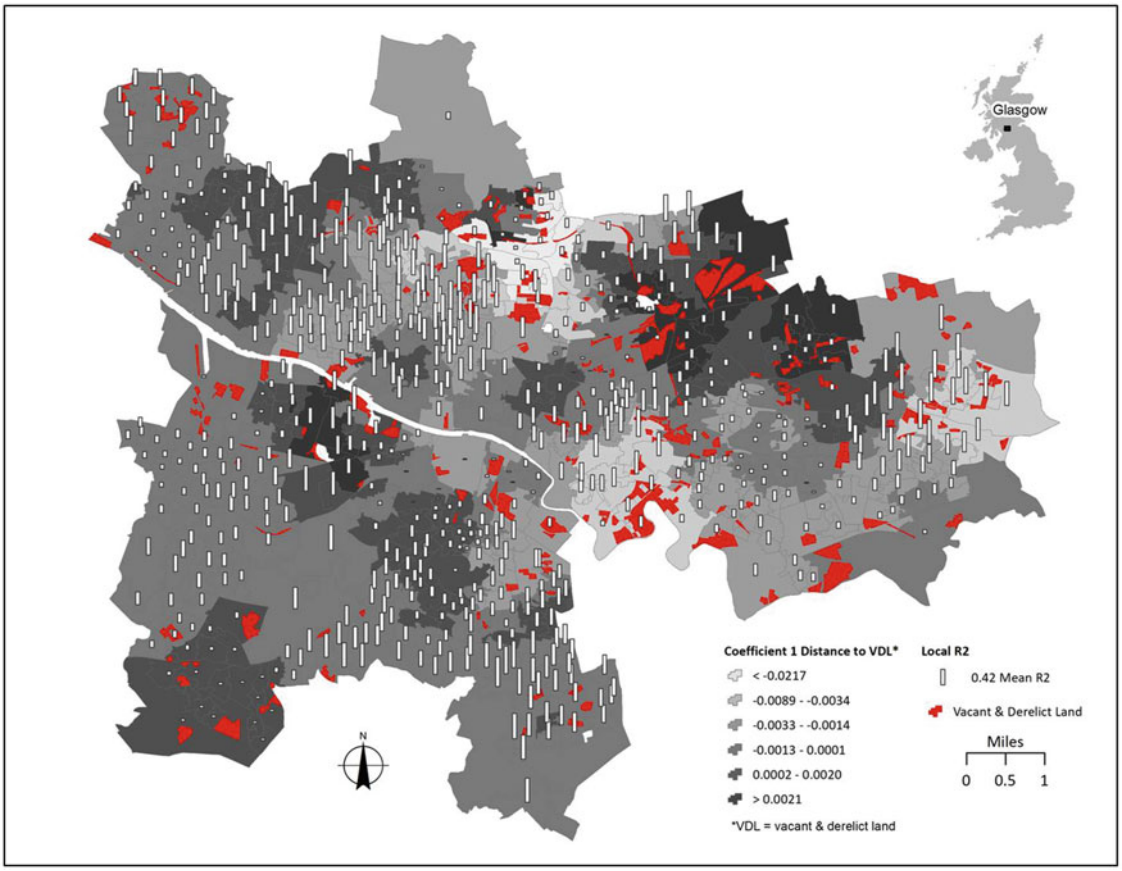


Fig. 9 Results from a geographically weighted regression analysis (GWR) testing the association between distance from VDL and the rate of prescription for drugs related to anxiety, depression, or psychosis (Rx rate) while adjusting for educational attainment and income, and the effects of vacant and derelict land (VDL), as well as deprivation, on mental health. A GWR represents a local ordinary least squares regression which allows the associations to vary over space. Living close to a VDL can lead to exposure to various environmental stressors (including the VDL itself), whereas having a low income and low educational attainment can be thought of as a proxy for exposure to social stressors. This means we are testing for the relative effect of exposure to environmental stressors on mental health while adjusting for social stressors for each data zone. Shown here is the regression coefficient 1 which represents the distance to VDL and the local R-squared value for the association in each data zone. A negative value of coefficient 1 means that if the distance to VDL increases, Rx prescription decreases, or if the distance to VDL decreases, Rx rates increase – this would indicate that exposure to VDLs is connected with higher rates of anxiety, depression, or psychosis; a positive value means that if distance to VDL increases, Rx prescription increases, or if distance to VDL decreases, Rx rates decrease, which would indicate that exposure to VDLs is connected with lower rates of anxiety, depression, or psychosis. For large part of Glasgow, the value of coefficient 1 is negative implying that in these areas, exposure to VDL has a negative effect on mental health. However, there are areas in the Northeast and the Southwest of Glasgow as well as in Central-Glasgow where coefficient 1 is positive implying that in these areas, exposure to VDL has a negative effect on mental health. When we take into consideration the R-squared values, it becomes clear that the areas with a positive coefficient have generally low R-squared values indicating that the association between VDL and mental health is not statistically significant in these areas. It is noteworthy that overall the values for coefficient 1 are close to zero implying a weak association between VDL exposure and mental health. For further details, see Maantay and Maroko 2015. (*Data sources*: Scottish Neighborhood Statistics 2011; Office for National Statistics 2018; Scottish Government Vacant and Derelict Land Report 2012. *Figure credit*: Angelika Winner)

countries, and urban dwellers tend to live longer than their rural counterparts, most likely due to better access to health care (Singh and Siahpush 2014). But these averages can conceal the differences within various geographies. In addition to the *inter*-area comparisons, there is also an important *intra*-area consideration – the variation that occurs within smaller geographic areas, such as within cities, as opposed to between them. EJ and health disparities can be mapped and analyzed to help sort out the relationships, make a strong case for the need to address these problems, and provide possible answers and recommendations to resolve them (See Fig. 10).

The earliest research on environmental health justice was rooted in the idea that less affluent people and communities of color were/are often subjected to disproportionate environmental burdens – things like air pollution, contaminated water, brownfields, hazardous waste sites, nuclear facilities, bus depots, highways, factories that store, use, or emit toxic substances in processing, waste transfer stations, coal-fired

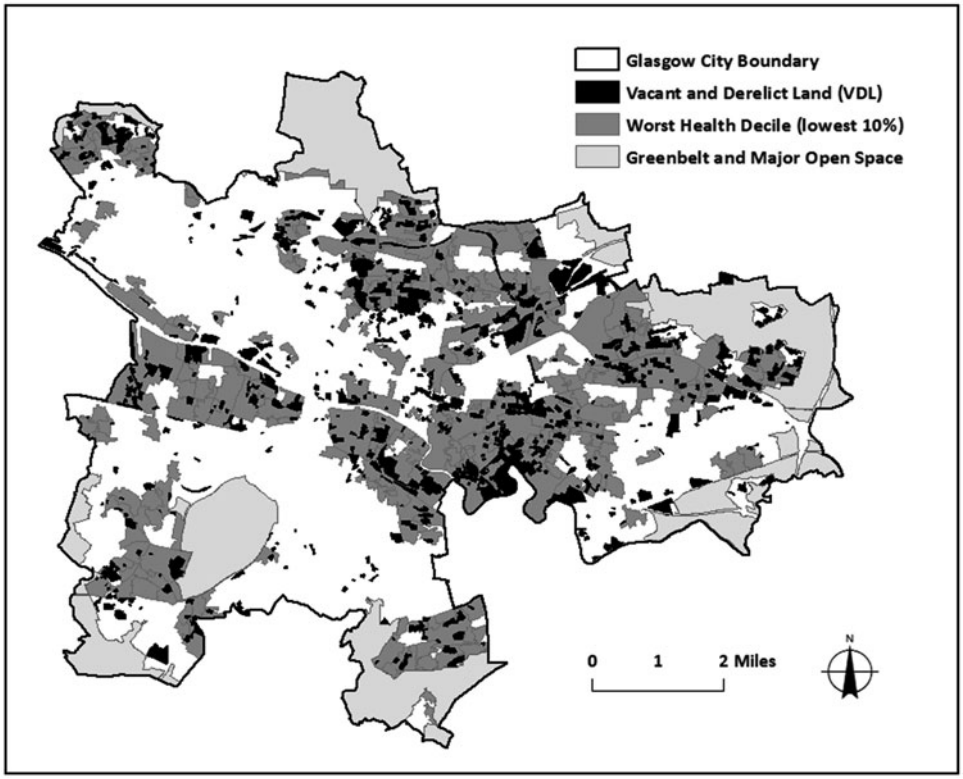


Fig. 10 Map of Glasgow, Scotland, showing a visualization of the spatial relationship between vacant and derelict land (or VDL, which is considered to be an environmental stressor) and the areas where the population has the poorest health, as measured by whether the data zone (DZ) is in the lowest decile (lowest 10%) of health measures, according to the Scottish Index of Multiple Deprivation (health domain). The strong spatial correspondence shown here is not meant as evidence of causality, but to indicate the potential risk associated with living near many of these VDL sites, given the history of industrial land use in Glasgow, and the likelihood that even land formerly used for housing might have originally been land contaminated by industry. Whether or not the actual risk of exposure or causality with health outcomes can be proved, populations in these areas are vulnerable physically and mentally to the adverse effects of being in close proximity to VDL. These proximate populations already suffer from higher than expected rates of many diseases, do not enjoy long life expectancy, and have to bear the stress of poverty and other forms of deprivation, and are therefore more vulnerable in general. Vacant land affects community well-being by overshadowing positive aspects of the community and impacting physical health through possible dermal contact or inhalation of contaminants, injury, the buildup of trash, and attraction of rodents, as well as mental health through anxiety and stigma of living among blight. Although this figure shows only a visualization of the spatial relationship, spatial statistics can be performed to quantify and explore the relationships among variables, for instance analyzing density and concentration and clustering of phenomena. For further details, see Maantay 2013. (*Data sources*: U.K. Ordnance Survey (basemap layers); Vacant and Derelict Land Survey, Scottish Government, 2012 (VDL); Scottish Neighbourhood Statistics, Scottish Government, 2010 (health data). *Figure credit*: Juliana Maantay)

power-generating stations, industrial land uses in general, poor quality, unsafe, or overcrowded housing, and blighted landscapes (Baden and Coursey 2002; Bullard et al. 2007; Fitos and Chakraborty 2010; Grineski and Collins 2008; Grineski et al. 2013; Kay and Katz 2012; Maantay and Maroko 2009; Mohai et al. 2009; Sicotte and Swanson 2007; Taquino et al. 2002; Tiefenbacher and Hagelman 1999) (See Fig. 11). In the previous section, we discussed how exposure to environmental burdens can be mapped and estimated.

Environmental health justice can also be examined from the perspective of access, or lack of access, to the environmental benefits that also affect health status and outcomes (Abercrombie et al. 2008; Boone et al. 2009; Galvez et al. 2008; Kirkpatrick and Tarasuk 2010; Miyake et al. 2010; Moore et al. 2008; Morland and Filomena 2007; Morland et al. 2002; Nicholls 2001; Smoyer-Tomic et al. 2008). This entails using a method that measures access or proximity in order to achieve a valid assessment of true "accessibility" (See Fig. 12).

Relative Inequality

The pattern of inequality that we see in most places reveals quite a lot about the dynamic role of place in health, but often the more significant metric is not the absolute differences between one place and the next, say, in overall mortality rates or rates of certain diseases, but rather the gap between the best and the worst within each area (Sasagawa et al. 2017; Wilkinson and Pickett 2008). This type of inequality is called

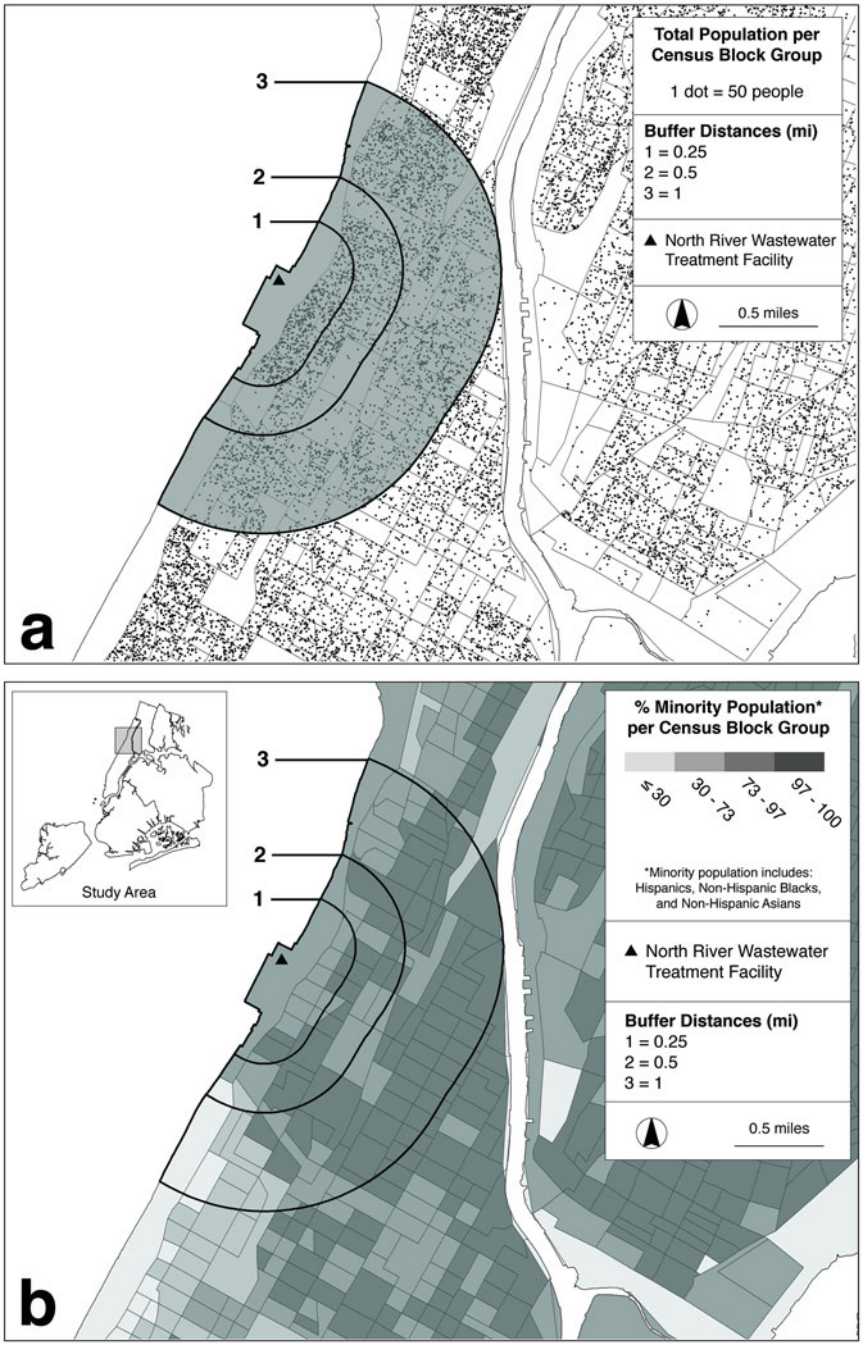


Fig. 11 The siting of the North River Waste Water Treatment Plant (WWTP) in Harlem, New York, was very controversial and contentious because it adversely affects a community comprised largely of a racial minority population. The WWTP had originally been planned for West 72nd Street, a location surrounded by a predominantly white community, but due to intense political pressure from the residents, it was re-envisioned for Harlem, at that time a much less powerful political force in the city. The area within 1 mile of the West 145th Street plant is about 70% minority (mainly African-Americans), with an average household income of $26,000 per year and 34% of its population below the federal poverty line, whereas the population within 1 mile of the original 72nd Street site was 84% white with an average household income of $123,000 per year, and only 8.5% of its people below the poverty line. Air quality problems stemming from the plant have resulted in health impacts felt a mile or more away, including a dramatic increase in respiratory ailments, nausea, headaches from the putrid odors, itchy and watering eyes, and shortness of breath. This example might be a precursor to a more detailed study involving air dispersion modeling, ambient air quality estimations, and the correspondence of air pollution to adverse health outcomes experienced by the community residents. For further details, see Maantay and Maroko 2015. (*Data sources*: US Census, 2010 (socio-demographics). *Figure credit*: Adam Jessup)

the equity divide and is measured by indices such as the Gini coefficient or ratio (See Fig. 13). The Gini coefficient is a measure of statistical dispersion – the inequality of a distribution – whether there is an equal or unequal distribution of values (income, education, etc.). The Atkinson Index and the Generalized Entropy Index also measure inequality but are considered by some to provide a more nuanced understanding of the distribution of inequity (De Maio 2007). These indices are important in analyzing the geography of health because many researchers have come to believe, for instance, that the absolute level of poverty of an area is less important in perceptions of well-being (and perhaps also in *actual* well-being) than the difference (gap) between the wealthiest and the poorest in that area, i.e., the differences in any given area between the "haves" and the "have-nots."

There are several other well-known indices, not necessarily measuring a population's inequality but quantifying a population's well-being, such as the Human Development Index (HDI), the Global Peace Index, Human Poverty Index, the Quality-of-Life Index, the Happiness Quotient (or Gross National Happiness – GNH), various global and national-scale deprivation indices, and other similar indices, such as

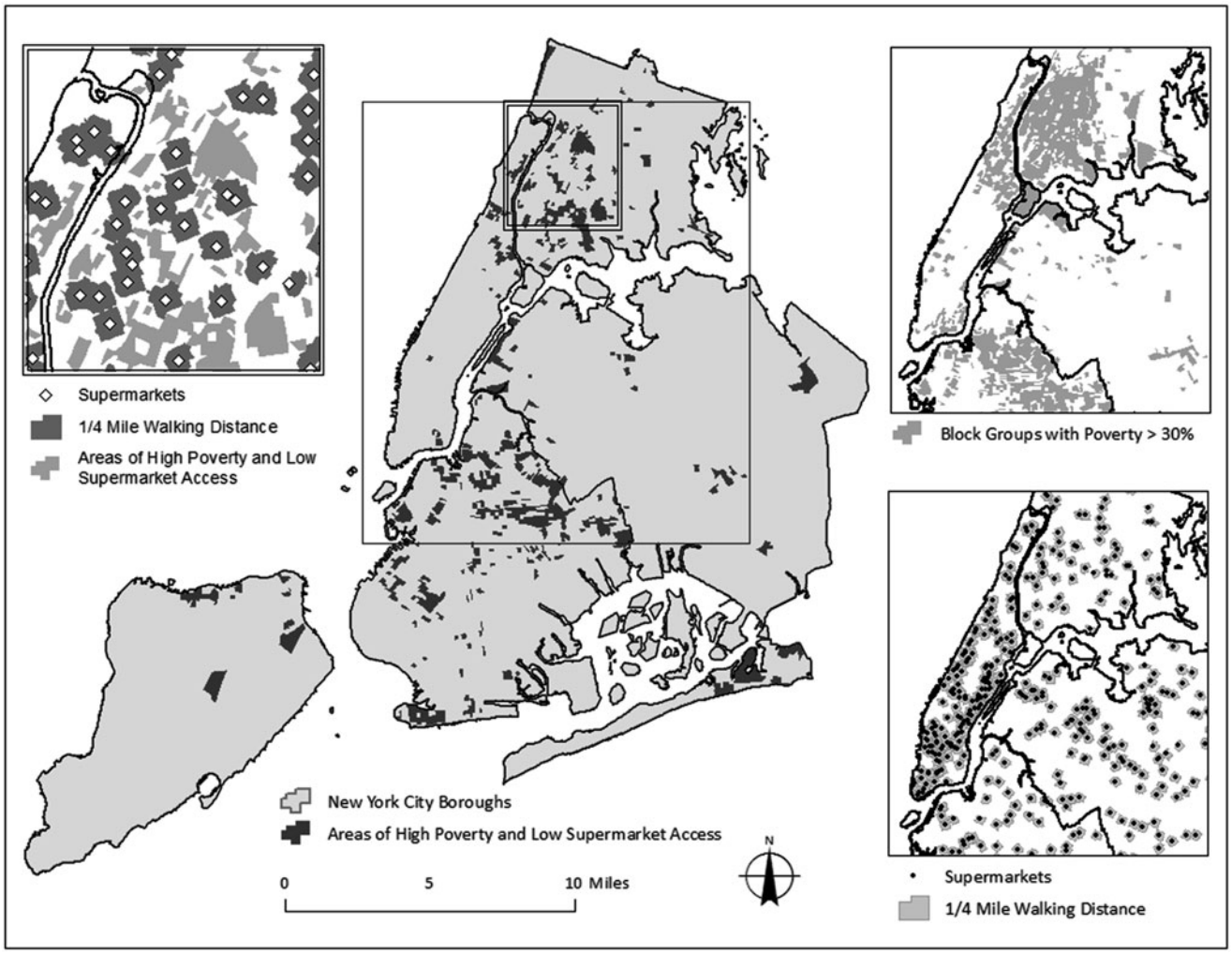


Fig. 12 In addition to proximity to environmentally harmful facilities and land uses, access to environmentally beneficial locations, such as healthy food options (in this case, supermarkets), is also a way to quantify environmental justice. The issue of access to healthy foods has implications for health outcomes such as obesity, diabetes, and cardiovascular disease. Many urban areas are essentially "food deserts," being too far from supermarkets or other healthy fresh food options to be convenient for normal shopping, while fast food restaurants or small shops with poor choices in terms of quality and selection are often the only readily available food sources. By using network analysis, which measures distances along an actual pedestrian street network, we can avoid the over-simplification and inaccuracies of a conventional fixed-distance circular buffer analysis, which measures distance "as the crow flies," without regard for the reality of how people are able to walk in cities to access different amenities. Network analysis can also be used to look at vehicular access, although in the case of New York City, walkability to supermarkets is a better indicator of access. This case study is a classic example of a multi-criteria analysis, using Boolean operators to find sites that meet various criteria, in this case, areas (census block groups) that have high poverty rates (>30%) and are also more than ¼ mile from a supermarket. An average nearest neighbor (ANN) analysis and z-score statistics were also conducted, showing that supermarkets in New York City are highly clustered. Of the 5733 census block groups in NYC, 1456 have more than 30% of their households below the poverty line and 625 block groups have more than 30% of households below the poverty line in addition to being more than ¼ mile walking distance from a supermarket. These areas without walkable access to a supermarket are potentially "food deserts," which may put the under-served populations at risk for diet-related adverse health outcomes, especially those areas with a high percentage of vulnerable households below the poverty line. It is important to keep in mind that there are several types of "access," all of which may play a part in hindering the availability of health-promoting facilities or services. In addition to geographic access, discussed above, we must also consider economic access and cultural access, any combination of which can block true access. (*Data Sources*: US Census, 2010 (socio-demographics); Dunn and Bradstreet, 2001 (supermarkets); NYC Department of City Planning, 2009 (streets). *Figure credit*: Juliana Maantay)

the Scottish Index of Multiple Deprivation that was mentioned in the caption to Fig. 10. Indices are of increasing importance in understanding and depicting the patterns of health and justice, both globally and locally.

Segregation

Race, ethnicity, income/class, and immigrant status are relevant factors in analyzing health disparities, since structural inequalities based on social and cultural characteristics exist

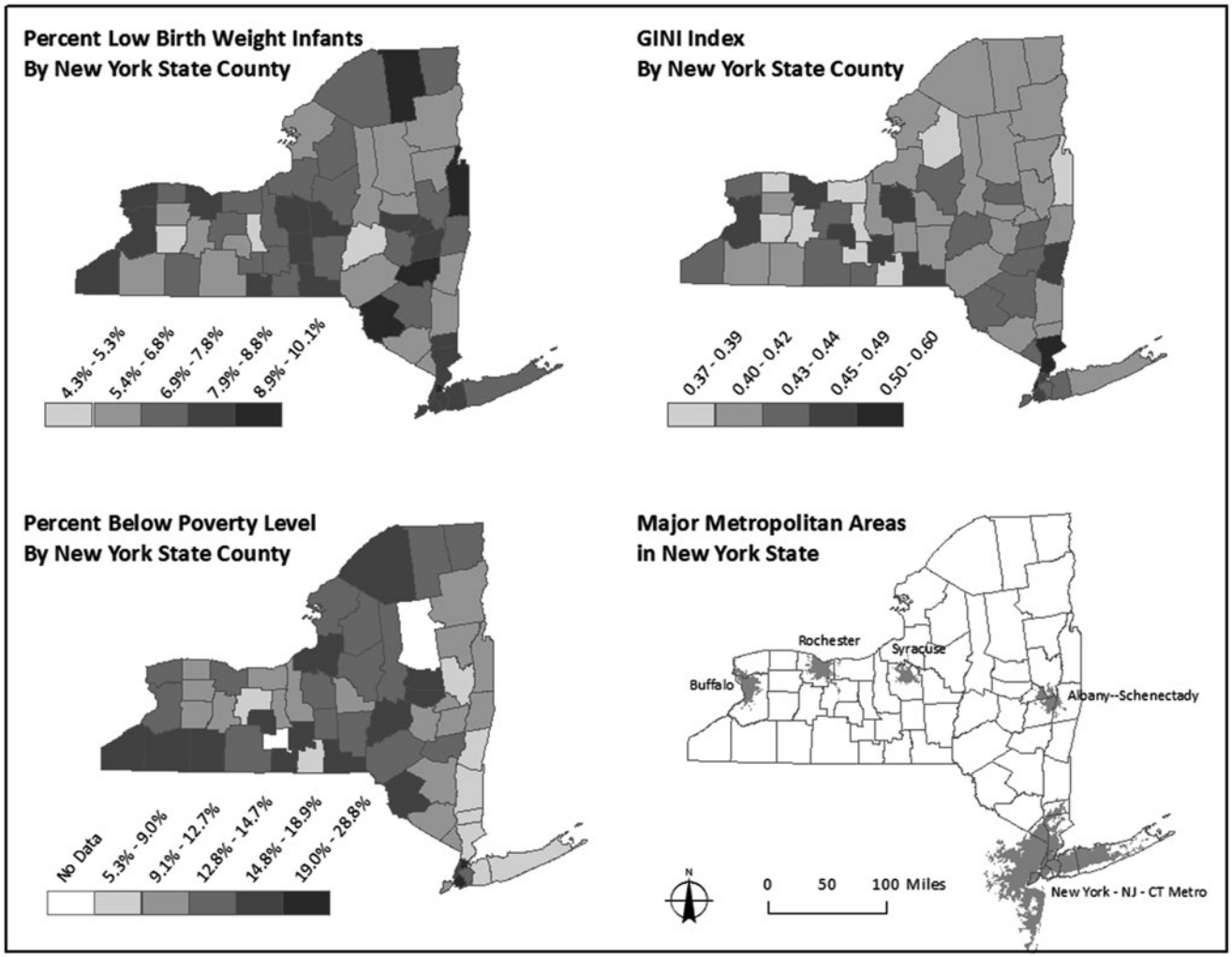


Fig. 13 The relationship between incidence of low birth weight (LBW) infants in New York State and the Gini Index of income inequality and the % of households below the poverty line. Low birth weight (<2500 grams or <5.5 pounds) is the single most important factor affecting neonatal mortality. Infants weighing less than 2500 grams are almost 40 times more likely to die during their first 4 weeks of life than are infants of normal birth weight. Low birth weight infants who survive are at increased risk for health problems ranging from neuro-developmental handicaps to lower respiratory tract ailments (Martinson and Reichman 2016; Paneth 1995). In this example, Gini measures income inequality where a value of "0" indicates complete equality and a value of "1" suggests complete inequality. In looking at New York State at the county level, we can see that the range of Gini values starts at about 0.37 at the low end and goes to 0.60 at the high, keeping in mind that this may mask income inequality for smaller geographic units of aggregation within the county. For comparison, the range of Gini income inequality values for the more affluent countries in the world is between 0.24 and 0.45. In this case study, the Pearson correlation between % low birth weight infants and Gini is 0.261, and the correlation between % LBW and % Below Poverty is 0.242, demonstrating that LBW is slightly more correlated with Gini than with % below poverty, but both are only weakly correlated with LBW. These correlations tend to be stronger in major metropolitan areas in NYS, and visual analysis suggests that urban areas correlate to high poverty rates and highly inequitable distribution of wealth, as well as tending to have higher rates of LBW infants (although some rural counties also have high rates of LBW). Bronx County in New York City appears in the top five worst counties among the 62 counties in the state for all three variables: % LBW, % below poverty, and Gini Index scores. For New York City counties, overall the correlation between % LBW and Gini, or between % LBW and % below poverty, is in the 0.80 range, much higher than that for the state as a whole. (*Data sources*: New York State Dept. of Health Vital Statistics, 2010, (LBW data); US Census, 2010 (Gini and socio-demographic). *Figure credit*: Juliana Maantay)

in our society. Many researchers, advocates, and activists believe that residential segregation by race and/or income class is at the root of many environmental justice and health equities problems (Laveist et al. 2011). When groups of people become "ghettoized" and forced to live in isolation from wider society in a non-integrated manner, because of either economic constraints or discrimination that reduces housing location choice (or both), the result is that often the wider society and those in decision-making positions of power view the ghettoized area as a convenient dumping ground for all sorts of unwanted facilities and land uses – the so-called LULUs, that is, locally unwanted land uses. Residential seg-

regation also frequently results in housing overcrowding (due to a smaller stock of housing available to the stigmatized or segregated group), poor housing conditions (due to landlords taking advantage of the fact that ghettoized populations often have little choice in where to live and therefore will be forced to put up with conditions that other populations would not have to tolerate), unsafe areas (due to high crime, poor policing, high volumes of vehicular traffic, etc.), general governmental dis-investment in community amenities, such as parks, playgrounds, and physical infrastructure, etc. (due to the perception that powerless people and the areas they live in do not have to be well taken care of, and that in times of reduced resources, it is foolish to throw good money after bad to improve areas that are already too far gone to help), and private capital dis-investment (reluctance to site and build full-service supermarkets, recreational and leisure activity venues, healthier eating establishments, etc., in less-affluent neighborhoods), thereby resulting in a downward spiral of neighborhoods going from bad to worse.

Residential segregation can be a direct cause of poor health outcomes and high rates of pre-mature mortality (due to unhealthy housing, environmental burdens, violence, lack of healthy food options, etc.), or it can be an indirect cause of poor health outcomes (due to stress from overcrowding and urban incivilities, lack of positive local amenities, the psycho-social stressors of poverty and marginalization). Residential segregation occurs in all types of locations, but it is particularly prevalent and pernicious in urban areas.

Another relevant factor is class, which in many places is correlated with race. For example: "It's long been known that children in poorer neighborhoods . . . are more likely to be exposed to lead, vehicle exhaust and other pollution. Now, scientists are beginning to suspect that these low-income children aren't just more exposed – they actually may be more biologically susceptible to contaminants, even at low levels. A growing body of research suggests that the chronic stressors of poverty may fundamentally alter the way the body reacts to pollutants, especially in young children. 'It's like having the fight or flight response turned on all the time,' said Harvard epidemiologist Rosalind Wright. Facing financial strain, racial tension and high crime rates can wear down immunity and disrupt hormones, making kids more vulnerable to everything around them, including the lead in their yards, water pipes, or paint, and the car exhaust in their neighborhood" (Konkel 2012).

There are a number of indices used to measure segregation (Massey and Denton 1988; Wong 2005). Two of the commonly used ones are the Isolation Index and the Dissimilarity Index. The *Isolation Index* tells us to what extent individuals are exposed only to other individuals of their racial/ethnic group within their residential area. The *Dissimilarity Index* tells us how evenly distributed a racial/ethnic group is across an area. Both New York City and New Orleans have high percentages of minority populations, and both are known to have a high degree of residential segregation among racial and ethnic groups. Comparing the two cities using two different indices of segregation might reveal some insights as to the structure and meaning of segregation and how it plays out in different locations (See Fig. 14).

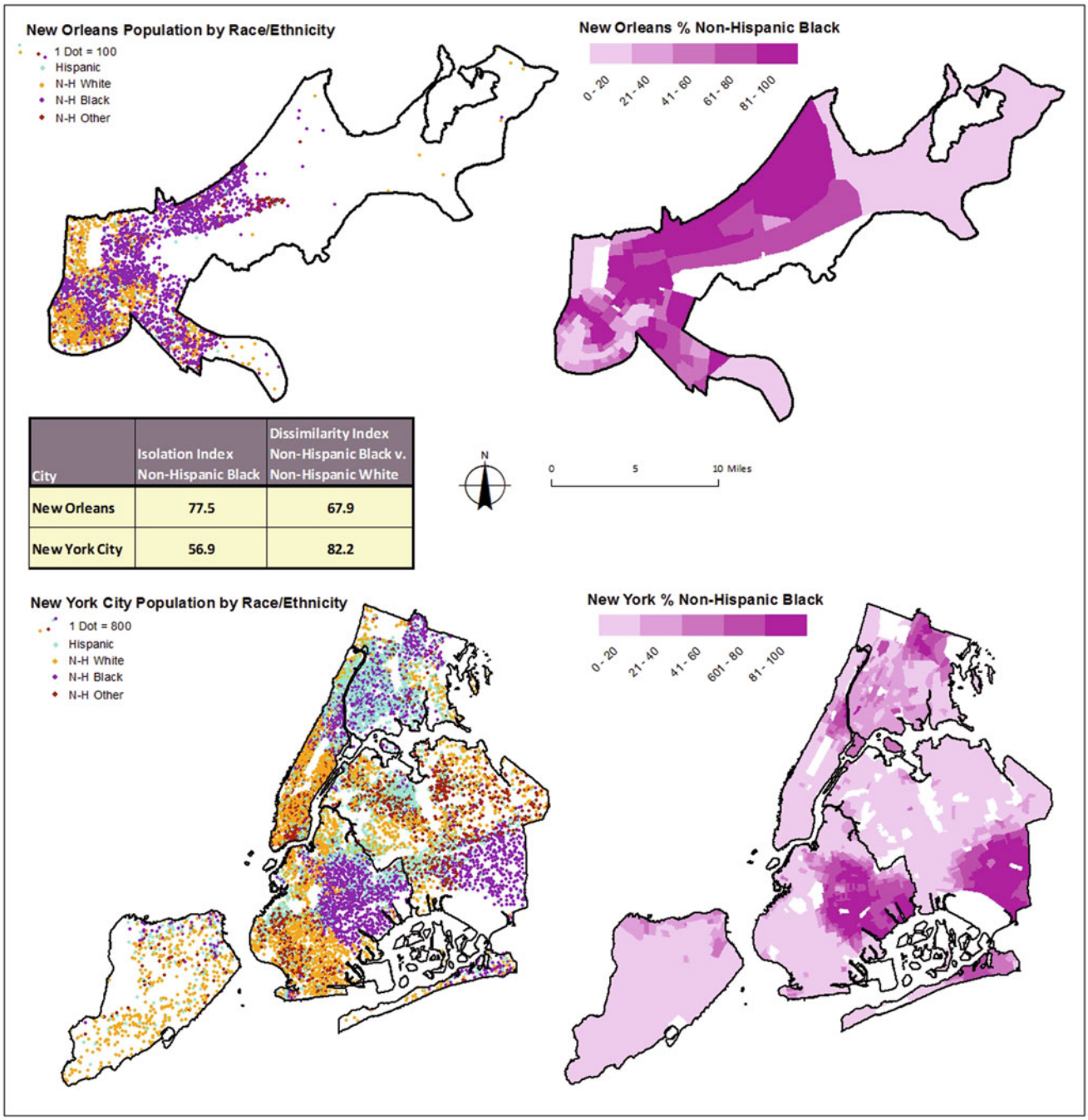
Both cities display a visible trend of segregation between two major groups, non-Hispanic Blacks (NHB) and non-Hispanic Whites (NHW), as evidenced by mapping the percentages of each group by census tract. But mapping percentages of racial/ethnic groups only tells us part of the story. This visual pattern can be further quantified using one or more of the segregation indices.

The Index of Dissimilarity measures the evenness with which two groups are distributed across tracts that make up the larger city, comparing, for instance, how dissimilarly two mutually exclusive demographic groups are dispersed across census tracts. This index ranges from 0 to 100, with higher values indicating greater separation. The index value for NYC between NHB and NHW is quite high, at 82.2, compared to slightly lower but still heavily segregated New Orleans, at 67.9, perhaps indicating that NYC census tracts are much more racially stratified than those in New Orleans and that NHB populations live quite segregated from NHW populations. Another way to think of this is that in NYC, 82% of the NHB population would have to move in order to achieve equal distribution of both NHB and NHW, whereas in New Orleans, only 67.5% of the NHB population would have to move to achieve equal distribution.

The Isolation Index represents the percentage of same population group in the census tract where the average group member lives. The index value also ranges from 0 to 100, with lower values indicating more integration and dispersion of that group within the census tract and higher values indicating greater isolation. For the NHB population of NYC, a value of 56.9 represents moderately high isolation, but significantly lower (less isolated) than the index value of 77.5 for NHB populations in New Orleans.

Social and Environmental Stressors

Social and environmental stressors often exert multiple and interacting influences on health. What we normally think of as exposure to pollution and other environmental stressors, and their consequent adverse health impacts, do not happen in isolation. The combination of toxins and social stressors has synergistic effects that may contribute to the development of, and exacerbate the effects of, diseases such as asthma, obesity, and behavioral disorders (Diez-Roux 2001; Croucher et al. 2007; Downey and Van Willigen 2005; Guite et al. 2006; Maantay 2013; Maantay and Maroko 2015; Maroko



City	Isolation Index Non-Hispanic Black	Dissimilarity Index Non-Hispanic Black v. Non-Hispanic White
New Orleans	77.5	67.9
New York City	56.9	82.2

Fig. 14 *Left*: Dot density maps of New York City and New Orleans showing the distribution of major racial and ethnic populations. *Right*: Choropleth maps of % non-Hispanic Black populations in both cities. These maps offer two ways to view residential segregation, while the segregation indices offer different perspectives. Areas on maps that are white have no or very low populations (<100). These are usually major open spaces, airports, or large industrial areas with no residential population. (*Data sources*: US Census, 2010 (demographic info). *Figure credit*: Juliana Maantay)

et al. 2013). The contribution of psycho-social stressors to negative health outcomes is known as "allostatic loads." In Glasgow, Scotland, for instance, people in high deprivation areas that have high concentrations of vacant and derelict land are more prone to being medicated for depression, anxiety, and psychosis (Maantay and Maroko 2015) (See Fig. 15). Likewise, in a Philadelphia, PA community, a high proportion of vacant and derelict land was found to affect community well-being, physical health, and mental health (Garvin et al. 2013). Similar effects of allostatic loads have been identified by studies on the negative health impacts resulting from psycho-social stressors of the Marcellus Shale

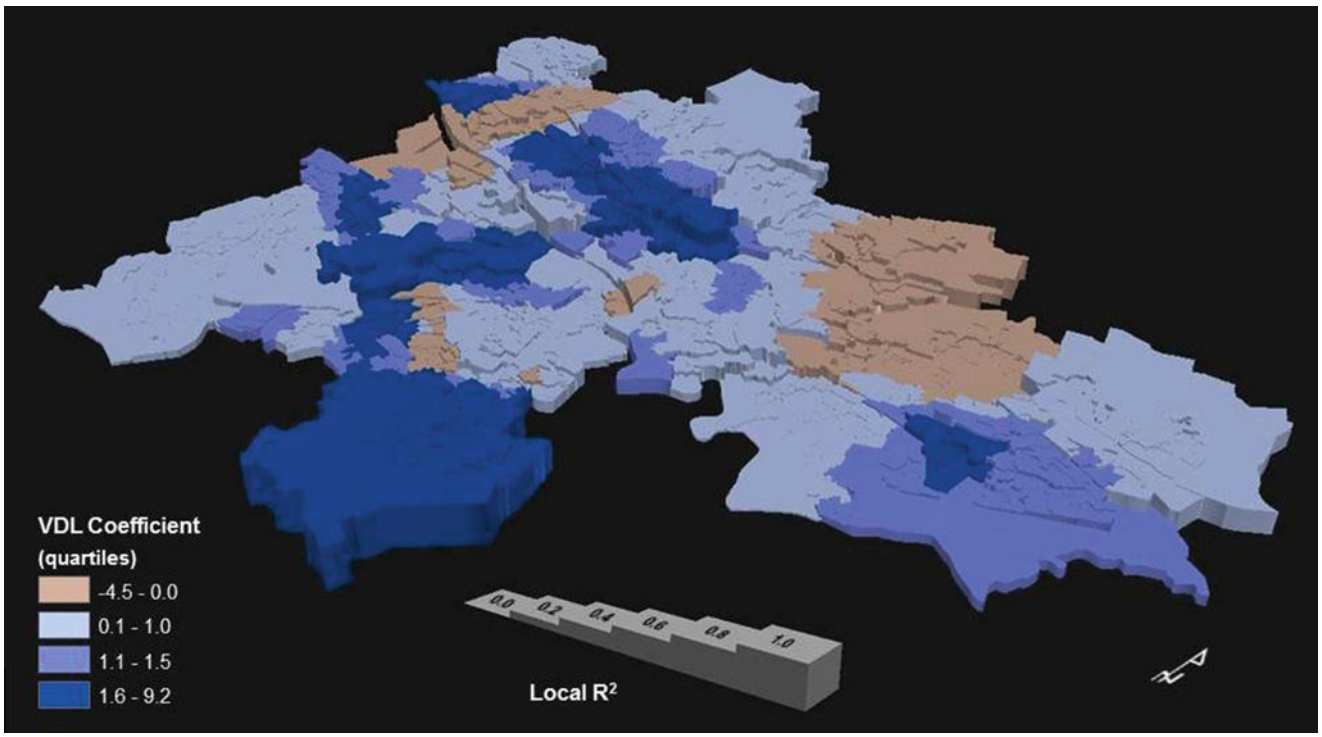


Fig. 15 This map is an example of a bivariate stepped choropleth map, and it depicts the outputs of a geographically weighted regression (GWR), showing the relationship between the density of vacant and derelict land (VDL), a potential environmental stressor, and prescription rates for mental health medications, in Glasgow, Scotland. The local R-squared shows how well the local model performs, whereas the VDL coefficient represents the direction and magnitude of the association between VDL density and mental health prescription rates, while adjusting for selected socio-demographic characteristics. The findings of this study demonstrate an inequity with respect to the distribution of vacant and derelict land, as confirmed by Pearson correlations between VDL density and deprivation. This suggests that many deprived communities are disproportionately burdened with environmental impacts and psycho-social stressors associated with this land use. Regression analyses show a significant positive association between the proportion of the population who were prescribed medication for anxiety, depression, or psychosis and the density of vacant and derelict land while adjusting for socio-demographic characteristics. This indicates that areas with higher VDL densities tend to exhibit higher rates of mental health issues. For further details, see Maantay and Maroko 2015. (*Data sources*: Scotland Census, SIMD 'Health Domain 2010,' 2012; Scottish Government, 2012; Glasgow DRS, 2012. *Figure credit*: Andrew Maroko and Ragnar Thrastarson)

hydraulic fracturing gas extraction process on the local residents (Ferrar et al. 2013). The mitigating effects of positive environmental factors must also be considered, such as access to parks and open space.

Vulnerability and Risk

Finally, vulnerability and resilience in the face of natural and/or man-made disasters and climate change are crucial areas of concern today, especially considering how different communities are able to respond to the impacts of global climate change. These two concepts are related to health inequities because the very same communities experiencing health disparities are often also more affected by disasters and hazards and typically have fewer resources with which to mitigate and recover from the disaster. Resilience can alleviate or ameliorate some of the negative impacts of natural and social hazards. The converse of resilience can be thought of as "vulnerability," and an understanding of how the vulnerability of individuals and populations plays into disaster preparation, planning, mitigation, and recovery is crucial in the quest to deal with wide-spread and unprecedented natural and man-made threats (Blaikie et al. 1994). Climate change is likely going to have one of the largest impacts on public health in this century, with densely settled coastal cities at high risk (Maantay and Becker 2012). This may also have an environmental justice aspect to it, since oftentimes poor communities are located in the parts of the city most susceptible to flooding and other climate change-related hazards (Maantay and Maroko 2009).

Various vulnerability indices have been developed to evaluate this aspect of communities (Cutter et al. 2003; Jones and Andrey 2007; Maantay et al. 2010; Tate 2012), similar in some respects to the residential segregation indices, and GISc has been used extensively in creating and using these indices to understand the potential spatial bias of the impacts of hazards and other burdensome conditions. Many in the

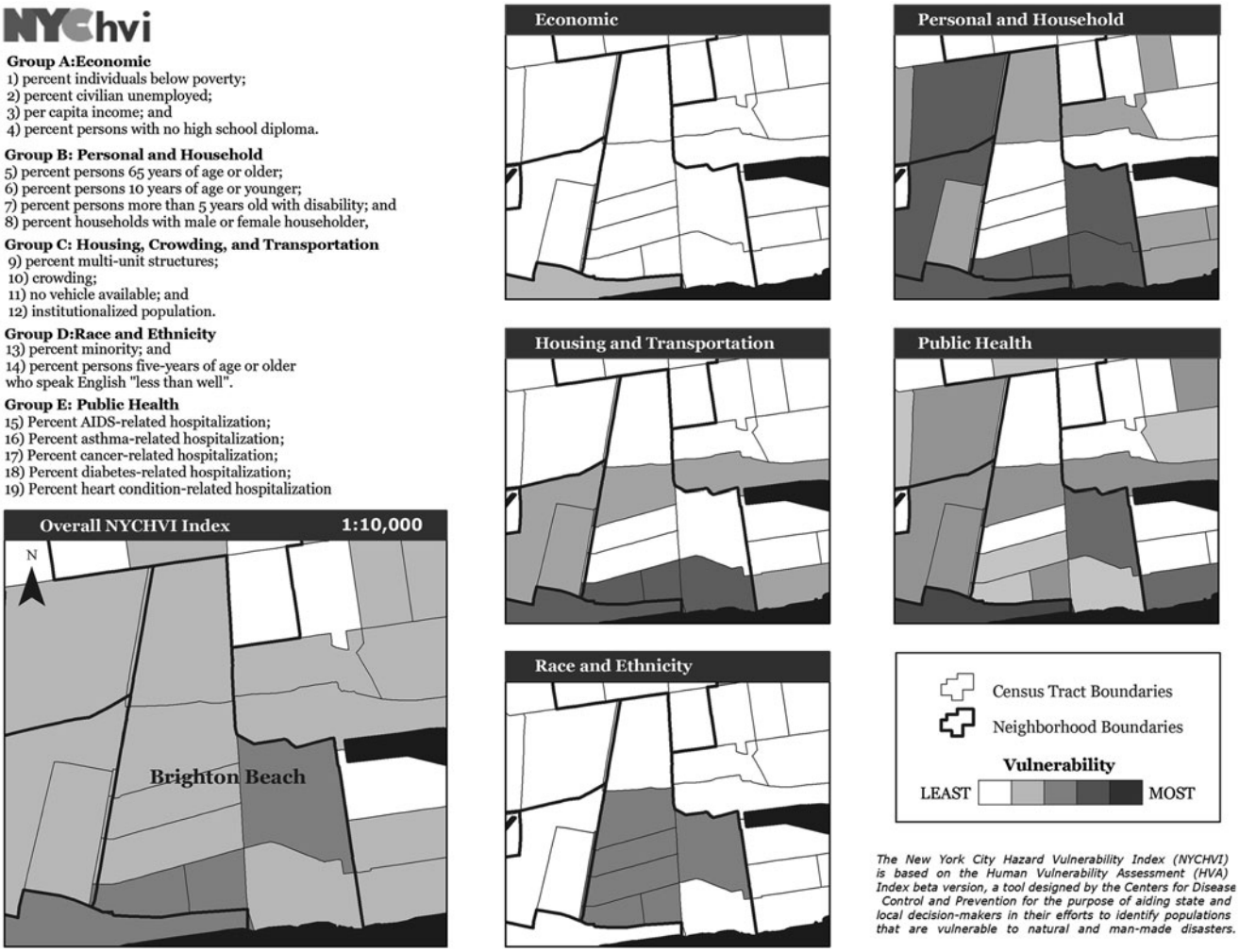


Fig. 16 The New York City Human Vulnerability Index (NYCHVI) is an unweighted index that was developed to assess the vulnerability of the residential population in New York City, on a census tract-by-tract basis. This case study example uses Brighton Beach, Brooklyn, to illustrate how the index works. The NYCHVI index is based on an existing national Center for Disease Control (CDC) index, but tailored to the specific conditions in NYC. It includes 19 indicators, covering socio-economic status, household structure, and disability; minority status and language; housing and transportation; and public health factors, which were selected to represent characteristics that could make people more vulnerable in the event of a flood. The vulnerability score can range from 0 (low vulnerability) to 19 (high vulnerability). Additionally, critical lifeline and special needs facilities and infrastructure were added to the maps to help identify areas likely to require additional assistance in the event of an emergency. The NYCHVI index, as applied to each census tract, and the ancillary locational information can support planning efforts for emergency management, preparation, prevention, mitigation and recovery planning, and encourage planning and response activities that can address the specific needs of the populations involved. Culturally- and linguistically-appropriate materials can be developed to improve disaster preparation by better informing affected communities and to better serve populations in the disaster's aftermath. Having precise knowledge of an at-risk populations' general health conditions (e.g., disabilities and mobility issues) that might complicate evacuation preparation or response activities could be critical. Such mapped info also facilitates targeted aid to areas with high proportion of elderly, disabled, or young. For further details, see Maantay et al. 2010. (*Data sources*: US Census, 2000 (socio-demographic); LotInfo, 2003 (Property lot data). *Figure credit*: Gretchen Culp.)

emergency management profession and scientists researching the prognosis of recovery from natural and man-made disasters agree that being able to assess vulnerability and identify the most vulnerable populations and their locations will help to minimize the impacts of natural and man-made disasters and protect those most at risk (See Figs. 16, 17, and 18).

Clustering and Spatiotemporal Analysis

One of the important concepts related to urban health is the way in which prevalence or incidence of a disease, or a factor related to a disease, manifests in space and time. For instance, if there is a group of census tracts within a city that has statistically higher rates of diabetes when comparedwith

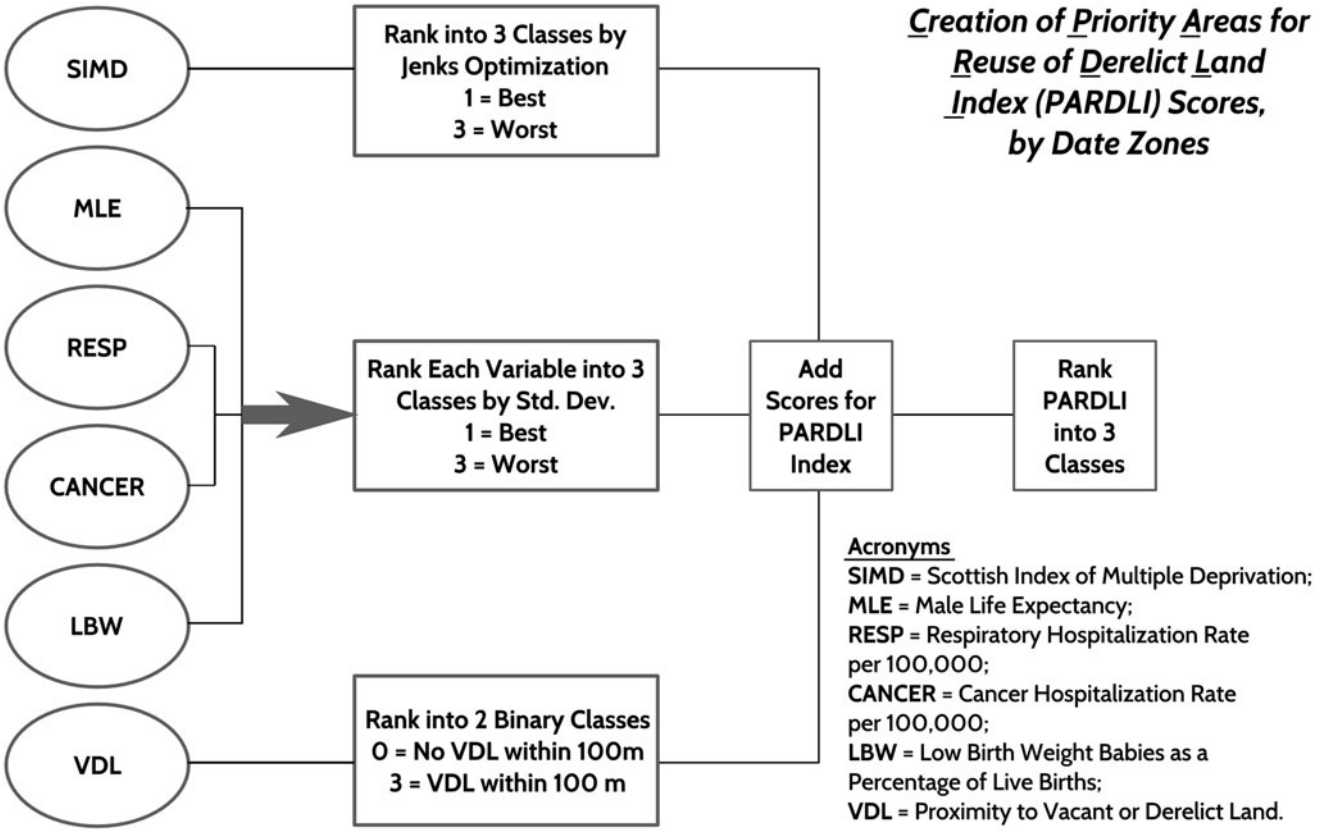


Fig. 17 Diagram showing the construction of the Priority Areas for Reuse of Derelict Lands Index (PARDLI). The PARDLI scores, intended to aid in the decision-making process of resource allocation, combine three aspects of vulnerability: overall high deprivation (need/social vulnerability), adverse health conditions (health vulnerability/need), and proximity to an environmental burden (exposure). The following outlines how the index was constructed: Health variables (LBW, RESP, CANCER, and MLE) were re-classed into three categories: high, medium, and low, by classifying the rates and percentages by standard deviation. Numerical scores of 1, 2, and 3 were used to represent low, medium, and high, respectively. Vacant and derelict land (VDL) was buffered with 100-meter buffer distance, and any data zone (DZ) that intersected one or more of these buffers was considered to be in proximity to a vacant and derelict land site. The 100-meter distance was used rather than a larger impact buffer since it is a more conservative estimation of impact, from the standpoint of both visual blight and quality-of-life factors, as well as any potential impact from contamination. This metric of potential exposure appears in the index as a binary feature (proximate/not proximate to VDL). In the absence of any compelling rationale for weighting one variable higher than the others, the index was created by simple addition of the scores of the six variables. Combined PARDLI scores ranged from a low of 4 (best, lowest priority area) to a high of 18 (worst, highest priority area). The scores were then divided into the three classes of low, medium, and high, as before for the individual variables. For further details, see Maantay 2013. (*Figure credit*: Juliana Maantay)

other areas, we often refer to this as clustering. However, diseases can also be spatially and temporally dynamic, meaning prevalence or incidence can change over time and space based on any number of factors. To capture this dynamism, or disease movement, it is often useful to employ space-time statistics both for quantitative analyses and cartographical purposes. As with many spatio-analytical techniques applied to urban health, outputs of cluster and space-time analyses can be valuable not only for the intrinsic information they provide (e.g., where there are elevated rates of a disease) but also for exploratory analysis and hypothesis generation (e.g., what built environment, natural, and social characteristics are present within a disease cluster that may explain the elevated rates).

Clustered, Dispersed, or Randomly Distributed?

To understand the spatial distribution of a disease or its related factors, we often first test for the nature of the distribution itself. For instance, identifying if a variable is spatially autocorrelated (clustered) in a study area can help us to become familiar with the nature of the data. Global statistics, such as Moran's I, can calculate a simple one-number summary of an entire study area such as a city, thus revealing if the variable of interest is clustered, randomly distributed, or dispersed (Goodchild 1986; Helbich et al. 2012). These data can then be visualized through a variety of cartographic techniques in order to explore their intra-urban spatial distribution, as part of the Exploratory Spatial Data Analysis

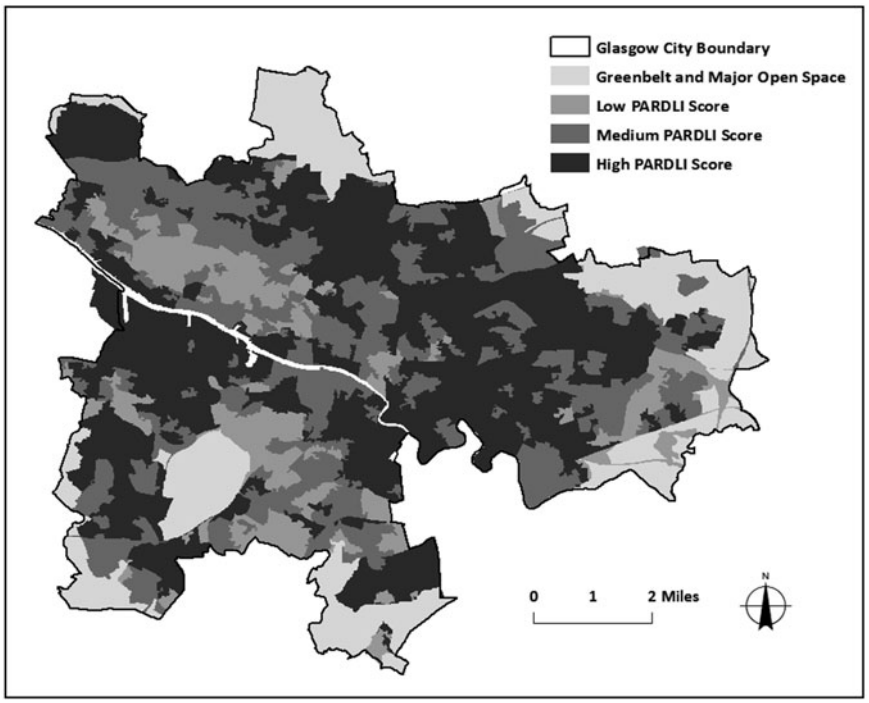


Fig. 18 Output map applying PARDLI index scores to each data zone (DZ) in Glasgow, Scotland. Higher PARDLI scores represent areas with high levels of deprivation, poor health outcomes, and proximity to environmental stressors (VDL). From this index, five case study areas were selected, each comprising three or four data zones, approximating a neighborhood, dispersed around the different sections of the city. For further detail, see Maantay 2013. (*Data sources*: U.K. Ordnance Survey (basemap layers); Vacant and Derelict Land Survey, Scottish Government, 2012 (VDL data); Scottish Index of Multiple Deprivation, General Report and Technical Report, Scottish Government Census, 2009 (SIMD data); Scottish Neighbourhood Statistics, Scottish Government, 2010 (health data). *Figure credit*: Juliana Maantay)

(ESDA) process. For instance, if there are point locations representing sampled observations of urban environmental stressors such as heavy traffic, litter, broken sidewalks, graffiti, or substandard housing, we can use a method such as kernel density estimation (KDE) to create a statistical surface of these stressors (Fig. 19). KDE, even though strictly a cluster detection technique, can smooth data based on either location alone (e.g., X/Y coordinates of where a cardiac death occurred) or based on some other attribute (e.g., number of cardiac deaths within a given distance from a focal point) (Pathak et al. 2011; Yang et al. 2006). This enables more intuitive visualization of areas with unusually high- or low-density values as well as the ability to restructure the data for further analyses while mitigating some common sources of error such as edge effect (when an imposed boundary does not properly consider data outside of that boundary) and the modifiable areal unit problem (MAUP, when the shape, size, or orientation of administrative boundaries can impact the findings of a study due to the distribution of the underlying variable of interest) (Carlos et al. 2010; Fotheringham and Wong 1991).

Hot Spot Analysis

To detect local clusters, various approaches are often used depending on the nature of the data (e.g., points, Fig. 20; or polygons, Fig. 21) and the research question of interest. Commonly used approaches include Nearest Neighbor Hierarchical Cluster (NNH), K-means clustering, Getis-Ord (GI*), Spatial Scan Statistic, and many others (Jacquez 2008). In general, these methods are able to statistically identify spatial groupings of unusually high or low values. Unlike global statistics such as Moran's I, these methods define regions within a study area that may be of concern or interest. For instance, these statistics may be able to identify a group of city blocks with unusually high rates of asthma (i.e., a hot spot), regions that may appear to have unusually low rates of asthma – a protective effect (i.e., cold spot), other areas of interest or outliers (e.g., a neighborhood with very high asthma rates surrounded by neighborhoods with very low rates or vice versa).

The utility of cluster analyses in urban health applications is broad. Aside from identification of the clusters themselves, it can also be used as model inputs for regression models (e.g., a binary variable of "in cluster"/"out cluster"), or as the

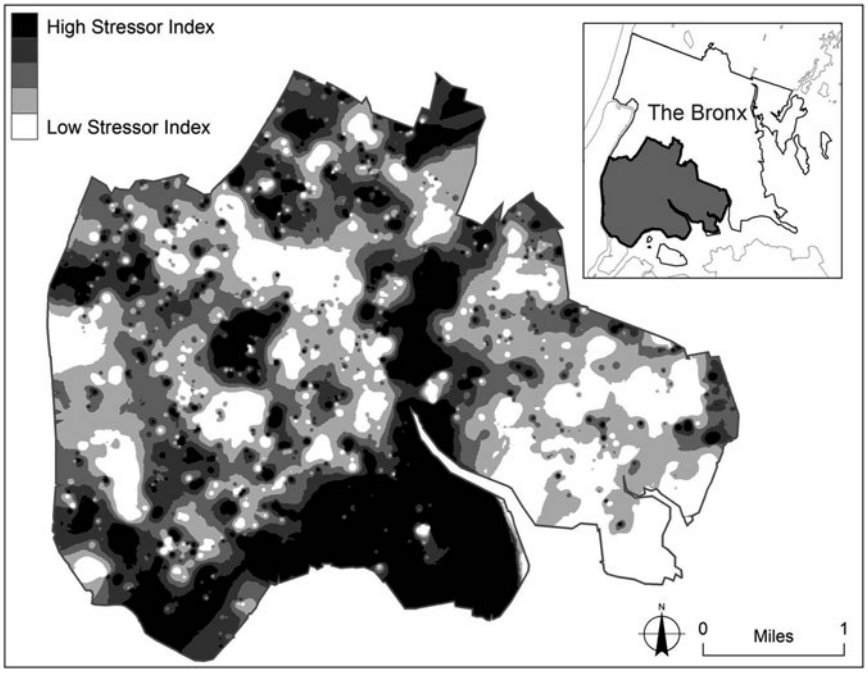


Fig. 19 South Bronx environmental stressors, 2011. This map depicts a kernel density estimation surface of built, natural, and social environmental stressors in the South Bronx, New York City. The KDE surface is constructed from over 2000 sample points that were audited by researchers in the Spring of 2011 using a stratified random sample of blocks in the study area. The audit survey, based on the Project on Human Development in Chicago Neighborhoods (PHDCN 1995), included 40 physical and social variables, both positive and negative, such as street condition, presence of empty alcohol containers, housing condition, presence of street trees, arguing/fighting/hostile adults, open drinking of alcohol, adults stopping to greet one another, loud music, and smoking behavior. Each variable was recoded as an environmental stressor or benefit and collapsed into an index. The KDE surface was then created and ultimately aggregated to census tracts in order to be compared with socio-demographic and economic variables to test for potential environmental injustices (Maroko et al. 2014). For further details, see Maroko et al. 2014. (*Data sources*: Administrative Boundaries – US Bureau of the Census, 2010. *Figure credit*: Andrew Maroko)

first step of a descriptive exploration where characteristics of the built, natural, or social environments within clusters are compared to those outside of the clusters. This enables us to have the ability to explore what may be driving the clustering (e.g., is there a relationship between public transportation hubs in an urban area and crime hot spot? Figure 20, or to examine the impact of the clusters on other variables (how does access to community gardens impact gentrification? Fig. 21).

Space-Time Analysis

The introduction of a temporal component to geographic analyses opens up a tremendous amount of opportunities for urban health research. For instance, at the individual level, mobility can be modeled over both space and time in order to better estimate exposures (e.g., the food environment) that may lead to detrimental health outcomes (e.g., obesity) (Wang and Kwan 2018). The "activity space" of individuals can cartographically describe, and enable analysis of, not only the everyday lives of participants but also their interactions with their environments in a spatiotemporal context (Kwan and Lee 2003). At a population level, it is often diffusion which is examined. Diffusion is a concept which can be applied to any number of phenomena such as cultural diffusion, diffusion of capital, diffusion of innovation, and of course disease diffusion. Conceptually, there are a variety of ways that diffusion often occurs. For instance, relocation diffusion describes when a disease appears in a new region as a result of a population with the disease relocating to a new area and as such brings the illness with them. Contagious diffusion is a result of direct contact between an infected individual to one who is not infected, thus spreading the disease spatially. When the disease is spread along transportation networks, it can be referred to as network diffusion. There is also hierarchical diffusion, by which a disease or other phenomena move from a location to another location not spatially proximate to the first, but one with similar characteristics such as movement from one large city to the next before moving to the next level of cities in size or some other factor. For instance, HIV/AIDS first appeared in the USA in large cities, like New York and San Francisco, having signifi-

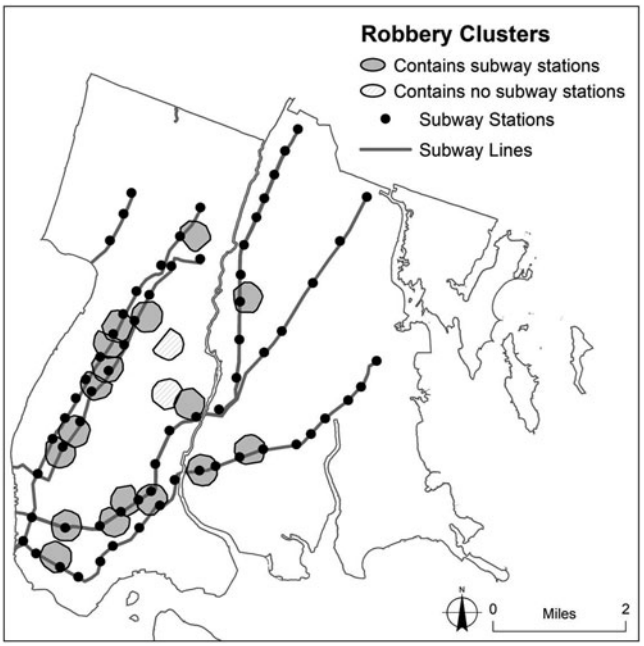


Fig. 20 Robbery clusters and subway stations: Transit stations may create ideal settings for robbery opportunity due to a steady flow of potential offenders and victims (Ceccato and Uittenbogaard 2014; Newton et al. 2015). Crime is often considered a public health issue in urban environments both in a direct sense (e.g., crimes resulting in injury) and in an indirect sense (e.g., decreased quality of life, increased stress). This map shows how nearest neighbor hierarchical clusters (NNH) of robbery point locations from 2009 to 2016, inclusive ($n = 35{,}867$) relate to subway stations in the Bronx, NY. Of the 19 clusters identified, 17 of them (89%) contain at least one subway station. Further analyses revealed more evidence that subway stations can act as crime generators or attractors by demonstrating that robbery rates in areas surrounding stations are significantly lower during complete station closures when compared to robbery rates before or after the closures (Herrmann et al. 2021). (*Data sources*: Transit data – GIS Lab at the Newman Library of Baruch College, CUNY; robbery data – New York City Police Department. *Figure credit*: Andrew Maroko)

cant at-risk populations, before moving to the next order of medium-sized cities, and thence eventually to smaller towns. Naturally, disease movements in urban areas may involve a combination of diffusion types as well, known as mixed diffusion (Cromley and McLafferty 2002). Similar to cluster analyses, spatiotemporal analyses can be used in a multitude of ways including space-time cluster identification (Fig. 23), descriptive statistics of characteristics in the areas of interest, and model inputs for further statistical analyses.

Limitations of Methods and Future Urban Health Geospatial Analytical Research

Despite improvements in exposure assessment techniques, urban health research remains constrained by several limitations. The most obvious limitations of urban health analyses pertain to the datasets themselves that are used. Datasets on health outcomes, due to their confidential and protected nature, are difficult and often impossible to acquire at the resolution needed to reliably establish connections between environmental conditions and socio-demographic characteristics of the resident population (Chakraborty and Maantay 2011). For instance, health outcome data at the state-wide or county-level will not be very useful in investigating impacts of local air pollution concentrations available for specific monitored point locations. The lack of individual-level health and socio-demographic data forces many researchers to conduct ecological studies, using spatially and temporally aggregated data which are based on pre-defined census units, postal codes, or health data collection units, but these boundaries typically do not define the impacted community well. Furthermore, using data aggregated by administrative units also forces researchers to rely on areal interpolation methods based on unrealistic assumption about the homogeneity of population distribution within the unit. Another problem arises from relying almost exclusively on census data and that has to do with the way the census is conducted: people are counted in their place of residency and not in their place of work, school, or other activity spaces. This means that many health studies focus exclusively on night-time exposure and that they assume people are non-mobile and are not exposed to pollution at non-residential locations.

Data accuracy issues also constitute possible limitations in achieving reliable results, and data accuracy takes two main forms: positional accuracy and attribute accuracy, both of which have substantial ramifications on the geospatial analysis of urban health. Positional accuracy, meaning the correct placement of polygon boundaries, and point and line features, is difficult to guarantee, and even small shifts or displacements of features can effectively invalidate an analysis. Positional accuracy can be compromised at many junctures in the data collection and acquisition process – survey measurement errors, image interpretation differences, map projection changes, generalization of spatial data classes, and data overlay operations and other geospatial functions such as clipping and masking.

Attribute accuracy, meaning the textual information contained in the database about spatial features, can also suffer from incompleteness, definitional discrepancies, changes in definitions over time, expanding, combining, or collapsing categories, any of which could possibly stem from incorrect data input and other forms of human error. Depending on the seriousness of the errors and how many types of inaccuracies are present, the deleterious effect on the believability of the analysis will be compounded. For this reason, some assessment of data uncertainty or data reliability is useful to include in the results and limitations section of the research.

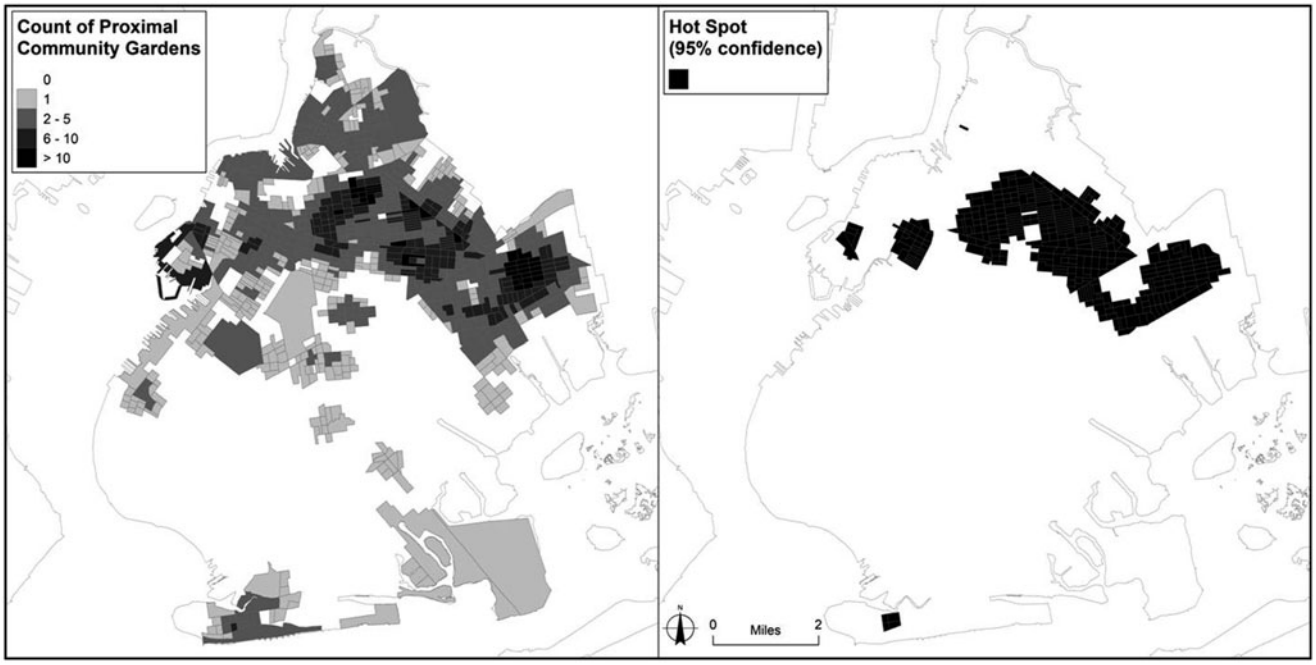


Fig. 21 Community gardens and gentrification: Many post-industrial cities have areas of vacant and derelict land (VDL) which can have negative health and environmental impacts on nearby residents. VDL is often located in poorer neighborhoods, posing a disproportionate risk upon these communities. Repurposing these areas into green spaces and community gardens may mitigate the risk of health and environmental hazards, but they may also result in unintended adverse impacts on the community such as gentrification, which may have its own associations with poor health outcomes (Huynh and Maroko 2014) in addition to displacement of current residents stemming from rises in property values. The maps above represent a portion of a larger study designed to examine the potential impact of proximity to community gardens to gentrification in lower-income areas and the associated implications for environmental justice and health (Maantay and Maroko 2018). *Left*: The map depicts the number of community gardens within a network walking distance of ¼ mile from each census block group in Brooklyn, NY. *Right*: Census block group-based hot spots (polygons) using the GI* statistic showing statistically significant clusters of community garden access. For further details, see Maantay and Maroko 2018. (*Data sources*: Community gardens – Open Accessible Space Information System (OasisNYC) online mapping service; administrative boundaries and demographics – US Bureau of the Census, 2015. *Figure credit*: Andrew Maroko)

Decisions during the development of the research design about the appropriate study area (the scale or geographic extent of the study), the appropriate unit of analysis (the spatial resolution), and the temporal resolution (the time periods to be studied) are almost always dictated by the available health, environmental, and socio-economic data. However, the implications of these decisions can be profound. The differences in the unit of analysis selected, for instance, can have dramatic impacts on the results of the study, with the use of geographic units at a finer resolution leading to more nuanced results than the use of coarser resolution units (Maantay 2007). Additionally, researchers by necessity must use the geographic unit that makes sense in terms of the available data, but these boundaries may have little to do with defining the actual or potential impacted area or populations. This is especially true when using administrative or jurisdictional boundaries such as census tracts, postal codes, or property lots for population and health outcome data, in order to ascertain potential environmental impacts or exposures pertaining to natural phenomena such as watersheds, air pollutant concentrations, vegetative cover, hazard zones, etc., which follow much different types of contours and do not reflect or coincide with the artificial boundaries of the administrative divisions.

The issue of the modifiable areal unit problem (MAUP) has relevance to the selection of appropriate geographic units of analysis (Openshaw 1984). Depending upon where the boundaries are drawn when aggregating data, the geographic pattern and statistical characterization exhibited (by the distribution of health events, noxious facilities, minority populations, and so forth) can change substantially.

Edge effects are another potential impediment to valid research design that must be taken into account. The delineation of study area extent, by necessity, needs to be established based on the research design and the focus area of interest. Study extent is typically defined in such a way that the areas adjacent to the outer boundaries of the area are dealt with as if all relevant data, impacts, and exposures stop at the boundary, which is not the case in reality for most study extents. What occurs on the other side of the outer boundaries of study areas can also have significant

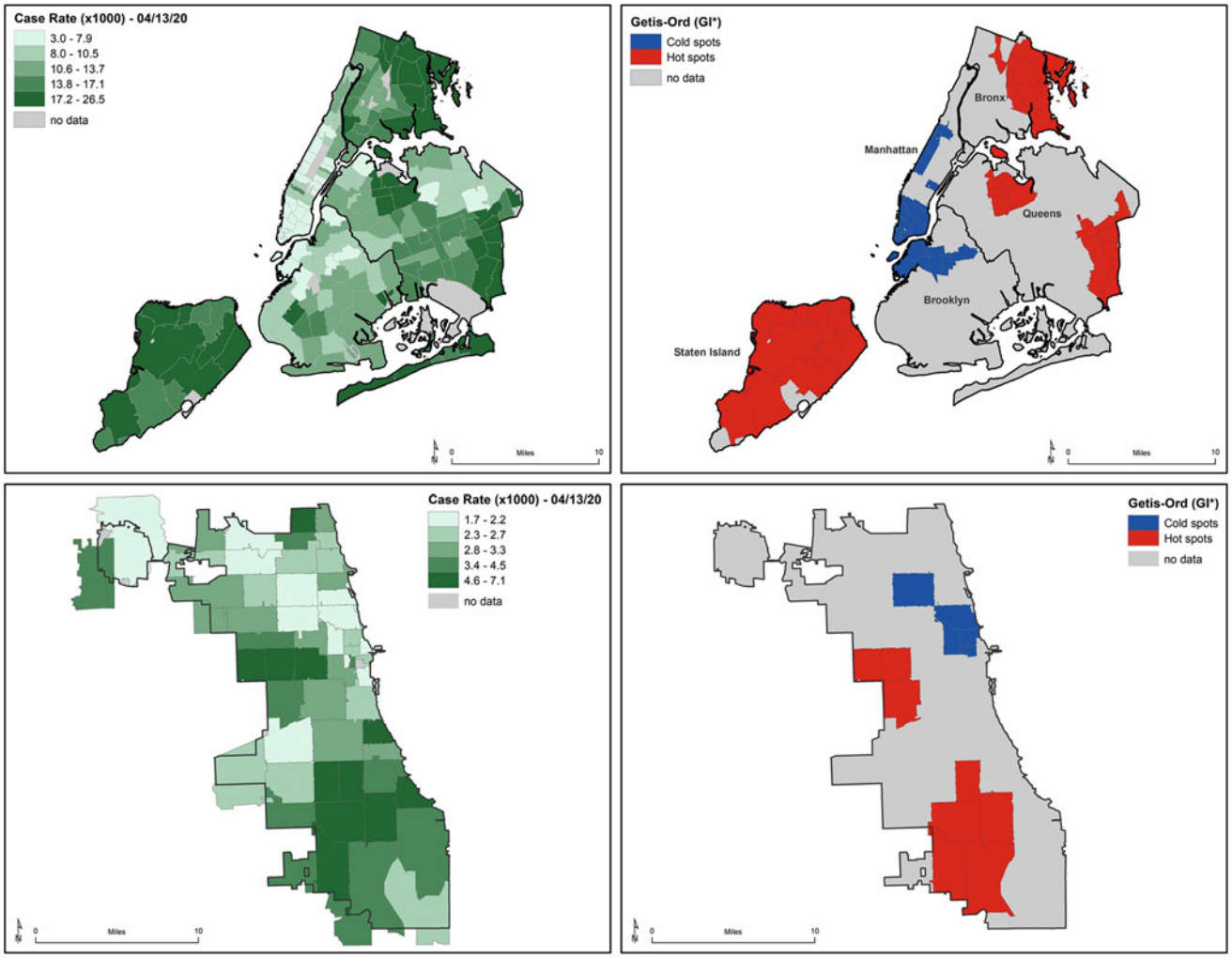


Fig. 22 The ongoing COVID-19 pandemic has disproportionately impacted traditionally vulnerable communities, including well-researched social determinants of health, such as racial and ethnic minorities, migrants, and lower income individuals and families. An early geographic ecological study sought to examine the economic and sociodemographic differences in hot and cold spots of SARS-CoV-2 rates in New York City and Chicago. It demonstrated that in both cities, hot spots (clusters of high rate ZIP code tabulation areas) tended to have lower proportions of college graduates and higher proportions of people of color. Larger households (more people per household), rather than overall population density, was also found to be more strongly associated with hot spots. The two choropleth maps on the left of Fig. 22 show the SARS-CoV-2 case rate (with New York City on the top left and Chicago on the bottom left). The two maps on the right depict the hot and cold spots based on the Getis-Ord (GI*) statistic, showing New York City on the top right and Chicago on the bottom right. For further details, see Maroko et al. 2020. (*Data Sources*: New York City Department of Health and Mental Hygiene's Incident Command System for COVID-19 Response, 2020; Illinois Department of Public Health, 2020; American Community Survey (ACS) 2018 5-year estimates via NHGIS.Org. Figure by: Andrew Maroko)

influence, but this is generally not taken into account in the analysis. Regardless of how the boundary is placed, the values outside the study extent will affect what is inside the study area, even though they are not taken into account as part of the study (Griffith 1983). Boundary problems (edge effects) are especially prevalent in spatial point pattern analysis, such as nearest neighbor statistic. It can be helpful to mitigate edge effects by the use of Epsilon bands (fuzzy boundaries), which include, within some distance threshold, the data on the other side of the outer boundaries in the analysis. Epsilon bands are most often used in accounting for positional inaccuracies of spatial data, but such buffer zones can also be employed so as to include data outside of the study area proper that might impact the analysis within the study area, so that edge effects are eliminated as much as possible from the study area itself. When assessing urban exposure to environmental hazards, several methodological limitations need to be considered. Most exposure assessment techniques are based on the assumption that everyone within a unit that hosts an environmental hazard will be equally impacted; however, it is well documented that pollution does not disperse equally in all directions from a source. Further-

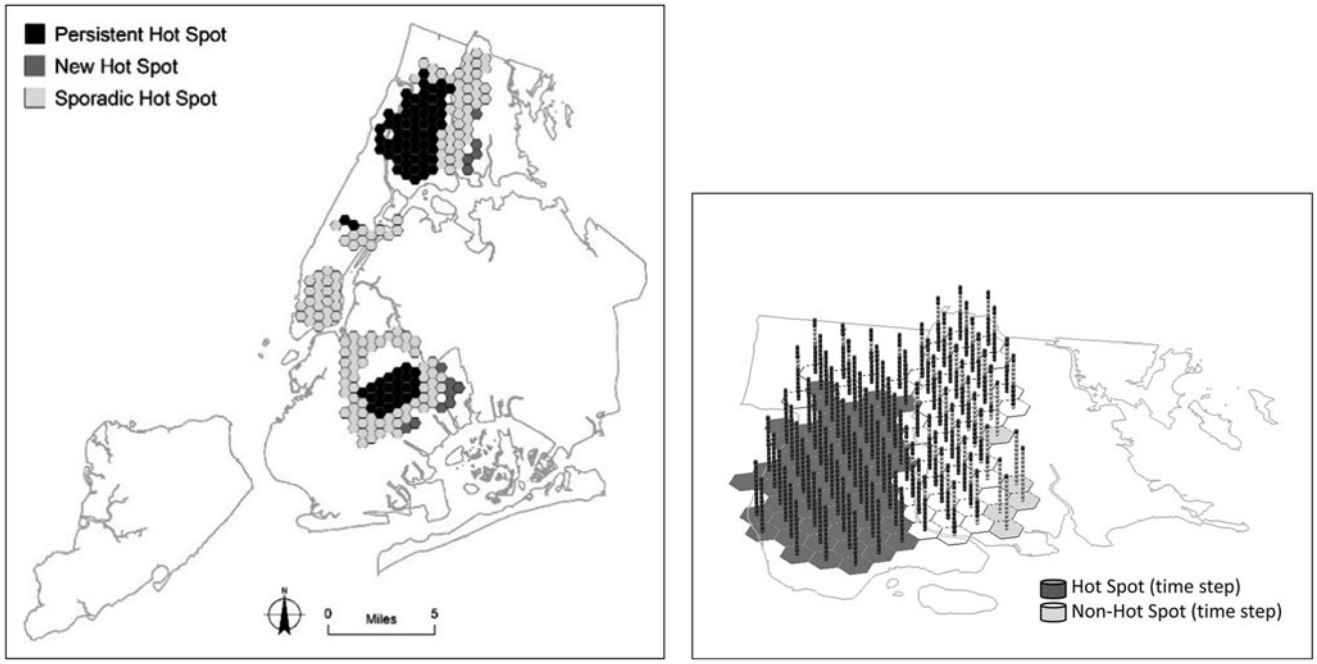


Fig. 23 Emerging hotspot analysis of circulatory system disease hospitalizations in New York City (2013–2016). Space-time clusters of hospitalization counts where the primary diagnosis was related to the circulatory system (ICD9 codes390–459). Using counts rather than rates enables the identification of areas where more resources may be needed (e.g., more hospitalized residents live in a certain neighborhood and as such those areas may be targeted for public health interventions or increased resources such as clinics or primary care providers). *Left*: This simplified map, with hospitalization count data aggregated by hex, identifies areas with hotspots which are (1) persistent (statistically significant hot spot for 90% of the time-step intervals without increasing or decreasing trends), new (statistically significant hot spot for only the final time step), and sporadic (less than 90% of the time steps have significant hot spots and none have been significant cold spots) (ESRI 2018). Note the large areas of persistent hotspots in the South Bronx and Central Brooklyn – regions which tend to be lower income with higher proportions of residents of color compared to other areas of the city. *Right*: An oblique view of the circulatory disease hospitalization count data in the Bronx space-time clusters. Time is represented in the vertical axis with more recent data at the top. Notice how the data in the "persistent" hot spot regions in the south-west of the study area show significant hot spots for each time step (darker symbols), whereas data in the "new" hot spot regions in the south-east only show hot spots in the most recent time step. (*Data sources*: Hospitalization data – Statewide Planning and Research Cooperative System (SPARCS) via Infoshare.org; administrative boundaries– US Bureau of the Census, 2010. *Figure credit*: Andrew Maroko)

more, distance-based exposure assessment methods are very sensitive to the often arbitrarily chosen distance that is used to estimate exposure. Also, some exposure assessment techniques such as the spatial coincidence method and LUR are susceptible to the edge effect. Additionally, many techniques do not take environmental hazard density in consideration. In other words, they do not differentiate between spatial units hosting only one hazard and those hosting several. However, exposure works cumulatively – generally speaking, the more the sources of pollution in a host unit, the higher the total exposure of its residents. The last limitation can be overcome by incorporating data on the types of pollutants and associated emission quantities in the exposure assessment.

One of the limitations of network analysis is the usually subjective choice of the distance metric to be used. Differences in network distance can have enormous effect on the analysis results, but it is generally difficult to be certain that the distance selected is the most correct one. For instance, can we be sure that ¼ mile represents the best estimate of walking distance for a certain area or population? Most prior research has used the ¼ mile distance, because based on experience, approximately five city blocks is considered to be the optimal maximum amount of walking in one direction of a two-way trip, especially for the elderly, children, and those carrying anything heavy, such as groceries. However, as analysts, we may need to discuss caveats with our distance decisions, and the uncertainty that this may bring to the results. Another potential stumbling block in terms of result believability is the accuracy of the network database itself. For instance, if the network analysis is investigating pedestrian access, then the correctness of how the streets and roadways have been designated will have ramifications for how accurately the walking distance access reflects reality. Additionally, realistic pedestrian access routes are influenced by things other than distance and thoroughfare designation, and databases may lack some salient info such as the presence of steep hills or pedestrian-unfriendly thoroughfares – information not likely to be represented in the database but which may

be critical is capturing the true routes and impediments encountered in "walking distances."

There are a number of obvious shortcomings when using indices. The construction of valid indices requires the correct and complete identification of the important factors to comprise the index. Often the important variables are determined by principal component analysis, or simply by expert judgment. If the index is a weighted one, expert judgment must be consulted on the assignment of the correct weights to the correct variables. If we are constructing an unweighted index, we must be able to justify why we believe that all variables are equally important, or at least explain why it is impossible to ascertain and quantify which variables are to be assessed as having greater importance, and that in the absence of being able to quantify weights for the variables, it is more defensible to assign no weights at all.

For either weighted or unweighted indices, some justification for selecting the variables and their importance and relevance to the purpose of the index must be provided. Does the index capture all major aspects of the issue? Is reducing a complex set of variables to essentially a one-number solution obscuring important details and nuance that would allow us to more clearly see relationships if looked at separately? The use of indices always entails some amount of compromise between a higher level of detail possible and the necessity of reducing complex information to a more easily understandable and analyzable set of values.

Cluster and spatiotemporal analyses, like many other GISc-based models, are extremely sensitive to parameterization and data aggregation. This issue is most clearly visible with respect to the conceptualization of the spatial relationship among units of analysis. For instance, when attempting to identify hot spots of asthma hospitalizations by census tract, one must first decide if the clustering will be strictly based on distance (e.g., threshold distance) or on contiguity (e.g., nearest neighbors). If distance-based, then at what scale should the clusters be analyzed? If a large distance value is used, one may identify regional clusters; however, if a smaller value is used, then perhaps local clusters will be identified. Additionally, one must consider options such as the use of Euclidean or network distances, binary distance thresholds or distance decay functions, etc. Both distance- and contiguity-based methods are sensitive to the size of the administrative units of analysis. This is most easily seen when using census units such as tracts which are designed to represent a certain number of residents rather than a fixed area. In more densely populated regions, the census tracts tend to be smaller, whereas less populated areas often have tracts which can be quite large. This can introduce vagaries into the clustering analyses based on the number of samples when using distance-based methods (e.g., densely populated urban areas will have many census tracts within a 2 mile threshold distance, whereas suburban or rural areas will have far fewer tracts in that same radius) as well as contiguity-based models (e.g., 10 nearest neighbors could represent a very small area in an urban region and a very large area in a suburban or rural region). Similar aggregation-related issues are present when examining temporal aspects of spatial data. For instance, if disease data are temporally aggregated by year, it may obfuscate seasonal hotspots. Conversely, if temporal data are too granular, it may lack the power to detect clusters due to small sample sizes in each time step. These issues should be addressed based on the nature of the phenomena being studied, the characteristics of the data being used, and the research question being explored.

Future urban health research will benefit from the acquisition of more robust data, including wherever possible the use of anonymized individual patient-level records, the incorporation of local household surveys to obtain address-level socio-economic and self-reported health and behavioral data, and use of techniques such as cadastral dasymetric mapping to estimate the characteristics of potentially exposed individuals more accurately. The development of accurate but less data-intensive environmental models, less complex models, and ones requiring less data processing would also be helpful, as would the availability of more environmental data in general. Additionally, future urban health studies need to consider the long-term mobility of the potentially exposed population as well as their daily mobility, for example by utilizing GPS technology to incorporate the daily movements of people in the exposure assessment.

References

Abercrombie, L.C., J.F. Sallis, T.L. Conway, L.D. Frank, B.E. Saelens, and J.E. Chapman. 2008. Income and racial disparities in access to public parks and private recreation facilities. *American Journal of Preventive Medicine* 34 (1): 9–15.

Baden, B.M., and D. Coursey. 2002. The locality of waste within the city of Chicago: A demographic, social, and economic analysis. *Resource and Energy Economics* 24: 53–93.

Blaikie, P., T. Cannon, I. Davis, and B. Wisner. 1994. *At risk: Natural hazards, people's vulnerability, and disasters*. London: Routledge.

Boer, J.T., M. Pastor Jr., J.L. Sadd, and L.D. Synder. 1997. Is there environmental racism? The demographics of hazardous waste in Los Angeles County. *Social Science Quarterly* 78 (4): 793–810.

Boone, C.G., G.L. Buckley, J.M. Grove, and C. Sister. 2009. Parks and people: An environmental justice inquiry in Baltimore, Maryland. *Annals of the Association of American Geographers* 99 (4): 767–787.

Bryant, B., ed. 1995. *Environmental justice: Issues, policies, and solutions*. Washington, DC: Island Press.

Bullard, R., ed. 1994. *Unequal protection: Environmental justice and communities of color*. San Francisco: Sierra Club Books.

Bullard, R., P. Mohai, R. Saha, and B. Wright. 2007. *Toxic waste and race at twenty, 1987–2007: A report prepared for the United Church of Christ, Justice & Witness Ministries*. Cleveland: United Church of Christ. Available at: https://www.nrdc.org/sites/default/files/toxic-wastes-and-race-at-twenty-1987-2007.pdf.

Carlos, H.A., X. Shi, J. Sargent, S. Tanski, and E.M. Berke. 2010. Density estimation and adaptive bandwidths: A primer for public

health practitioners. *International Journal of Health Geographics* 9 (1): 39.
Ceccato, V., and A.C. Uittenbogaard. 2014. Space-time dynamics of crime in transport nodes. *Annals of the Association of American Geographers* 104 (1): 131–150.
Chakraborty, J. 2021. Convergence of COVID-19 and chronic air pollution risks: Racial/ethnic and socioeconomic inequities in the U.S. *Environmental Research* 193 (2): 110586.
Chakraborty, J., and M.P. Armstrong. 1997. Exploring the use of buffer analysis for the identification of impacted areas in environmental equity assessment. *Cartography and Geographic Information Systems* 24 (3): 145–157.
Chakraborty, J., and J.A. Maantay. 2011. Proximity analysis for exposure assessment in environmental health justice research. In *Geospatial analysis of environmental health*, ed. J.A. Maantay and S. McLafferty, 111–138. Heidelberg London New York: Springer.
Cromley, E.K., and S.L. McLafferty. 2002. *GIS and public health*. Guilford Publications.
Croucher, K., L. Myers, R. Jones, A. Ellaway, and S. Beck. 2007. *Health and the physical characteristics of urban neighbourhoods: A critical review*. Glasgow: Glasgow Centre for Population Health.
Cutter, S.L., J. Boruff, and W.L. Shirley. 2003. Social vulnerability to environmental hazards. *Social Science Quarterly* 84 (2): 24–261.
De Maio, F.G. 2007. Income inequality measures. *Journal of Epidemiology and Community Health* 61 (10): 849–852.
Diez-Roux, A.V. 2001. Investigating neighborhood and area effects on health. *American Journal of Public Health* 91: 1783–1789.
Downey, L., and M. Van Willigen. 2005. Environmental stressors: The mental health impacts of living near industrial activity. *Journal of Health and Social Behavior* 46 (3): 289–305.
ESRI. 2018. How emerging hot spot analysis works. Retrieved from http://desktop.arcgis.com/en/arcmap/latest/tools/space-time-pattern-mining-toolbox/learnmoreemerging.htm
Ferrar, K.J., J. Kriesky, C.L. Christen, L.P. Marshall, S.L. Malone, R.K. Shama, D.R. Michanowicz, and B.D. Goldstein. 2013. Assessment and longitudinal analysis of health impacts and stressors perceived to result from unconventional shale gas development in the Marcellus shale region. *International Journal of Occupational and Environmental Health* 19 (2): 104–112.
Fitos, E., and J. Chakraborty. 2010. Race, class, and wastewater pollution. In *Spatial and environmental injustice in an American metropolis: A study of Tampa Bay, Florida*, ed. J. Chakraborty and M.M. Bosman, 145–151. Amherst: Cambria Press.
Fotheringham, A.S., and D.W.S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space* 23 (7): 1025–1044. https://doi.org/10.1068/a231025.
Freudenberg, N., S. Saegert, and S. Klitzman, eds. 2009. *Urban health and society: Interdisciplinary approaches to research and practice*. San Francisco: Jossey Bass.
Galvez, M.P., K. Morland, C. Raines, J. Kobil, E. Siskind, J. Godbold, and B. Brenner. 2008. Race and food store availability in an inner-city neighbourhood. *Public Health and Nutrition* 11: 624–631.
Garvin, E., C. Branas, S. Keddem, J. Sellman, and C. Cannuscio. 2013. More than just an eyesore: Local insights and solutions on vacant land and urban health. *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 90 (3): 412–426.
Gatrell, A.C., and S.J. Elliott. 2015. *Geographies of health*. 3rd ed. Hoboken: Wiley-Blackwell.
Goodchild, M.F. 1986. *Spatial autocorrelation*. Norwich: Geo Abstracts University of East Anglia.
Griffith, D. 1983. The boundary value problem in spatial statistics. *Journal of Regional Science* 23: 377–387.
Grineski, S., and T. Collins. 2008. Exploring environmental injustice in the global South: Maquiladoras in Ciudad Juárez. *Population and Environment* 29: 247–270.
Grineski, S.E., T.W. Collins, J. Chakraborty, and Y. McDonald. 2013. Environmental health injustice: Exposure to air toxics and children's respiratory hospital admissions in El Paso, Texas. *The Professional Geographer* 65 (1): 31–46.
Guite, H.F., C. Clark, and G. Ackrill. 2006. The impact of the physical and urban environment on mental well-being. *Public Health* 120: 1117–1126.
Helbich, M., M. Leitner, and N.D. Kapusta. 2012. Geospatial examination of lithium in drinking water and suicide mortality. *International Journal of Health Geographics* 11: 19–19.
Hennig, F., D. Sugiri, L. Tzivian, K. Fuks, S. Moebus, K. Jöckel, D. Vienneau, T.A.J. Kuhlbusch, K. de Hoogh, M. Memmesheimer, H. Jakobs, U. Quass, and B. Hoffmann. 2016. Comparison of land-use regression modeling with dispersion and chemistry transport modeling to assign air pollution concentrations within the Ruhr area. *Atmosphere* 7 (48), 1–9.
Herrmann, C.R., A.R. Maroko, and T.A. Taniguchi. 2021. Subway station closures and robbery hot spots in New York City—Understanding mobility factors and crime reduction. *European Journal on Criminal Policy and Research*. https://doi.org/10.1007/s10610-020-09476-x.
Hofrichter, R., ed. 1993. *Toxic struggles: The theory and practice of environmental justice*. Philadelphia: New Society Publishers.
Huynh, M., and A.R. Maroko. 2014. Gentrification and preterm birth in New York City, 2008-2010. *Journal of Urban Health* 91 (1): 211–220.
Jacquez, G. 2008. Spatial cluster analysis. In *The handbook of geographic information science*, ed. S. Fotheringham and J. Wilson, 395–416. Malden: Blackwell Publishing.
Johnson, S. 2007. *The Ghost Map: The story of London's most terrifying epidemic–and how it changed science, cities, and the modern world*. New York: Riverhead Books.
Johnston, B.R., ed. 1994. *Who pays the price? The sociocultural context of environmental crisis*. Washington, DC: Island Press.
Jones, B., and J. Andrey. 2007. Vulnerability index construction: Methodological choices and their influence on identifying vulnerable neighborhoods. *International Journal of Emergency Management* 4 (2): 269–295.
Kay, J., and C. Katz. 2012. Pollution, poverty, people of color: Living with industry. *Scientific American online*. https://www.scientificamerican.com/article/pollution-poverty-people-color-living-industry/
Kirkpatrick, S.I., and V. Tarasuk. 2010. Assessing the relevance of neighbourhood characteristics to the household food security of low-income Toronto families. *Public Health and Nutrition* 13: 1139–1148.
Kleinschmidt, I., M. Bagayoko, G.P.Y. Clark, M. Craig, and D.L. Sueur. 2000. A spatial statistical approach to malaria mapping. *International Journal of Epidemiology* 29: 355–361.
Koch, T. 2005. *Cartographies of disease: Maps, mapping, and medicine*. Redlands: ESRI Press.
Konkel, L. 2012. Pollution, poverty, and people of color: Children at risk. *Scientific American online*. https://www.scientificamerican.com/article/children-at-risk-pollution-poverty/
Kwan, M.-P., and J. Lee. 2003. *Geovisualization of human activity patterns using 3D GIS: A time-geographic approach*. In Michael F. Goodchild and Donald G. Janelle. Eds. 2003. *Spatially integrated social science: Examples in best practice, chapter 3*. Oxford, UK: Oxford University Press.
Laveist, T.A., D. Gaskin, and A.J. Trujillo. 2011. *The effects of racial segregation on health inequalities*. Baltimore: Joint Center for Political and Economic Studies, Hopkins Center for Health Disparities Solutions, Johns Hopkins Bloomberg School of Public Health.
London School of Hygiene and Tropical Medicine (LSHTM) Archives. 2013. Exhibit: The legacy of John Snow: Epidemiology yesterday,

today, and tomorrow. http://johnsnowbicentenary.lshtm.ac.uk/http://www.johnsnow.org.uk/mobile-map/gallery-map.html. Accessed 21 Jan 2019.

Maantay, J.A. 2002. Mapping environmental injustices: Pitfalls and potential of Geographic Information Systems in assessing environmental health and equity. *Environmental Health Perspectives* 110 (S.2): 161–171.

———. 2007. Asthma and air pollution in the Bronx: Methodological and data considerations in using GIS for environmental justice and health research. *Health and Place* 13: 32–56.

———. 2013. The collapse of place: Derelict land, deprivation, and health inequality in Glasgow, Scotland. *Cities and the Environment* 6 (1).

———. 2019. Environmental justice and fairness. In *Environmental planning and sustainability*, ed. R. Cowell, S. Davoudi, I. White, and H. Blanco. Oxford, UK: Taylor and Francis.

Maantay, J.A., and S. Becker. 2012. The health impacts of global climate change: A geographic perspective. *Applied Geography* 33: 1–4.

Maantay, J.A., and A.R. Maroko. 2009. Mapping urban risk: Flood hazards, race, and environmental justice in New York. *Applied Geography* 29 (1): 111–124.

———. 2015. 'At-risk' places: Inequities in the distribution of environmental stressors and prescription rates of mental health medications in Glasgow, Scotland. *Environmental Research Letters* 10: 1–16.

———. 2017. Assessing population at risk: Areal interpolation and dasymetric mapping. In *Handbook of environmental justice*, ed. G. Walker, R. Holifield, and J. Chakraborty. Abingdon: Routledge.

———. 2018. Brownfields to Greenfields: Environmental justice versus environmental gentrification. *International Journal of Environmental Research and Public Health* 15 (10): 2233.

Maantay, J.A., and S. McLafferty. 2011. Environmental health and geospatial analysis: An overview. In *Geospatial analysis of environmental health*, ed. J.A. Maantay and S. McLafferty, 3–38. Heidelberg London New York: Springer.

Maantay, J.A., A.R. Maroko, and H. Porter-Morgan. 2008. Research note—A new method for mapping population and understanding the spatial dynamics of disease in urban areas: Asthma in the Bronx, New York. *Urban Geography* 29 (7): 724–738.

Maantay, J.A., J. Tu, and A.R. Maroko. 2009. Loose-coupling an air dispersion model and a geographic information system (GIS) for studying air pollution and asthma in the Bronx, New York City. *International Journal of Environmental Health Research* 19 (1): 59–79.

Maantay, J.A., A.R. Maroko, and G. Culp. 2010. Using Geographic Information Science to estimate vulnerable urban populations for flood hazard and risk assessment in New York City. In *Geotechnical contributions to urban hazard and disaster analysis*, ed. P. Showalter and Y. Lu, 71–97. Dordrecht: Springer.

Maheswaran, R., and M. Craglia. 2004. Using GIS for exposure assessment: Experience from the small area health statistics unit. In *GIS in public health practice*, ed. K. de Hoogh, D. Briggs, S. Cookings, A. Bottle, R. Maheswaran, and M. Craglia, 109–124. CRC Press, Boca Raton, FL.

Maroko, A.R. 2010. Chronic exposure to fine particulate matter and heart failure in New York City: A methodological exploration of environmental justice and health (Doctoral dissertation). Retrieved from Graduate Center Retrospective Dissertations, 1965–2013.

Maroko, A.R., J.A. Maantay, N. Sohler, K. Grady, and P. Arno. 2009. The complexities of measuring access to parks and physical activity sites in New York City: A quantitative and qualitative approach. *International Journal of Health Geographics* 34: 1–24.

Maroko, A.R., J.A. Maantay, and K. Grady. 2011. Using geovisualization and geospatial analysis to explore respiratory disease and environmental health justice in New York City. In *Geospatial analysis of environmental health*, ed. J.A. Maantay and S. McLafferty, 39–66. Dordrecht: Springer.

Maroko, A.R., R. Weiss-Riley, M. Reed, and M. Malcolm. 2014. Direct observation of neighborhood stressors and environmental justice in the South Bronx, New York City. *Population and Environment* 35 (4): 477–496.

Maroko, A.R., D. Nash, and B.T. Pavilonis. 2020. COVID-19 and inequity: A comparative spatial analysis of New York City and Chicago hot spots. *Journal of Urban Health* 97: 462–470.

Martinson, M.L., and N.E. Reichman. 2016. Socioeconomic inequalities in low birth weight in the United States, the United Kingdom, Cancada, and Australia. *American Journal of Public Health* 106 (4): 748–754.

Massey, D., and N. Denton. 1988. The dimensions of residential segregation. *Social Forces* 67 (2): 281–315.

Miyake, K.K., A.R. Maroko, K. Grady, J.A. Maantay, and P.S. Arno. 2010. Not just a walk in the park: Methodological improvements for determining environmental justice implications of park access in New York City for the promotion of physical activity. *Cities and the Environment 3* (1): article 8, 1–17. http://escholarship.bc.edu/cate/vol3/iss1/8.

Mohai, P., and R. Saha. 2006. Reassessing racial and socioeconomic disparities in environmental justice research. *Demography* 43 (2): 383–399.

Mohai, P., P.M. Lantz, J. Morenoff, J.S. House, and R.P. Mero. 2009. Racial and socioeconomic disparities in residential proximity to polluting industrial facilities: Evidence from the Americans' Changing Lives Study. *American Journal of Public Health* 99 (S3): 649–655.

Moore, L.V., A.V. Diez Roux, K.R. Evenson, A.P. McGinn, and S.J. Brines. 2008. Availability of recreational resources in minority and low socioeconomic status areas. *American Journal of Preventive Medicine* 34 (1): 16–22.

Morland, K., and S. Filomena. 2007. Disparities in the availability of fruits and vegetables between racially segregated urban neighbourhoods. *Public Health and Nutrition* 10: 1481–1489.

Morland, K., S. Wing, A.V. Diez Roux, and C. Poole. 2002. Neighborhood characteristics associated with the location of food stores and food service places. *American Journal of Preventative Medicine* 22: 23–29.

Newton, A., H. Partridge, and A. Gill. 2015. In and around: Identifying predictors of theft within and near to major mass underground transit systems. In *Safety and security in transit environments: An interdisciplinary approach*, ed. V. Ceccato and A. Newton, 99–115. London: Palgrave Macmillan UK.

Nicholls, S. 2001. Measuring the accessibility and equity of public parks: A case study using GIS. *Managing Leisure* 6: 201–219.

Openshaw, S. 1984. Ecological fallacies and the analysis of areal census data. *Environment and Planning A* 16: 17–31.

Paneth, N.S. 1995. The problem of low birth weight. *The Future of Children* 5 (1): 19–34.

Pathak, E.B., S. Reader, J.P. Tanner, and M.L. Casper. 2011. Spatial clustering of non-transported cardiac decedents: The results of a point pattern analysis and an inquiry into social environmental correlates. *International Journal of Health Geographics* 10 (1): 46.

Perry, R. 1844. *Facts and Observations on the Sanitory State of Glasgow* with statistical tables and maps of the late epidemic shewing the connection existing between poverty, disease, and crime. Glasgow: Gartnavel Press. http://special.lib.gla.ac.uk/exhibns/month/feb2006.html. Accessed 21 Jan 2019.

PHDCN. 1995. Project on human development in Chicago neighborhoods. Retrieved from http://www.icpsr.umich.edu/icpsrweb/PHDCN/instruments.jsp

Pulido, L. 2000. Rethinking environmental racism: White privilege and urban development in Southern California. *Annals of the Association of American Geographers* 90 (1): 12–40.

Sasagawa, M., P.S. Amieux, and M.R. Martzen. 2017. Health equity and the Gini index in the United States. *Journal of Clinical Medicine and Therapeutics* 2 (2): 15.

Setton, E.M., R. Allen, P. Hystad, and C.P. Keller. 2011. Outdoor air pollution and health – A review of the contributions of geotechnologies to exposure assessment. In *Geospatial analysis of environmental health*, ed. J.A. Maantay and S. McLafferty, 67–92. Heidelberg London New York: Springer.

Sicotte, D., and S. Swanson. 2007. Whose risk in Philadelphia? Proximity to unequally hazardous industrial facilities. *Social Science Quarterly* 88: 515–534.

Singh, G.K., and M. Siahpush. 2014. Widening rural-urban disparities in life expectancy, U.S., 1969-2009. *American Journal of Preventative Medicine* 46 (2): 19–29.

Smoyer-Tomic, K.E., J.C. Spence, and K.D. Raine. 2008. The association between neighborhood socioeconomic status and exposure to supermarkets and fast food outlets. *Health and Place* 14: 740–754.

Talen, E. 1997. The social equity of urban service distribution: An exploration of park access in Pueblo, Colorado, and Macon, Georgia. *Urban Geography* 18 (6): 521–541.

Taquino, M., D. Parisi, and D.A. Gill. 2002. Units of analysis and the environmental justice hypothesis: The case of industrial hog farms. *Social Science Quarterly* 83 (1): 298–316.

Tate, E. 2012. Social vulnerability indices: A comparative assessment using uncertainty and sensitivity analysis. *Natural Hazards* 63: 325–347.

Thomson, M.C., S.J. Connor, P. Milligan, and S.P. Flasse. 1997. Mapping malaria risk in Africa: What can satellite data contribute? *Parasitology Today* 13 (8): 313–318.

Tiefenbacher, J.P., and R.R. Hagelman. 1999. Environmental equity in urban Texas: Race, income, and patterns of acute and chronic toxic air releases in metropolitan counties. *Urban Geography* 20: 516–533.

Tobler, W.A. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.

United Church of Christ's Commission for Racial Justice. 1987. *Toxic wastes and race in the United States: A national report on the racial and socio-economic characteristics of communities with hazardous waste sites*. New York: United Church of Christ.

Wang, J., and M.-P. Kwan. 2018. An analytical framework for integrating the spatiotemporal dynamics of environmental context and individual mobility in exposure assessment: A study on the relationship between food environment exposures and body weight. *International Journal of Environmental Research and Public Health* 15 (9): 2022.

Wilkinson, R.G., and K.E. Pickett. 2008. Income inequality and socioeconomic gradients in mortality. *American Journal of Public Health* 98: 699–704.

Wilson, S.M., H. Fraser-Rahim, E. Williams, H. Zhang, L. Rice, E. Svendsen, and W. Abara. 2012. Assessment of the distribution of toxic release inventory facilities in metropolitan Charleston: An environmental justice case study. *American Journal of Public Health* 102 (10): 1974–1980.

Wolch, J., J. Wilson, and J. Fehrenback. 2005. Parks and park funding in Los Angeles: An equity mapping analysis. *Urban Geography* 25: 4–35.

Wong, D. 2005. Formulating a general spatial segregation measure. *The Professional Geographer* 57 (2): 285–294.

Yang, D.-H., R. Goerge, and R. Mullner. 2006. Comparing GIS-based methods of measuring spatial accessibility to health services. *Journal of Medical Systems* 30 (1): 23–32.

Geospatial Tools for Social Medicine: Understanding Rural-Urban Divide

Steven A. Cohen, Mary L. Greaney, Elizabeth Erdman, and Elena N. Naumova

Introduction

Place impacts population health. Increasing evidence suggests that one's place of residence plays a substantial role in determining one's health status in the USA and many other nations across the globe. As a result, health disparities based on geography can and do occur. Among the multitude of studies that have demonstrated geographic health disparities, examples include, but are not limited to, cancer (Krieger et al. 2002), physical activity and obesity (Gordon-Larsen et al. 2006), and healthcare quality and access (Baicker et al. 2005; Stiel et al. 2017; Walker and Crotty 2015). The examination of the causes of place-based health disparities has focused primarily on social determinants of health, such as wealth, education, environmental factors, crime, and others, on a defined geographic level, such as the region, state, or county (Woolf and Braveman 2011). Recently, there has been increasing interest in assessing smaller geographic areas to examine how small-area, place-based neighborhood characteristics influence health. Policies, demographics, natural resources, and economic conditions on the local level may affect availability and quality of resources, development, and economic opportunities (Braveman et al. 2011). A growing body of research suggests that understanding if and how small-area social determinants, including education, wealth, crime, environmental factors, and housing, influence population health is critical to reducing health disparities that may often occur within these areas (Beck et al. 2017; Benach et al. 2001; Diez-Roux 1998; Grow et al. 2010; Kruger et al. 2007; Kulkarni et al. 2011; Lippert et al. 2017; Marmot and Bell 2011).

To identify, understand, and address any potential mechanisms through which place-based factors influence population health and lead to geographic health disparities, it is important to understand how the notion of "place" is conceptualized in population health research. In a seminal paper by Macintyre and colleagues, the authors suggest that there are three categories of geographic variation in health—compositional, contextual, and collective—and that these categories are not mutually exclusive (Macintyre et al. 2002). Compositional factors are attributes of the individuals living in a particular area, such as socioeconomic status, race/ethnicity, and other factors. Contextual factors refer to those in the local environment with emphasis on sociocultural and historical features of the community, such as changing demographics, business, and crime. Collective factors are the collective norms, traditions, values, and needs of the community (Macintyre 1997).

Understanding place-based drivers of health is critical to address health disparities. Increasing research suggests that much of the variability in population health is not due to medical-related factors but to geographic differences in non-medical factors, including social, economic, and demographic factors related to geography. For example, a recent analysis found that social, economic, and physical factors account for nearly 54% of population health (Park et al. 2015), factors that are explicitly linked or related to place-based factors (Fig. 1) (Minnesota Department of Health). Similarly,

S. A. Cohen (✉) · M. L. Greaney
Department of Health Studies, University of Rhode Island, Kingston, RI, USA
e-mail: steven_cohen@uri.edu; mgreaney@uri.edu

E. Erdman
Department of Civil and Environmental Engineering, School of Engineering, School of Engineering, Tufts University, Medford, MA, USA
e-mail: lizerdman505@gmail.com

E. N. Naumova
Division of Nutrition Epidemiology and Data Science, Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, USA
e-mail: elena.naumova@tufts.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_10

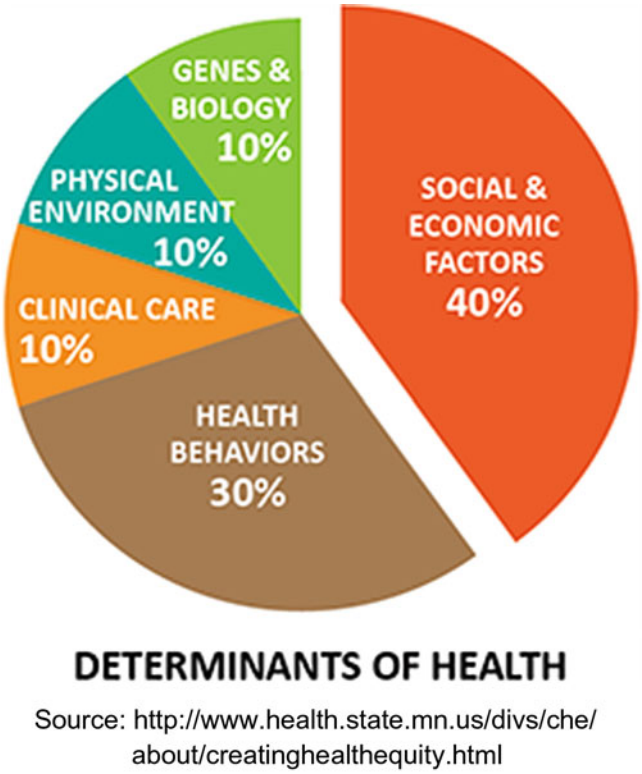


Fig. 1 Determinants of health. (Source: http://www.health.state.mn.us/divs/che/about/creatinghealthequity.html)

analyses from the County Health Rankings and Roadmaps model suggest that approximately 50% of population health is attributable to social and economic factors and the physical environment (Remington et al. 2015; Tarlov 1999; Adler and Newman 2002; McGinnis et al. 2002). Therefore, addressing population-level, place-based drivers of health disparities may have a substantially greater impact on population health than medical and medical-related interventions alone.

Given that place-based factors are associated with population health, a logical question to ask is how do place-based factors influence health. Many such potential mechanisms have been discussed in the literature. The oldest and most well-known example is the Broken Windows theory developed initially by James Wilson that suggests that the appearance of a community's physical environment influences individual behaviors, which ultimately impacts the individual's health and the collective health of the population (Wilson 1987). This theory suggests a dynamic relationship between the environment, health behaviors, and health status. Therefore, based on this theory, as neighborhoods deteriorate in physical appearance, so-called social buffers that could reduce high-risk behaviors may gradually disappear. As the overall health behaviors of a community start to worsen, population health declines, and this is why population health outcomes are often worse in areas of substantial neighborhood degradation (Cohen et al. 2000).

In addition to the Broken Windows theory, other mechanisms by which place may influence population health have been hypothesized. While it is unlikely that place-based factors, such as socioeconomic status (SES), directly impact health, these types of factors shape conditions that ultimately impact population health (Adler and Rehkopf 2008). It is posited that a place-based factor may impact population health through indirect pathways across the lifespan. For example, low SES conditions may contribute to poor nutrition, exposure to harmful environmental contaminants, discrimination, and other aspects of life that may impact health behaviors, access to healthcare, and health status (Adler and Ostrove 1999; Williams and Collins 1995). Such potentially harmful place-based conditions may lead to increased allostatic load throughout life (Seeman et al. 2010) so that stress accumulates over time upon continual exposure to harmful living conditions (McEwen 1998; Juster et al. 2010; Lupien et al. 2015).

The biological, psychological, and sociological pathways of place-based factors which influence population health are just beginning to be understood. Cummins et al. argue that much of the existing research on how "place" influences population health through contextual factors, focusing predominantly on deprivation, has not focused on the contextual and compositional factors (Cummins et al. 2007). Adler and Rehkopf suggest that appropriately combining multiple types of data on SES, demographic, psychosocial, and biological factors on multiple levels will facilitate the creation of causal models that identify direct and indirect pathways that lead to critical but addressable health disparities (Williams and Collins 1995; Adler and Rehkopf 2008).

Exploration of place-based contextual factors has currently centered on a fixed population in a clearly delineated geographic area, such as a county, state, or neighborhood, at fixed points in time. The current view of how place-based characteristics influence population health centers on readily quantifiable, often static measures, such as SES, availability of resources, existence of and proximity to resources, segregation, etc. (Diez-Roux 1998). Some argue that this conventional approach to understanding specific mechanisms through which "place" affects health is integral to population health not just for strengthening causal inferences about place-based risk factors but also for identifying potential avenues for intervention on the population level (Cummins et al. 2007). They suggest moving from the "contextual and compositional" approach to a "relational" approach. In a relational approach, place-based characteristics are viewed somewhat differently. For instance, this approach relies more on socio-relational distances and networks than on physical distance and boundaries and uses area definitions that are more dynamic and fluid. Additionally, a relational approach tends to focus on the cultural aspects of place rather than

the resource or deprivation-based aspects of place. There are, however, numerous methodological challenges in utilizing a relational approach to understand place-based influences on population health which make undertaking such studies more difficult. In this chapter, we focus on the more conventional approach toward understanding the influence of place on health while identifying potential areas where researchers can integrate the more relational approach to understanding these complex pathways.

Why Focus on Defining Place-Based Characteristics in Geospatial Models?

Defining place-based characteristics is integral for population health and assessment of health disparities in geospatial models. Geospatial models have been used extensively for a wide variety of research investigating geographic variation in population health and healthcare, health inequalities, and healthcare service needs. Wennberg and colleagues were among the first to conduct a comprehensive geospatial analysis of variation in Medicare services across the USA (Wennberg et al. 2002). Numerous other studies have followed and have examined critical geographic variability in healthcare service using GIS and geospatial models (Gilmer and Kronick 2011; Matlock et al. 2013; Nicholas et al. 2011; Newhouse and Garber 2013a; Hanchate et al. 2017; Chui et al. 2011). Geospatial models have been widely used to assess health inequalities across populations (Krieger et al. 2002; Weich et al. 2003; McDonald et al. 2012; Suzuki et al. 2012; Newhouse and Garber 2013b; Dwyer-Lindgren et al. 2017) and the need for healthcare services (Black et al. 2004; Padilla et al. 2016). Geospatial modeling provides a critical tool for the analysis and planning of health services and infrastructure to reduce inequalities and promote population health, regardless of geography (McLafferty 2003). Geospatial modeling provides insights into the spatial organization of health services and population health, which allows researchers to incorporate multiple dimensions of place-based factors and potentially multiple levels of observation (e.g., individual and different levels of geographic influence) into the analysis. Examples include calculating travel time to health facilities (Branas et al. 2000; Pearce et al. 2006), assessing optimal locations for healthcare centers (Jia et al. 2007), health behaviors and social capital (Mohnen et al. 2012), and obesity prevalence (Prince et al. 2012).

To optimize the effectiveness and utility of any geospatial model, researchers should delve into what and how place-based factors truly drive and mold potential relationships between health and place. Many geospatial models have focused on assessing specific place-based factors that exist in the geographic regions (e.g., state) and how they contribute to spatial disease patterns and health behaviors. Geospatial models of health outcomes were originally developed to quantify spatial patterns across different geographies by incorporating place-based characteristics (Waller and Gotway 2004; Banerjee et al. 2014). However, incorporating place-based factors in geospatial models requires a multitude of considerations that are often overlooked in such modeling procedures. Therefore, the remainder of this chapter will focus on discussing several important challenges to properly identifying and utilizing place-based spatial characteristics in geospatial modeling for applied population health research, with an emphasis on measuring rural-urban gradients. This chapter also will provide insight into strengths and limitations of different approaches, as well as opportunities for future research into these potentially powerful analytical tools to maximize their utility in research and policy.

Challenges in Determining Place-Based Spatial Characteristics

Selection of Characteristics

In geospatial models of health outcomes, the researchers' assumptions related to a spatial distribution of a sociodemographic indicator of interest, specifically its homogeneity, or consistency within the study area, are critical to proper interpretation and use in developing policies and interventions to address health issues. Building geospatial models starts with the selection of place-based characteristics. Characteristics commonly used in such geospatial models may include socioeconomic status (e.g., wealth and income), demographic composition (e.g., racial/ethnic composition, age, and gender), environmental factors (e.g., climate, air/water/soil quality), and education. These characteristics and how they are measured are typically dynamic, location-specific, multidimensional, and potentially highly correlated and could be costly. Thus, rarely does a universally accepted, singular measure of place-based characteristics suffice for geospatial modeling. This raises several issues, some of which are addressed in the subsequent sections.

Spatial Heterogeneity

An important challenge in determining which place-based spatial characteristics to use, particularly when analyzing spatially aggregated data, is spatial heterogeneity of a selected characteristic. Spatial heterogeneity generally refers to uneven and often heavily skewed distributions of characteristics within an area or between adjacent areas. This section focuses on two specific aspects of spatial heterogeneity,

heterogeneity both within and between observational units in geospatial models of aggregated data.

Heterogeneity Within Observational Units

Place-based characteristics can be defined on multiple geographic levels. Oftentimes, there can be notable heterogeneity in terms of these place-based characteristics, especially when aggregating on a large geographic level such as a county, state, or country both in terms of central tendency and variability. Consider the case of two adjacent counties on the western side of the San Francisco, California Bay Area: San Mateo County and the City and County of San Francisco. Both counties are highly urbanized and are part of the metropolitan statistical area of San Francisco. Both are among the wealthiest counties in the USA in terms of per capita income. San Mateo County has the thirteenth-highest per capita income, $50,262, while the City and County of San Francisco has the sixth-highest per capita income of $55,567 (based on the most recent data from the US Census Bureau 2016 American Community Survey 5-year estimates).

Functionally, in studies examining county-level effects on population health, both counties would be considered as individual units of observation of equal importance. However, a deeper examination of the counties themselves reveals substantial differences between the two. First, there are notable differences in the size and composition. San Francisco has a total land area of 46.9 square mile, while San Mateo County has a total land area of 448.4 square miles. The population size of the counties is comparable, with San Francisco (805,235) having a slightly higher population than San Mateo County (718,451). Due to their differences in land area, the population density of San Francisco is considerably higher (8042 per square mile) than that of San Mateo County (604 per square mile). A simple comparison of housing units reveals further distinctions between the two counties. In San Francisco, there are 376,942 total housing units, whereas in San Mateo County, there are 271,031 total housing units. The resultant data represent notable differences between the two county units in terms of the number of people per household, with the average number of people per household being 24.1% higher in San Mateo County than in San Francisco (2.65 vs 2.14 people per household, respectively). Although this difference appears small, San Francisco has one of the lowest average household sizes for all US counties, whereas San Mateo's average household size is above the US average of 2.54 people per household.

The heterogeneity of these two counties also can be compared using smaller geographic units. Although both county-equivalent units of San Mateo County and San Francisco are similar in terms of per capita income (as shown above), a closer examination of the census tracts within each county reveals stark differences between the counties. It should be noted that San Francisco has 195 census tracts while San Mateo County has 156. There is more variability in per capita income among the census tracts of San Francisco than among the census tracts of San Mateo County (Fig. 2). Although San Francisco's per capita income is just over 10% higher than that of San Mateo County, San Francisco has a wider

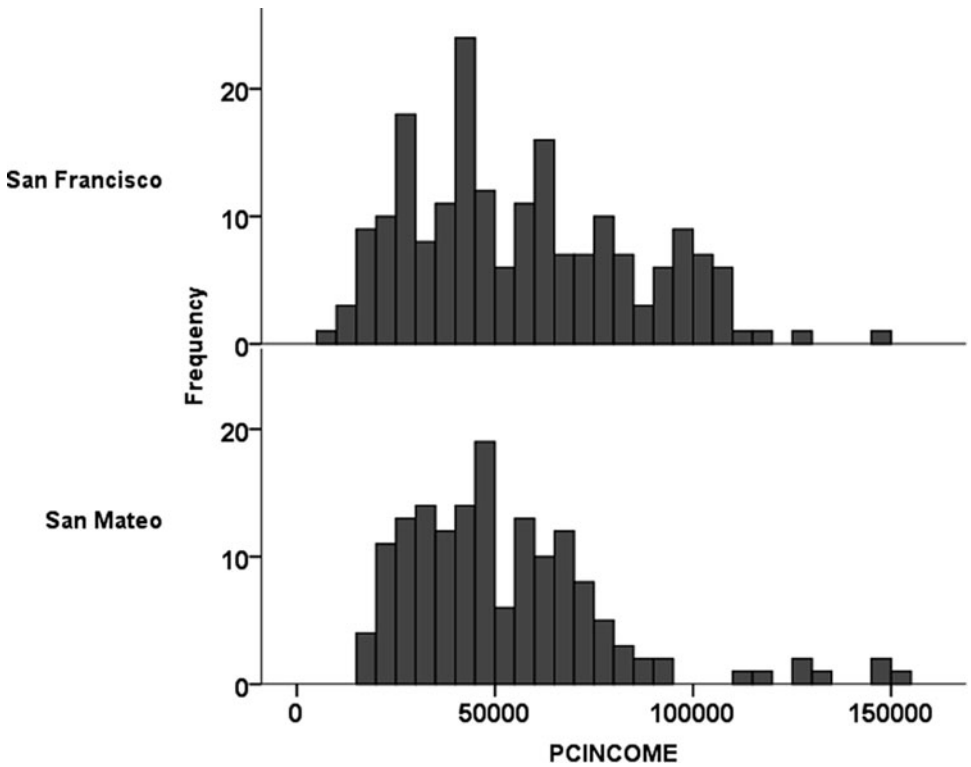


Fig. 2 Distribution of per capita income by census tract in San Francisco City County and San Mateo County (2010)

range of census tract-level per capita income values than San Mateo County does. Despite having the sixth-highest county-equivalent per capita income in the USA, San Francisco contains some of the poorest census tracts in the nation, with the lowest census tract per capita income of $7355. The lowest tract-level per capita income in San Mateo County is $16,732. Considering all census tracts in both counties, the six lowest ones in terms of per capita income are all found in San Francisco. Thirteen of San Francisco's 195 census tracts (6.7%) have a per capita income below $20,000, compared to just 4 of San Mateo's 156 census tracts (2.6%). On the opposite end of the income spectrum, six of the top seven census tracts in terms of per capita income are found in San Mateo County.

This San Francisco and San Mateo County example illustrates that the sum of the parts does not necessarily represent the whole due to spatial heterogeneity within the units of observation. In terms of average per capita income, the two adjacent county-equivalent units appear similar. However, a comparison of additional sociodemographic factors reveals that despite their geographic proximity and similarity in per capita income, these two counties have notably different distributions of the study characteristics. Examining the census tracts within these two county-equivalent units with respect to per capita income reveals further spatial heterogeneity that is masked when examining per capita income on the county level in isolation. It is important to consider such spatial heterogeneity when conducting spatial analyses. In such analyses, where possible, identifying and perhaps quantifying such spatial heterogeneity and underling spatial distributions would enrich the analysis of key socioeconomic, demographic, and other place-based characteristics.

Level of Aggregation

A closely related topic linked to spatial heterogeneity specific to geospatial analyses of aggregated data is the choice of the geographic level of aggregation such as the state or province, county, census tract, or any other unit. There are strengths and drawbacks to each level of aggregation. In this section we discuss a few important issues I to consider when selecting an appropriate level of aggregation in a geospatial analysis. Importantly, this discussion is not exhaustive assessment of the strengths and drawbacks of different levels of aggregation but rather a "jumping-off point" to consider how the selected level of aggregation may influence the results and interpretation of findings and their implications for research and policy.

Data Quality and Confidentiality

There are a variety of valid geographic scales, and the choice of geographic level can lead to different but equally valid results that emphasize different data features (Elliott and Wartenberg 2004). The challenge of selecting a proper scale and aggregation method is referred to as the modifiable area unit problem (Openshaw 1984). The goal of selecting variables to use in geospatial modeling using aggregated data is to choose the smallest geographic units possible to simultaneously maximize sample size and minimize spatial heterogeneity. Yet, the choice is often dictated by available data. Due to limited available data, there is a trade-off between homogeneity within selected geographic units and precision of the estimated associations or disease frequencies. Therefore, a key issue in in geospatial analyses is selecting the most reasonable geographic unit of observation while recognizing its limitation with respect to accuracy and bias this aggregation may introduce.

If they are found to be valid and reliable, a large number of units of observation such as county, state, or region typically increases the likelihood that the geographic information contained within the study area has broad coverage. However, it may reduce the ability to detect potentially critical small-scale trends and associations. Since there is no singular industry standard in terms of data source or protocol for evaluating data quality at different levels of spatial aggregation, the data user must assess the benefits and drawbacks of each potential level of aggregation and, at minimum, identify and discuss the drawbacks in any publically disseminated research project.

When analysis includes records on a fine scale, issues of confidentiality and privacy may also arise, especially when a research question addresses vulnerable populations or people with unique demographic characteristics. Such issues are most pronounced in spatial analysis using small geographic units, such as street address, the census tract, or block group (Clapp and Wang 2006). Methods that attempt to address data confidentiality and privacy include geographic masking, the process of altering the coordinates of geographic data to limit the risk of re-identification in the released data to make it difficult to accurately reverse geocode the released data (Zandbergen 2014). Masking techniques are especially useful in non-aggregated data and also apply to aggregated data (Armstrong et al. 1999). It is worth noting that aggregation itself may mask problematic issues of confidentiality that occur with point-source data (Kounadi and Leitner 2014). Nonetheless, data confidentiality and privacy issues remain a highly debated issue in geospatial modeling (Fefferman et al. 2005; O'Keefe and Rubin 2015), and there is an ongoing need to develop and test statistical methods to address this issue.

Policy Relevance

Another issue to consider when using geospatial models for health research is the utility of the geographic level of aggregation in terms of informing policy. Many health policies are set on the state level by state governments, which

make analyses comparing state-level differences appealing and useful for this purpose. The results of state-level research can immediately inform individual states as to which states are better and which are worse in terms of whatever health outcome or risk factor is examined. In the USA, however, this approach is often limited simply by the limited number of units of observation (50, or slightly more if the District of Columbia and US territories are included), which greatly reduces statistical power, especially in geospatial models where some of the error is explained by spatial correlations.

Counties offer more granularity and greatly increase the number of units of observation (3142 counties and county-equivalents in the USA). There are several important caveats to consider when using county as the level of spatial aggregation. First, the function of county governments varies by state. Some states do not have an active county government system, and all governance is done on the state or municipality level (e.g., New England states). Second, counties and county-equivalent units vary in terms of size and structure within and between states. For example, all independent cities in Virginia, regardless of population size or area, are treated as county-equivalents. Consider the case of Norton, Virginia, an independent city in the rural western part of the state with a population of under 4000 as of the 2010 US Census and a geographic area of 7.5 square miles. County-equivalent units such as Norton and other small, independent cities are considered on the same level of observation as actual counties in Virginia that may be orders of magnitude larger, either based on geographic size, population size, or both, such as Fairfax County, with a population of 1.14 million and a land area of 396 square miles. In many other states, all municipalities, regardless of size, are considered to be part of a county. In Massachusetts, for example, the major city of Boston is part of the larger Suffolk County. All municipalities in Massachusetts, even minor cities with population sizes over 100,000, are part of a larger county. Similarities are found in states such as California, where only large cities, such as San Francisco, are considered county-equivalent units while all 1 smaller cities and towns are part of the California county system. Numerous other similar examples can be found across the USA. Collectively, these are just a few examples and illustrations of the heterogeneity in terms of function, size, and composition, within and between US counties, especially evident when comparing counties across different states.

Smaller units of observation, such as the census tract, municipality, and block group, offer additional gains in terms of the number of units of observation and offer an increasing amount of granularity and the ability to detect key neighborhood and other area-level differences. Geographic levels of observation created by the Census Bureau, such as the census tract and block group, are designed to be relatively homogeneous with respect to population size and function across the small areas they represent. Nonetheless, data aggregated to fine levels of geography such as these may be subject to issues of data reliability, privacy, and confidentiality, as noted previously. Furthermore, policies and interventions designed to address population health issues assessed at a fine geographic level may be difficult to implement due to a variety of factors, including, but not limited to, spillover effects from one area to another and population migration and movement among these small geographic units.

Example: The Swiss Paradox

A key example of how the level of aggregation can affect the findings of geospatial models and therefore impact downstream policies and programs is known as the "Swiss paradox" (Clough-Gorr et al. 2015). It has been widely established in the public health and social medicine and public health literature that higher income inequality is generally associated with worse population health outcomes (Kawachi and Kennedy 1999; Krieger et al. 2002; Lynch et al. 2000). Examples of this association are numerous and include obesity (Zhang and Wang 2004; Wilkinson and Pickett 2006), self-reported health (Kondo et al. 2009), and overall mortality (Kennedy et al. 1996; Vincens and Stafström 2015). Although there are a variety of theories and empirical evidence to support these associations, the precise reasons for them are not entirely clear. A seminal article by Kawachi and Kennedy (1997) suggested that income inequality promotes poorer health outcomes by reducing social cohesion. Further studies have suggested other potential complementary mechanisms through which income inequality affects health outcomes. One hypothesis is that income inequality is a correlate of other structural, demographic inequalities, such as racial segregation, whereby spatial concentrations of race and poverty influence individual and population health outcomes (Subramanian and Kawachi 2004).

The term "Swiss paradox" was coined by Clough-Gorr and colleagues in a 2015 article, one of the first studies to formally investigate how level of spatial aggregation may influence the associations between income inequality and health. When measured on the state level, income inequality is associated with poorer health outcomes (Kahn et al. 2000; Kennedy et al. 1998; Subramanian and Kawachi 2004; Subramanian and Kawachi 2003). However, with lower levels of aggregation, such as the census tract and county, the findings are mixed (Fiscella and Franks 1997; Soobader and LeClere 1998; Eckenrode et al. 2014). Clough-Gorr and colleagues observed that higher income inequality in Swiss municipalities was consistently associated with lower mortality risk, except for certain health outcomes, even after accounting for sex, marital status, nation of origin, rural-urban status, and other potential confounding factors. Their results challenge current beliefs about the effect of income inequality on health

Table 1 Parameter estimates from the association between Gini index and each of the five listed health outcomes on the county and state levels

Model type and Gini level	Obesity	Diabetes	Current smoker	Poor/fair SRH	Sedentary lifestyle
Unadjusted					
County	−0.33 (−0.54, −0.13)	−0.08 (−0.12, 0.27)	0.05 (−0.10, 0.20)	0.82 (0.59, 1.04)	0.19 (−0.04, 0.42)
State	−0.01 (−0.25, 0.23)	0.55 (0.32, 0.78)	0.24 (0.07, 0.41)	1.11 (0.84, 1.38)	0.66 (0.39, 0.94)
Income-adjusted					
County	−0.39 (−0.59, −0.19)	0.03 (−0.16, 0.23)	0.01 (−0.14, 0.15)	0.66 (0.45, 0.88)	0.05 (−0.17, 0.28)
State	−0.09 (−0.33, 0.15)	0.50 (0.27, 0.73)	0.18 (0.01, 0.36)	0.89 (0.64, 1.15)	0.48 (0.21, 0.75)
Fully adjusted					
County	−0.42 (−0.63, −0.20)	−0.10 (−0.31, 0.10)	0.01 (−0.14, 0.17)	0.63 (0.41, 0.86)	0.23 (−0.01, 0.47)
State	−0.25 (−0.50, 0.01)	0.30 (0.06, 0.55)	0.12 (−0.06, 0.30)	0.71 (0.44, 0.98)	0.58 (0.30, 0.85)

on a fine geographic scale. The reasons for such findings, however, remain unclear and merit further research.

A direct comparison between income inequalities based in the USA that examined the effect of aggregating at the state level versus the county level further corroborates the Swiss paradox. To illustrate the challenges, data from the 2012 Behavioral Risk Factor Surveillance System (BRFSS) were utilized. The BRFSS is a nationally representative phone survey of nearly 500,000 US residents in all 50 states, plus districts and overseas territories. The 2012 BRFSS sample was selected because it was the last year in which county of residence was publicly available in the data set. The association between income inequality and the county-level prevalence of five representative health behaviors and outcomes—obesity, diabetes, current smoking, sedentary lifestyle, and fair/poor self-reported health—was assessed using generalized linear models. The analysis was conducted using income inequality on two levels of spatial aggregation, the state and county, adjusting for income and other sociodemographic factors. Findings identified three distinct patterns of associations (Table 1). First, for fair/poor self-reported health, higher income inequality on both the state and county levels was associated with an increase in the prevalence of this health outcome, which is what might be expected. Second, higher income inequality was associated with a higher prevalence of both diabetes and having a sedentary lifestyle when income inequality was measured on the state level, but not when measured on the county level. Similar results were obtained for current smoking status, except the association between state-level income inequality and prevalence of current smoking became nonsignificant in the fully adjusted models. Third, and perhaps most interestingly, for obesity, higher income inequality on the county level was actually associated with a decreased prevalence of obesity, while there were no significant associations observed when income inequality was measured on the state level. In this analysis, several challenges are apparent. This analysis considered each geographic unit as spatially independent and did not test for potential spatial dependency among geographic units using Moran's I or other statistic. This is likely a more important problem for counties than for states (Manley et al. 2006) in terms of ability to distinguish local patterns of spatial autocorrelation. Additionally, there is a considerable difference in sample size and the number of units of observation between states (51, including DC) and counties (3143), resulted from spatial aggregation. Related to this caveat, the distribution of Gini index is notably different when measured on the state and county levels (Table 2).

Table 2 Descriptive statistics for Gini index on the state and county levels (2012, source: US Census Bureau)

Statistic	States	Counties
N	51	3143
Mean (SD)	0.4552	0.4350
Median	0.4559	0.4325
Min	0.4132	0.3161
Max	0.5315	0.5994
Skewness	0.7190	0.3573
Kurtosis	1.9461	0.3150

Study findings underscore the notion that level of aggregation matters. Why the association between income inequality and health varied based on the level of aggregation is not entirely clear. Uncovering some of the potential mechanisms through which these social characteristics affect health on these and other geographic levels is integral to creating effective policies and programs designed to reduce health inequalities and improve population health, regardless of geography.

Case Study: Rural-Urban Status

Examining place-based factors that drive population health and promote health disparities requires careful attention to the place-based factors and characteristics studied. There are many instances in which there is no scientific consensus as to the best measure of a certain social, demographic, environmental, or economic factors as each measure may

have its own benefits and drawbacks. One key example of this is the measure of rural-urban status (Cohen et al. 2018b). Whether describing rural health issues, assessing rural-urban health disparities, and examining the process of urbanization or any other issues pertaining to the rural-urban divide, it is essential to understand, utilize, and interpret appropriate measures of rural-urban status to properly characterize the place-based characteristics the researcher seeks to address. Furthermore, it is valuable to note that place-based characteristics, especially those concerning measures such as rural-urban status, depend heavily on environmental factors that have a meaningful impact on health both in and surrounding the areas of study. Such factors include, but are not limited to, the use of agricultural land, roads, landfills, presence of bodies of water, forests, national preserves, parks, and even concentrations of man-made structures such as buildings (Erdman et al. 2015; Jagai et al. 2010). The following exemplar case studies illustrate some of the many options and considerations of measuring rural-urban status in geospatial and other related models.

Defining Rural-Urban Status

One basic issue to consider when using rural-urban status in geospatial models is which definition of rural-urban status to use. As is the case of many sociodemographic measures, there is no scientific consensus as to the "best" measure of rural-urban status, and each one has unique strengths and weaknesses that should be taken into account (Hart et al. 2005). Furthermore, each measure requires a unique interpretation and may reflect different aspects of the geographies under study. It may be useful to note that some measures are only defined and available on a certain geographic level of aggregation.

Commonly used measures or rural-urban status in social medicine and public health studies include, but are not limited to, population density, percent urban population, Urban Influence Codes (UIC), Rural-Urban Continuum Codes (RUCC), and Rural-Urban Commuting Areas (RUCA). Population density and percent urban population are available from the US Bureau of the Census and have the flexibility to be used at the state, county, census tract, and block group. The UIC, RUCC, and RUCA are produced and maintained by the US Department of Agriculture. These three measures—UIC, RUCC, and RUCA—are only available on certain geographic level of aggregation. The UIC and RUCC are available only at the county level, while the RUCA is available on the census tract level, which could be scaled up to other geographic levels with appropriate weighting schemes.

Consider the case of the percent urban variable that is used in many studies of rural-urban health and health disparities. This variable, defined as the percentage of the area population that is deemed by the Census Bureau to live in an urbanized place, has far-reaching research caveats. Take, for example, an in-depth examination of the 29 counties or county-equivalent places with a 100% urban population. Among those 29 counties are large metropolitan counties, such as Denver County, Colorado, one of the largest cities and counties in the country, with a population of 285,797, as well as far smaller counties and county-equivalents, such as Covington, Virginia, with a population of just 3067. Covington, Virginia, is situated in a highly rural, mountainous area with no major population centers within several hundred miles. Yet, both Denver and Covington would be considered to be equally "urban" according to the percent urban variable. For comparison, San Diego County, California, which comprises the majority of the second-largest city in California with just over 1.1 million county residents, would be considered less urban (96.5%) than Covington, Virginia (100%), using percent urban as the measure of rural-urban status.

As a result of this limitation, composite indices of rural-urban status take into account multiple aspects of the rural-urban gradient and are gaining traction in public health and biomedical research (Naumova et al. 2009). An example of a composite measure is the Index of Relative Rurality (IRR) (Waldorf 2007), which is a continuous measure (0 to 1) of rural-urban status that takes into account population density, population size, proximity to metropolitan areas, and percent urban population. This measure was originally used at the county level but can easily be calculated for other geographic levels, such as the census tract or block group. The IRR and other related measures have clear strengths, such as they are continuous, take into account multiple aspects of the rural-urban gradient, and are flexible on different geographic scales. The central drawback of using this type of measure is in its interpretation. As in the example of the IRR, since, by definition, the measure is a relative measure of rurality, differences between geographic units have no immediate, obvious, and easy-to-comprehend interpretation. For example, the difference in IRR between San Diego County, California (0.24), and Covington, Virginia (0.31), is 0.07 IRR units. The scale of the IRR ranges from 0.04 for New York City Manhattan Borough to 0.89 for Northwest Arctic Borough in Alaska.

The choice of how to measure rural-urban status affects the potential associations observed between rural-urban status and health outcomes. While several of the individual rural-urban status measures are strongly correlated to each other, others are not. Further complicating this issue is that the magnitude of some of the correlations varies substantially by geographic region. For instance, the rank correlation among the RUCC, UIC, population density, percent urban, and IRR was as high as 0.917 ($p < 0.001$) for the RUCC-UIC correlation, to as low as 0.521 ($p < 0.001$) for the UIC-percent urban correlation for US counties (Cohen et al.

2015). When stratified into nine Census divisions, the range of the UIC-percent urban correlation varied substantially from 0.802 in the Pacific states to as low as 0.384 in the West South Central states. The same study found that, as a result, the magnitude and direction of the association between rural-urban status and the health outcome of obesity varied considerably by the choice of rural-urban status measure as well as the geographic region of analysis.

Consideration of Variable Type in Assessing Rural-Urban Differences

Another important element of assessing and using rural-urban status in geospatial models is what type of rural-urban measurement to use (i.e., dichotomous, ordinal, discrete, or continuous). There is no scientific consensus as to which type of variable to use (Hart et al. 2005). One type of variable commonly used is a rural-urban dichotomy (Haque and Telfair 2000; Dahly and Adair 2007) which has several advantages. Perhaps the most obvious advantage to dichotomizing rural-urban status is the ease of interpretation and dissemination in research and practice. When a dichotomous measure, such as metropolitan vs nonmetropolitan, or when a continuous measure of rural-urban status, such as population density or percent urban, is dichotomized, it is straightforward to interpret in the context of disparities and can facilitate easy comparison. The concepts of "rural" and "urban" can be directly compared for interpretation, statistical analysis, and subsequent dissemination in research and to the general public.

As would be the case for converting any continuous measure to a dichotomous measure, there is a critical issue of deciding which cut point to use when delineating "rural" from "urban." Consider the example of population density and obesity among a sample of older adults aged 65 and above abstracted from the 2012 BRFSS. In this analysis, nine different cut points are used to delineate "rural" from "urban" counties in the USA at each decile of population density (Fig. 3). If the tenth decile is used, which indicates the lowest 10% of population density (extremely rural) versus all other counties, the prevalence of obesity in those counties considered to be rural is significantly lower (24.2%) than in those counties considered to be urban (27.7%). However, if the 90th percentile of population density is used, which would separate counties into highly urban (top 10%) versus all others, the prevalence of obesity in the rural counties is significantly higher (27.6%) than that of the urban counties (25.2%). Similar results are observed when using the 80th percentile of population density as the cutoff value: the prevalence of obesity is significantly higher in the rural counties (27.6%) than in urban counties (26.0%). Using the median county population density or any of the surrounding deciles as cutoffs (20th through 70th), there would be no significant differences between rural and urban counties in the prevalence of obesity. Therefore, in this example, it is evident that when dichotomizing a continuous variable to obtain a measure of rural-urban status, the choice of cutoff value makes a substantial difference in the conclusions reached about the health outcome of study. In this case, the selection of two different cutoff values—at 10% and 90%—to delineate "rural" from "urban" results in completely op-

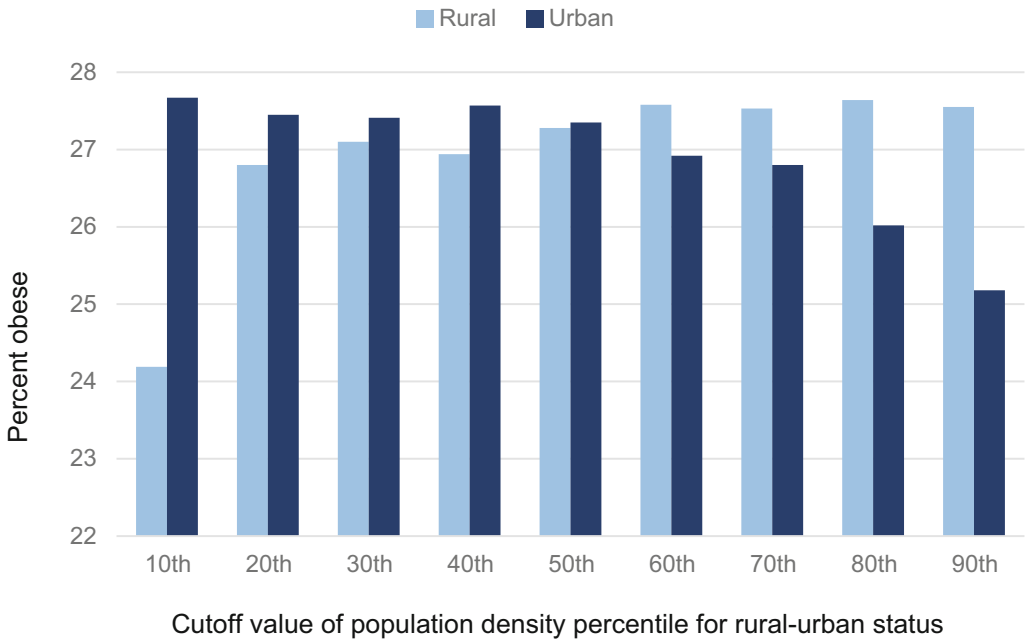


Fig. 3 Percent obese using ten different cut points for rural-urban status based on population density decile

posite findings. These results are simply an application or extension of the problem of dichotomization used in other, non-geospatial models.

Ordinal variables, such as the RUCC, RUCA, and UIC, discussed previously, are advantageous for a variety of reasons, but they also have several drawbacks that should be taken into consideration. Ordinal variables may be preferred over dichotomous variables because of their ability to distinguish finer gradations of rural-urban status. For example, the RUCC is a classification scheme that delineates metropolitan counties from nonmetropolitan counties. Metropolitan counties are classified by the population size of their metro area and nonmetropolitan counties by degree of urbanization and proximity to a metro area. The RUCC ranks counties on a scale from 1 to 9 based on these characteristics. An advantage of using an ordinal variable, such as the RUCC, is the flexibility to treat it either as a continuous or discrete predictor variable or as a series of dummy or indicator variables if there is enough statistical power to do so. The advantage of the former is to assess to see if there is a quasi-linear association between rural-urban status and the outcome, while the advantage of the latter is to assess potential nonlinear associations between rural-urban status.

Nonetheless, there are some inherent drawback to using ordinal variables, some unique to variables such as the RUCC, UIC, and RUCA. The first pertains to using RUCC, for example, as a continuous or discrete predictor in models. This assumes that there is a linear association between the RUCC and the outcome of interest (whether continuous, ordinal, or dichotomous). If there is a nonlinear component to the relationship, i.e., curvilinear, j-shaped, etc., the model may not adequately quantify this association. A more serious issue may be is the construction of the measure itself. While the RUCC is presented as an ordinal variable (1 to 9), the gradations between each unit do not reflect an ordinal process. Consider a RUCC value of 3, which represents "counties in metro areas of fewer than 250,000 population," whereas a RUCC value of 4 represents "counties with an urban population of 20,000 or more, adjacent to a metro area." A RUCC value of 3 is considered more urban than a value of 4. Yet, there are numerous examples of counties with population levels well below the threshold of 250,000 that lie in a metropolitan area (considered a 3), while much more populous counties lie just outside and immediately adjacent to one or two metropolitan areas with large urban populations (considered a 4). Although the county considered a 3 on the RUCC scale might appear less urban than the county considered a 4, the former would be considered to be more urban than the latter. Similar issues exist with the RUCA and UIC measures as well.

It is important to note that no classification system, dichotomous, ordinal, discrete, continuous, or other, is free of issues and caveats. Ordinal measures such as the RUCC, UIC, and RUCA provide a robust array of options for assessing rural-urban status above and beyond many traditional unidimensional measures, such as population density or percent urban population. Given that there is no universal, standard measure of rural-urban status, there are a variety of available measures and variable types to use to suit the needs of researchers interested in assessing place-based characteristics, such as rural-urban status. There is value in understanding the strengths and weaknesses of each one, but if they are used properly, their use will not render an analysis invalid.

Assessment of Nonlinearity in the Rural-Urban Gradient

Example: Rural-Urban Status and Health Outcomes

As an example, we consider linear measures of rural-urban status and assess potential nonlinearity of an association between rural-urban status and health outcomes. In the case study highlighted here, we assessed the associations between rural-urban status and multiple health outcomes from a national survey of older adults using seven commonly used measures of rural-urban status: RUCC, UIC, RUCA, Euclidean distance to nearest metropolitan area, population size, population density, and percent of the population that is urban, with each measure being stratified into quintiles. The association between quintile of rural-urban status measures and the examined health outcomes (obesity and missing annual medical checkup) was assessed through logistic regression modeling, accounting for complex sampling and controlling for confounding variables. We also examined linear trends by treating quintile of each rural-urban status measure as a discrete variable. Details are outlined in the article (Cohen et al. 2018b).

Study results emphasize some of the points made previously. First, compared to the most urban quintile of each measure (reference group), generally speaking, the odds of each outcome—obesity and missing an annual checkup—were significantly higher in the more rural areas (Fig. 4), with some key exceptions. For population density, the odds of obesity were significantly lower in the most rural quintile and significantly higher in the third and fourth quintiles compared to the most urban quintile. Analyses revealed a significant monotonic association and population density quintile (increasing urbanity was associated with an increased likelihood of obesity). However, a linear or monotonic association was not evident for any of the other six measures (RUCC, UIC, RUCA, Euclidian distance, population size, and percent urban), likely to the curvilinear relationship between rural-urban status and obesity for many of the measures.

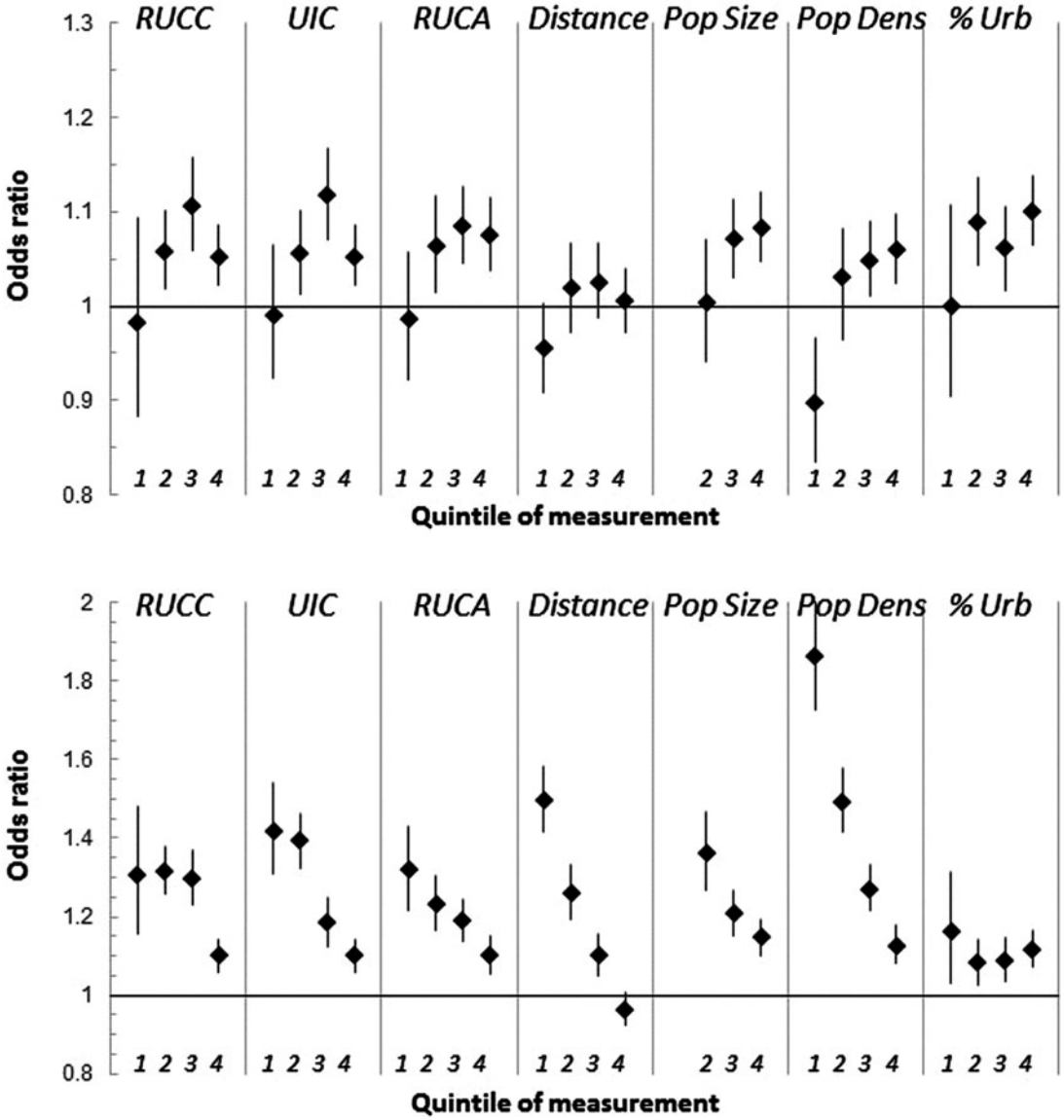


Fig. 4 Odds ratio of obesity (top panel) and missed annual checkup (bottom panel) for seven measures of rural-urban status in quintiles: RUCC, UIC, RUCA, distance to nearest metropolitan area, population size, population density, and percent urban population. Reference group is the highest (most urban) quintile. (*Adapted from Cohen et al. 2018a)

Therefore, using a non-dichotomous measure of rural-urban status revealed a nuanced, U- or J-shaped association between rural-urban status and obesity that might have been masked had a dichotomous measure of rural-urban status been used. Also, the associations depended upon the specific measure of rural-urban status. Had population density quintile been a discrete variable, we would have been able to assess a potential monotonic relationship between it and the log odds of obesity. A monotonic relationship would have been observed: increasing population density is associated with greater odds of obesity. However, the results show that this is not entirely true, based on the data. There may be a positive monotonic relationship between population density and obesity but only among the four most rural quintiles of population density. In other words, the monotonic association does not hold for the most urban quintile. The odds ratios of the association between both the third and fourth quintiles of population density and obesity were above 1. Therefore, the risk of obesity is highest in the intermediate (third and fourth quintiles) of population density, and not in the highest (most urban) quintile, and the association between obesity and rural-urban status was curvilinear and non-monotonic.

Analogous findings also were observed for missing an annual checkup. In the case of this measure, six of the seven measures of rural-urban status were inversely and monotoni-

cally related to the likelihood of missing an annual checkup: as rurality increased, respondents were significantly more likely to have missed an annual checkup for all measures, except percent urban population. For percent urban, although respondents in each of the four most rural quintiles were significantly more likely to have missed their annual checkup, the associations varied in magnitude, which likely precluded a monotonic association.

Example: Rural-Urban Status, Vegetation, and Asthma in Older Adults

Here we further illustrate the effect of nonlinearity in the rural-urban gradient in exploring the relationship between hospitalizations among older adults due to asthma in the New England states (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont) and New York State (Erdman et al. 2015). These associations can be clearly affected by population density in many ways. The seven states included in this study range from densely populated New York City and southern Connecticut to rural Maine and New Hampshire. All of the seven states have relatively similar climates and other environmental factors. In 2006, the population of these states was 11% of the entire US population (33,576,172), 13% (4,439,893) of whom were over 65 years of age (census.gov). To assess the associations between the rates of disease and environmental characteristics, it was necessary to spatially and temporally align multiple data sets. We used satellite imagery to assess degree of greenness and created a measure for the percent of green cover for 16-day time periods during 2005–2006. Census data (2010) were used to abstract population size and calculated hospitalization rates based on patients' zip code of residence.

The first step of the analysis was creating the data set to match the hospitalization data to the satellite imagery and detecting alignment and identify misclassification issues. For example, hospitalization records were arranged by residential zip code, whereas for imagery data, we used a shape file to align with zip codes boundaries. The shape file generally had fewer pixels of data on smaller areas that tend to also have larger population densities. The hospitalization records listed 3109 zip codes; after merging those with the census data resulted in a data set of 2864 zip codes, a net loss of 245 zip codes with low population sizes. To reduce spuriously high rates, zip codes with older adult populations fewer than 100 residents were merged with adjoining zip codes providing these were in the same state. Neighboring zip codes with the most similar population size were merged until the joined county population exceeded 100 residents aged 65 + .

Linking medical claims and satellite imagery along with spatial alignment should consider temporal alignment as well. While medical records were complete for the study period, some satellite images were missing or unusable during that period. When we linked imagery data by zip code to the hospitalization records, we lost additional 424 zip codes resulting in 2876 complete matches. We then merged zip codes that had missing values with neighbors that are likely to be similar in environmental exposures. Both spatial and temporal alignments within and between data sets are time-consuming, but the effect of missing data may compound across multiple data sets and may influence the final analysis and its findings. Thus far, the data linkage procedures are rarely described in the epidemiological literature, and a system of checks and balances to identify data discrepancies does not yet exist.

As we explored the associations between degree of greenness and asthma rates, we noted that the relationships were influenced by population density and that association was not monotonic. We applied simple cutoffs and marked a zip code as urban if a zip code had >830.7 persons per square mile, rural if a zip code had <107 persons per square mile, and semi-urban otherwise. In the studied seven states, the average elderly population was 13.1% and an average log population density of 2.1 (131 people per square mile). The number of zip code falling in a rural category was high for Vermont, New Hampshire, and Maine while Connecticut, New York, Massachusetts, and Rhode Island had almost equal mix of rural and urban zip codes. Overall the relationship between hospitalization rates and population density was U-shaped with a marked increase at both extremes: for heavily populated and the least populated zip codes (Figs. 5 and 6). This nonlinearity requires exploring the relationship separately for urban, rural, and semi-urban zip codes. After adjusting for income and percent elderly population, higher evergreen vegetation in urban areas demonstrated a small yet protective effect.

Summary of Examples

The provided or included examples are not intended to imply that all measures of rural-urban status are invalid and inconsistent. Rather, they highlight the need to consider the specific rural-urban status measure being used and what aspect or aspects of the rural-urban continuum the selected measure is intended to emphasize. Moreover, as with any predictor variable used in modeling health outcomes, whether it is geospatial or traditional, non-spatial models, it is important to consider the trade-offs of using one variable type over another. For example, as discussed, treating rural-urban status as a continuous or discrete variable reduces a model degrees of freedom and may optimize statistical power. However, this use assumes a monotonic association between rural-urban status and the health outcome(s) under study. Using indicator variables, as illustrated, can allow for the assessment of non-monotonic associations but require additional model degrees of freedom. There is no one valid way to use rural-urban

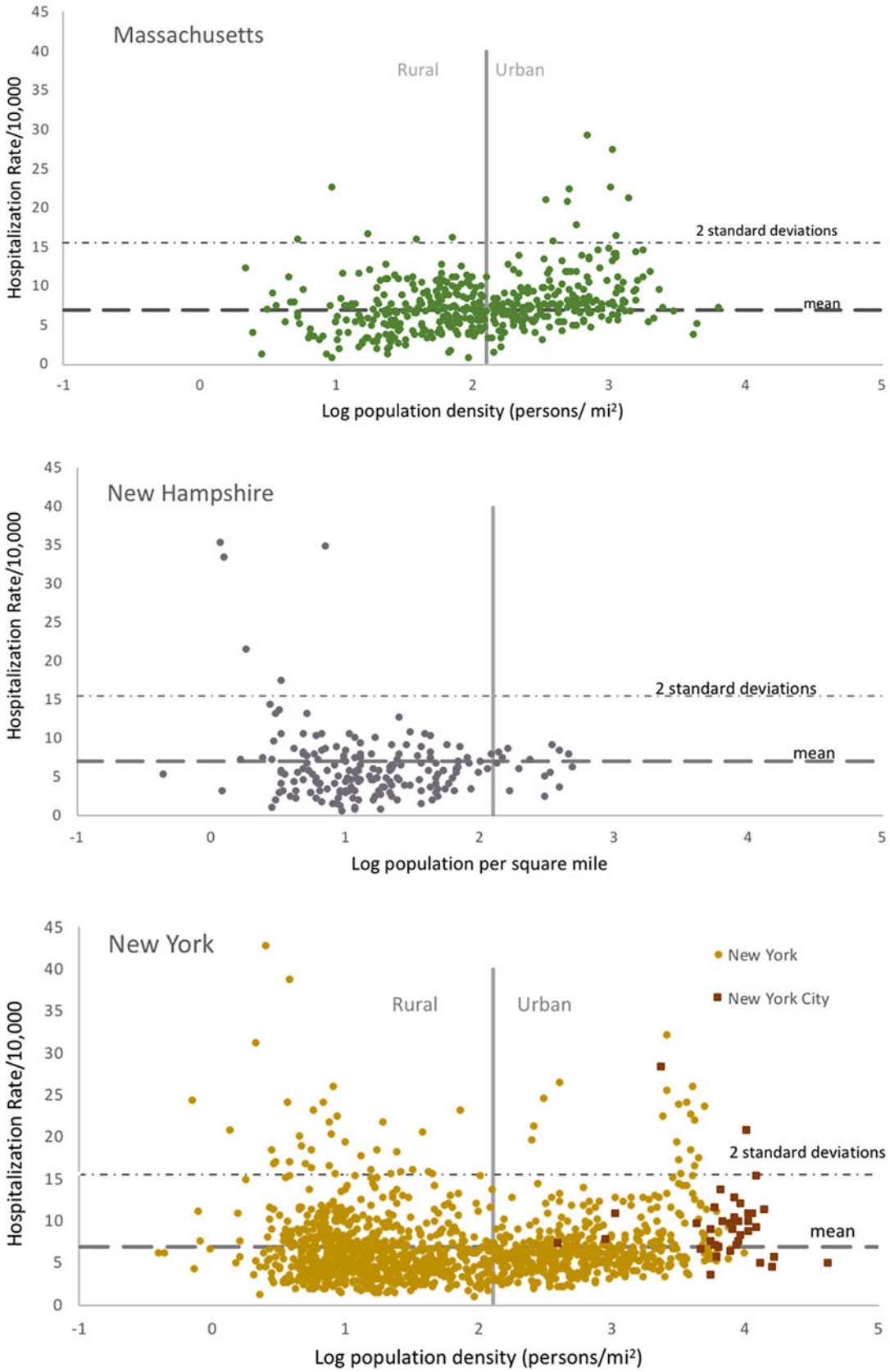


Fig. 5 Relationship between population density of asthma hospitalization rates among older adults in selected states, 2005–2006

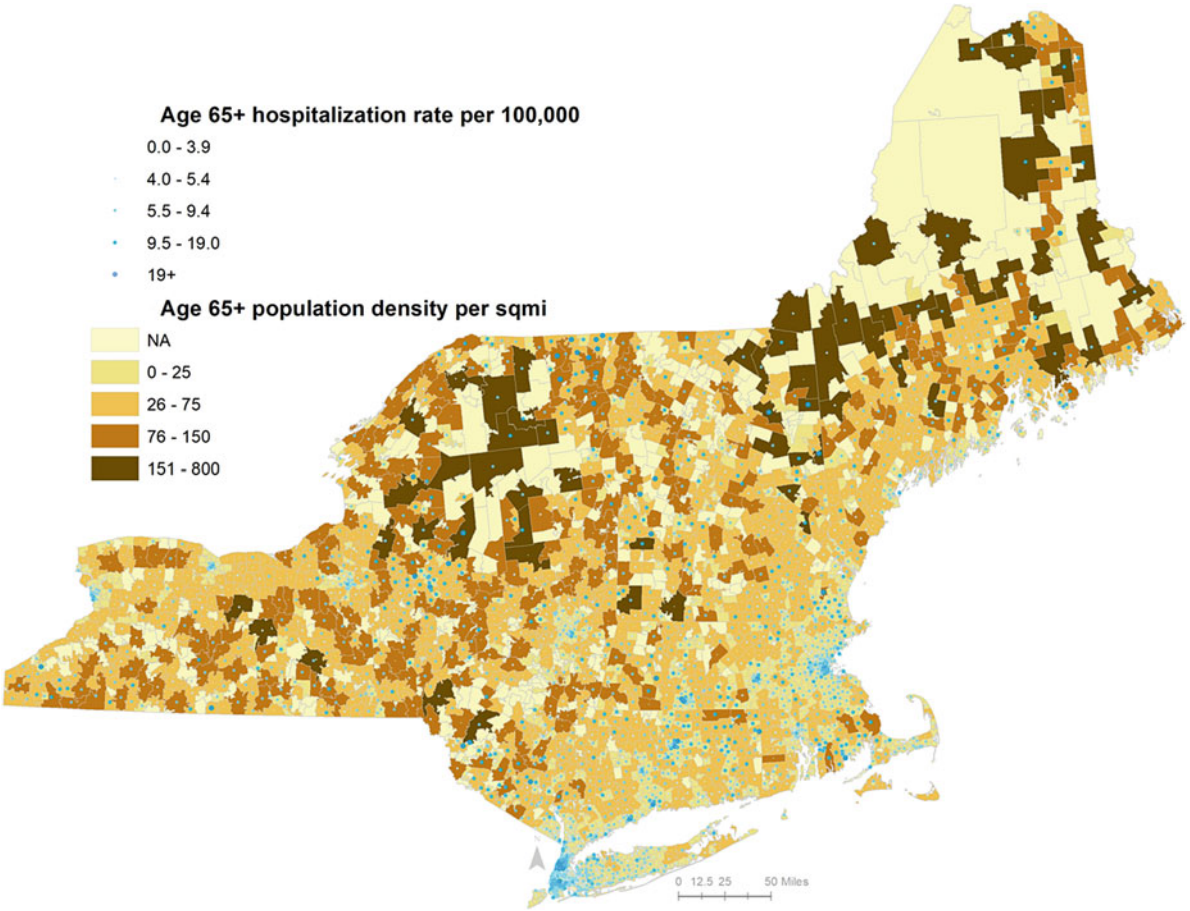


Fig. 6 Relationship between population density of asthma hospitalization rates among older adults in New York and six New England states, 2005–2006

status, but its use requires attention to these and other issues to properly characterize the relationship between rural-urban status and the health outcome under study.

Recommendations and Conclusions

This chapter addresses a handful of the many issues associated with selecting and utilizing variables to address key place-based social determinants of health, with examples to the rural-urban gradient. This chapter may raise more questions than it answers regarding decisions around selection of measures to use and how to use them. There remains no scientific consensus as to best practices, and it is left to the researcher to decide which measures to use based on study questions, on what geographic scale or scales to use them, and interpretation of findings.

When using place-based characteristics in geospatial models and, by extension, in non-spatial models, it is important to consider the following questions: First, on what geographic scale will the characteristic be measured and analyzed? Different scales provide certain benefits and drawbacks in terms of statistical stability, policy relevance, sample size, availability of data, and other considerations. Second, what factors or determinants are most relevant for answering the research question? This question raises the issue of policy relevance, ability to take action upon significant findings, accuracy of the measure, and numerous other issues. Many measures are multidimensional, and selecting one over the other may have meaningful implications for the directionality, magnitude, and overall nature of any observed association. Third, what type of variable will be used in the analysis? This question is relevant to all types of models, not just geospatial models. Different types of variables offer trade-offs in terms of modeling the type of association, interpretability of findings, and statistical power. In the example of rural-urban status, there is a need to use the concept of a power law to incorporate rural-urban metrics that take into account population distribution and population density measures that could be sustainable and valid across different geographic aggregation schemes.

This chapter discussed the rural-urban gradient as an example of a social determinant of health explored in a growing

body of population health studies that is intrinsically linked to geography. Although the discussion, particularly the case studies, focuses on issues pertaining to measuring rural-urban health status specifically, the broader concepts of geographic scale, policy relevance, statistical power, the modifiable area unit problem, and many of the other issues described above extend to other social determinants of health, such as SES, household composition, education, income inequality, and demographic structure (e.g., age, race/ethnicity, gender, etc.). Furthermore, for rural-urban status, income inequality and other social determinants of health are intrinsically tied to the concept of place. The processes we are trying to capture are dynamic, yet we are limited by the preponderance of static tools and measures available to researchers. Therefore, all of these measures have a critical temporal component that may be challenging to address in standard geospatial modeling. What we observed today with respect to these social determinants is not necessarily can be observed in the past yet quite likely reflects the consequences of the past, including historical nature-made and man-made events and other perhaps ongoing measurable and unmeasurable processes.

The place-based factors discussed in this chapter and other social determinants of health often represent the ultimate or distal causes of disease and health disparities. On the other hand, they also provide opportunities on which base policies, programs, and interventions can be designed to promote healthy behaviors, improve population health, and ultimately reduce health disparities. Awareness of the issues surrounding measurement of these determinants is integral to conduct meaningful and impactful research through which population health can be improved.

References

Adler, N.E., and K. Newman. 2002. Socioeconomic disparities in health: Pathways and policies. *Health Affairs* 21 (2): 60–76.

Adler, N.E., and J.M. Ostrove. 1999. Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences* 896 (1): 3–15.

Adler, N.E., and D.H. Rehkopf. 2008. US disparities in health: Descriptions, causes, and mechanisms. *Annual Review of Public Health* 29: 235–252.

Armstrong, M.P., G. Rushton, and D.L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18: 497–525.

Baicker, K., A. Chandra, and J. Skinner. 2005. Geographic variation in health care and the problem of measuring racial disparities. *Perspectives in Biology and Medicine* 48 (1): 42–S53.

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2014. *Hierarchical modeling and analysis for spatial data*. CRC press.

Beck, A.F., M.T. Sandel, P.H. Ryan, and R.S. Kahn. 2017. Mapping neighborhood health geomarkers to clinical care decisions to promote equity in child health. *Health Affairs* 36 (6): 999–1005.

Benach, J., Y. Yasui, C. Borrell, M. Sáez, and M.I. Pasarin. 2001. Material deprivation and leading causes of death by gender: Evidence from a nationwide small area study. *Journal of Epidemiology & Community Health* 55 (4): 239–245.

Black, M., S. Ebener, P.N. Aguilar, M. Vidaurre, and Z. El Morjani. 2004. Using GIS to measure physical accessibility to health care. *World Health Organization*: 3–4.

Branas, C.C., E.J. MacKenzie, and C.S. ReVelle. 2000. A trauma resource allocation model for ambulances and hospitals. *Health Services Research* 35 (2): 489.

Braveman, P.A., S.A. Egerter, S.H. Woolf, and J.S. Marks. 2011. When do we know enough to recommend action on the social determinants of health? *American Journal of Preventive Medicine* 40 (1): S58–S66.

Chui, K.K., S.A. Cohen, and E.N. Naumova. 2011. Snowbirds and infection–new phenomena in pneumonia and influenza hospitalizations from winter migration of older adults: A spatiotemporal analysis. *BMC Public Health* 11 (1): 444.

Clapp, J.M., and Y. Wang. 2006. Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics* 59 (2): 259–284.

Clough-Gorr, K.M., M. Egger, and A. Spoerri. 2015. A Swiss paradox? Higher income inequality of municipalities is associated with lower mortality in Switzerland. *European Journal of Epidemiology* 30 (8): 627–636.

Cohen, D., S. Spear, R. Scribner, P. Kissinger, K. Mason, and J. Wildgen. 2000. "Broken windows" and the risk of gonorrhea. *American Journal of Public Health* 90 (2): 230.

Cohen, S.A., L. Kelley, and A.E. Bell. 2015. Spatiotemporal discordance in five common measures of rurality for US counties and applications for health disparities research in older adults. *Frontiers in Public Health* 3: 267.

Cohen, S.A., S.K. Cook, T.A. Sando, and N.J. Sabik. 2018a. What aspects of rural life contribute to rural-urban health disparities in older adults? Evidence from a national survey. *Journal of Rural Health* 34: 293–303.

Cohen, S.A., M.L. Greaney, and N.J. Sabik. 2018b. Assessment of dietary patterns, physical activity and obesity from a national survey: Rural-urban health disparities in older adults. *PLoS One* 13 (12): e0208268.

Cummins, S., S. Curtis, A.V. Diez-Roux, and S. Macintyre. 2007. Understanding and representing 'place' in health research: A relational approach. *Social Science & Medicine* 65 (9): 1825–1838.

Dahly, D.L., and L.S. Adair. 2007. Quantifying the urban environment: A scale measure of urbanicity outperforms the urban–rural dichotomy. *Social Science & Medicine* 64 (7): 1407–1419.

Diez-Roux, A.V. 1998. Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health* 88 (2): 216–222.

Dwyer-Lindgren, L., A. Bertozzi-Villa, R.W. Stubbs, C. Morozoff, J.P. Mackenbach, F.J. van Lenthe, and C.J. Murray. 2017. Inequalities in life expectancy among US counties, 1980 to 2014: Temporal trends and key drivers. *JAMA Internal Medicine* 177 (7): 1003–1011.

Eckenrode, J., E.G. Smith, M.E. McCarthy, and M. Dineen. 2014. Income inequality and child maltreatment in the United States. *Pediatrics*. peds-2013.

Elliott, P., and D. Wartenberg. 2004. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* 112 (9): 998.

Erdman, E., A. Liss, D. Gute, C. Rioux, M. Koch, and E.N. Naumova. 2015. Does the presence of vegetation affect asthma hospitalizations among the elderly? A comparison between rural, suburban, and urban areas. *International Journal of Environment and Sustainability* 4 (1).

Fefferman, N., E. O'Neil, and E.N. Naumova. 2005. Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of Public Health Policy* 26 (3): 430–449.

Fiscella, K., and P. Franks. 1997. Poverty or income inequality as predictor of mortality: Longitudinal cohort study. *BMJ* 314 (7096): 1724.

Gilmer, T.P., and R.G. Kronick. 2011. Differences in the volume of services and in prices drive big variations in Medicaid spending among US states and regions. *Health Affairs* 30 (7): 1316–1324.

Gordon-Larsen, P., M.C. Nelson, P. Page, and B.M. Popkin. 2006. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics* 117 (2): 417–424.

Grow, H.M.G., A.J. Cook, D.E. Arterburn, B.E. Saelens, A. Drewnowski, and P. Lozano. 2010. Child obesity associated with social disadvantage of children's neighborhoods. *Social Science & Medicine* 71 (3): 584–591.

Hanchate, A.D., M.K. Paasche-Orlow, K.S. Dyer, W.E. Baker, C. Feng, and J. Feldman. 2017. Geographic variation in use of ambulance transport to the emergency department. *Annals of Emergency Medicine* 70 (4): 533–543.

Haque, A., and J. Telfair. 2000. Socioeconomic distress and health status: The urban-rural dichotomy of services utilization for people with sickle cell disorder in North Carolina. *The Journal of Rural Health* 16 (1): 43–55.

Hart, L.G., E.H. Larson, and D.M. Lishner. 2005. Rural definitions for health policy and research. *American Journal of Public Health* 95 (7): 1149–1155.

Jagai, J.S., J.K. Griffiths, P.H. Kirshen, P. Webb, and E.N. Naumova. 2010. Patterns of protozoan infections: Spatiotemporal associations with cattle density. *EcoHealth* 7 (1): 33–46.

Jia, H., F. Ordóñez, and M. Dessouky. 2007. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions* 39 (1): 41–55.

Juster, R.P., B.S. McEwen, and S.J. Lupien. 2010. Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews* 35 (1): 2–16.

Kahn, R.S., P.H. Wise, B.P. Kennedy, and I. Kawachi. 2000. State income inequality, household income, and maternal mental and physical health: Cross sectional national survey. *BMJ* 321 (7272): 1311.

Kawachi, I., and B.P. Kennedy. 1997. Socioeconomic determinants of health: Health and social cohesion: Why care about income inequality? *BMJ* 314 (7086): 1037.

———. 1999. Income inequality and health: Pathways and mechanisms. *Health Services Research* 34 (1 Pt 2): 215.

Kennedy, B.P., I. Kawachi, and D. Prothrow-Stith. 1996. Income distribution and mortality: Cross sectional ecological study of the Robin Hood index in the United States. *BMJ* 312 (7037): 1004–1007.

Kennedy, B.P., I. Kawachi, R. Glass, and D. Prothrow-Stith. 1998. Income distribution, socioeconomic status, and self rated health in the United States: Multilevel analysis. *BMJ* 317 (7163): 917–921.

Kondo, N., G. Sembajwe, I. Kawachi, R.M. van Dam, S.V. Subramanian, and Z. Yamagata. 2009. Income inequality, mortality, and self rated health: Meta-analysis of multilevel studies. *BMJ* 339: b4471.

Kounadi, O., and M. Leitner. 2014. Why does geoprivacy matter? The scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics* 9 (4): 34–45.

Krieger, N., J.T. Chen, P.D. Waterman, M.J. Soobader, S.V. Subramanian, and R. Carson. 2002. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. *American Journal of Epidemiology* 156 (5): 471–482.

Kruger, D.J., T.M. Reischl, and G.C. Gee. 2007. Neighborhood social conditions mediate the association between physical deterioration and mental health. *American Journal of Community Psychology* 40 (3–4): 261–271.

Kulkarni, S.C., A. Levin-Rector, M. Ezzati, and C.J. Murray. 2011. Falling behind: Life expectancy in US counties from 2000 to 2007 in an international context. *Population Health Metrics* 9 (1): 16.

Lippert, A.M., C.R. Evans, F. Razak, and S.V. Subramanian. 2017. Associations of continuity and change in early neighborhood poverty with adult cardiometabolic biomarkers in the United States: Results from the National Longitudinal Study of Adolescent to Adult Health, 1995–2008. *American Journal of Epidemiology* 185 (9): 765–776.

Lupien, S.J., I. Ouellet-Morin, A. Hupbach, M.T. Tu, C. Buss, D. Walker, et al. 2015. Beyond the stress concept: Allostatic load—A developmental biological and cognitive perspective. *Developmental Psychopathology: Volume Two: Developmental Neuroscience*: 578–628.

Lynch, J.W., G.D. Smith, G.A. Kaplan, and J.S. House. 2000. Income inequality and mortality: Importance to health of individual income, psychosocial environment, or material conditions. *BMJ* 320 (7243): 1200–1204.

Macintyre, S. 1997. What are spatial effects and how can we measure them? In *Exploiting national survey data: The role of locality and spatial effects*, ed. A. Dale, 1–17. Manchester: Faculty of Economic and Social Studies, University of Manchester.

Macintyre, S., A. Ellaway, and S. Cummins. 2002. Place effects on health: How can we conceptualise, operationalise and measure them? *Social Science & Medicine* 55 (1): 125–139.

Manley, D., R. Flowerdew, and D. Steel. 2006. Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems* 30 (2): 143–160.

Marmot, M., and R. Bell. 2011. Social determinants and dental health. *Advances in Dental Research* 23 (2): 201–206.

Matlock, D.D., P.W. Groeneveld, S. Sidney, S. Shetterly, G. Goodrich, K. Glenn, et al. 2013. Geographic variation in cardiovascular procedure use among Medicare fee-for-service vs Medicare advantage beneficiaries. *JAMA* 310 (2): 155–161.

McDonald, D.C., K. Carlson, and D. Izrael. 2012. Geographic variation in opioid prescribing in the US. *The Journal of Pain* 13 (10): 988–996.

McEwen, B.S. 1998. Protective and damaging effects of stress mediators. *New England Journal of Medicine* 338 (3): 171–179.

McGinnis, J.M., P. Williams-Russo, and J.R. Knickman. 2002. The case for more active policy attention to health promotion. *Health Affairs* 21 (2): 78–93.

McLafferty, S.L. 2003. GIS and health care. *Annual Review of Public Health* 24 (1): 25–42.

Mohnen, S.M., B. Völker, H. Flap, and P.P. Groenewegen. 2012. Health-related behavior as a mechanism behind the relationship between neighborhood social capital and individual health-a multilevel analysis. *BMC Public Health* 12 (1): 116.

Naumova, E.N., S.M. Parisi, D. Castronovo, M. Pandita, J. Wenger, and P. Minihan. 2009. Pneumonia and influenza hospitalizations in elderly people with dementia. *Journal of the American Geriatrics Society* 57 (12): 2192–2199.

Newhouse, J.P., and A.M. Garber. 2013a. Geographic variation in Medicare services. *New England Journal of Medicine* 368 (16): 1465–1468.

———. 2013b. Geographic variation in health care spending in the United States: Insights from an Institute of Medicine report. *JAMA* 310 (12): 1227–1228.

Nicholas, L.H., K.M. Langa, T.J. Iwashyna, and D.R. Weir. 2011. Regional variation in the association between advance directives and end-of-life Medicare expenditures. *JAMA* 306 (13): 1447–1453.

O'Keefe, C.M., and D.B. Rubin. 2015. Individual privacy versus public good: Protecting confidentiality in health research. *Statistics in Medicine* 34 (23): 3081–3103.

Openshaw, S. 1984. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*.

Padilla, C.M., W. Kihal-Talantikit, V.M. Vieira, and S. Deguen. 2016. City-specific spatiotemporal infant and neonatal mortality clusters: Links with socioeconomic and air pollution spatial patterns in France. *International Journal of Environmental Research and Public Health* 13 (6): 624.

Park, H., A.M. Roubal, A. Jovaag, K.P. Gennuso, and B.B. Catlin. 2015. Relative contributions of a set of health factors to selected health outcomes. *American Journal of Preventive Medicine* 49 (6): 961–969.

Pearce, J., K. Witten, and P. Bartie. 2006. Neighbourhoods and health: A GIS approach to measuring community resource accessibility. *Journal of Epidemiology & Community Health* 60 (5): 389–395.

Prince, S.A., E.A. Kristjansson, K. Russell, J.M. Billette, M.C. Sawada, A. Ali, et al. 2012. Relationships between neighborhoods, physical activity, and obesity: A multilevel analysis of a large Canadian city. *Obesity* 20 (10): 2093–2100.

Remington, P.L., B.B. Catlin, and K.P. Gennuso. 2015. The county health rankings: Rationale and methods. *Population Health Metrics* 13 (1): 11.

Seeman, T., E. Epel, T. Gruenewald, A. Karlamangla, and B.S. McEwen. 2010. Socio-economic differentials in peripheral biology: Cumulative allostatic load. *Annals of the New York Academy of Sciences* 1186 (1): 223–239.

Stiel, L., S. Soret, and S. Montgomery. 2017. Geographic patterns of change over time in mammography: Differences between Black and White US Medicare enrollees. *Cancer Epidemiology* 46: 57–65.

Subramanian, S.V., and I. Kawachi. 2003. The association between state income inequality and worse health is not confounded by race. *International Journal of Epidemiology* 32 (6): 1022–1028.

———. 2004. Income inequality and health: What have we learned so far? *Epidemiologic Reviews* 26 (1): 78–91.

Suzuki, E., S. Kashima, I. Kawachi, and S.V. Subramanian. 2012. Social and geographic inequalities in premature adult mortality in Japan: A multilevel observational study from 1970 to 2005. *BMJ Open* 2 (2): e000425.

Tarlov, A.R. 1999. Public policy frameworks for improving population health. *Annals of the New York Academy of Sciences* 896 (1): 281–293.

United States Bureau of the Census. Decennial Census 2010. Website: https://data.census.gov/cedsci/advanced. Accessed August 22, 2021.

Vincens, N., and M. Stafström. 2015. Income inequality, economic growth and stroke mortality in Brazil: Longitudinal and regional analysis 2002-2009. *PLoS One* 10 (9): e0137332.

Waldorf, B. 2007. What is rural and what is urban in Indiana. *Purdue Center for Regional Development Report* 4.

Walker, K.E., and S.M. Crotty. 2015. Classifying high-prevalence neighborhoods for cardiovascular disease in Texas. *Applied Geography* 57: 22–31.

Waller, L.A., and C.A. Gotway. 2004. *Applied spatial statistics for public health data*. Vol. 368. Wiley, Hoboken, New Jersey, USA.

Weich, S., G. Holt, L. Twigg, K. Jones, and G. Lewis. 2003. Geographic variation in the prevalence of common mental disorders in Britain: A multilevel investigation. *American Journal of Epidemiology* 157 (8): 730–737.

Wennberg, J.E., E.S. Fisher, and J.S. Skinner. 2002. Geography and the debate over Medicare reform. *Health Affairs* 21 (2): 10–10.

Wilkinson, R.G., and K.E. Pickett. 2006. Income inequality and population health: A review and explanation of the evidence. *Social Science & Medicine* 62 (7): 1768–1784.

Williams, D.R., and C. Collins. 1995. US socioeconomic and racial differences in health: Patterns and explanations. *Annual Review of Sociology* 21 (1): 349–386.

Wilson, W.J. 1987. *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press.

Woolf, S.H., and P. Braveman. 2011. Where health disparities begin: The role of social and economic determinants—And why current policies may make matters worse. *Health Affairs* 30 (10): 1852–1859.

Zandbergen, P.A. 2014. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in Medicine*: 2014.

Zhang, Q., and Y. Wang. 2004. Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Social Science & Medicine* 58 (6): 1171–1180.

Identifying and Visualizing Space-Time Clusters of Vector-Borne Diseases

Michael Desjardins, Alexander Hohl, Eric Delmelle, and Irene Casas

Introduction

Globally, vector-borne diseases (VBDs) are responsible for over 700,000 annual deaths (malaria alone kills more than 400,000 people), accounting for approximately 17% of infectious diseases, and over half of the world's population are at risk of infection (WHO 2017). Mosquitoes are the most common vector and transmit a variety of diseases, such as dengue fever, chikungunya, Zika, malaria, yellow fever, and West Nile fever. During the last few decades, there have been a global increase in VBDs (especially mosquito-borne diseases) due to climate change, increases in globalization and urbanization, human movement, and a general decline in vector control programs. Furthermore, endemic areas have experienced increases in infections, while the ever-expanding geographic range of VBDs has resulted in novel outbreaks in various regions around the world.

For example, before 2013, chikungunya (CHIK) was mostly found in Southeast Asia, Africa, and India. However, CHIK was introduced to the Americas and the Caribbean in 2013, resulting in over a million reported cases within 1 year (Yactayo et al. 2016). Notably, dengue fever (DENF) is the world's most widespread VBD, infecting more than 390 million people per year, while over a third of the world's population are susceptible to transmission (Bhatt et al. 2013; Wilson and Chen 2015). Zika was first discovered in 1947 in Uganda and was relatively rare until the 2014–2016 outbreaks in the South Pacific and Brazil (Dick et al. 1952; Duffy et al. 2009; Campos et al. 2015; Hennessey et al. 2016). Since 2015, over 90 countries around the world are at risk of Zika transmission (CDC 2018). CHIK, DENF, and Zika are spread by the peridomestic container-breeding *Aedes aegypti* and *Aedes albopictus* mosquitoes, which also transmit yellow fever.

It is critical to implement surveillance strategies that can improve the understanding of VBD transmission. VBD surveillance may involve the examination of disease incidence in human populations, including the (spatial) variations among socioeconomic groups, age, and sex; the geographic distribution of vector populations capable of transmitting various VBDs, especially identifying suitable habitats (e.g., environmental variables); and analyzing human movement and interaction with their environment that may facilitate disease transmission (Palaniyandi et al. 2017). Furthermore, identifying significant (space-time) clusters of disease cases is typically the primary stage of surveillance, while the domain of geographic information science can greatly facilitate the monitoring of VBDs.

There is an inherent link between place and health outcomes, and geographic information science (GIScience) plays a vital role in VBD surveillance (Eisen and Eisen 2011; Blatt 2015). For example, mapping the spatial variation in disease rates and risk is vital for formulating etiological hypotheses (Delmelle et al. 2016). GIScience can facilitate the detection and visualization of VBD outbreaks in space and time, which can improve targeted interventions to mitigate outbreaks, such as improving healthcare accessibility and vector control strategies (Delmelle et

M. Desjardins (✉) · E. Delmelle
Spatial Science for Public Health Center, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Department of Geography and Earth Sciences, University of North Carolina at Charlotte, Charlotte, NC, USA
e-mail: mdesjar3@jhmi.edu; Eric.Delmelle@uncc.edu

A. Hohl
Department of Geography, The University of Utah, Salt Lake City, UT, USA
e-mail: u6025895@utah.edu

I. Casas
School of History and Social Sciences, Louisiana Tech University, Ruston, LA, USA
e-mail: icasas@latech.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_11

al. 2011, 2014a, b; Kienberger et al. 2013). Space-time approaches in GIScience can increase the timeliness of public health decision-making by examining the severity and duration of outbreaks (Duncombe et al. 2012), seasonality, and risk of diffusion (Khormi and Kumar 2015) and identify populations with an elevated risk of VBD transmission (Kitron 2000).

Disease data that can be used for space-time analyses is available at the disaggregated or aggregated level (Cromley and McLafferty 2011). Disaggregated data is represented by points, such as the location of individual disease cases. Aggregated data at a geographic unit level (e.g., counties, towns, or neighborhoods) typically reflect rates. This chapter focuses on exploratory space-time cluster detection approaches for both disaggregated and aggregated levels and 3D visualization techniques to improve the understanding of space-time dynamics of disease clusters for VBD surveillance. However, the methodologies and concepts presented in this chapter can also be applied to other infectious diseases and other domains such as criminology.

The remainder of this chapter is as follows: section "Spatiotemporal Methods for Vector-Borne Disease Surveillance: Strengths and Limitations" describes common approaches in GIScience to detect space-time clusters of disease for both disaggregated and aggregated data. Section "Spatiotemporal Methods for Vector-Borne Disease Surveillance: Strengths and Limitations" explains the mechanisms of some of the most widely used exploratory space-time clustering approaches in the literature. Strengths and limitations of each approach are also discussed. Section "Visualizing Space-Time Clusters" sheds light on techniques to visualize space-time clusters in both 2D and 3D. Section "Case Study: Chikungunya and Dengue Outbreaks in Colombia (2015–2016)" provides a case study of space-time clusters of VBDs in Colombia (aggregated at the municipality level), which were detected using the univariate and multivariate space-time scan statistic (Kulldorff et al. 2005). The resulting space-time clusters are also visualized in 2D and 3D using the techniques described in section "Visualizing Space-Time Clusters". Finally, section "Conclusions" provides concluding remarks.

Spatiotemporal Methods for Vector-Borne Disease Surveillance: Strengths and Limitations

Space-Time Ripley's K Function

Many spatial analysis methods solely focus on the geographic distribution of the phenomenon under study while neglecting its temporal aspects: They either ignore or collapse the temporal dimension (Bach et al. 2017) or discretize time to a number of time slices for which again time is collapsed (Boyandin et al. 2012). However, such approaches fail to represent time as a continuous dimension, which is crucial for analyzing spatiotemporal patterns of disease outbreaks. The space-time Ripley's K function estimates the second-order property (variance) of a set of spatiotemporal points, i.e., disease cases. The resulting statistic depends on (1) the number and (2) distance between the points and returns the degree by which the observed point pattern deviates from randomness for multiple spatiotemporal scales (Bailey and Gatrell 1995; Dixon 2013). In theory, the K function is calculated by Eq. 1, i.e., the division of E, the expected number of points within spatial and temporal bandwidth (d and t, respectively), by the intensity λ (first-order property) of a set of points S.

$$K\left(d,t\right) = E\left(d,t\right)/\lambda \tag{1}$$

The spatial and temporal bandwidths form cylinders of radius d and height t, centered on each data point. Dividing the total number of observed points n by the product of the study area A and the study period T results in estimated intensity λ. Computing Ripley's K function equates to counting all points within the cylinders and repeating this process with increasing spatial and temporal bandwidths d and t (cylinders of increasing size). Thereby, we expect $K(d,t) = \pi d^2 t$ if the point pattern exhibits complete spatiotemporal randomness (CSTR), $K(d,t) > \pi d^2 t$ if the pattern shows clustering within spatial and temporal distances d and t, and $K(d,t) < \pi d^2 t$ if the pattern is regular. In practice, Eq. (2) is used to compute the space-time Ripley's K function:

$$K\left(d,t\right) = \frac{L * R}{n^2} \sum_{i}^{n} \sum_{j}^{n} \frac{I_{h,t}\left(d_{ij}, t_{ij}\right)}{w_{ij}} \tag{2}$$

where d_{ij} is the distance between events i and j. The term w_{ij} is a factor to correct for edge effects, which potentially bias the outcome of the K function, when cylinders intersect the boundary of the study area or period. Methods for edge correction are well studied (Yamada and Rogerson 2003; Gabriel 2012). $I_h(d_{ij})$ indicates whether a point i locates within the cylinder or not (Eq. 3):

$$I_{h,t}\left(d_{ij}, t_{ij}\right) = \begin{cases} 1 \text{ if } d_{ij} \leq h \text{ AND } t_{ij} \leq t, \\ 0 \quad \text{otherwise} \end{cases} \tag{3}$$

Since it is a cumulative measure, the space-time K function values increase with increasing spatial and temporal bandwidths (d and t, respectively). Statistical confirmation to distinguish regular, clustered, or random patterns may require Monte Carlo simulation: The space-time K function is

evaluated for a large number (M) of simulated point sets. For each simulation, N points (equal to the number of observed points) are randomly generated within the study area/period. For given values of d and t, if the observed K value is larger than the upper simulation envelope (the largest value of K among all simulations, at given values of d and t), clustering for the corresponding spatiotemporal bandwidths is statistically significant. Observed K values below than the lower simulation envelope indicate regularity. Hence, if the K function value is above, between, or below the upper and lower simulation envelopes, the point pattern is clustered, random, or regular, respectively, for given values of d and t.

Using Eq. (4), the K function can be transformed to the L function to obtain constant variance with respect to a benchmark of zero, facilitating the comparison of values across all d and t:

$$L(d,t) = [K(d,t)/\pi t]^{1/2} - d \tag{4}$$

where $L(d, t) = 0$ under CSTR, $L(d, t) > 0$ for clustered patterns, and $L(d, t) < 0$ for regular patterns. Ripley's K function exists in global and local forms (Anselin 1995): The global form produces one graph that indicates the scale at which the point pattern is significant for the entire study area/period, while its local form allows to pinpoint where exactly such a pattern occurs (Hohl et al. 2017).

Many different studies have employed the space-time Ripley's K function for point pattern analysis: outbreaks of dengue fever (Hohl et al. 2016), patterns of legionnaire's disease (Diggle et al. 1995), and human *Campylobacter* infections (Gabriel and Diggle 2009) and interactions between forest fire and spruce budworms (Lynch and Moorcroft 2008), among others. Recent methodological advances that allow for computing Ripley's K function include flow data (Tao and Thill 2016), analyzing network-constrained point patterns (Yamada and Thill 2007), four-dimensional data (3D + time, Hohl et al. 2018), and handling massive datasets (Tang et al. 2015). These examples illustrate the wide realm of its applicability, making Ripley's K function one of the most important methods for characterizing space-time patterns of any geospatial phenomena, including VBDs.

Space-Time Knox Test

In the study of epidemiology, Knox test for space-time interaction (Knox 1963, 1964) is used to evaluate whether there is space-time clustering of disease cases within a given study area. It is of interest if cases are more clustered than what would be expected based on the underlying geographical population distribution or by a purely temporal trend (Kulldorff and Hjalmars 1999). The null hypothesis of Knox test states that spatial proximity of two cases is independent of their temporal proximity. In other words, no space-time interaction is observed in the case data.

The test statistic X is the number of pairs of cases that are near to one another in both space and time, given some user-specified spatial and temporal distance thresholds (s and t, Eq. 5).

$$X = \sum_{i=1}^{N} \sum_{j=1}^{i-1} a_{ij}^{s} a_{ij}^{t} \tag{5}$$

where N is the total number of cases; a_{ij}^{s} is an indicator function that evaluates to 1, if i and j are close in space (their distance is less than threshold s); and a_{ij}^{t} is an indicator function that evaluates to 1, if i and j are close in time (their distance is less than threshold t). One can use either Poisson approximation or Monte Carlo simulation for significance testing (Mantel 1967).

Knox test for space-time interaction has been employed to study cleft lip and cleft palate birth defects (Knox 1963), childhood leukemia in the UK (Knox 1964), and the 1991–1992 outbreak of dengue fever in Florida (Morrison et al. 1998), among many other applications. However, Knox method is limited due to its arbitrary definition of closeness, as users have to specify the spatial and temporal distance thresholds (Robertson et al. 2010). This may not be an issue for well-studied diseases, where closeness could be defined by disease transmission distance. However, a poor choice of such parameters may invalidate the findings of the analysis (Aldstadt 2007). Preliminary analysis of the space-time Ripley's K function to define the distance thresholds for Knox test is a practicable workaround for this issue (Delmelle et al. 2011).

Mantel's Test

The shortcomings of Knox test are addressed by Mantel's test (Mantel 1967), which allows to introduce the notion of distance decay, where pairs of disease cases within close space-time proximity are more important than pairs of far proximity. Mantel's test generalizes the indicator function terms a_{ij}^{s} and a_{ij}^{t}, used by Knox method (Eq. 5), to any suitable distance measures, including raw or transformed Euclidean distances (Meyer et al. 2016). For instance, Jacquez (1996) introduces a standardized form of Mantel's test (Eq. 6):

$$r = \frac{1}{(N^2 - N - 1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\left(d_{ij}^{s} - \overline{d}^{s}\right)}{S^{s}} \frac{\left(d_{ij}^{t} - \overline{d}^{t}\right)}{S^{t}} \tag{6}$$

where N is the total number of cases, d_{ij}^{s} and d_{ij}^{t} are the (spatial and temporal) distances between cases i and j, $\overline{d}^{s}$ and $\overline{d}^{t}$ are the average distances, and S^{s} and S^{t} are their standard deviations. The null hypothesis states that spatial and temporal distances are independent of each other (no interaction). It is tested by creating a reference distribution using a Monte Carlo approach, where either the spatial distances are repeatedly permuted while the temporal distances are left untouched or vice versa. The test statistic r is calculated for each permutation, and the p-value for significance is the rank of the observed r among the simulated ones. Mantel's test has been applied to study psychiatric inpatient admissions in Switzerland (Meyer et al. 2016), human immunodeficiency virus (HIV) infections and the development of aids (Shankarappa et al. 1999), fire outbreaks in boreal black spruce forests in northern Quebec (Jacquez 1996), and vegetation studies related to the concept of climax (McCune and Allen 1985) and the environmental control model (Burgman 1987).

Space-Time Kernel Density Estimation

Space-time kernel density estimation (STKDE) is a method that explicitly incorporates the temporal dimension and is popular for visualizing patterns of point events characterized with spatial information (x, y) together with a timestamp t. STKDE generates density values estimated along a 3D grid of voxels (**vo**lumetric pi**xels**), by weighting surrounding events that also have spatial and temporal coordinates, by a distance-decay relationship (Hohl et al. 2016). STKDE may be used in conjunction with the space-time cube framework (Nakaya and Yano 2010a, b), which employs a 3D geographic space, where the vertical dimension is substituted for time, meaning cases that were observed early (late) during the study period are displayed at lower (higher) altitude. This combination allows for visualization of density estimates to detect spatiotemporal patterns of disease occurrence. The space-time kernel density is estimated by Eq. (7):

$$\hat{f}(x, y, t) = \frac{1}{n h_s^2 h_t} \sum_i k_s\left(\frac{d_i}{h_s}\right) k_t\left(\frac{d_t}{h_t}\right) \qquad (7)$$

Density estimates $\hat{f}(x, y, t)$ are calculated for voxels with coordinates *(x, y, t)* and depend on the spatiotemporal distribution of disease cases i. Each data case i contributes to density at surrounding voxels, based on case-voxel distance (spatial and temporal components d_i and d_t, respectively). These distances are plugged into spatial (k_s) and temporal (k_t) kernel functions to obtain the density contribution of case i to voxel s: close proximity yields large contribution, distance proximity yields small contribution, and voxels that are located further away from i than the spatial (h_s) and temporal (h_t) bandwidths receive zero weight. This process is best explained by the metaphor of the moving cylinder: A cylinder is defined by the radius of the circle at its base (d_i) and its height (d_t). Such a cylinder is centered on a case within the space-time cube, and only voxels inside it receive contribution to their density from the respective disease case. The cylinder keeps moving from case to the case until it has "visited" all of them, which terminates the process. The grid of voxels is now complete: every voxel now holds a density estimate that is determined by neighboring disease cases.

Popular kernel functions include Epanechnikov, Quartic, and Gaussian (Silverman 2018). h_s and h_t have a profound influence on the characteristics of the resulting visualization: Large bandwidths yield smooth density volumes, while small bandwidths result in rough density volumes (see Saule et al. 2017, Fig. 1). Optimal values of h_s and h_t, i.e., the spatial and temporal scales at which clustering is strongest, are obtained by prior analysis of the space-time Ripley's K function (see section "Space-Time Knox Test"). STKDE is a widespread method for characterizing space-time patterns and has been employed in many different settings: for the study of dengue fever (Delmelle et al. 2014a, b; Hohl et al. 2016), crime (Nakaya and Yano 2010b), and cell phone activity (Sagl et al. 2014). It has been improved and extended, allowing to go beyond merely analyzing point datasets: Recent methodological advances allow for analyzing network-constrained events (Xie and Yan 2008), animal movement (Demšar and Virrantaus 2010), patterns of vessel activity (Scheepens et al. 2011), and trajectories of hurricanes (Eaglin et al. 2017) and to enable high-performance parallel processing for big data handling (Saule et al. 2017). This is an incomplete collection of STKDE examples, and applications are widespread in many different domains.

The Space-Time Scan Statistic (STSS)

Within the context of spatial epidemiology, scan statistics are one of the most common approaches to identify statistically significant clusters of disease. For the purpose of this chapter, we are referring to discrete scan statistics (see Kulldorff 2018 for continuous models). Scan statistics compare the number of observed and expected cases in a defined area, while the expected cases are typically proportional to the at-risk host population in a given area (Kulldorff 1997). STSS (Kulldorff et al. 2005) were developed to examine the space-time dynamics of disease transmission (e.g., size and duration). The STSS systematically move cylinders of different space-time dimensions across the geographic and temporal space, while the cylinders are typically centered on the centroid of an areal unit (e.g., municipality), but disaggregated, individual-level observations may be used. Furthermore, the base of

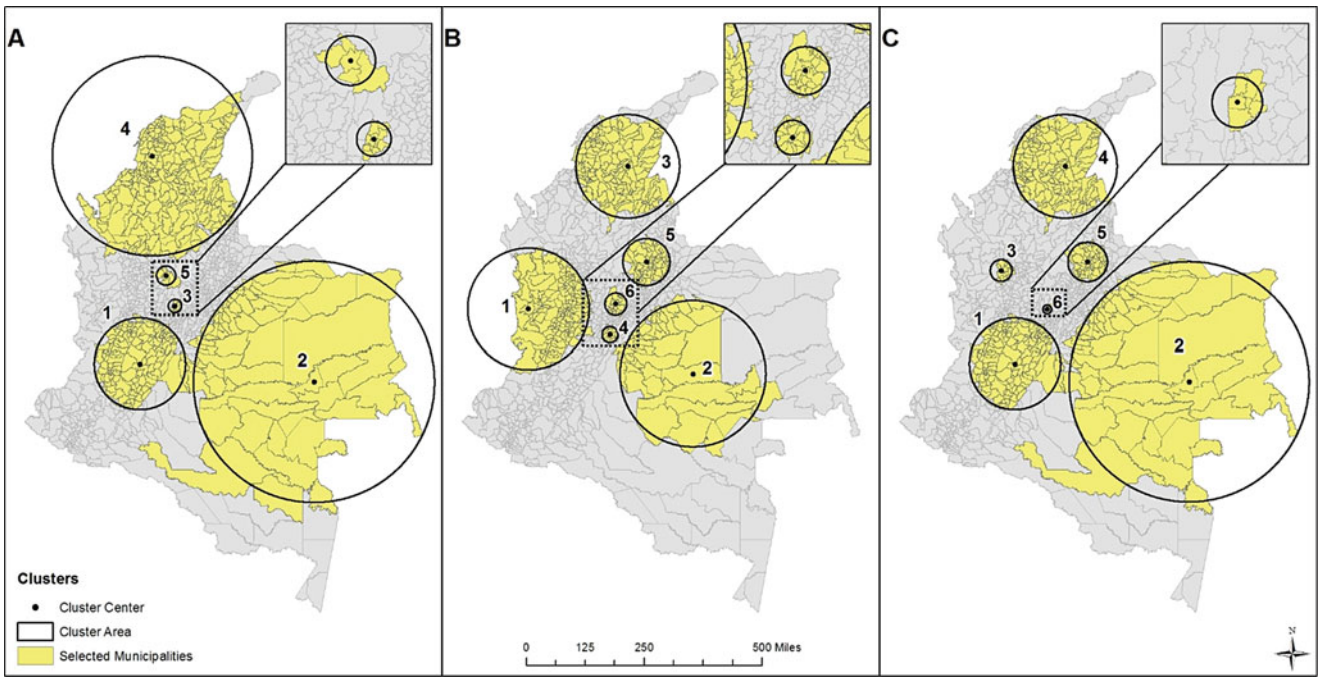


Fig. 1 Significant space-time clusters of (**a**) CHIK, (**b**) DENF, and (**c**) Multivariate

Table 1 Probability models for STSS (Kulldorff 2018)

Data	Example	Model	Reference
Count	(1) Case only (2) Cases/controls (3) Cases/total population	(1) Space-time permutation (2) Bernoulli (3) Discrete Poisson	(1) Kulldorff et al. (2005) (2) and (3) Kulldorff (1997)
Categorical	Disease with multiple types (dengue)	Multinomial	Jung et al. (2010)
Ordered categorical	Cancer stages	Ordinal	Jung et al. (2007)
Survival time	AIDS cases over a 10-year period	Exponential	Huang et al. (2007)
Other – Continuous	Weight	Normal	Kulldorff et al. (2009)

each cylinder is the spatial scan, while the height reflects the temporal scan. The number of observed and expected cases are counted within each cylinder and compared to the observed and expected outside of the cylinder. Conceptually, an infinite number of overlapping cylinders are produced until the entire study region and period is covered, and a user-defined maximum spatial and temporal scan is reached (each cylinder is a potential space-time cluster). Although cylinders are the most common space-time scanning methods, it is possible to employ irregularly shaped windows (see Duczmal and Assuncao 2004; Tango and Takahasi 2005; Ullah et al. 2017).

STSS use a variety of probability models, depending on the characteristics of the dataset. Table 1 provides the appropriate model that should be employed for different types of data. For this chapter, the discrete Poisson probability model employing cylindrical windows will be used.

The discrete Poisson probability model assumes that the disease cases follow a Poisson distribution according to an area's population. The null hypothesis (H_0) states that the model reflects an inhomogeneous Poisson process with an intensity μ, which is proportional to the at-risk population. The alternative hypothesis (H_A) states that the number of observed disease cases exceeds the number of expected cases derived from the null model (elevated risk within a cylinder). A maximum likelihood ratio test evaluates H_0 and H_A and is defined in Eq. 8, while the parameters for Eq. 8 are defined in Table 2.

$$\frac{L(Z)}{L_0} = \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{N-n_Z}{N-\mu_{(Z)}}\right)^{N-n_Z}}{\left(\frac{N}{\mu(A)}\right)^{N}} \tag{8}$$

A cylinder Z with a likelihood ratio greater than 1 denotes an elevated risk compared to the outside of the cylinder, that is, $\frac{n_Z}{\mu(Z)} > \frac{N-n_Z}{N-\mu_{(Z)}}$. As the STSS is repeated over different cylinder sizes, the one with the maximum likelihood ratio constitutes the most likely cluster (the cluster that is least likely to have occurred by chance). Secondary clusters are

Table 2 Parameters for the maximum likelihood ratio test for the discrete Poisson model

Parameters	Definition
$L(Z)$	Likelihood function for cylinder Z
L_0	Likelihood function for H_0
n_Z	Observed cases in cylinder Z
$\mu(Z)$	Expected cases in cylinder Z
N	Total number of observed cases in the study area across all time periods
$\mu(A)$	Total number of expected cases in the study area across all time periods

also reported if they are statistically significant. The higher the likelihood ratio, the higher relative risk the cluster can be described as having, as it displays the strongest statistical evidence of clustering. To assess the statistical significance, Monte Carlo testing returns a p-value for each candidate cluster, essentially comparing simulated datasets to the real dataset (recommended minimum of 999 simulations).

The relative risk of the locations belonging to a statistically significant cluster can also be reported and is recommended to identify the highest-risk areas inside the cluster. Relative risk is derived in Eq. 9 and is defined as the risk of infection in a target location compared to all surrounding locations in a study area.

$$\mathrm{RR} = \frac{c/e}{(C-c)/(C-e)} \tag{9}$$

The total observed cases in a target location is c; the total expected cases in the target location is e; and C is the total observed cases in the entire study area and period.

Since many VBDs can be transmitted by the same vector (e.g., dengue and chikungunya – *A. aegypti* and *A. albopictus*), identifying space-time clusters where multiple VBDs co-occur can be beneficial for targeted intervention programs. Conversely, the multivariate STSS can identify the simultaneous excess incidence of two or more diseases (Kulldorff et al. 2007). The multivariate STSS follows the same procedure as the univariate STSS and then sums the log likelihood ratio (LLR) for each disease within a scanning window, producing a new LLR for each candidate space-time cluster. The maximum LLR is the most likely multivariate cluster, while secondary clusters are also reported if they are statistically significant. Within the context of VBD research, the univariate STSS has been utilized to examine outbreaks of malaria (Gaudart et al. 2006; Coleman et al. 2009), Lyme disease (Li et al. 2014), chikungunya (Nsoesie et al. 2015), West Nile (Mulatti et al. 2015), and dengue (de Melo et al. 2012; Li et al. 2012; Banu et al. 2012), for example. The multivariate STSS has been used to examine simultaneous clusters of both chikungunya and dengue in Colombia (Desjardins et al. 2018a).

Despite the strengths of STSS, there are a variety of limitations worth mentioning. First, although the cylindrical search window is most commonly used, it does not preserve the true shape of the outbreaks. As previously mentioned, it is possible to implement irregularly shaped scanning windows. However, the STSS is an exploratory technique to detect space-time clusters, and the cylindrical search window is widely accepted as a valid approach and can still detect noncylindrical outbreaks (Kulldorff 2005). Second, there can be uncertainty in the expected counts especially with a large number of temporal observations. Specifically, the population data that influences the expected counts at each data location will likely be static throughout the study period. Therefore, population dynamics such as seasonal trends (e.g., tourism) and migration are not considered and may result in higher or lower relative risk estimates. Third, relative risk will likely vary during the study period, while the STSS reports the total relative risk for the entire study period. Finally, the STSS can be computationally demanding due to the high number of Monte Carlo simulations and the number of data locations and temporal observations and is further exacerbated when the multivariate approach is utilized.

Visualizing Space-Time Clusters

Space-time clusters of diseases are generally visualized in three different manners. First, and this is the most common approach, clusters are highlighted on a map, and information on the beginning and end dates of the clusters are indicated on the same map or in a separate table. Examples of such approach are provided in Norström et al. (2000), Sheehan and DeChello (2005), Nagar et al. (2014), Iftimi et al. (2015), Nsoesie et al. (2015), and Xu and Wu (2018). The second approach consists of producing small multiples, where each map reflects a particular time period (e.g., weeks, months, year). In this sequential approach, maps with cluster information are arranged in a mosaic framework, side-by-side (Dorling 1992; Brunsdon et al. 2007). Examples of the sequential approach are provided in Gaudart et al. (2006), Onozuka and Hagihara (2007), Coleman et al. (2009), Banu et al. (2012), Pereira et al. (2015), Mulatti et al. (2015), and Scripcaru et al. (2017). The third approach – and this is the one we use in this paper – visualizes clusters in a space-time framework, following Hägerstrand's time geography concept. Using a 3D framework, the Y-axis is used to reflect the temporal dimension. Explicit examples of space-time clusters in a 3D environment include Nakaya and Yano (2010) and Desjardins et al. (2018a) mapping crime in Kyoto (Japan) and dengue fever and chikungunya epidemics in Colombia. Although the first approach is relatively straightforward, the temporal component is not explicitly represented, and it remains challenging to visualize (1) the duration of space-time clusters

and (2) whether specific clusters occur at the same time. The second approach addresses these shortcomings, yet the reader must cognitively reconstruct the temporal dimension to better understand the dynamics of the clusters. The users must move from one image to another and reconstruct the movement of the clusters, although maps can be animated to reconstruct cluster duration. The third approach can improve our understanding of the space-time patterns of a disease including duration and size and also how clusters may move through time.

Case Study: Chikungunya and Dengue Outbreaks in Colombia (2015–2016)

Background

Colombia is located in the northwest corner of South America with a population of over 41 million people as of 2018 (DANE 2018). Colombia has various altitudinal zones, but around 80% of the country is classified as Tierra Caliente with temperatures above 24 °C (World Mosquito Program 2018). As a result, more than 90% of the country is below 2200 m, resulting in the perfect habitat for the *Aedes Aegypti* and *Aedes Albopictus* mosquitoes (vectors of DENF and CHIK). Therefore, the majority of Colombia's population is at risk of contracting DENF and CHIK, as well as other VBDs transmitted by *Aedes* such as yellow fever and Zika. In the case of DENF, during the 1950s and 1960s, several measures were adopted to eradicate the presence of mosquitoes including fumigation and the elimination of mosquito foci (Dick et al. 2012). However, a re-infestation occurred in the 1970s, and DENF has since been endemic in certain areas of the country. Periodic DENF outbreaks have occurred in the last 20 years in 2001, 2006, 2010, 2013, and 2016 (Restrepo et al. 2014; Ocampo et al. 2014; Cali 2010; Villegas et al. 2010). CHIK first emerged in Colombia in 2014, with the first reported cases in late July with 45 notified cases, while 16 classified as suspicious (INS 2014). By the end of 2014, there were 106,763 reported CHIK cases that were in found in 30 departments. This marked the beginning of an epidemic, which extended to mid-2016 (INS 2016). In 2015, there were 43,787 cases according to data retrieved from the National Institute of Health, covering 30 departments and several urban centers. Most of the cases for 2016 were reported in the first half of the year. By the end of 2016, a total of 12,187 cases were reported, while almost half of the cases occurred in four departments (Valle del Cauca, Santander, Tolima, and Risaralda), and more than 60% of the cases were woman.

Data

Data for the case study corresponds to the number of DENF and CHIK cases per municipality in Colombia for 2015 and 2016. The data was obtained from SIVIGILA (Sistema Nacional de Vigilancia en Salud Pública – National Public Health Surveillance System).[1] SIVIGILA is a system administered by the National Institute of Health of Colombia (Instituto Nacional de Salud – INS) which has as a primary goal to provide information regarding events that can affect the health of the Colombian population in a timely manner. Data is uploaded into the system by the UPGDs (from their acronym in Spanish: Unidades Primarias Generadoras de Datos – Data Generating Primary Units) on a weekly basis. UPGDs are defined as any private or public entity that diagnoses the occurrence of a public health event of interest (INS 2018a).

The INS makes SIVIGILA data available at the aggregate level through their Routine Surveillance webpage (INS 2018b). The aggregate summaries contain weekly disease cases for each municipality including suspicious, probable, and confirmed cases. A probable DENF case is identified as exhibiting fever with two or more of the following symptoms: headache, retroocular pain, myalgia, arthralgia, and rash (INS 2018c). A suspicious CHIK case is identified as a patient residing or visiting a healthcare facility 8–15 days prior to the onset of symptoms in a municipality where there have not been laboratory cases of CHIK confirmed, including the following symptoms: running a fever over 38 °C, arthralgia or arthritis, uniform erythema, or symptoms that cannot be explained by other medical conditions (INS 2018d). The data stored in the SIVIGILA system is described as dynamic, subject to analysis, and adjustment (this means data is revised as more information becomes available). Population data was obtained from the Geographic Information System for Planning and Land Use Ordering of Colombia (SIGOT: Sistema de Información Geográfica para la Planeación y Ordenamiento Territorial).[2] The data contains population totals for each municipality in 2015 and 2016.

The dataset includes 43,452 CHIK cases in 2015 and 11,964 CHIK cases in 2016. For DENF, there are 94,856 cases in 2015 and 99,703 cases in 2016. The CHIK and DENF data for this case study is similar to Desjardins et al. (2018a). However, this study received updated CHIK and DENF case counts from INS, resulting in a discrepancy between the number of reported total CHIK and DENF cases during 2015 and 2016. This discrepancy can be explained by the dynamic reporting of cases by the national health surveil-

[1] http://www.ins.gov.co/lineas-de-accion/Subdireccion-Vigilancia/sivigila/Paginas/sivigila.aspx. Last accessed 29 Sept 2018

[2] http://sigotn.igac.gov.co/sigotn/frames_pagina.aspx. Last accessed 29 Sept 2018

Table 3 Space-time clusters of CHIK (RR: relative risk)

Cluster	Duration (weeks)	*p*-value	Observed	Expected	*RR*	Municipalities	Population in cluster
1	1–26	<0.01	20,407	2172.03	14.29	144	7,671,766.9
2	15–24	<0.01	5520	341.45	17.84	74	1,575,080.1
3	4–29	<0.01	984	23.99	41.75	10	84,561.4
4	1–5	<0.01	2292	660.00	3.58	255	12,151,474.5
5	7–10	<0.01	328	5.02	65.78	5	115,174.4

lance system (SIVIGILA). For example, CHIK, DENF, and Zika have similar symptomology, which can make initial diagnosis difficult. Furthermore, there can be uncertainty and delays in reporting to SIVIGILA from the UPGDs, especially since CHIK was a novel disease in Colombia at the time of the epidemic. The data for this paper was retrieved in December of 2017, a year after the epidemic was over, while the data in Desjardins et al. (2018a) was retrieved shortly after the year was over for 2015 and 2016.

Methods

For this case study, a discrete Poisson univariate and multivariate STSS with cylindrical scanning windows was used to identify individual and simultaneous space-time clusters of CHIK and DENF in Colombia, during the 2015 and 2016 outbreaks. For both the univariate and multivariate STSS, the parameters were set as followed: (1) a maximum spatial scan of 25% of the at-risk population; (2) a maximum temporal scan of 25% of the study period; (3) a minimum temporal duration of 2 weeks; and (4) 999 Monte Carlo simulations. Clusters with a *p*-value <0.05 were reported. The relative risk of Colombian municipalities that belong to a significant space-time cluster is also reported to facilitate targeted interventions. Finally, the space-time clusters are also visualized in 3D using the space-time cube approach, which was implemented in ArcScene™ 10.6.

CHIK Results

Five space-time clusters of CHIK were reported, which included 490 of the 1125 contiguous Colombian municipalities (Fig. 1a). All five clusters occurred in 2015, with cluster centers in Ataco, Tolima Department (cluster 1: weeks 1–26); Mapiripana, Guainía Department (cluster 2: weeks 15–24); La Peña, Cundinamarca Department (cluster 3: weeks 4–29); San Jacinto, Bolivar Department (cluster 4: weeks 1–5); and Puerto Nare, Antioquia Department (cluster 5: weeks 7–10). The characteristics of the five CHIK clusters are provided in Table 3.

Clusters 1 (most likely cluster) and 3 had the longest duration of 25 weeks each, while cluster 5 had the highest relative risk of 65.78. It is more informative to report the relative risk of the locations that belong to a cluster, and Fig. 2a shows the relative risk of CHIK for the 490 selected municipalities. Notably, 194 of the 490 municipalities reported a relative risk greater than 1, which indicates that there were more CHIK cases than expected. Conversely, 296 municipalities had more expected than observed cases (RR between 0 and 1); and 73 municipalities had no observed cases of CHIK (RR = 0). Cali, Valle del Cauca Department (cluster 1), had the most observed cases (4178) during the study period with a relative risk of 1.59. Roldanillo, Valle del Cauca Department (cluster 1), had the highest relative risk of 87.6, with 3078 observed and 37.17 expected cases.

DENF Results

Six space-time clusters of DENF were reported, which included 474 of 1125 Colombian municipalities (Fig. 1b). Clusters 2 (weeks 1–24) and 3 (weeks 35–53) occurred in 2015, with centers in Mapiripán, Meta Department, and Astrea, Cesar Department, respectively. The last week of cluster 3 occurred during the first week of January 2016. Clusters 1 (weeks 54–79), 4 (weeks 52–77), 5 (weeks 52–77), and 6 (weeks 52–74) occurred in 2016, with centers in Medio Baudó, Chocó Department; Tibacuy, Cundinamarca Department; Hato, Santander Department; and El Peñón, Antioquia Department, respectively. Table 4 provides detailed characteristics of the six space-time DENF clusters.

Clusters 1 (most likely cluster), 4, and 5 had the longest duration of 25 weeks each, while cluster 4 had the highest relative risk of 6.95. Figure 2b depicts the relative risk for each of the 474 municipalities belonging to a space-time DENF cluster. Out of the 474 municipalities, 211 contained a relative risk greater than 1, 1263 had more expected that observed cases (RR between 0 and 1), and 18 had no observed DENF cases (RR =0). Cali, Valle del Cauca Department (cluster 1), contained the most observed cases of DENF during the study period ($n = 33{,}748$), with a relative risk of 4.08. Soatá, Boyacá Department, had the highest relative risk (RR = 17.53) with 500 observed and 28.59 expected cases. Notably, Medellín, Antioquia Department, belongs to cluster 1 with 20,990 observed cases (RR = 2.25).

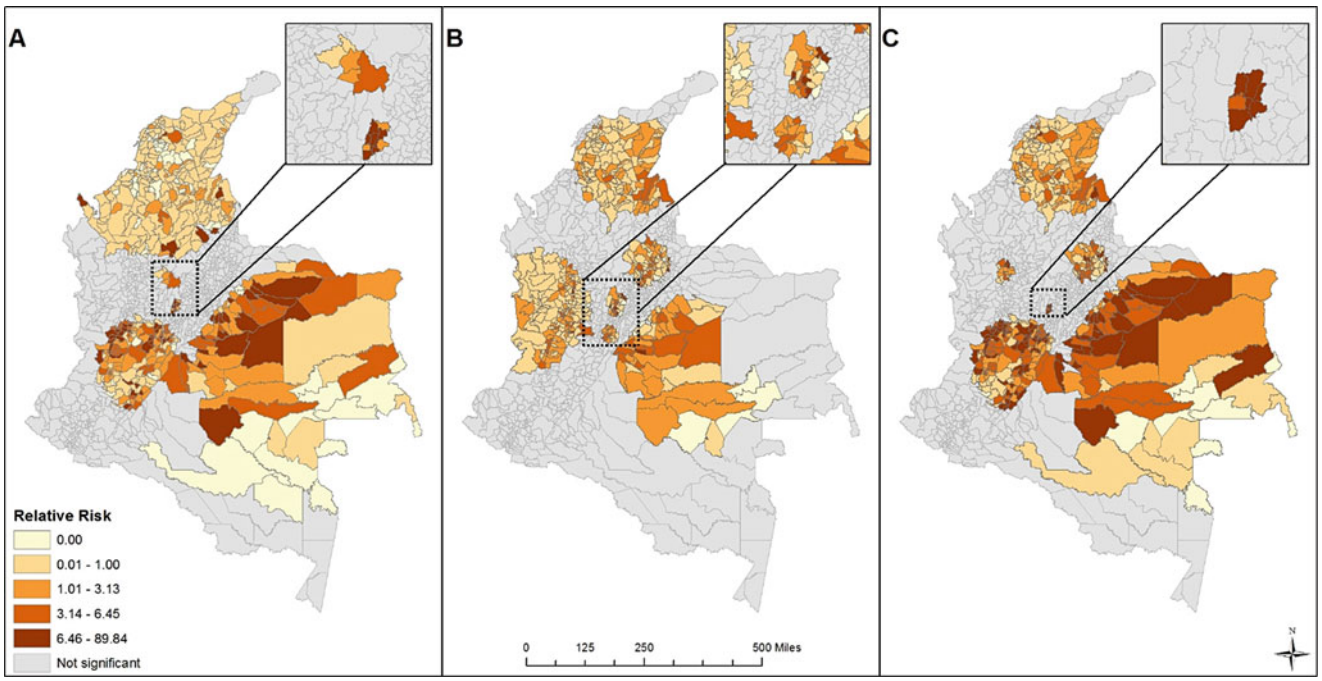


Fig. 2 Relative risk per municipality for (**a**) CHIK, (**b**) DENF, and (**c**) Multivariate

Table 4 Space-time clusters of DENF (RR: relative risk)

Cluster	Duration (weeks)	*p*-value	Observed	Expected	RR	Municipalities	Population in cluster
1	54–79	<0.01	39,363	11,638.94	3.99	165	11,610,411.86
2	1–24	<0.01	5903	1239.48	4.88	52	1,357,165.98
3	35–53	<0.01	8971	3169.72	2.92	140	4,362,901.13
4	52–77	<0.01	2452	356.42	6.95	16	355,191.67
5	52–77	<0.01	5527	1894.43	2.97	74	1,893,281.25
6	52–74	<0.01	1118	208.43	5.39	25	235,570.56

Multivariate Results

Six space-time clusters were reported after running the multivariate STSS, which included 440 of the 1125 municipalities (Fig. 1c). Four of the clusters occurred in 2015, cluster 1 (weeks 1–26), cluster 2 (weeks 4–24), cluster 4 (weeks 35–53), and cluster 6 (weeks 2–27), with centers in Ataco, Tolima Department; Mapiripana, Guainía Department; Astrea, Cesar Department; and Quebradanegra, Cundinamarca Department, respectively, while two occurred in 2016, cluster 3 (weeks 65–90) and cluster 5 (weeks 54–77), with centers in Anzá, Antioquia Department, and Cabrera, Cundinamarca Department, respectively. Furthermore, the last week of cluster 4 occurred in this first week of January 2016. Table 5 sheds light on the characteristics of the six multivariate space-time clusters of CHIK and DENF.

The multivariate STSS will report significant clustering for one or more datasets; therefore, the multivariate results may include clusters that only contain CHIK or DENF. For this study, four of the clusters included simultaneous clustering of both CHIK and DENF (clusters 1, 2, 5, and 6), while two of the clusters only included significant clustering of DENF (clusters 3 and 4). Clusters 1, 3, and 6 had the longest duration of 25 weeks each. Cluster 1 had the highest observed cases for both CHIK (n = 20,407) and DENF (n = 21,808), while cluster 6 had the highest relative risk of CHIK (RR = 55.33) and DENF (RR = 7.19). Cali, Valle del Cauca Department (cluster 1), contained the most combined observed cases with n = 37,926 (CHIK = 4178; DENF = 33,748; combined RR = 5.67). Figure 2c shows the relative risk for the 440 municipalities belonging to the multivariate clusters. Furthermore, 13 municipalities had no observed cases of CHIK nor DENF (RR = 0); 145 municipalities had less observed than expected cases (RR between 0 and 1); and 295 had more observed than expected cases (RR > 1). Roldanillo, Valle del Cauca Department (cluster 1), had the highest combined relative risk (RR = 89.84; CHIK = 87.6 & DENF = 2.23). Notably, Medellín, Antioquia Department (cluster 1), had a combined relative risk of 2.49, with 675 observed cases of CHIK and 20,990 observed cases of DENF.

Table 5 Multivariate space-time clusters of CHIK and DENF (RR: relative risk)

Cluster	Duration (weeks)	Municipalities	p-value	VBD	Observed	Expected	RR
1	1–26	144	<0.01	CHIK	20,407	2172.03	14.29
				DENF	21,808	7625.74	3.09
2	4–24	71	<0.01	CHIK	5569	355.84	17.29
				DENF	5320	1249.30	4.35
3	65–90	18	<0.01	CHIK	0	0	0
				DENF	15,739	3479.14	4.83
4	35–53	140	<0.01	CHIK	0	0	0
				DENF	8971	3169.73	2.92
5	54–77	59	<0.01	CHIK	1517	404.54	3.83
				DENF	4608	1420.3	3.30
6	2–27	6	<0.01	CHIK	877	16.10	55.33
				DENF	406	56.52	7.19

Visualizing CHIK and DENF Clusters in 3D

Figures 3, 4, and 5 visualize the space-time clusters of CHIK, DENF, and co-occurrence of CHIK and DENF in a 3D environment, respectively. The design of the 3D visualizations includes the following elements: (1) cylinders representing the size, location, and duration of the cluster; (2) black rings around each cluster representing a particular week during the study period; (3) a 2D layer of the municipalities belonging to a cluster, which is superimposed on Colombia; (4) a 2D layer of the radii of the clusters superimposed on Colombia; (5) labels that denote a cluster's ID; and (6) two temporal axis with labels to denote the start and end dates of each cluster. The 3D visualizations improve the conceptualization of the space-time dynamics of the reported clusters.

For example, Fig. 3 shows that the five space-time clusters of CHIK began and ended during the first half of 2015. Clusters 1–3 lasted the longest while affecting the south-central portions of Colombia. Clusters 4 and 5 occurred in the north-central portions of the country, while they had very short durations between January and March of 2015, respectively. Figure 4 shows that two DENF clusters occurred in 2015 (2 and 3), while four (1, 4–6) occurred in 2016. Cluster 2 in the central region of Colombia began in January 2015 and lasted until late June 2015. The next cluster (3) appeared in August 2015, which lasted until January 2016. The four clusters of DENF in 2016 all began in January and lasted until June and July, while they affected the central and western portions of the country. Figure 5 clearly indicates that four out of the six multivariate clusters occurred during 2015, with two occurring in 2016. Again, clusters 3 and 4 only include significant clustering of DENF, not significant co-occurrence of both DENF and CHIK (Table 5). Therefore, 2015 was a more severe epidemic year regarding the co-occurrence of DENF and CHIK, since clusters 1, 2, and 6 occurred in the first half of 2015, while cluster 5 was the only cluster displaying significant co-occurrence in 2016.

Discussion

The results of the case study highlight statistically significant space-time clusters of DENF, CHIK, and regions of simultaneous excess incidence of both diseases (see multivariate results). The reported space-time clusters of the univariate and multivariate cases correspond to regions of suitable habitat ranges of *A. aegypti* and *A. albopictus*. However, due to the cylindrical scanning window of the statistic, there are municipalities found in a cluster that are above 1.7 kilometers (*Aedes* rarely found above this threshold). To circumvent this issue of selecting municipalities where transmission is rare, relative risk was reported for each municipality belonging to a cluster. Many of the municipalities with a relative risk of 0 are found in regions with an elevation greater than 1.7 km. Reporting and visualizing the relative risk for each municipality also facilitates targeted interventions by identifying the municipalities that have statistically significant excess cases of each disease (i.e., RR > 1), reducing the uncertainty of solely reporting the space-time clusters.

The multivariate STSS reported four clusters of space-time co-occurrence of both DENF and CHIK. Since CHIK just recently appeared in Colombia, it is important to identify areas of co-circulation with DENF, which is hyperendemic in many regions of the country. Since the clinical manifestations of DENF and CHIK (also Zika) are similar, identifying the correct disease via clinical diagnosis is challenging in regions of co-circulation (Silva Jr et al. 2018). Unlike DENF, chronic complications following a CHIK infection are common (de Andrade et al. 2010), which may last for weeks, months, and even years. Therefore, it is critical to implement timely and effective diagnostic methods (e.g., laboratory testing) to confirm the viral etiology between DENF, CHIK, and Zika. Reducing misdiagnosis is especially important in areas of co-circulation, and identifying areas that experience simultaneous outbreaks of DENF, CHIK, or Zika (e.g., via multivariate STSS) can facilitate targeted interventions. Co-

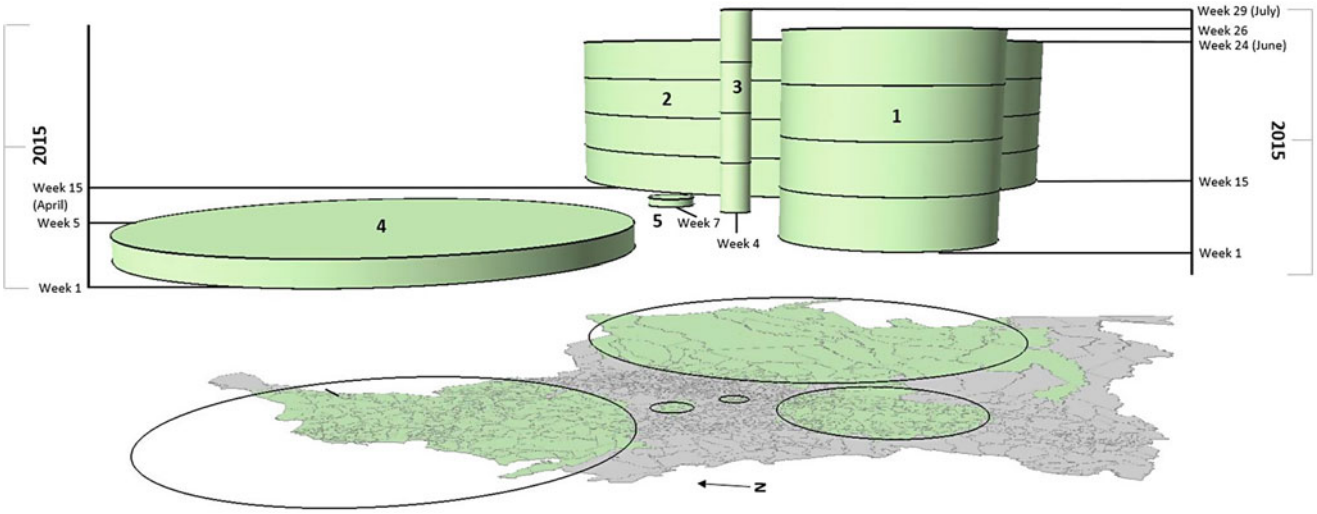


Fig. 3 3D visualization of the CHIK space-time clusters in Colombia (2015–2016)

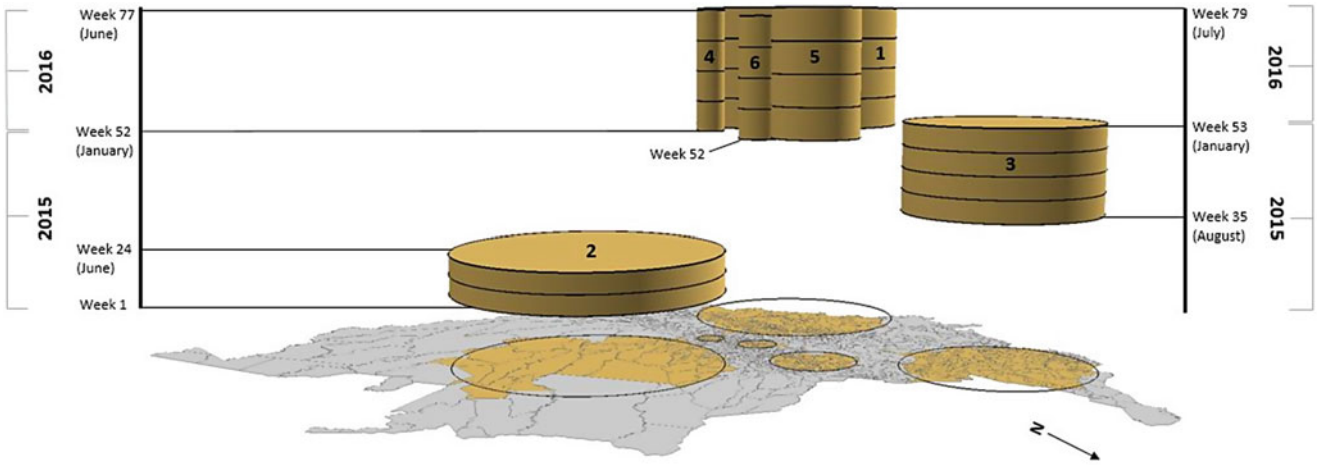


Fig. 4 3D visualization of the DENF space-time clusters in Colombia (2015–2016)

infection of DENF and CHIK is also possible; however, there has not been any observable clinical significance, such as exacerbated symptoms (Furuya-Kanamori et al. 2016).

The 3D visualizations (Figs. 3, 4, and 5) can improve the understanding of the size, duration, and movement of space-time clusters of disease (Desjardins et al. 2018a). 3D visualizations should supplement traditional 2D approaches (Desjardins et al. 2018b), especially for space-time analyses that include a large number of temporal observations. Otherwise, key space-time patterns can be masked by solely using 2D techniques. However, the 3D visualizations provided are static, and crowding and occlusion could have been an issue if there were a larger number of reported space-time clusters. Integrating the 3D visualizations in an interactive environment (e.g., web-GIS platform) can improve their effectiveness by allowing the user to move around the image, for example.

The univariate and multivariate STSS approaches coupled with the 2D and 3D visualizations are an example of exploratory VBD surveillance. The results can be used to improve targeted interventions by identifying statistically significant space-time clusters while shedding light on which regions experienced the greatest burden of DENF and CHIK (i.e., reporting relative risk per municipality). Further research can examine the risk factors that influence VBD incidence and risk in the reported space-time clusters, while analysis at fine geographic scales (e.g., neighborhoods) is necessary for local prevention and mitigation of DENF and CHIK.

Conclusions

Disease surveillance has become a vibrant field of research at the intersection of statistics, computing, and health geog-

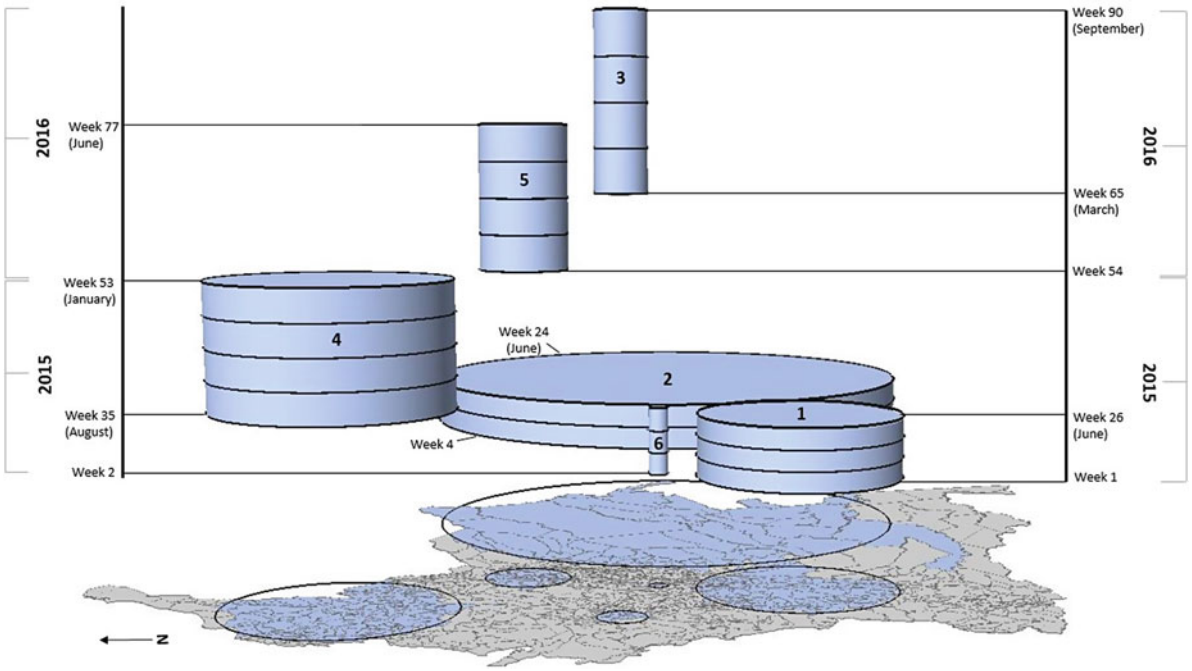


Fig. 5 3D visualization of the multivariate space-time clusters in Colombia (2015–2016)

raphy. The space-time clustering methods and visualization approaches described in this chapter are not an exhaustive list, but rather an example of some of the most commonly used exploratory techniques in geospatial health. Overall, exploratory space-time cluster approaches should be used to shed light on the space-time dynamics of epidemics and outbreaks and highlight the areas that experienced the greatest burden of disease. Subsequent research is necessary to understand the factors that influence disease transmission, while fine-level analysis (e.g., neighborhoods) can uncover local variations of disease incidence within at-risk areas. More research efforts should focus on evaluating the effectiveness of 3D visualization approaches for space-time clusters, such as user studies. 3D visualizations can also benefit from interactive environments that allow the user to navigate freely, rather than static images (such as Figs. 3, 4, and 5). Software that specializes in space-time clustering techniques, such as SaTScan, may not allow visualization of the results, which requires familiarity and training with a GIS and other visualization software. Future developments in software should integrate visualization functionality to streamline subsequent analysis. As novel technologies emerge and data becomes available, new epidemiological questions will arise requiring to investigate additional facets of space-time analytics. For instance, population data become increasingly detailed with respect to their spatial and temporal resolutions, which will enable us to adjust clustering methods for spatially and temporally inhomogeneous background populations. In addition, as techniques for tracking or inferring individual people's location are already available at large scales, research about space-time disease clustering may shift focus from the point- and polygon-based paradigms to trajectory-based methods.

References

Aldstadt, J. 2007. An incremental Knox test for the determination of the serial interval between successive cases of an infectious disease. *Stochastic Environmental Research and Risk Assessment* 21 (5): 487.

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27 (2): 93–115.

Bach, B., P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale. 2017. A descriptive framework for temporal data visualizations based on generalized space-time cubes. *Computer Graphics Forum* 36 (6): 36–61.

Bailey, T.C., and A.C. Gatrell. 1995. *Interactive spatial data analysis*. Vol. 413. Essex: Longman Scientific & Technical.

Banu, S., W. Hu, C. Hurst, Y. Guo, M.Z. Islam, and S. Tong. 2012. Space-time clusters of dengue fever in Bangladesh. *Tropical Medicine & International Health* 17 (9): 1086–1091.

Bhatt, S., P.W. Gething, O.J. Brady, J.P. Messina, A.W. Farlow, C.L. Moyes, et al. 2013. The global distribution and burden of dengue. *Nature* 496 (7446): 504.

Blatt, A.J. 2015. Using geographic information for disease surveillance at mass gatherings. In *Health, science, and place*, 25–37. Cham: Springer.

Boyandin, I., E. Bertini, and D. Lalanne. 2012. A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. In *Computer Graphics Forum*, vol. 31(3pt2), 1005–1014. Oxford, UK: Blackwell Publishing Ltd.

Brunsdon, C., J. Corcoran, and G. Higgs. 2007. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems* 31 (1): 52–75.

Burgman, M.A. 1987. An analysis of the distribution of plants on granite outcrops in southern Western Australia using Mantel tests. *Vegetatio* 71 (2): 79–86.

Cali, S. 2010. *Historia del dengue en Cali. Endemia o una continua epidemia*. Cali: Secretaria de Salud Publica Municipal de Cali.

Campos, G.S., A.C. Bandeira, and S.I. Sardi. 2015. Zika virus outbreak, Bahia, Brazil. *Emerging Infectious Diseases* 21 (10): 1885.

Centers for Disease Control and Prevention. 2018. World Map of areas with risk of Zika. https://wwwnc.cdc.gov/travel/page/world-map-areas-with-zika. Last Accessed 9 Sept 2018.

Coleman, M., M. Coleman, A.M. Mabuza, G. Kok, M. Coetzee, and D.N. Durrheim. 2009. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal* 8 (1): 68.

Cromley, E.K., and S.L. McLafferty. 2011. *GIS and public health*. Guilford Press, New York, NY.

DANE. 2018. Censo Nacional de Población y Vivienda 2018. Departamento Administrativo Nacional de Estadística. https://sitios.dane.gov.co/cnpv-presentacion/src/#cuantos00. Last accessed 1 Oct 2018.

de Andrade, D.C., S. Jean, P. Clavelou, R. Dallel, and D. Bouhassira. 2010. Chronic pain associated with the Chikungunya fever: Long lasting burden of an acute illness. *BMC Infectious Diseases* 10 (1): 31.

de Melo, D.P.O., L.R. Scherrer, and A.E. Eiras. 2012. Dengue fever occurrence and vector detection by larval survey, ovitrap and MosquiTRAP: A space-time clusters analysis. *PLoS One* 7 (7): e42125.

Delmelle, E., E.C. Delmelle, I. Casas, and T. Barto. 2011. HELP: A GIS-based health exploratory analysis tool for practitioners. *Applied Spatial Analysis and Policy* 4 (2): 113–137.

Delmelle, E., C. Dony, I. Casas, M. Jia, and W. Tang. 2014a. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science* 28 (5): 1107–1127.

Delmelle, E.M., H. Zhu, W. Tang, and I. Casas. 2014b. A web-based geospatial toolkit for the monitoring of dengue fever. *Applied Geography* 52: 144–152.

Delmelle, E., M. Hagenlocher, S. Kienberger, and I. Casas. 2016. A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta Tropica* 164: 169–176.

Demšar, U., and K. Virrantaus. 2010. Space–time density of trajectories: Exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science* 24 (10): 1527–1542.

Desjardins, M.R., A. Whiteman, I. Casas, and E. Delmelle. 2018a. Space-time clusters and co-occurrence of chikungunya and dengue fever in Colombia from 2015 to 2016. *Acta Tropica* 185: 77–85.

Desjardins, M.R., A. Hohl, A. Griffith, and E. Delmelle. 2018b. A space–time parallel framework for fine-scale visualization of pollen levels across the Eastern United States. *Cartography and Geographic Information Science* 46(5): 428–440.

Dick, G.W.A., S.F. Kitchen, and A.J. Haddow. 1952. Zika virus (I). Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 46 (5): 509–520.

Dick, O.B., J.L. San Martín, R.H. Montoya, J. del Diego, B. Zambrano, and G.H. Dayan. 2012. The history of dengue outbreaks in the Americas. *The American Journal of Tropical Medicine and Hygiene* 87 (4): 584–593.

Diggle, P.J., A.G. Chetwynd, R. Häggkvist, and S.E. Morris. 1995. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research* 4 (2): 124–136.

Dixon, P.M. 2013. Ripley's K function. Encyclopedia of Environmetrics.

Dorling, D. 1992. Stretching space and splicing time: From cartographic animation to interactive visualization. *Cartography and Geographic Information Systems* 19 (4): 215–227.

Duczmal, L., and R. Assuncao. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 45 (2): 269–286.

Duffy, M.R., T.H. Chen, W.T. Hancock, A.M. Powers, J.L. Kool, R.S. Lanciotti, et al. 2009. Zika virus outbreak on Yap Island, federated states of Micronesia. *New England Journal of Medicine* 360 (24): 2536–2543.

Duncombe, J., A. Clements, W. Hu, P. Weinstein, S. Ritchie, and F.E. Espino. 2012. Geographical information systems for dengue surveillance. *The American Journal of Tropical Medicine and Hygiene* 86 (5): 753–755.

Eaglin, T., I. Cho, and W. Ribarsky. 2017. Space-time kernel density estimation for real-time interactive visual analytics. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Eisen, L., and R.J. Eisen. 2011. Using geographic information systems and decision support systems for the prediction, prevention, and control of vector-borne diseases. *Annual Review of Entomology* 56: 41–61.

Furuya-Kanamori, L., S. Liang, G. Milinovich, R.J.S. Magalhaes, A.C. Clements, W. Hu, et al. 2016. Co-distribution and co-infection of chikungunya and dengue viruses. *BMC Infectious Diseases* 16 (1): 84.

Gabriel, E. 2012. Estimating second-order characteristics of inhomogeneous spatio-temporal point processes: Influence of edge correction methods and intensity estimates. 25. <hal-00818145>.

Gabriel, E., and P.J. Diggle. 2009. Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica* 63 (1): 43–51.

Gaudart, J., et al. 2006. Space-time clustering of childhood malaria at the household level: A dynamic cohort in a Mali village. *BMC Public Health* 6 (1): 286.

Hennessey, M., M. Fischer, and J.E. Staples. 2016. Zika virus spreads to new areas—Region of the Americas, May 2015–January 2016. *American Journal of Transplantation* 16 (3): 1031–1034.

Hohl, A., E. Delmelle, W. Tang, and I. Casas. 2016. Accelerating the discovery of space-time patterns of infectious diseases using parallel computing. *Spatial and Spatio-Temporal Epidemiology* 19: 10–20.

Hohl, A., M. Zheng, W. Tang, E. Delmelle, and I. Casas. 2017. Spatiotemporal point pattern analysis using Ripley's K function. In *Geospatial data science: techniques and applications*. Boca Raton, FL: CRC Press.

Hohl, A., A.D. Griffith, M.C. Eppes, and E. Delmelle. 2018. Computationally enabled 4D visualizations facilitate the detection of rock fracture patterns from acoustic emissions. *Rock Mechanics and Rock Engineering*: 1–14.

Huang, L., M. Kulldorff, and D. Gregorio. 2007. A spatial scan statistic for survival data. *Biometrics* 63 (1): 109–118.

Iftimi, A., F. Martínez-Ruiz, A.M. Santiyán, and F. Montes. 2015. Spatio-temporal cluster detection of chickenpox in Valencia, Spain in the period 2008-2012. *Geospatial Health* 10 (1).

INS. 2014. Boletín epidemiológico semanal número 31, 27 Julio – 2 de Agosto, 2014, Dirección de Vigilancia y Análisis del Riesgo en Salud Pública.

———. 2016. Boletín epidemiológico semanal número 25, 19 Julnio – 25 Junio 2016, Dirección de Vigilancia y Análisis del Riesgo en Salud Pública.

———. 2018a. In: Sivigila, E. (Ed.), Manual del usuario sistema aplicativo SIVIGILA. INS, Colombia. http://portalsivigila.ins.gov.co/sivigila/documentos/manuales_2018/Manual_SIVIGILA_2018.pdf

———. 2018b. *Vigilancia Rutinaria por Evento Municipal*. Colombia: INS. http://portalsivigila.ins.gov.co/sivigila/documentos/Docs_1.php.

———. 2018c. Protocolo para la vigilancia en salud pública del dengue. PAHO. http://www.paho.org/col/index.php?option=com_docman&view=download&category_slug=publicaciones-ops-oms-colombia&alias=1216-protocolo-para-la-vigilancia-en-salud-publica-del-dengue&Itemid=688.

———. 2018d. Protocolo de vigilancia en salud pública Chikunguña. https://www.ins.gov.co/Direcciones/Vigilancia/sivigila/Protocolos/PRO%20Chikungunya.pdf#search=Protocolo%20para%20la%20vigilancia%20en%20salud%20p%C3%BAblica%20del%20dengue%20PAHO%2E

Jacquez, G.M. 1996. A k nearest neighbour test for space–time interaction. *Statistics in Medicine* 15 (18): 1935–1949.

Jung, I., M. Kulldorff, and A.C. Klassen. 2007. A spatial scan statistic for ordinal data. *Statistics in Medicine* 26 (7): 1594–1607.

Jung, I., M. Kulldorff, and O.J. Richard. 2010. A spatial scan statistic for multinomial data. *Statistics in Medicine* 29 (18): 1910–1918.

Khormi, H.M., and L. Kumar. 2015. *Modelling interactions between vector-borne diseases and environment using GIS*. CRC Press, Boca Raton, FL.

Kienberger, S., M. Hagenlocher, E. Delmelle, and I. Casas. 2013. A WebGIS tool for visualizing and exploring socioeconomic vulnerability to dengue fever in Cali, Colombia. *Geospatial Health* 8 (1): 313–316.

Kitron, U. 2000. Risk maps: Transmission and burden of vector-borne diseases. *Parasitology Today* 16 (8): 324–325.

Knox, G. 1963. Detection of low intensity epidemicity: Application to cleft lip and palate. *British Journal of Preventive & Social Medicine* 17 (3): 121.

Knox, G.E. 1964. The detection of space-time iterations. *Journal of the Royal Statistical Society* 13: 25–29.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26 (6): 1481–1496.

———. 2018. SaTScan™ user guide for version 9.6. https://www.satscan.org/.

Kulldorff, M., and U. Hjalmars. 1999. The Knox method and other tests for space-time interaction. *Biometrics* 55 (2): 544–552.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. 2005. A space–time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2 (3): e59.

Kulldorff, M., F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt. 2007. Multivariate scan statistics for disease surveillance. *Statistics in Medicine* 26 (8): 1824–1833.

Kulldorff, M., L. Huang, and K. Konty. 2009. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 8 (1): 58.

Li, Z., et al. 2012. Spatiotemporal analysis of indigenous and imported dengue fever cases in Guangdong province, China. *BMC Infectious Diseases* 12 (1): 132.

Li, J., et al. 2014. Spatial and temporal emergence pattern of Lyme disease in Virginia. *The American Journal of Tropical Medicine and Hygiene* 91: 1166–1172.

Lynch, H.J., and P.R. Moorcroft. 2008. A spatiotemporal Ripley's K-function to analyze interactions between spruce budworm and fire in British Columbia, Canada. *Canadian Journal of Forest Research* 38 (12): 3112–3119.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27 (2 Part 1): 209–220.

McCune, B., and T.F.H. Allen. 1985. Will similar forests develop on similar sites? *Canadian Journal of Botany* 63 (3): 367–376.

Meyer, S., I. Warnke, W. Rössler, and L. Held. 2016. Model-based testing for space–time interaction using point processes: An application to psychiatric hospital admissions in an urban area. *Spatial and Spatio-Temporal Epidemiology* 17: 15–25.

Morrison, A.C., A. Getis, M. Santiago, J.G. Rigau-Perez, and P. Reiter. 1998. Exploratory space-time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991-1992. *The American Journal of Tropical Medicine and Hygiene* 58 (3): 287–298.

Mulatti, P., M. Mazzucato, F. Montarsi, S. Ciocchetta, G. Capelli, L. Bonfanti, and S. Marangon. 2015. Retrospective space–time analysis methods to support West Nile virus surveillance activities. *Epidemiology & Infection* 143 (1): 202–213.

Nagar, R., Q. Yuan, C.C. Freifeld, M. Santillana, A. Nojima, R. Chunara, and J.S. Brownstein. 2014. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research* 16 (10).

Nakaya, T., and K. Yano. 2010a. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* 14 (3): 223–239.

———. 2010b. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* 14 (3): 223–239.

Norström, M., D.U. Pfeiffer, and J. Jarp. 2000. A space–time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. *Preventive Veterinary Medicine* 47 (1–2): 107–119.

Nsoesie, E.O., et al. 2015. Spatial and temporal clustering of Chikungunya virus transmission in dominica. *PLOS Neglected Tropical Diseases* 9: e0003977.

Ocampo, C.B., N.J. Mina, M. Carabalí, N. Alexander, and L. Osorio. 2014. Reduction in dengue cases observed during mass control of Aedes (Stegomyia) in street catch basins in an endemic urban area in Colombia. *Acta Tropica* 132: 15–22.

Onozuka, D., and A. Hagihara. 2007. Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-time scan statistic. *BMC Infectious Diseases* 7 (1): 26.

Palaniyandi, M., P.H. Anand, and T. Pavendar. 2017. Environmental risk factors in relation to occurrence of vector borne disease epidemics: Remote sensing and GIS for rapid assessment, picturesque, and monitoring towards sustainable health. *International Journal of Mosquito Research* 4 (3): 09–20.

Pereira, M.G., L. Caramelo, C.V. Orozco, R. Costa, and M. Tonini. 2015. Space-time clustering analysis performance of an aggregated dataset: The case of wildfires in Portugal. *Environmental Modelling & Software* 72: 239–249.

Restrepo, A.C., P. Baker, and A.C. Clements. 2014. National spatial and temporal patterns of notified dengue cases, Colombia 2007–2010. *Tropical Medicine & International Health* 19 (7): 863–871.

Robertson, C., T.A. Nelson, Y.C. MacNab, and A.B. Lawson. 2010. Review of methods for space–time disease surveillance. *Spatial and Spatio-Temporal Epidemiology* 1 (2–3): 105–116.

Sagl, G., E. Delmelle, and E. Delmelle. 2014. Mapping collective human activity in an urban environment based on mobile phone data. *Cartography and Geographic Information Science* 41 (3): 272–285.

Saule, E., D. Panchananam, A. Hohl, W. Tang, and E. Delmelle. 2017. Parallel space-time kernel density estimation. In *Parallel Processing (ICPP), 2017 46th International Conference*, 483–492. IEEE.

Scheepens, R., N. Willems, H. van de Wetering, and J.J. Van Wijk. 2011. Interactive visualization of multivariate trajectory data with density maps. In *Visualization Symposium (PacificVis), 2011 IEEE Pacific*, 147–154. IEEE.

Scripcaru, G., C. Mateus, and C. Nunes. 2017. A decade of adverse drug events in Portuguese hospitals: Space-time clustering and spatial variation in temporal trends. *BMC Pharmacology and Toxicology* 18 (1): 34.

Shankarappa, R.A.J., J.B. Margolick, S.J. Gange, A.G. Rodrigo, D. Upchurch, H. Farzadegan, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology* 73 (12): 10489–10502.

Sheehan, T.J., and L.M. DeChello. 2005. A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. *International Journal of Health Geographics* 4 (1): 15.

Silva, J.V., Jr., L.F. Ludwig-Begall, E.F. de Oliveira-Filho, R.A. Oliveira, R. Durães-Carvalho, T.R. Lopes, et al. 2018. A scoping review of Chikungunya virus infection: Epidemiology, clinical characteristics, viral co-circulation complications, and control. *Acta Tropica*.

Silverman, B.W. 2018. *Density estimation for statistics and data analysis*. Routledge.

Tang, W., W. Feng, and M. Jia. 2015. Massively parallel spatial point pattern analysis: Ripley's K function accelerated using graphics processing units. *International Journal of Geographical Information Science* 29 (3): 412–439.

Tango, T., and K. Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4 (1): 11.

Tao, R., and J.C. Thill. 2016. Spatial cluster detection in spatial flow data. *Geographical Analysis* 48 (4): 355–372.

Ullah, S., H. Daud, S.C. Dass, H.N. Khan, and A. Khalil. 2017. Detecting space-time disease clusters with arbitrary shapes and sizes using a co-clustering approach. *Geospatial Health* 12 (2): 210–216.

Villegas, A.V., E.G. Aristizabal, and J.H. Rojas. 2010. *Analisis epidemiologico de dengue en Cali*. Cali: Secretaria de Salud Publica Municipal.

Wilson, M.E., and L.H. Chen. 2015. Dengue: Update on epidemiology. *Current Infectious Disease Reports* 17 (1): 457.

World Health Organization. 2017. Vector-borne diseases. Factsheet number 387. http://www.who.int/mediacentre/factsheets/fs387/en/. Last Accessed 23 Apr 2018.

World Mosquito Program. 2018. Dengue, Zika y chikungunya en Colombia. http://www.eliminatedengue.com/colombia/encolombia. Last Accessed 1 Oct 2018.

Xie, Z., and J. Yan. 2008. Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems* 32 (5): 396–406.

Xu, W., and C. Wu. 2018. Detecting spatiotemporal clusters of dementia mortality in the United States, 2000–2010. *Spatial and Spatio-Temporal Epidemiology* 27: 11–20.

Yactayo, S., J.E. Staples, V. Millot, L. Cibrelus, and P. Ramon-Pardo. 2016. Epidemiology of Chikungunya in the Americas. *The Journal of Infectious Diseases* 214 (suppl_5): S441–S445.

Yamada, I., and P.A. Rogerson. 2003. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35 (2): 97–109.

Yamada, I., and J.C. Thill. 2007. Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis* 39 (3): 268–292.

Machine Learning, Big Data, and Spatial Tools: A Combination to Reveal Complex Facts That Impact Environmental Health

David J. Lary, Lakitha Omal Harindha Wijeratne, Gebreab K. Zewdie, Daniel Kiv, Daji Wu, Fazlay S. Faruque, Shawhin Talebi, Xiaohe Yu, Yichao Zhang, Estelle Levetin, Rebecca J. Allee, Nabin Malakar, Annette Walker, and Hamse Mussa

Introduction

Beyond environmental studies, machine learning has already proved immensely useful in a wide variety of applications in science, business, healthcare, and engineering. Machine learning allows us to *learn by example* and to *give our data a voice*. It is particularly useful for those applications for which we do *not* have a complete theory, yet which are of significance. Machine learning is an automated implementation of the scientific method (Domingos 2015), following the same process of generating, testing, and discarding or refining hypotheses. While a scientist or engineer may spend his entire career coming up with and testing a few hundred hypotheses, a machine-learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is therefore not surprising that machine learning is currently revolutionizing many areas of science, technology, business, and medicine (Lary et al. 2016a).

Machine learning is now being routinely used to work with large volumes of data in a variety of formats, such as image, video, sensor, health records, etc. Machine learning can be used in understanding this data and creating predictive and classification tools. When machine learning is used for regression, empirical models are built to predict continuous data, facilitating the prediction of future data points, e.g., algorithmic trading and electricity load forecasting. When machine learning is used for classification, empirical models are built to classify the data into different categories, aiding in the more accurate analysis and visualization of the data. Applications of classification include facial recognition, credit scoring, and cancer detection. When machine learning is used for clustering, or unsupervised classification, it aids in finding the natural groupings and patterns in data. Applications of clustering include medical imaging, object recognition, and pattern mining. Object recognition is a process for identifying a specific object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. These technologies are being used for applications such as driver-less cars, automated skin cancer detection, etc.

There are now a variety of open-source tools that can greatly facilitate the use of machine learning, such as scikit-learn,[1] TensorFlow,[2] Caffe,[3] and Spark Mlib.[4] Common programming environments used for machine learning include R,[5] Julia,[6] Python,[7] and MATLAB.[8] All of the applications shown in this chapter used MATLAB.

The original version of this chapter was revised: Updated chapter has been uploaded to Springerlink. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-71377-5_21.

D. J. Lary (✉) · L. O. H. Wijeratne · G. K. Zewdie · D. Kiv · D. Wu
S. Talebi · X. Yu · Y. Zhang · E. Levetin · R. J. Allee · N. Malakar
A. Walker · H. Mussa
Hanson Center for Space Sciences, The University of Texas at Dallas, Richardson, TX, USA
e-mail: david.lary@utdallas.edu; gebreab.zewdie@utdallas.edu; drk150030@utdallas.edu; Shawhin.Talebi@utdallas.edu; estelle-levetin@utulsa.edu; nmalakar@worcester.edu

F. S. Faruque
Department of Preventive Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA
e-mail: ffaruque@umc.edu

[1] http://scikit-learn.org/stable/.
[2] https://www.tensorflow.org.
[3] http://caffe.berkeleyvision.org.
[4] http://spark.apache.org/mllib/.
[5] https://cran.r-project.org.
[6] https://julialang.org/#tab-math
[7] https://www.python.org.
[8] https://www.mathworks.com/solutions/machine-learning.html.

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*, https://doi.org/10.1007/978-3-030-71377-5_12

Fig. 1 From data to predictions and insights: A flow chart showing the steps used in preparing and using data with machine learning

In this paper, we will give an overview of several remote sensing applications of machine learning made over the last two decades and then take a look ahead to some likely future applications.

What Is Machine Learning?

Machine learning is an automated approach to building empirical models from the data *alone*. Figure 1 shows a flow chart showing the steps used in preparing and using data with machine learning. A key advantage of this is that we make *no* a priori assumptions about the data, it's functional form, or probability distributions. It is an empirical approach, so we do not need to provide a theoretical model. However, it also means that for machine learning to provide the best performance, we do need a *comprehensive representative set of examples*, which spans as much of the parameter space as possible. This comprehensive set of examples is referred to as the *training data* (Fig. 2).

So, for a successful application of machine learning, we have *two* key ingredients, both of which are essential, a machine learning algorithm and a comprehensive training data set. Then, once the training has been performed, we should test its efficacy using an independent validation data set to see how well it performs when presented with data that the algorithm has *not* previously seen, i.e., test its *generalization*. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.

It should be noted, that with a given machine learning algorithm, the performance can go from poor to outstanding with the provision of a progressively more complete training data set. Machine learning really is learning by example, so it is critical to provide as complete a training data set as possible. At times, this can be a labor-intensive endeavor.

When using machine learning, we are typically performing one of three tasks:

1. Multivariate nonlinear non-parametric regression
2. Supervised classification
3. Unsupervised classification

Each of these tasks can be achieved by a variety of different algorithms. Some of the commonly used algorithms include neural networks (McCulloch and Pitts 1943; Haykin 2001, 2007, 1994, 1999; Demuth et al. 2014; Bishop 1995), support vector machines (Vapnik 1982, 1995; Cortes and Vapnik 1995; Vapnik 2000, 2006), decision trees (Safavian and Landgrebe 1991), and random forests (Ho 1998; Breiman 1984, 2001). Our goal in this chapter is to present a set of examples of applying machine learning to spatial datasets.

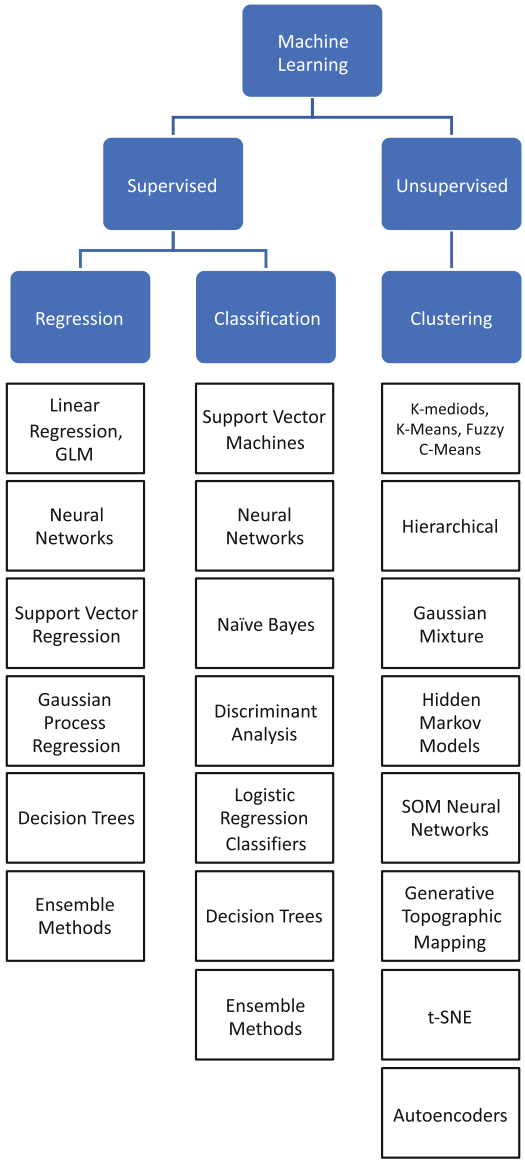


Fig. 2 A schematic giving an overview of some of the types of machine learning

Some Existing Machine Learning Applications

There have been many previous machine learning studies (Lary et al. 2016b; Brown et al. 2008; Lary et al. 2009; Lary and Aulov 2008; Lary et al. 2004; Malakar et al. 2013; Lary 2010; Malakar et al. 2012a; Lary 2013, 2007; Albayrak et al. 2011; Brown et al. 2006; Lary et al. 2003; Malakar et al. 2012b; Lary 2014; Lary et al. 2015; Kneen et al. 2016; Lary et al. 2010; Medvedev et al. 2016; Lary et al. 2016c; O et al. 2017; Wu et al. 2017; Nathan and Lary 2019; Lary et al. 2019, 2018; Wu et al. 2019; Alavi et al. 2016; Ahmad et al. 2016; Zewdie and Lary 2018; Malakar et al. 2018; Zewdie et al. 2019a,b; Chang et al. 2019; Choi et al. 2019). Let us start by looking at several examples of bias correction. Bias identification and correction is of particular importance for every single remote sensing instrument. Bias correction can also prove to be a particularly challenging issue, one which involves multiple factors.

Multivariate Nonlinear Non-parametric Regression

The ubiquitous issue of inter-instrument biases is an obvious example of where we do *not* have a complete theoretical understanding, and so machine learning can be of particular use.

In many areas of remote sensing, we have multiple instruments simultaneously observing the earth on a variety of platforms. Many of these sensors may be providing data on the same parameters, such as the surface vegetation or the composition of the atmosphere or ocean. A ubiquitous issue faced is inter-instrument bias between the contemporaneously observing instruments. This inter-instrument bias can be due to a variety of known reasons that may include different instruments, different observing geometries and orbits, etc., as well as some causes that we do not know.

This is an important issue, as we routinely need to provide data fusion of multiple datasets, datasets which are inevitably biased relative to each other, sometimes even after the mandatory calibration/validation process. When we are seeking to construct a long-term record spanning many decades, this inevitably will often involve a large number of instruments, a matter very relevant for climate variables. In addition, data assimilation has become an important part of effectively utilizing remotely sensed data. However, data assimilation is a *best linear unbiased estimator* (BLUE), and fusing biased data can cause serious issues.

This data fusion typically involves large teams of scientists and engineers. On the one hand, the instrument teams have a keen sense of faithfully reporting the data, as it is, warts and all. They are naturally loath to empirically correct biases; they would like to theoretically understand the cause of the bias and data issues from first principles. However, as the Earth System is so complex, with many interacting processes, and often the instruments are also complex, this is not always possible. Residual data issues can, and usually do, remain. On the other hand, the modelers know that data bias exist, but are very reticent to make changes to data products that they did not collect, so we therefore have *a problem of closure*.

Biases are ubiquitous, not all of them can be explained theoretically. Yet, we typically need to fuse multiple datasets to construct long-term time series and/or improve global coverage. If the biases are not corrected before data fusion, we

introduce further problems, such as spurious trends, leading to the possibility of unsuitable policy decisions. When data assimilation is involved, any use of biased observations can lead to the suboptimal use of the observations, nonphysical structures in the analysis, biases in the assimilated fields, and extrapolation of biases due to multivariate background constraints. To compound matters further, the instruments whose data we would like to fuse are often not making coincident measurements in time or space. It is imperative to inter-compare observations in their appropriate context and be able to address the pernicious issue of inter-instrument bias, an issue where machine learning has proved to be most useful. Let us now take a look at some examples.

Machine Learning for New Product Creation

Let us know turn our attention to an example of creating new data products through the holistic use of satellite and in situ data, a new data product that is of societal significance.

Airborne Particulates

There is an increasing awareness of the health impacts of particulate matter and a growing need to quantify the spatial and temporal variations of the global abundance of ground-level airborne particulate matter ($PM_{2.5}$). In March 2014, the World Health Organization (WHO) released a report that in 2012 alone, a staggering 7 million people died as a result of air pollution exposure (WHO), one in eight of the total global deaths. A major component of this pollution is airborne particulate matter (e.g., $PM_{2.5}$ and PM_{10}).

The recent study by Lary et al. (2014) used machine learning to provide daily global estimates of airborne $PM_{2.5}$ from 1997 to 2014. This was achieved by utilizing a massive amount of data (40 TB) from a suite of about 100 remote sensing and meteorological data products together with ground-based observations of $PM_{2.5}$ from 8329 measurement sites in 55 countries taken between 1997 and 2014. This data was used to train a machine learning algorithm to estimate the daily distributions of $PM_{2.5}$ from 1997 to 2014. This allowed the creation of a new global $PM_{2.5}$ product at 10 km resolution from August 1997 up to the present (Lary et al. 2014). This new dataset is specifically designed to support health impact studies. Lary et al. (2014) showed some examples of this global $PM_{2.5}$ dataset and examined its associations with mental health emergency room admissions in Baltimore, MD. They demonstrate that the new $PM_{2.5}$ data product can reliably represent global observations of $PM_{2.5}$ for epidemiological studies. They showed that airborne particulates can have some surprising associations with health outcomes. As an example of this, (Lary et al. 2014) presented an analysis of Baltimore schizophrenia emergency room admissions in the context of the levels of ambient pollution. $PM_{2.5}$ had a statistically significant association with some aspects of mental health.

A useful validation of the new $PM_{2.5}$ data product is to survey the key features of the global $PM_{2.5}$ distribution and see if they capture what we expect to find and what has been reported in the literature. In Fig. 3a, we see that the eastern half of the USA has a higher average abundance of $PM_{2.5}$ than the western half with the exception of California. This is consistent with the overlaid EPA observations shown as color-filled circles. The color fill for the observations uses the same color scale as the machine learning estimate depicted using the background colors. There are persistently high levels of $PM_{2.5}$ in Mexico's dusty and desolate Baja California Sur. The particularly high values are in Mulege Municipality close to Guerrero Negro (marked A in panel (a) of Fig. 3). Straddling the region close to the Mexico, Arizona, and California borders is the Sonoran Desert. This is a region characterized by a high-average $PM_{2.5}$ abundance (marked B) and haboobs, massive dust storms. The Sonoran desert has an area of 311,000 square kilometers and is one of the hottest and dustiest parts of North America. This is clearly evident in the high 16-year average $PM_{2.5}$ abundance in this region. The persistently high $PM_{2.5}$ abundance associated with Los Angeles is visible (marked C). The regions of high population density usually coincide with the region of high particulate abundance. California's heavily agricultural Central Valley has a high $PM_{2.5}$ loading (marked D); note the good agreement of our estimates with the 16-year average observations. The EPA has designated Central Valley as a non-attainment area for the 24-h $PM_{2.5}$ National Ambient Air Quality Standards (NAAQS). The high $PM_{2.5}$ abundance associated with the Great Salt Lake Desert in northern Utah close to the Nevada border is clearly visible (marked E). There is a nearby measurement supersite at Salt Lake City recording a particulate abundances consistent with our estimates. Mexico City is known for its high levels of particulates and is clearly visible (marked F) as a localized hot spot. Close to the Mexico/Texas border, we see the elevated $PM_{2.5}$ abundance associated with the Chihuahuan Desert and the Big Bend Desert (marked G). Dust storms in this area often impact El Paso in Texas and Ciudad Juarez in Mexico. The Ohio River Valley (marked H) encompasses several states and is home to numerous coal-fired power plants, chemical plants, and industrial facilities, leading to high levels of ambient particulates. The Ohio River Valley has a higher average abundance of $PM_{2.5}$ than the rest of the East Coast. Our analysis agrees closely with the in situ observations for the Athens super-site. The Piura desert in Northern Peru (marked I) on the coast and western slopes of the Andes is a region of high particulate abundances. The region in South America from the high Andean semi-arid Altiplano basin in the north, coming down through the Salar de Uyuni Desert (the world's largest salt flats), passing by Santiago in Chile and San Miguel de Tucumn, San Juan and

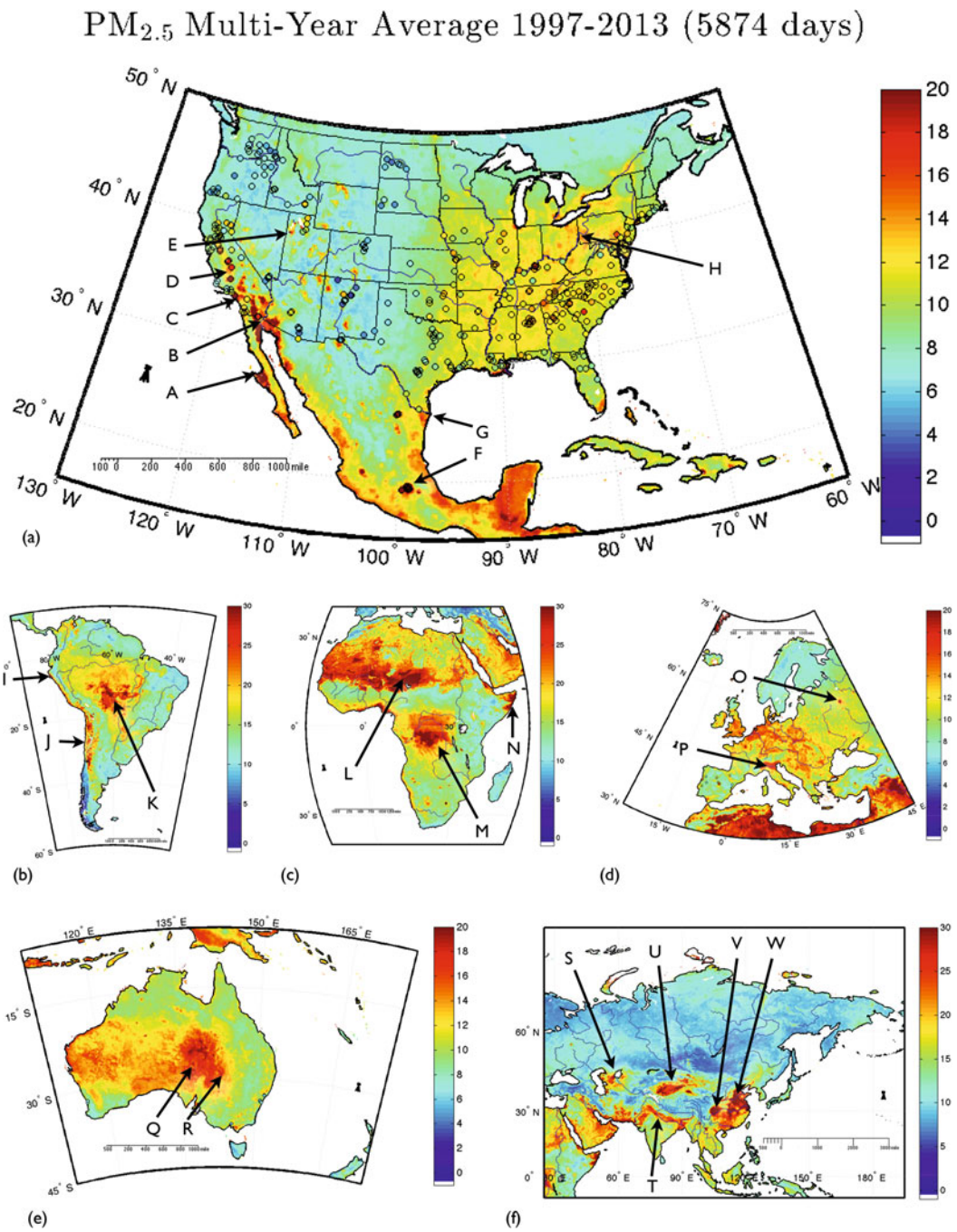


Fig. 3 The average estimated surface $PM_{2.5}$ abundance of the 5874 daily estimates from August 1, 1997, to August 31, 2013, in $\mu g/m^3$ for (**a**) the USA, (**b**) South America, (**c**) Africa, (**d**) Europe, (**e**) Australia, and (**f**) Asia

Mendoza in Argentina, and down to the Neuquén Basin in the south, is characterized by a high abundance of particulates from a combination of dust, salt, and pollution (marked J). The southern Amazon in Bolivia and the surrounding region has a lot of burning leading to persistently high particulate abundances (labeled K).

The Bodélé depression is Chad's lowest point on the Sahara's southern edge that supplies the Amazon forest with the majority of its mineral dust. The high abundance of $PM_{2.5}$ over the Bodélé is clearly visible (marked L). Typically, there are dust storms originating from the Bodélé depression on around 100 days a year. The low flat desert in the North African Western Sahara is some of the most inhospitable and arid land on earth and a substantial dust source, clearly visible in the high abundance of $PM_{2.5}$. Burning in the Democratic Republic of the Congo (marked M) leads to high levels of particulates. Much of coastal Somalia is desert characterized by high levels of particulates (marked N).

The Italian Po valley (marked P in Fig. 3) has some of the highest average abundance of particulates in Europe. Industrial emissions coupled with persistent fog lead to heavy smog. High levels of $PM_{2.5}$ are found in the Netherlands and North-west Germany. An example of a local pollution hotspot in Europe is Moscow (marked O).

Lake Eyre is Australia's largest lake and lowest point (marked Q). When the lake has dried out, a salt crust remains. When Lake Eyre is dry, it is typically Australia's largest dust source, Lake Eyre usually only fills with water after the heavy rains that typically occur once every 3 years; during these periods, the $PM_{2.5}$ abundance in the vicinity of Lake Eyre is lower than usual. Just east of the Lake Eyre Basin is the Strzelecki Desert, another major Australian dust source (marked R). The arid region just south of the Hamersley Range in Western Australia, the Gibson Desert, Great Victoria Desert, and MacDonnell Ranges are also dusty environments with elevated average abundances of $PM_{2.5}$.

Asia has some of the highest particulate abundances anywhere on earth. The Aral Sea (marked S) lying across the border of Kazakhstan and Uzbekistan is heavily polluted with major public health problems. The Ganges Valley is home to 100 million people and is highly polluted (marked T). The cold Taklimakan Desert of northwest China is a major source of $PM_{2.5}$ (marked U). Particularly high levels of particulates are found in the Sichuan Basin (marked V) and in western China in the region from Beijing in the North down to Guangxi in the south (marked W).

Asthma Health

The strongest risk factors for asthma exacerbation are a combination of genetic predisposition and environmental exposure to inhaled substances that provoke allergic reactions or irritate the airways (Faruque et al. 2014). Particulate matters (PM) are known to be a major contributing factor exacerbating asthma. While medication can control asthma, avoiding the exposures to asthma triggers can significantly reduce the severity of symptoms.

The desired standard for fine PM has been an issue of controversy since its establishment; while the fact that lowering the abundance of airborne PM can reduce the burden of asthma and other health problems has been supported by many studies (Etchie et al. 2018; Giannadaki et al. 2016; Mirabelli et al. 2016; Guarnieri and Balmes 2014; Weinhold 2013; Esworthy 2012; Esworthy and McCarthy 2013; Berman et al. 2012; Johnson and Graham 2005). Based on the growing evidences of health problems due to even lower levels of fine PM than previously thought, the EPA has been making the standard stricter for $PM_{2.5}$, once in 2006 for daily and once in 2012 for annual (EPA 2019). However, scientists and physicians are not yet sure about the exact level of PM below which health conditions will not be negatively impacted.

In a study, funded by the National Institute of Environmental Health Sciences, researchers attempted to examine the impact of $PM_{2.5}$ on asthma in a region where the ambient $PM_{2.5}$ in general stays below the current annual National Ambient Air Quality Standards (NAAQS) of 12 $\mu g/m^3$. The associations between asthma morbidity and local $PM_{2.5}$ in the Jackson, Mississippi, area were examined in this epidemiological study (Chang et al. 2019).

Through implementing machine learning, the first daily global estimates of ground-level $PM_{2.5}$ for the period of 1997 to 2014 were developed by Lary et al. (2014) through a project funded in 2011 by the National Institute of Environmental Health Sciences. Since the error value for every estimated value per grid was calculated (Lary et al. 2014, 2015, 2016c), the usability of these estimated data for health studies can be readily assessed. These estimated $PM_{2.5}$ were used in an epidemiological study to examine the associations between asthma morbidity and local $PM_{2.5}$ in the Jackson, Mississippi, area (Chang et al. 2019). Because of the availability of seamless $PM_{2.5}$ data over a long period, this population-based time-series study was possible to conduct a 9-year period of asthma morbidity. The findings of this health study support a relationship between air quality and asthma morbidity even in a region of relatively low levels of $PM_{2.5}$ exposure, which is an important information regarding respiratory health for many parts of the world.

In another study (Lary et al. 2019), we found that machine learning was able to effectively estimate student learning outcomes geospatially across all the campuses in a large urban independent school district. The machine learning showed that key factors in estimating the student learning outcomes included the number of days students were absent from school. In turn, one of the most important factors in estimating the number of days a student was absent was whether or not the student had asthma. This highlights the significant impact of asthma on student learning outcomes.

Pollen Estimation

Pollen is known to be a trigger for allergic diseases, e.g., asthma, hay fever, and allergic rhinitis (Oswalt and Marshall 2008; Howard and Levetin 2014). It is interesting that a variety of non-respiratory issues such as strokes (Low et al. 2006) and, surprisingly, even suicide and attempted suicide (Matheson et al. 2008) have an association with the daily concentration of atmospheric particulates. However, so far, there is no defined threshold amount of pollen known to trigger allergy for sensitive individuals (Voukantsis et al. 2010). One of the factors for the lack of knowledge of the threshold amount of pollen is the absence of an accurate estimation on a fine spatial scale of the hourly, bi-hourly, or daily amount of pollen. Individual physiological differences such as gender and age among sensitive people also adversely affect in knowing the threshold amount of pollen in the surrounding (Britton et al. 1994; Ernst et al. 2002).

Of all plants, weeds, and particularly those of the *Ambrosia* species, e.g., *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed), are major producers of large amounts of pollen. For example, a common ragweed can produce up to about 2.5 billion pollen grains per plant per day (Laaidi et al. 2003). *Ambrosia artemisiifolia* and *Ambrosia trifida* combined can produce more allergens than all other plants combined (Lewis et al. 1983). Grasses (e.g., *rye grass*) are also known to trigger an allergic response. Following *Ambrosia artemisiifolia*, grass pollen are known for their high allergic potency than most weeds (Esch et al. 2001; Lewis et al. 1983). Tree pollen can cause an allergic response, but one that is typically less than that of weeds and grasses, although in some regions, tree pollen can trigger a significant allergic response. For instance, the airborne concentration of mountain cedar pollen grains can reach tens of thousands of pollen grains per cubic meter and trigger a significant allergic response in central Texas during winter, known as cedar fever (Andrews et al. 2013; Ramirez 1984).

Both global climate change and air pollution affect the abundance of airborne pollen and, consequently, its allergic impact (Kinney 2008; Wayne et al. 2002; Voukantsis et al. 2010). For example, the abundance of pollutants such as CO_2 (Wayne et al. 2002) and NO_2 (Zhao et al. 2016) can affect the extent of growing season of major pollen producing plants and thereby also affect the airborne pollen concentration as well as altering the onset and end dates of seasonal allergies. Overall, more people are exposed to pollen, and sensitive individuals become exposed to large amounts of pollen for a longer period of time over larger areas.

Globally millions of people are affected by seasonal allergies, and the number of people affected is increasing each year. In North America alone, as of 2008, about 50 million adult Americans and 9% of children aged below 18 have experienced pollen-caused allergies (Howard and Levetin 2014). Similarly, in Europe, about 15 million people are affected by hay fever, asthma, and rhinitis (D'amato and Spieksma 1991). Hence, pollen allergies are becoming an increasingly significant environmental health issue. Furthermore, just as accurate daily weather forecasts are of significant use, accurate daily pollen forecasts are likely to become increasingly important.

Remote sensing has been employed to study atmospheric pollen concentrations. For example, the polarization of LIDARs has been used to observe the airborne tree pollen abundance at Fairbanks Alaska (Sassen 2008). In this case, the pollen produces a depolarization of the LIDAR backscattering signals from the lower atmosphere. The light scattering properties of pollen is also manifested in the shape of the solar corona they create. The shape of the solar corona associated with pollen depends on the shape of the pollen grains and their atmospheric concentration (Tränkle and Mielke 1994). However, this approach can be complicated since atmospheric light scattering is also caused by other airborne particulates.

Common pollen estimation techniques, particularly those made in Europe, stress the importance of meteorologic variables (Kasprzyk 2008). Usually forecasting the amount of airborne pollen is based on the interaction of atmospheric weather and pollen (Arizmendi et al. 1993). Meteorologic variables such as the daily mean, maximum, change in temperature, and dew point variables show positive correlation with the pollen concentration (Kasprzyk 2008). Kasprzyk (2008) found that atmospheric humidity shows a negative correlation to the pollen concentration. Other studies show that temperature, precipitation, and wind speed are significant meteorologic parameters in estimating pollen concentration (Stark et al. 1997).

Most of these meteorologic variable-based forecasting methods employed statistical methods such as linear regression, the polynomial method, and time series analysis (Sánchez-Mesa et al. 2002). Only few studies used advanced machine learning methods such as neural network (Sánchez-Mesa et al. 2002; Rodríguez-Rajo et al. 2010; Puc 2012; Voukantsis et al. 2010) and random forest (Nowosad 2015) for pollen forecasting, and support vector machines are applied for related environmental studies (Voukantsis et al. 2010; Osowski and Garanty 2007).

Predicting Pollen Abundance

Over the past decade, neural networks have been applied to study pollen of different species over the European region. For example, (Csépe et al. 2014) used different computational intelligence (CI) methods to predict the *Ambrosia* pollen at two different places in Hungary and France. Castellano-Méndez et al. (2005) and Puc (2012) have employed the neural network to predict *Betula* pollen over Spain and Poland, respectively. Recently, (Nowosad 2015) used the random forest method to forecast different tree pollen species.

In this study, we used random forests, neural networks, and support vector machines to estimate daily *Ambrosia* pollen concentration at Tulsa, Oklahoma (location, 36.1511°N, 95.9446°W). We used a combination of environmental parameters and NEXRAD radar measurements. The combined parameters are listed in Table 1. The daily pollen concentration used in the training of our machine learning algorithms was obtained using a Burkhard spore trap at the University of Tulsa, Oklahoma.

After pollen is produced in the plant anthers, its emission, dispersion, and deposition are influenced by meteorological variables such as the temperature, wind speed and direction, and pressure (Kasprzyk 2008; Csépe et al. 2014; Howard and Levetin 2014). Other meteorological parameters such as dew point, humidity, rainfall, and sunshine duration are also

Table 1 Name and type of predictors (input variables) used for our machine learning training

Parameter	Unit	Type
Vegetation greenness fraction	fraction	Env.
Leaf area index	m^2	Env.
Roughness length, sensible heat	m	Env.
Displacement height	m	Env.
Energy stored in land	Jm^{-2}	Env.
Mean reflectivity	dB	NEXRAD
Mean Doppler velocity	ms^{-1}	NEXRAD
Mean spectral width	ms^{-1}	NEXRAD
Reflectivity [10–10 dB]	dB	NEXRAD
Velocity [10–10 dB]	ms^{-1}	NEXRAD
Spectral width [10–10] dB	ms^{-1}	NEXRAD
Reflectivity [20–20 dB]	dB	NEXRAD
Velocity [20–20 dB]	ms^{-1}	NEXRAD
Spectral width [20–20 dB]	ms^{-1}	NEXRAD
Reflectivity [40–40 dB]	dB	NEXRAD
Velocity [40–40 dB]	ms^{-1}	NEXRAD
Spectral width [40–40 dB]	ms^{-1}	NEXRAD
Wind direction at altitude 50 m	Degree	NEXRAD
Wind speed at altitude 50 m	ms^{-1}	NEXRAD

Parameters consist of environmental and NEXRAD radar measurements

known to affect pollen emission and distribution (Kasprzyk 2008).

We used a set of environmental and NEXRAD radar parameters (Table 1) in our machine learning training. Environmental parameters such as vegetation greenness fraction, roughness length (sensible heat), energy stored in all land reservoirs, and displacement height and leaf area index are selected. The other set of data we used are the NEXRAD measurements which consist of the reflectivity, Doppler velocity, and spectral width which represent, respectively, the amount of back scattered signals from a scattering volume, the velocity of the scatterer along the radar line of sight, and the width of the power spectrum. All NEXRAD measurements are taken at the lowest elevation. Additionally the NEXRAD provides measurements of the vertical profile of the direction and speed of the wind from about near the surface of the Earth. The dual polarization measurements, differential reflectivity, differential phase, and correlation coefficient, use the horizontal and vertical polarization signals and are particularly suited for particle identification. In this study, we do not use the dual polarization (polarimetric) NEXRAD measurements as we have only few days of the measurements in contrary to the ideal high-dimensional data requirement for machine learning.

The three machine learning methods were trained on the entire data set to assess their performance in predicting the *Ambrosia* pollen. The scatter diagrams are shown in Fig. 4. We also used a Newton-Raphson-based recursive Random Forest technique that has been developed in order to improve the accuracy. The method includes error estimation and correction. In order to evaluate the performance of the machine learning methods independently, 10% of the data are randomly selected and withdrawn for validation from the training, and the remaining 90% of the data is then used for training the model. After developing the model, its performance is tested using the independent validation dataset that was not used in training the machine learning regression. These results are shown in Fig. 4. Panels (a), (b), and (c) in Fig. 4 show scatter plots of predictions made by the support vector machine, neural network, and random forest machine learning methods, respectively, using the training data (black circles) and the validation data (red squares). Results of the iterative method applied to the random forest method are given by panels (d), (e), and (f). The random forest machine learning is trained using 200 decision trees.

From the top three panels of Fig. 4, we observe that the neural network and random forest methods produced better predictions than the support vector machine. The random forest method produced the best independent validation results (correlation coefficient, 0.62) of all the three methods. The high correlation value of neural network found using the training data (correlation coefficient 0.99) is not reproduced in the independent validation test which had a correlation coefficient of only 0.46. Error bar plots for the training and the validation data for the first iteration of the random forest are given by panel (d) in Fig. 4. We see that predictions using both the training and validation data exhibit large errors and a low correlation coefficient. Interestingly, after a few iterations, the random forest produced results with significantly reduced errors and correlation values close to 1 (panel (e) in Fig. 4). Panel (f) in Fig. 4 shows the correlation coefficient values between the normalized estimated and actual pollen for the training (blue curve) and validation data (red curve) sets for ten iterations. We observe that the iterative of the random forest method has reduced the error significantly, and the correlation coefficient values converge to one for both training and validation data sets.

The upper panel of Fig. 5 shows a comparison of the actual and predicted pollen using the recursive random forest. Another important application of machine learning methods is the selection of the best features (variables) that contribute most to the prediction and ranking them in order of the importance. In this way, we can determine the most important predictor variables and estimate the output leaving features that contribute less. The random forest provides such a ranking based on criteria attributed to the splitting variable in the data sampling to form a decision tree (Genuer et al. 2010; Kotsiantis et al. 2007; Friedman et al. 2001).

The lower panel of Fig. 5 shows the ranking of the relative importance of the variables provided by the random forest with 200 trees. The most important factors in estimating the pollen were the leaf area index, vegetation greenness function, and displacement height.

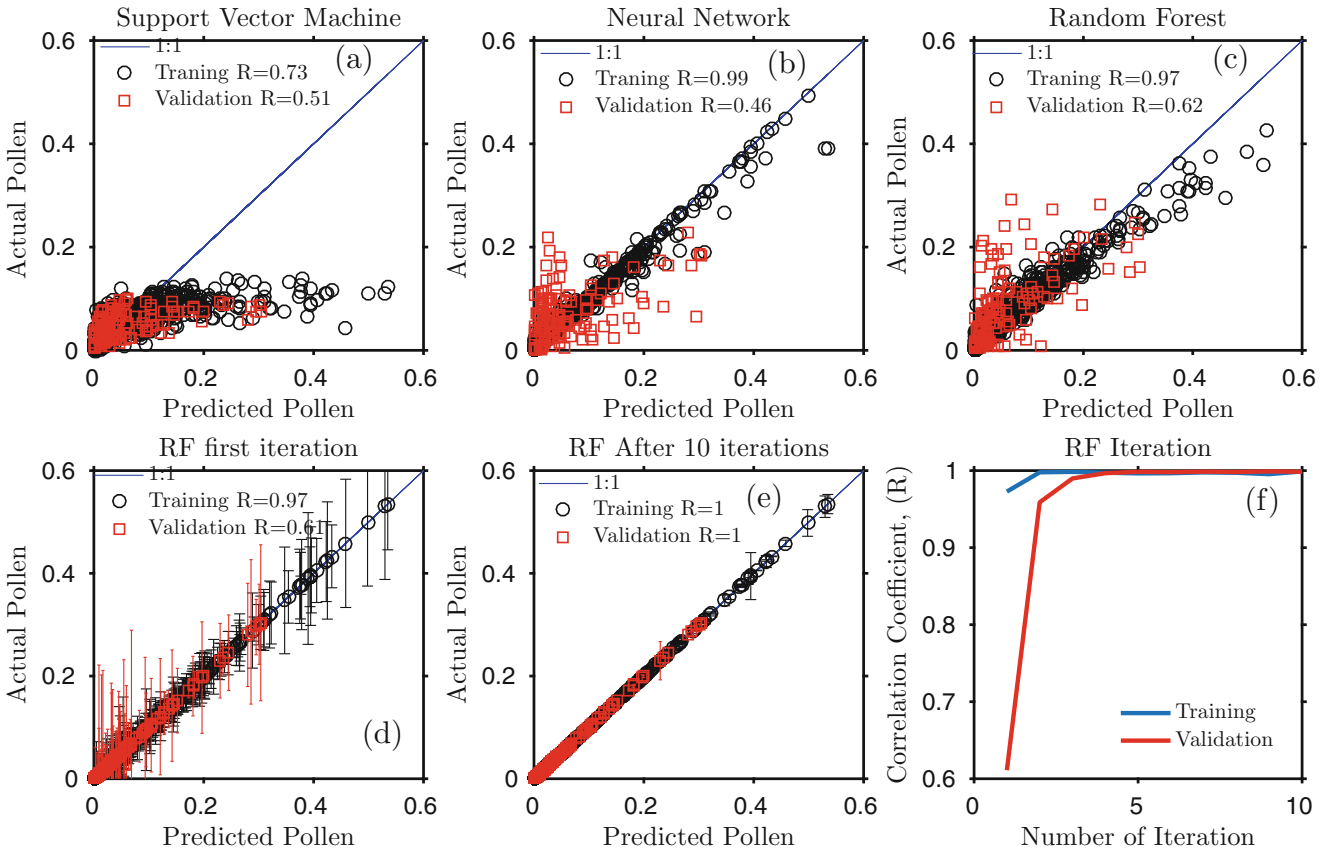


Fig. 4 Showing scatter plots of actual and predicted pollen for the support vector machine (panel **a**), neural network (panel **b**), and random forest (panel **c**). Panels (**d**), (**e**), and (**f**) show results of an iteration method applied to the random forest. Panels (**d**) and (**e**) show results of the first and 10th iteration for the training (black circles) and validation data (red squares). Panel (**f**) depicts plots of the correlation coefficient for the training and validation data versus iteration number

Dust Source Identification Using Unsupervised Classification

Unsupervised classification can be very useful when we would like to objectively split up our data into different regimes. A good example of this is a study to characterize dust sources (e.g., Fig. 6) (Lary et al. 2016a).

Dust sources of many kinds are found globally. One of the most salient features of dust sources is that they are often very localized. For example, in Figs. 6 and 8, we can clearly see that the source of the dust plumes are best described as an ensemble of many point sources, not broad dust emitting regions. Realistically capturing this very localized nature of dust sources has so far largely eluded automated diagnosis and, consequently, description in global models. Invariably current models describe dust sources as rather large-scale features, even when vegetation indices and similar approaches are used. This is in marked contrast to what we consistently see in the satellite imagery across the planet (e.g., Figs. 6 and 8).

Identifying dust sources is a critical yet challenging task for the accurate simulation of atmospheric particulate distributions relevant to air quality and climate change.

We take a new and radically different approach to any previous studies that have sought to identify global dust sources on a routine basis. We demonstrate that this new approach employing machine learning is very effective. The approach uses multi-wavelength spectral reflectivity signatures to characterize land surfaces, naturally paving the way for a new class of algorithms ideally suited to fully exploit the next generation of hyperspectral instruments. The production of thematic maps, such as those depicting land cover, using an image classification is one of the most common applications of remote sensing. New in our approach is that we can both operate at very high spatial resolution and distinguish between types of dust sources. For example, we can easily distinguish between the edge of salt flats (Fig. 8), dried-up wadis or lakes, and agricultural sources to name just three of many examples. The only limiting factor for the resolution is the resolution of the satellite imagery.

We employ machine learning to objectively provide an unsupervised multivariate and nonlinear classification into a very large number of surface types (in our demonstration study presented below, 1000 classes are used) using multispectral satellite data. In other words, we do not impose any

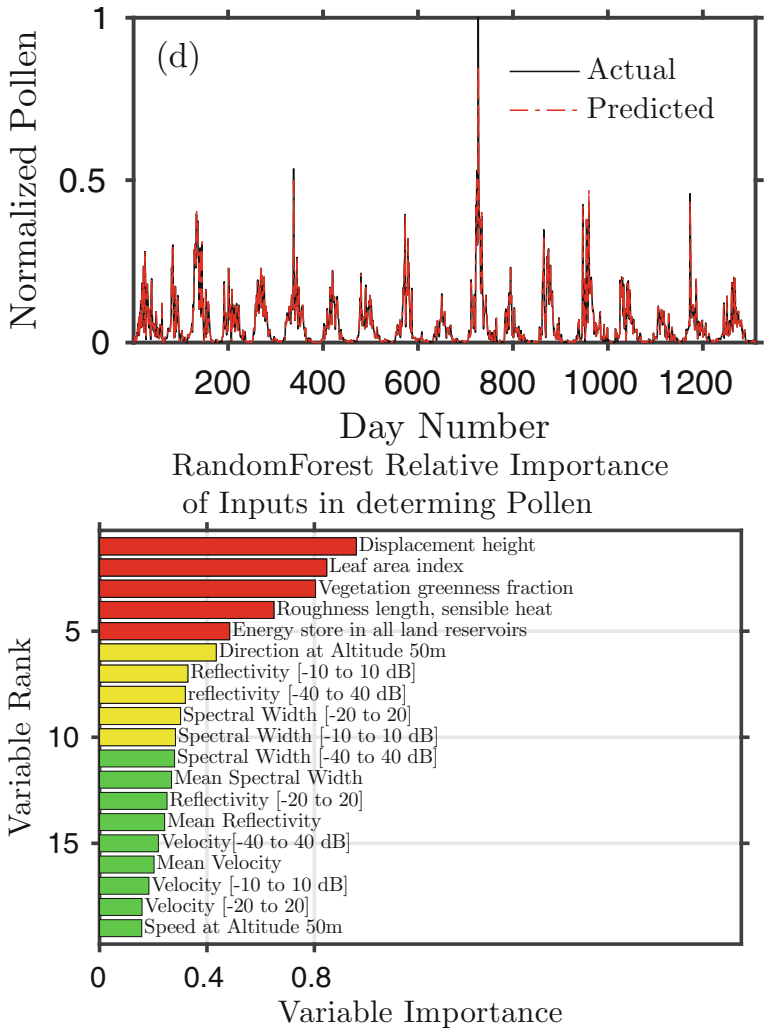


Fig. 5 The upper panel shows the comparison of actual and predicted pollen time series for Tulsa, OK. The lower panel shows the ranking of the relative importance of the variables provided by the random forest with 200 trees

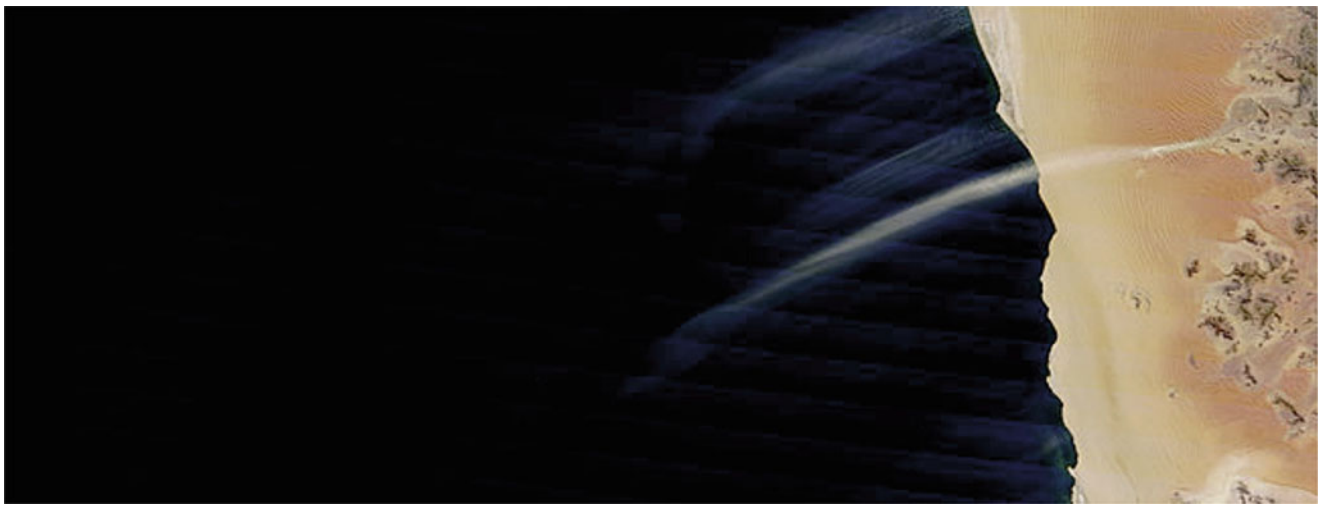


Fig. 6 Dust sources are typically localized point sources

a priori assumptions, but rather, we let the data speak for itself as to how we should classify surface types. Self-organizing maps (SOMs) are a data visualization and unsupervised classification technique invented by Professor Teuvo Kohonen that reduces the dimensions of data through the use of self-organizing neural networks.

SOMs help us address the issue that humans simply cannot visualize high-dimensional data unaided. The way SOMs go about reducing dimensionality is by producing a feature map, usually with two dimensions, that objectively plots the similarities of the data by grouping similar data items together. SOMs learn to classify input vectors according to how they are grouped in the input space. The SOM learns to recognize neighboring sections of the input space. Thus, SOMs learn both the distribution and topology of the input vectors they are trained on. This approach allows SOMs to display similarities and reduce the dimensionality. A SOM does not assume a priori a functional form for the analyzed data. A noteworthy enhancement of an SOM over principal component analysis is SOMs' ability to represent nonlinear functions or mappings.

The premise is that there are very many types of dust sources, from the diatom-rich sediments of the Bodélé depression in Chad to those at the edge of salt flats in Bolivia and Chile (Fig. 8) to those in the coastal Green Mountains of Libya. Each of these dust sources have distinct physical characteristics and therefore a distinct reflectance signature. If we are able to identify these signatures, then we can map the temporal and spatial evolution of each of these distinct dust sources. Once we have the surface-type classification, we then seek to identify which small subset of surface classes correspond to various kinds of dust sources. Once we have identified the signature of a wide variety of dust sources, we can precisely pick out these locations globally and how their distribution changes with time. This is particularly useful as dust sources are very localized and some dust sources have a significant seasonal time evolution. Having a methodology to identify the signature of these small-scale regions is invaluable.

The machine learning approach to dust source identification was first conceived in 2010 to face a very practical challenge that the Navy has in producing real-time visibility forecasts. If the standard type of dust sources is used [131], it was found that very poor regional visibility forecasts result. However, the quality of the Navy visibility forecasts drastically improved with an analyst (Annette Walker) manually identifying individual dust sources at the heads of plumes by examining sequences of satellite images such as those shown in Fig. 8 and also the EUMETSAT RGB Composites Dust images available online (http://oiswww.eumetsat.org/IPPS/html/MSG/RGB/DUST/). This methodology is very labor-intensive and does not lend itself to easy automation. The first prototype dust sources using the machine learning approach described here were devised specifically to automate the dust source identification and also allow for the accurate diagnosis of the time evolution in the spatial extent of the dust sources. Beyond the applications of accurate dust sources for visibility and air quality forecasts, the radiative forcing (RF) due to dust is a key concept in climate change calculations considered by the IPCC for the quantitative comparison of the strength of different human and natural agents causing climate change. Radiative forcing can be categorized into direct and indirect effects. A significant part of the direct effect is the mechanism by which aerosols scatter and absorb shortwave and longwave radiation, thereby altering the radiative balance of the Earth—atmosphere system. Mineral dust is a major component of global aerosols that exert a significant direct radiative forcing. Mineral dust aerosols are produced both naturally (≈70%) and anthropogenically (≈30%).

Our ultimate goal is to identify all the surface locations on the planet that are dust sources. To do this, we use a SOM to classify all the land surface locations into a very large set of n categories. In the examples shown here, $n = 1000$. A small subset of these 1000 categories will be regions that are dust sources. Naturally, there are a variety of distinct types of dust sources (e.g., dry river beds, agricultural sources, edge of salt flats, etc.) that we would like to delineate.

To achieve a comprehensive classification, we want to consider the conditions present throughout the year, so in the demonstration, we took an entire year of the 0.05° resolution MCD43C3 data product (Fig. 7). For this entire year of data, we then calculate the mean, μ, for each grid point. This is a massive dataset, and the computational time and memory required to perform the SOM classification increase with the number of data records. For the examples shown here, we therefore first restricted our attention to those broad MODIS surface types that may include dust sources, namely, barren or sparsely vegetated surfaces, croplands, grasslands, and open and closed shrublands. These are MODIS surface types 16, 12, 10, 7, and 6 respectively. For each of these surface types, we then constructed an input vector that contains 7 values, namely, for each of the seven bands provided in the MCD43C3 MODIS product, the mean, μ, of the directional and bihemispherical reflectance. When training the SOM, we use the Euclidean distance to compare the input vectors (each containing seven values).

In order to provide a fine gradation of classification, we use the SOM to group together the surface locations into 1000 classes, only a small subset of which correspond to regions that are dust sources. Once the classes that correspond to dust sources have been successfully identified, we have an automated method with which we can identify dust sources that can be routinely executed to provide a regular dust source data product that captures the spatial and temporal evolution of dust sources globally. We utilized the extensive hand classification of very localized dust sources produced by the Navy for the Middle East and South West Asia to guide our initial determination of which of the 1000 classes are dust sources. It is worth noting that the SOM classes are unique and distinct, and this will be seen below with the example of

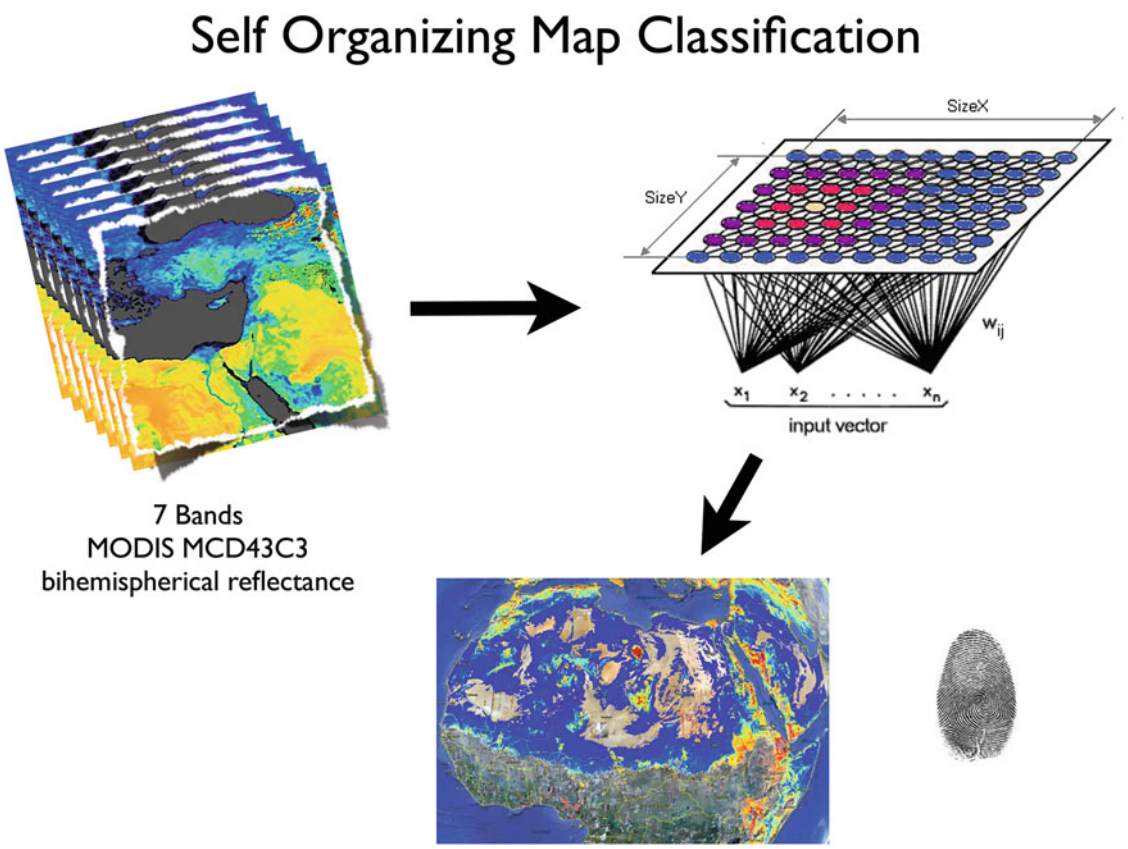


Fig. 7 Schematic of how self-organizing maps have been used in this study to classify land surface pixels into 1000 classes. Then a small subset of these classes are identified as dust sources

the Bodélé depression. Classes near each other are similar, but distinct.

Bolivia and Chile Salt Flats Dust Event

Figure 8 shows the dust event of July 18, 2010, in the Bolivian Altiplano. This event can be seen clearly in the MODIS Aqua True Color image where dust plumes emanate from fluviolacustine deposits and fluviodeltaic sediments around the Salars de Coipasa and Uyuni, Lake Poopo, and other smaller salt flats and lakes. Overlaid are the SOM classes that coincide with active dust sources on the Altiplano. Notice that the salt flats themselves are not dust sources; rather, we see the plumes forming around the edges of the flats and lakes. SOMs are very successful in identifying the unique spectral signatures of dust sources. A set of papers is in preparation describing an exhaustive atlas of the global dust sources.

Bodélé Depression Dust Event

Figure 9 shows the Bodélé depression dust event of January 9, 2011, at 09Z that started at 7Z and ceased at 18Z. The Bodélé depression is Chad's lowest point on the Sahara's southern edge. Typically, there are dust storms originating from the Bodélé depression on around 100 days a year that supplies the Amazon forest with the majority of its mineral dust (Washington and Todd 2005; Koren et al. 2006; Washington et al. 2006a; Todd et al. 2007; Bouet et al. 2012). The right panel shows the NRL processed EUMETSAT MSG/RGB satellite product. The two left panels show the dust sources identified by our approach with (lower) and without (upper) SOM class 137. The SOM had automatically determined that the sediment in the Bodélé depression was distinct from the surrounding dust sources and put it in a class all of its own, class 137. Indeed it is different; the Bodélé depression was once filled with a fresh water lake that has long since dried up (Washington et al. 2006b). This has left behind diatoms that now make up the surface of the depression. The two key points being, first that the dust source of the Bodélé is distinct from the surrounding dust sources and second that it consists of diatoms. This is interesting as if we could devise a way of distinguishing dust sources with containing certain biological materials, it would have significant applications for public health issues.

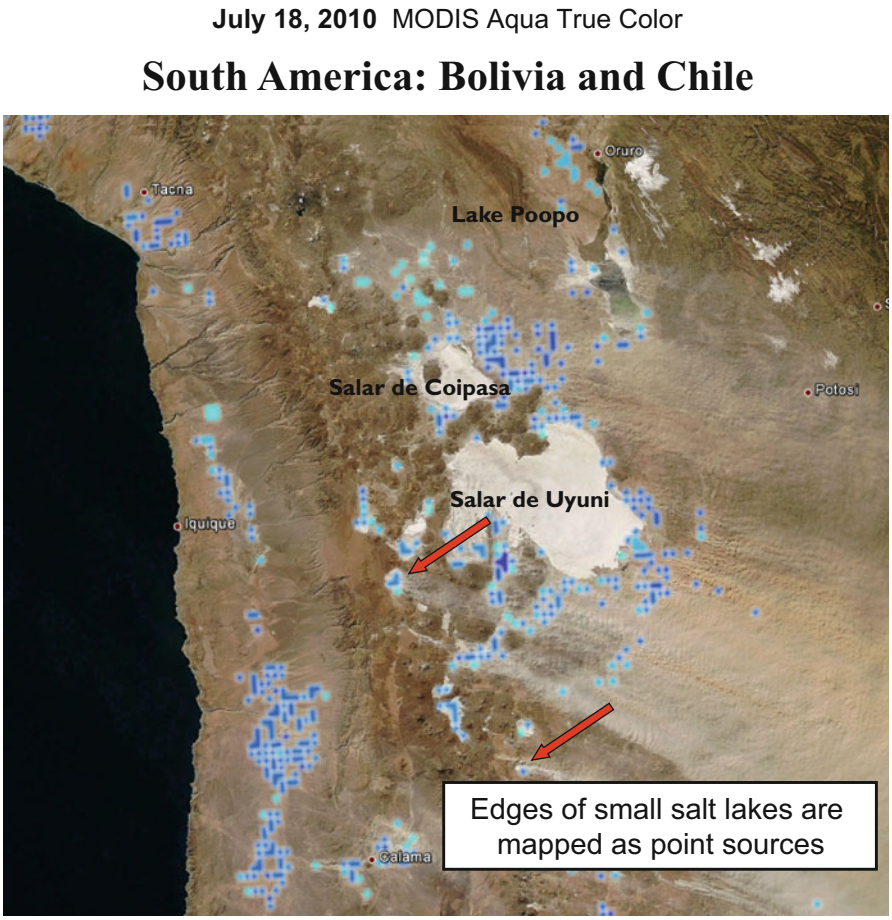


Fig. 8 Example of our machine learning approach correctly identifying very localized point sources around the edge of salt flats in Bolivia and Chile. Notice the narrow dust plumes originating from precisely the identified source regions that have been highlighted in blue and cyan

Some Likely Future Machine Learning Applications

Two recent advances are likely to open up a large number of new applications: first the improvement, size reduction, and cost reduction of hyperspectral imagery and secondly small embedded (credit card sized) GPU systems such as the NVIDIA Jetson TX1 with its 256 GPU cores.

Hyperspectral Imaging and Machine Learning for Real-Time Embedded Processing and Decision Support

So what is hyperspectral imaging? The human eye perceives the color of visible light in three bands using the cones, the photoreceptor cells in the retina (Fig. 10). These three bands are red (centered on 564 nm), green (centered on 534 nm), and blue (centered on 420 nm). By contrast, instead of using just three broad bands, hyperspectral cameras divide the spectrum into a very large number of narrow bands. Sometimes, as many as two to four hundred bands are used to create a hyperspectral datacube (Fig. 11). This technique of dividing images into bands can extend beyond the visible, into both the infrared and thermal infrared and into the ultraviolet (Fig. 12).

Hyperspectral imaging systems are used around the world in a variety of medical, laboratory, industrial, agricultural, and airborne applications. To illustrate the broader significance, let us briefly review just some of these (Fig. 13). Hyperspectral imaging (HSI) is used in various medical applications, especially in disease diagnosis and image-guided surgery. The disease diagnosis applications (e.g., skin examination) naturally lend themselves to telemedicine applications for rural communities where the network connectivity can drastically improve rural community medical care. For each snapshot in time, HSI acquires a three-dimensional dataset called a datacube (Fig. 11), with two spatial dimensions (just like a regular camera) and one spectral dimension, and there is a separate collocated image/layer for each wavelength band (Fig. 10).

Spatially resolved spectral imaging obtained by HSI can provide diagnostic information about the tissue physiology, morphology, and composition. With the advantage of acquiring two-dimensional images across a wide range of electromagnetic spectrum, HSI has been applied to numerous areas, including archaeology and art conservation (Angeletti

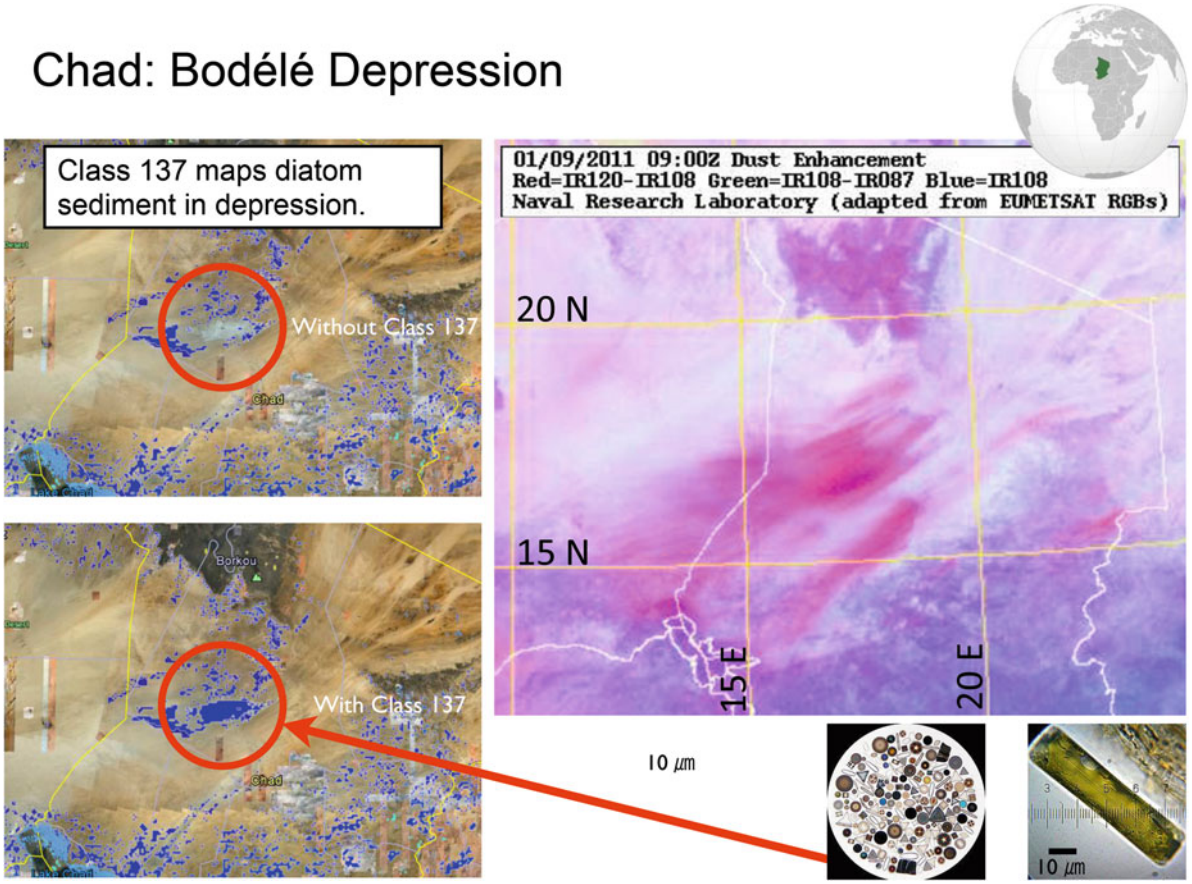


Fig. 9 The Bodélé depression dust event of January 9, 2011 (7Z–18Z). The right panel shows the NRL processed EUMETSAT MSG/RGB satellite product for January 9, 2011, 09Z. The two left panels show the dust sources identified by our approach with (lower) and without (upper) SOM Class 137. Lower right insets show microscopic images of Bodélé diatoms from soil samples taken from the depression

et al. 2005; Liang 2012), vegetation and water resource control (Govender et al. 2007), food quality and safety control (Gowen et al. 2007; Feng and Sun 2012), forensic medicine (Malkoff and Oliver 2000; Edelman et al. 2012), crime scene detection (Muller et al. 2003), biomedicine (Afromowitz et al. 1988; Carrasco et al. 2003), agriculture, security and defense, thin films, etc.

Figure 13 shows some of the many HSI applications. For example, using an airborne HSI, an invasive weed ('leafy spurge', Euphorbia esula) infestation could be clearly identified (Jay et al. 2010) and a weed coverage map generated (Fig. 13a). A study of seed germination (Nansen et al. 2015) using HSI showed that although viable and nonviable seeds appear identical to the human eye, they can be clearly distinguished using full reflectance spectra (Fig. 13b). Analysis of wound healing (La Fontaine et al. 2014) (Fig. 13c). Mapping hydrological formations (Fig. 13d). Fluorescent dye imaging (Fig. 13e). Examining the effect of surface pollution (Keith et al. 2009; Spangler et al. 2010) from leaking pipelines on vegetation (Fig. 13f). Checking food quality and fruit bruising (Fig. 13g). Classification of walnuts and shells (Fig. 13h). Automated analysis of cooked meats (Fig. 13i).

This diversity of examples demonstrates the general usefulness and applicability of HSI in a very broad range of contexts, in research, health, agriculture, industry, and more. We already saw in section "Dust Source Identification Using Unsupervised Classification" that combining the spectral signature in just seven wavelengths with machine learning was invaluable in uniquely identifying global dust sources with remarkable accuracy. So it can readily be seen that using more detailed hyperspectral signatures with on-board embedded processing can provide incredibly powerful insights in a very compact package. Figure 12 shows an example of some hyperspectral imagery we obtained using our aerial vehicles (Ramirez 2015). This approach is useful for many applications in smart agriculture, land surface classification, petrochemical surveying, disaster response (such as oil spills), etc. Let us take a closer look at the example of oil spill response.

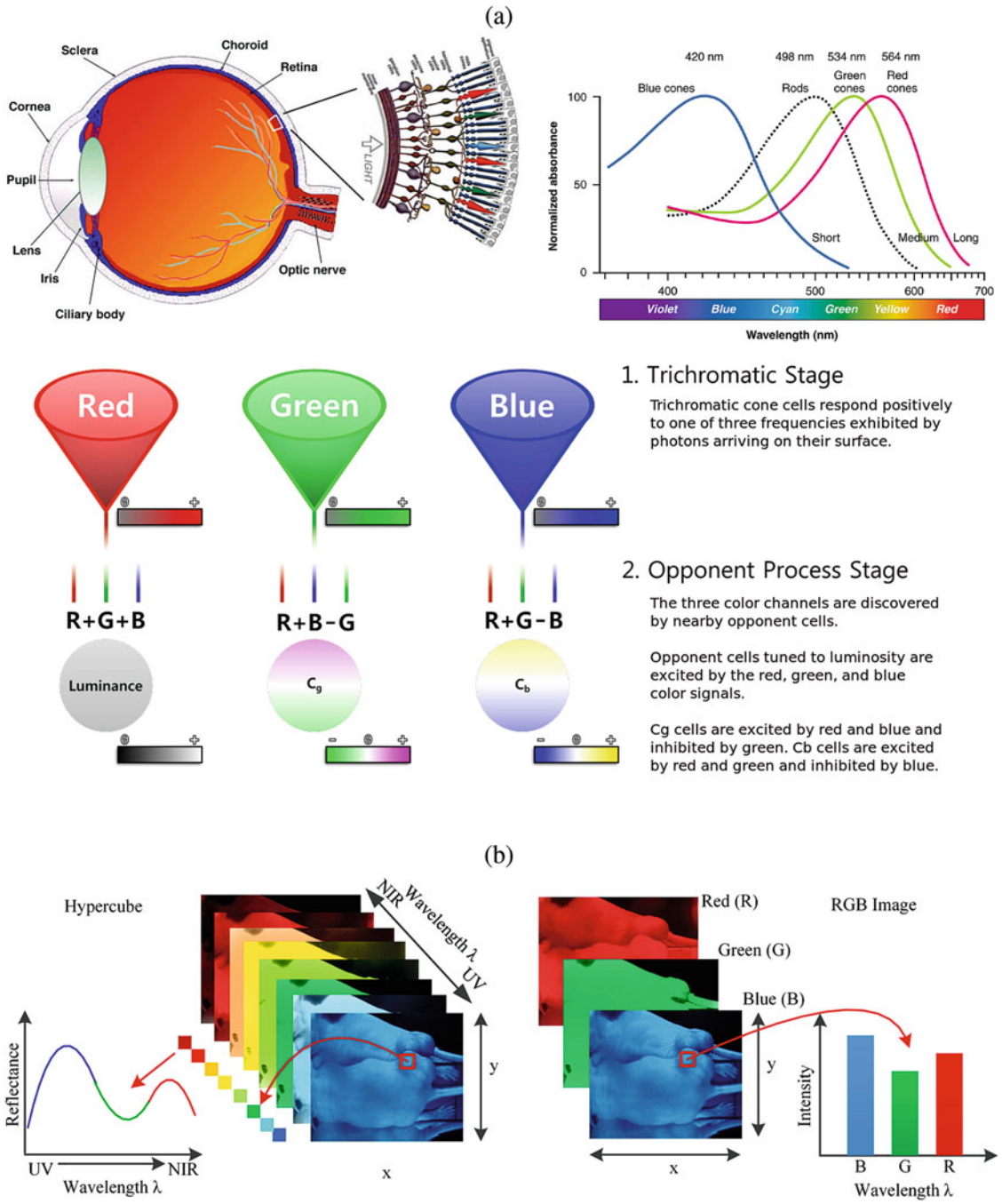


Fig. 10 Panel (**a**) Trichromatic cone cells in the eye respond to one of three wavelength ranges (RGB). These three bands are red (centered on 564 nm), green (centered on 534 nm), and blue (centered on 420 nm). Panel (**b**) shows a comparison between a hyperspectral cube and RGB images. A hyper-cube is a three-dimensional dataset consisting of two-dimensional image layers each for a different wavelength. So for each pixel in the image, we have a multi-wavelength spectra (spectral signature). This is shown schematically in the lower left. On the right, we see a conventional RGB color images with only three bands, images for red, green, and blue wavelengths. The lower right shows an example 3 wavelength broad band spectra from a conventional RGB color image

Oil Spills

The National Academy of Sciences estimates 1.7–8.8 million tons of oil are released into global waters every year. More than 70% of this release is related to human activities. The effects of these spills include dead wildlife, oil-covered marshlands, and contaminated water (Fingas and Brown 1997; Fingas 2010; Liu et al. 2013; Cornwall 2015). Spills of national significance (SONS), such as Deepwater Horizon (DWH), challenge response capabilities. In such

Fig. 11 Hyperspectral cube

large spills, *optimizing a coordinated response is a challenge*. There are always competing mission needs for aerial response resources such as helicopters and observer aircraft. Wildlife reconnaissance, oil observation overflights, and targeting chemical dispersant application are a few examples. If we consider just one aspect, i.e., the spill itself, the challenges include both characterizing the continual temporal and spatial evolution of the spill extent and the evolution of the oil itself as it weathers and emulsifies. Characterizing the oil spill can be made even more challenging due to the variable spill illumination and the weather. *Trained* observers are required, and their deployment needs can include a wide area, which is also challenging. Further, what is the optimal flight path(s) that should be used by the observers on each deployment to best meet the current needs and *anticipate the future evolution of the oil spill* to put in place any required preemptive measures or contingencies, such as shoreline precleaning or protective boom deployment? During the DWH oil spill operational trajectory forecasting, maps of key areas for aerial observations to improve trajectory modeling were produced daily by NOAA for the overflight teams.

The DWH oil spill and the associated impact monitoring was aided by extensive airborne and spaceborne passive and active remote sensing (Fingas and Brown 1997; Leifer et al. 2012; Liu et al. 2013; Fingas and Brown 2014). A good review of these remote sensing activities is provided by Leifer et al. (2012). During DWH, remote sensing was used to derive oil thickness (see Fig. 14) quantitatively for thick (>0.1 mm) slicks from AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) that measured 224 contiguous spectral bands with wavelengths from 400 to 2500 nanometers (nm) using a spectral library approach based on the shape and depth of near-infrared spectral absorption features (Kokaly et al. 2013; Leifer et al. 2012). MODIS (Moderate Resolution Imaging Spectroradiometer) satellite, visible-spectrum broadband data of surface-slick modulation of sunglint reflection, allowed extrapolation to the total slick. A multispectral expert system used a neural network approach to provide rapid response thickness class maps (Sveykovsky and Muskat 2006; Svejkovsky et al. 2009).

Airborne and satellite synthetic aperture radar (SAR) provides synoptic data under all-sky conditions (Liu et al. 2011; Leifer et al. 2012); however, SAR generally cannot discriminate thick (>100 μm) oil slicks from thin sheens (to 0.1 μm). The UAVSAR's (unmanned aerial vehicle SAR) significantly greater signal-to-noise ratio and finer spatial resolution allowed successful pattern discrimination related to a combination of oil slick thickness, fractional surface coverage, and emulsification.

Further, in situ burning and smoke plumes were studied with AVIRIS and corroborated spaceborne CALIPSO (Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observation) observations of combustion aerosols. CALIPSO and bathymetry lidar data documented shallow subsurface oil, although ancillary data were required for confirmation.

Airborne hyperspectral, thermal infrared data have nighttime and overcast collection advantages and were collected as well as MODIS thermal data. However, interpretation

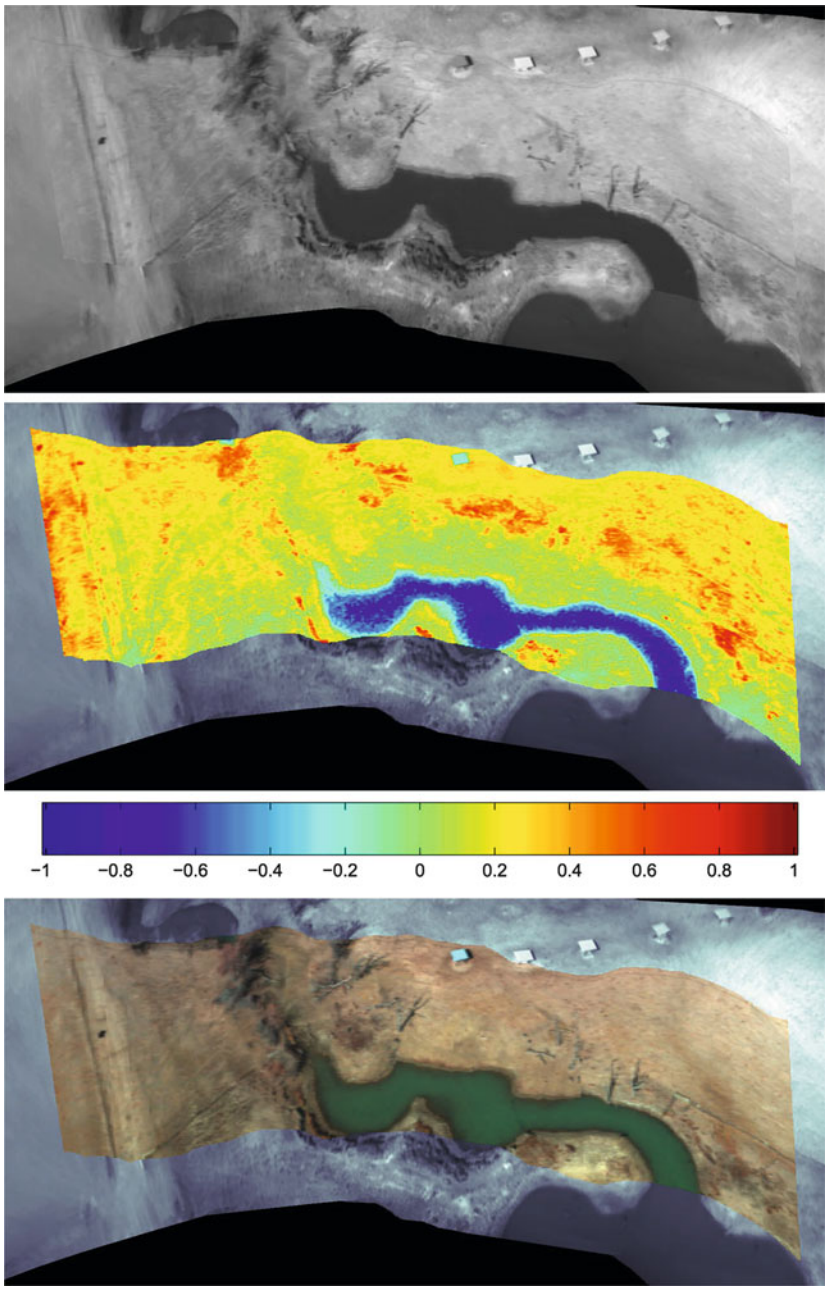


Fig. 12 Hyperspectral (HS) imaging of a rural landscape. Top image: sum of every spectral channel from the HS image, overlaid on top of the visible camera mosaic. Middle image: normalized difference vegetation index. Bottom image: pseudocolor from red, green, and blue channels (Ramirez 2015)

challenges and a lack of rapid response products prevented significant use. Rapid response products were key to response utilization—*data needs are time critical*; thus, a high technological readiness level is vital to the operational use of remote sensing products. The DWH oil spill experience demonstrated that development and operationalization of new near-real-time spill response remote sensing tools must precede the next major oil spill (Leifer et al. 2012).

Cleanup of a SONS involve *multiple* skimmer ships, vessels collecting oil for in situ burning, and chemical dispersant operations. Typically, these **slow**-moving response ships are spread over a **large** area and guided by air support (e.g., helicopter), as the vessel bridge is too low to see the variation in thickness of the oil. Typically, the manned air support will inform each ship of the location of recoverable oil ahead and then leave to overfly the next ship.

The cost of manned air support is significant, so each ship does *not* usually have dedicated continuous manned air support. In smaller spills, a report of oil location is given to the responding ship; these ships move slowly, so by the time they reach the location provided by manned air support, the oil has *moved*! For oil cleanup to be optimal (quickest)

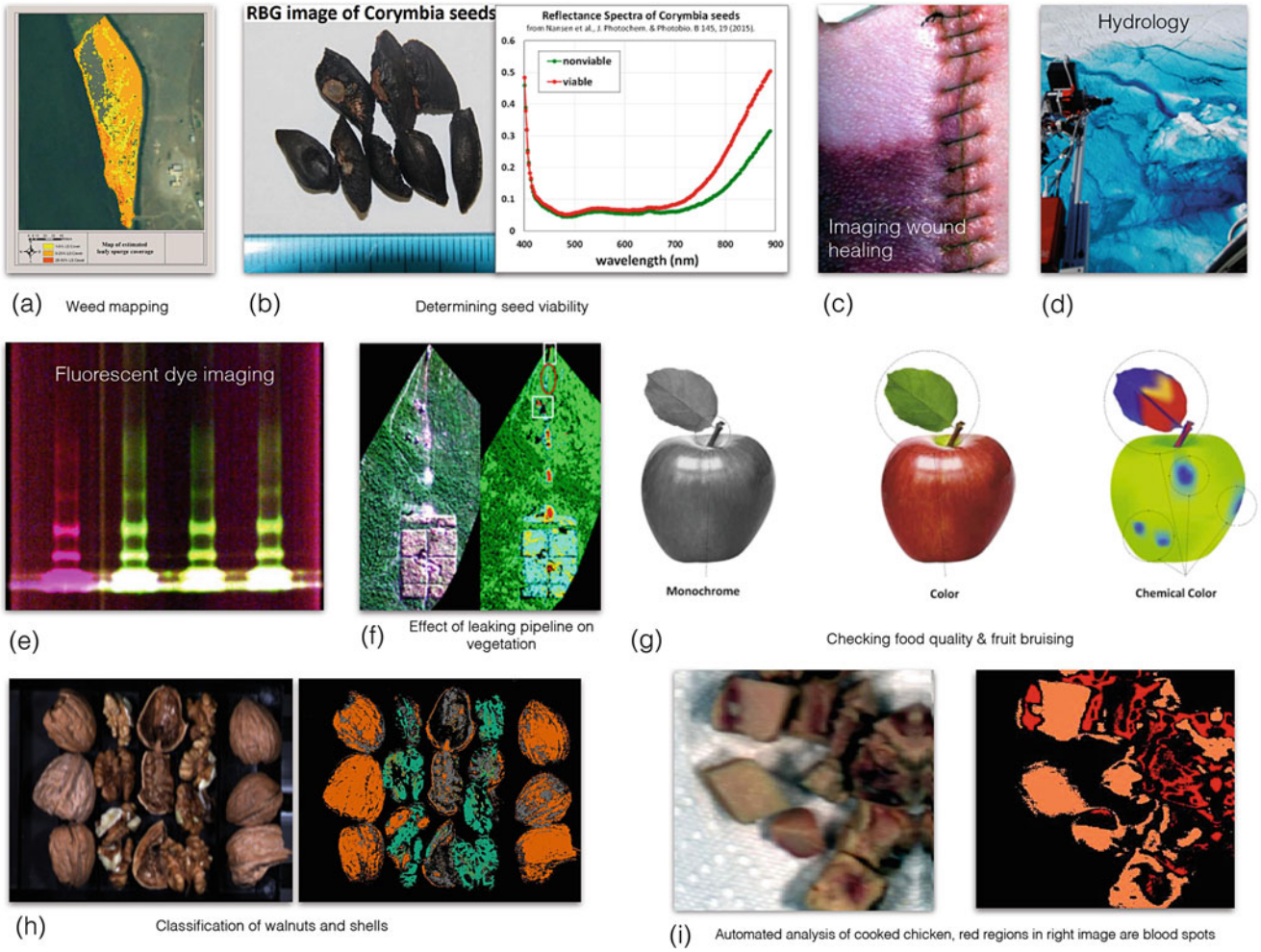


Fig. 13 Some examples of hyperspectral imaging applications

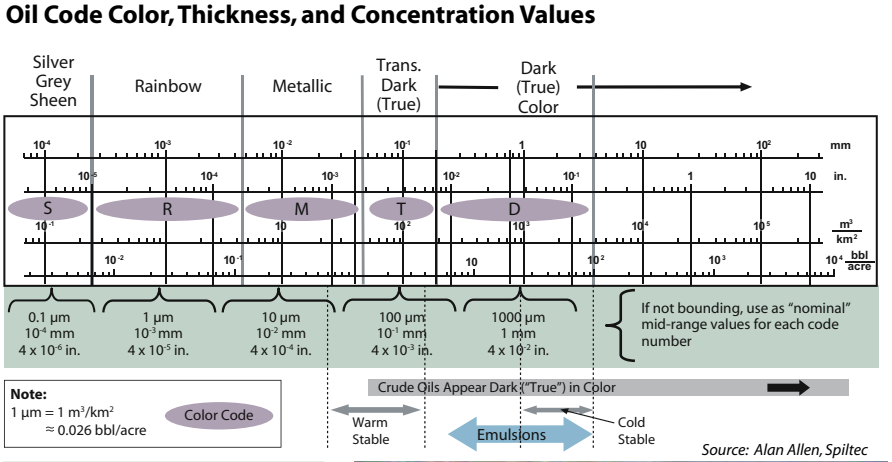
and effective (most oil recovered), the skimmer ships need to first focus on the regions of thickest oil. From the visual perspective of the ships' crew, relatively close to the water and with a shallow viewing angle, it is not easy to know where the thickest oil is. Accurately discerning the gradations in black oil thickness is a challenging task.

Rich information on the thickness of the oil and the degree of weathering is contained in the detailed hyperspectral signature of the oil spill. This information can be utilized through the use of machine learning to provide real-time response tools.

Summary

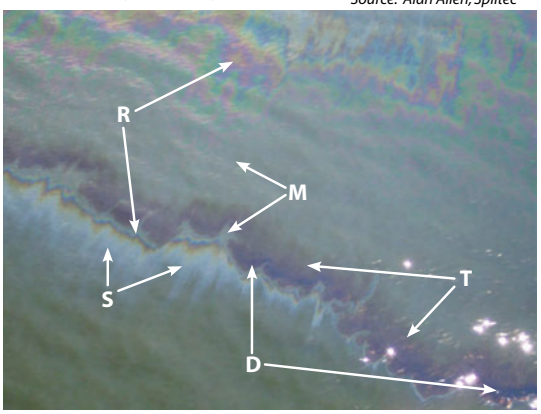
This chapter has given an overview of some examples that illustrate the usefulness of machine learning. Machines havebeen used to learn and make data-driven predictions for decades now. However, it is only relatively recently that machine learning has received widespread notoriety. The full potential of machine learning has yet to be reached. Machine learning is an automated approach to building empirical models from the data alone. Machine learning gives "computers the ability to learn without being explicitly programmed." Just as humans learn by experience, machine learning algorithms let computers learn from data. Machine learning also provides tools to give our data "a voice" (such as identifying characteristic patterns and signatures) and insights, such as which parameters are most important for accurately estimating a parameter of interest. A variety of tools exist that allow even a novice to readily utilize the power of machine learning. Machine learning enhances the readily available tool set to make data-driven decisions in a wide variety of scientific and societally relevant applications.

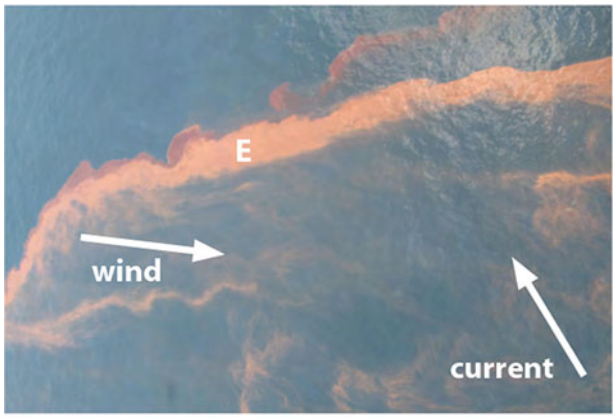
Fig. 14 Oil thickness chart and appearance from NOAA open water oil identification job aid



Common Descriptors	Code
Silver Sheen	S
Rainbow	R
Metallic	M
Transition	T
Dark	D
Emulsified	E

Note: "Structure" uses two lower-case letters, and "Color Codes" use single-letter capitals (R, S, M, T, D, E).





References

Afromowitz, M.A., J.B. Callis, D.M. Heimbach, L.A. Desoto, and M.K. Norton. 1988. Multispectral imaging of burn wounds—a new clinical instrument for evaluating burn depth. *IEEE Transactions on Biomedical Engineering* 35(10): 842–850. ISSN 0018-9294. https://doi.org/10.1109/10.7291.

Ahmad, Z., W. Choi, N. Sharma, J. Zhang, Q. Zhong, D.-Y. Kim, Z. Chen, Y. Zhang, R. Han, D. Shim, et al. 2016. Devices and circuits in CMOS for thz applications. In *Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM)*, pp. 29–8. New York: IEEE.

Alavi, Amir H., Amir H. Gandomi, and David J. Lary. 2016. *Progress of Machine Learning in Geosciences*.

Albayrak, A., J.C. Wei, Maksym Petrenko, D.J. Lary, and G.G. Leptoukh. 2011. Modis aerosol optical depth bias adjustment using machine learning algorithms. In *AGU Fall Meeting Abstracts*.

Andrews, Charles P., Paul H. Ratner, Benjamin R. Ehler, Edward G. Brooks, Brad H. Pollock, Dan A. Ramirez, and Robert L. Jacobs. 2013. The mountain cedar model in clinical trials of seasonal allergic rhinoconjunctivitis. *Annals of Allergy, Asthma and Immunology* 111(1): 9–13.

Angeletti, C., N.R. Harvey, V. Khomitch, A.H. Fischer, R.M. Levenson, and D.L. Rimm. 2005. Detection of malignancy in cytology specimens using spectral-spatial analysis. *Laboratory Investigation* 85(12): 1555–1564. ISSN 0023-6837. https://doi.org/10.1038/labinvest.3700357.

Arizmendi, C.M., J.R. Sanchez, N.E. Ramos, and G.I. Ramos. 1993. Time series predictions with neural nets: application to airborne pollen forecasting. *International Journal of Biometeorology* 37(3): 139–144.

Berman, Jesse D., Patrick N. Breysse, Ronald H. White, and Frank C. Curriero. 2012. Health-related benefits of attaining the daily and annual pm2. 5 air quality standards and stricter alternative standards. In *A103: Health Services Research and Administrative Databases*, pp. A2317–A2317. New York: American Thoracic Society.

Bishop, Christopher M. 1995. *Neural networks for pattern recognition*. Oxford: Oxford University.

Bouet, Christel, Guy Cautenet, Gilles Bergametti, Beatrice Marticorena, Martin C. Todd, and Richard Washington. 2012. Sensitivity of desert dust emissions to model horizontal grid spacing during the Bodele dust experiment 2005. *Atmospheric Environment* 50: 377–380.

Breiman, Leo. 1984. *Classification and regression trees*. In *The Wadsworth Statistics/Probability Series*. Belmont: Wadsworth International Group.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5–32. Times Cited: 3621.

Britton, J., I Pavord, K Richards, A Knox, A Wisniewski, I Wahedna, W Kinnear, A Tattersfield, and S Weiss. 1994. Factors influencing the occurrence of airway hyperreactivity in the general population: the importance of atopy and airway calibre. *European Respiratory Journal* 7(5): 881–887.

Brown, M.E., D.J. Lary, and H. Mussa. 2006. Using neural nets to derive sensor-independent climate quality vegetation data based on AVHRR, spot-vegetation, SeaWiFS and MODIS. In *AGU Spring Meeting Abstracts*.

Brown, Molly E., David J Lary, Anton Vrieling, Demetris Stathakis, and Hamse Mussa. 2008. Neural networks as a tool for constructing continuous ndvi time series from avhrr and modis. *International Journal of Remote Sensing* 29(24): 7141–7158.

Carrasco, O., R. Gomez, A. Chainani, and W. Roper. 2003. *Hyperspectral imaging applied to medical diagnoses and food safety*. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Spie)*, vol. 5097, pp. 215–221. ISBN 0277-786X 0-8194-4957-1. https://doi.org/10.1117/12.502589.

Castellano-Méndez, M., M.J. Aira, I. Iglesias, V. Jato, and W. González-Manteiga. 2005. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. *International Journal of Biometeorology* 49(5): 310–316.

Chang, Howard H., Anqi Pan, David J. Lary, Lance A. Waller, Lei Zhang, Bruce T. Brackin, Richard W. Finley, and Fazlay S. Faruque. 2019. Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS. *Environmental Monitoring and Assessment* 191(280): 1–10.

Choi, Wooyeol, Qian Zhong, Navneet Sharma, Yaming Zhang, Ruonan Han, Z Ahmad, Dae-Yeon Kim, Sandeep Kshattry, Ivan R Medvedev, David J Lary, et al. 2019. Opening terahertz for everyday applications. *IEEE Communications Magazine* 57(8): 70–76.

Cornwall, Warren. 2015. Deepwater horizon: After the oil. *Science* 348(6230): 22–29. https://doi.org/10.1126/science.348.6230.22.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3): 273–297. Times Cited: 3429.

Csépe, Z., László Makra, Dimitris Voukantsis, István Matyasovszky, Gábor Tusnády, Kostas Karatzas, and Michel Thibaudon. 2014. Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. *Science of the Total Environment* 476: 542–552.

D'amato, Gennaro, and Frits Th M. Spieksma. 1991. Allergenic pollen in Europe. *Grana* 30(1): 67–70.

Demuth, Howard B., Mark H. Beale, Orlando De Jess, and Martin T. Hagan. 2014. *Neural network design*. USA: Martin Hagan, 2nd ed. ISBN 0971-7321-16, 978-0-97-173211-7.

Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. London: Basic Books.

Edelman, G.J., E. Gaston, T.G. van Leeuwen, P.J. Cullen, and M.C.G. Aalders. 2012. Hyperspectral imaging for non-contact analysis of forensic traces. *Forensic Science International* 223(1–3): 28–39. ISSN 0379-0738. https://doi.org/10.1016/j.forsciint.2012.09.012.

EPA. 2019. Table of historical particulate matter (pm) national ambient air quality standards (NAAQS)—US EPA, history of the NAAQS for particulate matter, from 1971 to 2012. https://www.epa.gov/pm-pollution/table-historical-particulate-matter-pm-national-ambient-air-quality-standards-naaqs.

Ernst, P., H. Ghezzo, and M.R. Becklake. 2002. Risk factors for bronchial hyperresponsiveness in late childhood and early adolescence. *European Respiratory Journal* 20(3): 635–639.

Esch, Robert E., Cecelia J. Hartsell, Rodger Crenshaw, and Robert S. Jacobson. 2001. Common allergenic pollens, fungi, animals, and arthropods. *Clinical Reviews in Allergy and Immunology* 21(2): 261–292.

Esworthy, Robert. 2012. 2006 national ambient air quality standards (NAAQS) for fine particulate matter (pm2. 5): Designating nonattainment areas. In *Library of Congress, Congressional Research Service*.

Esworthy, Robert and James E. McCarthy. 2013. The national ambient air quality standards (NAAQS) for particulate matter (pm): EPA's 2006 revisions and associated issues. In Library of Congress, Congressional Research Service.

Etchie, Tunde O., Ayotunde T. Etchie, Gregory O. Adewuyi, Ajay Pillarisetti, Saravanadevi Sivanesan, Kannan Krishnamurthi, and Narendra K. Arora. 2018. The gains in life expectancy by ambient pm2. 5 pollution reductions in localities in Nigeria. *Environmental Pollution* 236: 146–157.

Faruque, Fazlay S., Hui Li, Worth B. Williams, Lance A. Waller, Bruce T. Brackin, Lei Zhang, Kim A. Grimes, and Richard W Finley. 2014. Geomedstat: An integrated spatial surveillance system to track air pollution and associated healthcare events. *Geospatial Health*, pp. S631–S646.

Feng, Y.Z., and D.W. Sun. 2012. Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition* 52(11): 1039–1058. ISSN 1040-8398. https://doi.org/10.1080/10408398.2011.651542.

Fingas, Marvin. 2010. *Oil spill science and technology*. London: Gulf Professional Publishing. ISBN: 978-1-85617-943-0.

Fingas, Mervin F., and Carl E. Brown. 1997. Review of oil spill remote sensing. *Spill Science and Technology Bulletin* 4(4): 199–208. ISSN 1353-2561. https://doi.org/10.1016/S1353-2561(98)00023-1. http://www.sciencedirect.com/science/article/pii/S1353256198000231. The Second International Symposium on Oil Spills.

Fingas, Merv, and Carl Brown. 2014. Review of oil spill remote sensing. *Marine Pollution Bulletin* 83(1): 9–23. ISSN 0025-326X. https://doi.org/10.1016/j.marpolbul.2014.03.059. http://www.sciencedirect.com/science/article/pii/S0025326X14002021.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*, vol. 1. Berlin: Springer.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. Variable selection using random forests. *Pattern Recognition Letters* 31(14): 2225–2236.

Giannadaki, Despina, Jos Lelieveld, and Andrea Pozzer. 2016. Implementing the US air quality standard for pm 2.5 worldwide can prevent millions of premature deaths per year. *Environmental Health* 15(1): 88.

Govender, M., K. Chetty, and H. Bulcock. 2007. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa* 33(2): 145–151. ISSN 0378-4738. <Go to ISI>://WOS:000246960100001.

Gowen, A.A., C.P. O'Donnell, P.J. Cullen, G. Downey, and J.M. Frias. 2007. Hyperspectral imaging—an emerging process analytical tool for food quality and safety control. *Trends in Food Science and Technology* 18(12): 590–598. ISSN 0924-2244. https://doi.org/10.1016/j.tifs.2007.06.001.

Guarnieri, Michael and John R. Balmes. 2014. Outdoor air pollution and asthma. *The Lancet* 383(9928): 1581–1592.

Haykin, Simon S. 1994. *Neural networks: A comprehensive foundation*. New York: Macmillan. 93028092 Simon Haykin. ill.; 26 cm. Includes bibliographical references (p. 635–690) and index.

Haykin, Simon S. 1999. *Neural networks: A comprehensive foundation*. Upper Saddle River: Prentice Hall, 2nd ed. 98007011 Simon Haykin. ill.; 25 cm. Includes bibliographical references (p. 796–836) and index.

Haykin, Simon S. 2001. *Kalman filtering and neural networks*. In *Adaptive and learning systems for signal processing, communications, and control*. New York: Wiley. 2001049240 ed. by Simon Haykin. ill.; 24 cm. "A Wiley Interscience publication." Includes bibliographical references and index.

Haykin, Simon S. 2007. *New directions in statistical signal processing: From systems to brain*. In *Neural Information Processing Series*. Cambridge: MIT Press.

Ho, T.K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832–844.

Howard, Lauren Eileen and Estelle Levetin. 2014. Ambrosia pollen in Tulsa, Oklahoma: aerobiology, trends, and forecasting model development. *Annals of Allergy, Asthma and Immunology* 113(6): 641–646.

Jay, Steven C., Rick L. Lawrence, Kevin S. Repasky, Lisa J. Rew, and IEEE. 2010. Detection of leafy spurge using hyper-spectral-spatial-temporal imagery. *2010 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4374–4376. https://doi.org/10.1109/igarss.2010.5652580.

Johnson, Philip R.S., and John J. Graham. 2005. Fine particulate matter national ambient air quality standards: Public health impact on populations in the northeastern united states. *Environmental Health Perspectives* 113 (9): 1140–1147.

Kasprzyk, Idalia. 2008. Non-native ambrosia pollen in the atmosphere of rzeszów (se poland); evaluation of the effect of weather conditions on daily concentrations and starting dates of the pollen season. *International Journal of Biometeorology* 52(5): 341–351

Keith, Charlie J., Kevin S. Repasky, Rick L. Lawrence, Steven C. Jay, and John L. Carlsten. 2009. Monitoring effects of a controlled subsurface carbon dioxide release on vegetation using a hyperspectral imager. *International Journal of Greenhouse Gas Control* 3(5): 626–632. ISSN 1750-5836. https://doi.org/10.1016/j.ijggc.2009.03.003.

Kinney, Patrick L. 2008. Climate change, air quality, and human health. *American Journal of Preventive Medicine* 35(5): 459–467.

Kneen, M.A., David J. Lary, William A. Harrison, Harold J. Annegarn, and Tom H. Brikowski. 2016. Interpretation of satellite retrievals of pm2.5 over the Southern African interior. *Atmospheric Environment* 128: 53–64.

Kokaly, Raymond F., Brady R. Couvillion, JoAnn M. Holloway, Dar A. Roberts, Susan L. Ustin, Seth H. Peterson, Shruti Khanna, and Sarai C. Piazza. 2013. Spectroscopic remote sensing of the distribution and persistence of oil from the Deepwater Horizon spill in Barataria Bay marshes. *Remote Sensing of Environment* 129(0): 210–230. ISSN 0034-4257. https://doi.org/10.1016/j.rse.2012.10.028. http://www.sciencedirect.com/science/article/pii/S0034425712004166.

Koren, Ilan, Yoram J. Kaufman, Richard Washington, Martin C. Todd, Yinon Rudich, J. Vanderlei Martins, and Daniel Rosenfeld. 2006. The bodele depression: A single spot in the Sahara that provides most of the mineral dust to the amazon forest. *Environmental Research Letters* 1(1): 014005.

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. 2007. *Supervised Machine Learning: A Review of Classification Techniques*.

Laaidi, Mohamed, Karine Laaidi, Jean-Pierre Besancenot, and Michel Thibaudon. 2003. Ragweed in France: an invasive plant and its allergenic pollen. *Annals of Allergy, Asthma and Immunology* 91(2): 195–201.

La Fontaine, Javier, Lawrence Lavery, and Karel Zuzak. 2014. *The use of hyperspectral imaging (HSI) in wound healing*. In *Proceedings of SPIE*, vol. 8979. ISBN 0277-786X 978-0-8194-9892-2. https://doi.org/10.1117/12.2041841.

Lary, D. 2007. Using neural networks for instrument cross-calibration. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2010. *Artificial intelligence in geoscience and remote sensing*. London: INTECH Open Access Publisher.

Lary, David J. 2013. Using multiple big datasets and machine learning to produce a new global particulate dataset: A technology challenge case study. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2014. Bigdata and machine learning for public health. In *Proceedings of the 142nd APHA Annual Meeting and Exposition 2014*. Washington: APHA.

Lary, D.J., and O. Aulov. 2008. Space-based measurements of HCL: Intercomparison and historical context. *Journal of Geophysical Research: Atmospheres* 113(D15).

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2003. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics Discussions* 3(6): 5711–5724.

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2004. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics* 4(1): 143–146.

Lary, David J., L.A. Remer, Devon MacNeill, Bryan Roscoe, and Susan Paradise. 2009. Machine learning and bias correction of modis aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* 6(4): 694–698.

Lary, D.J., A. Nikitkov, D. Stone, and Alexey Nikitkov. 2010. Which machine-learning models best predict online auction seller deception risk. In *American Accounting Association AAA Strategic and Emerging Technologies*.

Lary, David J., Fazlay S. Faruque, Nabin Malakar, Alex Moore, Bryan Roscoe, Zachary L. Adams, and York Eggelston. 2014. Estimating the global abundance of ground level presence of particulate matter (pm2. 5). *Geospatial Health* 8(3): 611–630.

Lary, D.J., T. Lary, and B. Sattler. 2015. Using machine learning to estimate global pm2. 5 for environmental health studies. *Environmental Health Insights* 9(Suppl 1): 41.

Lary, David J., Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. 2016a. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7(1): 3–10.

Lary, David J., Amir H Alavi, Amir H Gandomi, and Annette L Walker. 2016b. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7(1): 3–10.

Lary, David J., Tatiana Lary, and B. Sattler. 2016c. Using machine learning to estimate global particulate matter for environmental health studies. *Geoinformatics and Geostatistics: An Overview* 4(4): doi–10.

Lary, David J., Gebreab K. Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J. Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, et al. 2018. Machine learning applications for earth observation. In *Earth Observation Open Science and Innovation. ISSI Scientific Report Series*, vol. 15, pp. 165–218. Berlin: Springer.

Lary, Maria-Anna, Leslie Allsop, and David John Lary. 2019. Using machine learning to examine the relationship between asthma and absenteeism. *Environmental Modeling and Assessment* 191(332), 1–9.

Leifer, Ira, William J. Lehr, Debra Simecek-Beatty, Eliza Bradley, Roger Clark, Philip Dennison, Yongxiang Hu, Scott Matheson, Cathleen E. Jones, Benjamin Holt, Molly Reif, Dar A. Roberts, Jan Svejkovsky, Gregg Swayze, and Jennifer Wozencraft. 2012. State of the art satellite and airborne marine oil spill remote sensing: Application to the BP deepwater horizon oil spill. *Remote Sensing of Environment* 124(0): 185–209. ISSN 0034-4257. https://doi.org/10.1016/j.rse.2012.03.024. http://www.sciencedirect.com/science/article/pii/S0034425712001563.

Lewis, Walter Hepworth, Prathibha Vinay, and Vincent E. Zenger. 1983. *Airborne and allergenic pollen of North America*. New York: Johns Hopkins University Press.

Liang, H. 2012. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Applied Physics a-Materials Science and Processing* 106(2): 309–323. ISSN 0947-8396.

Liu, Peng, Xiaofeng Li, John J. Qu, Wenguang Wang, Chaofang Zhao, and William Pichel. 2011. Oil spill detection with fully polarimetric {UAVSAR} data. *Marine Pollution Bulletin* 62(12): 2611–2618. ISSN 0025-326X. https://doi.org/10.1016/j.marpolbul.2011.09.036. http://www.sciencedirect.com/science/article/pii/S0025326X11005248.

Liu, Y., A. MacFadyen, Z.G. Ji, and R.H. Weisberg. 2013. *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record Breaking Enterprise*. Geophysical Monograph Series. New York: Wiley. ISBN 978-1-11-867182-5.

Low, Ronald B., Leonard Bielory, Adnan I. Qureshi, Van Dunn, David F.E. Stuhlmiller, and David A. Dickey. 2006. The relation of stroke admissions to recent weather, airborne allergens, air pollution, seasons, upper respiratory infections, and asthma incidence, September 11, 2001, and day of the week. *Stroke* 37(4): 951–957.

Malakar, Nabin K., David J Lary, A Moore, D Gencaga, Bryan Roscoe, Arif Albayrak, and Jennifer Wei. 2012a. Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing. In *Proceedings of the 2012 Conference on Intelligent Data Understanding*, pp. 24–30. New York: IEEE.

Malakar, N.K., D.J. Lary, R. Allee, R. Gould, and D. Ko. 2012b. Towards automated ecosystem-based management: A case study of northern gulf of Mexico water. In *AGU Fall Meeting Abstracts*.

Malakar, N.K., D.J. Lary, D. Gencaga, A. Albayrak, and J. Wei. 2013. Towards identification of relevant variables in the observed aerosol optical depth bias between modis and aeronet observations. In *AIP Conference Proceedings*, vol. 1553, pp. 69–76. College Park: AIP.

Malakar, Nabin K., D.J. Lary, and B. Gross. 2018. Case studies of applying machine learning to physical observation. In *AGU Fall Meeting Abstracts*.

Malkoff, D.B., and W.R. Oliver. 2000. *Hyperspectral imaging applied to forensic medicine*. In *Progress in Biomedical Optics*, vol. 1, pp. 108–116. ISBN 1017-2661 0-8194-3536-8. <Go to ISI>://WOS:000086469300013.

Matheson, Eric M., Marty S. Player, Arch G. Mainous, Dana E. King, and Charles J. Everett. 2008. The association between hay fever and stroke in a cohort of middle aged and elderly adults. *The Journal of the American Board of Family Medicine* 21(3): 179–183.

McCulloch, W.S., and W. Pitts. 1943. *Bulletin of Mathematical Biophysics*, vol. 5, p. 115. https://doi.org/10.1007/BF02478259.

Medvedev, Ivan R., Robert Schueler, Jessica Thomas, O. Kenneth, Hyun-Joo Nam, Navneet Sharma, Qian Zhong, David J Lary, and Philip Raskin. 2016. Analysis of exhaled human breath via terahertz molecular spectroscopy. In *Proceedings of the 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz)*, pp. 1–2. New York: IEEE.

Mirabelli, Maria C., Ambarish Vaidyanathan, W. Dana Flanders, Xiaoting Qin, and Paul Garbe. 2016. Outdoor pm2. 5, ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. *Environmental Health Perspectives* 124(12): 1882–1890.

Muller, M.G., T.A. Valdez, I. Georgakoudi, V. Backman, C. Fuentes, S. Kabani, N. Laver, Z.M. Wang, C.W. Boone, R.R. Dasari, S.M. Shapshay, and M.S. Feld. 2003. Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma. *Cancer* 97(7): 1681–1692. ISSN 0008-543X. https://doi.org/10.1002/cncr.11255.

Nansen, Christian, Genpin Zhao, Nicole Dakin, Chunhui Zhao, and Shane R. Turner. 2015. Using hyperspectral imaging to determine germination of native Australian plant seeds. *Journal of Photochemistry and Photobiology B-Biology* 145: 19–24. ISSN 1011-1344. https://doi.org/10.1016/j.jphotobiol.2015.02.015.

Nathan, Brian J., and David J. Lary. 2019. Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale. *Environmental Modeling and Assessment* 191(337): 1–17.

Nowosad, Jakub. 2015. Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus, and Betula. *International Journal of Biometeorology*, 1–13.

O, K.K., Q. Zhong, N. Sharma, W. Choi, R. Schueler, I. R. Medvedev, H.-J. Nam, P. Raskin, F. C. De Lucia, J. P. McMillan, et al. 2017. Demonstration of breath analyses using CMOS integrated circuits for rotational spectroscopy. In *Proceedings of the International Workshop on Nanodevice Technologies, Hiroshima, Japan*.

Osowski, Stanislaw and Konrad Garanty. 2007. Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence* 20(6): 745–755.

Oswalt, Matthew L., and Gailen D. Marshall. 2008. Ragweed as an example of worldwide allergen expansion. *Allergy, Asthma and Clinical Immunology* 4(3): 130.

Puc, Małgorzata. 2012. Artificial neural network model of the relationship between Betula pollen and meteorological factors in Szczecin (Poland). *International Journal of Biometeorology* 56(2): 395–401.

Ramirez, Daniel A. 1984. The natural history of mountain cedar pollinosis. *Journal of Allergy and Clinical Immunology* 73(1): 88–93.

Ramirez, J.P., D.J. Lary, N. Gans. 2015. Low-altitude terrestrial spectroscopy from a pushbroom sensor. *Journal of Field Robotics*, 1–16. https://doi.org/10.1002/rob.21624.

Rodríguez-Rajo, F.J., Gonzalo Astray, J.A. Ferreiro-Lage, M.J. Aira, M.V. Jato-Rodriguez, and Juan C. Mejuto. 2010. Evaluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. *Neural Networks* 23(3): 419–425.

Safavian, S.R., and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 21(3): 660–674. ISSN 0018-9472. https://doi.org/10.1109/21.97458.

Sánchez-Mesa, J.A., C. Galán, J.A. Martínez-Heras, and C. Hervás-Martínez. 2002. The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula. *Clinical and Experimental Allergy* 32(11): 1606–1612.

Sassen, Kenneth. 2008. Boreal tree pollen sensed by polarization lidar: Depolarizing biogenic chaff. *Geophysical Research Letters* 35(18).

Spangler, Lee H., Laura M. Dobeck, Kevin S. Repasky, Amin R. Nehrir, Seth D. Humphries, Jamie L. Barr, Charlie J. Keith, Joseph A. Shaw, Joshua H. Rouse, Alfred B. Cunningham, Sally M. Benson, Curtis M. Oldenburg, Jennifer L. Lewicki, Arthur W. Wells, J. Rodney Diehl, Brian R. Strazisar, Julianna E. Fessenden, Thom A. Rahn, James E. Amonette, Jon L. Barr, William L. Pickles, James D. Jacobson, Eli A. Silver, Erin J. Male, Henry W. Rauch, Kadie S. Gullickson, Robert Trautz, Yousif Kharaka, Jens Birkholzer, and Lucien Wielopolski. 2010. A shallow subsurface controlled release facility in Bozeman, Montana, USA, for testing near surface co2 detection techniques and transport models. *Environmental Earth Sciences* 60(2): 227–239. ISSN 1866-6280. https://doi.org/10.1007/s12665-009-0400-2.

Stark, Paul C., Louise M. Ryan, James L. McDonald, and Harriet A. Burge. 1997. Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia* 13(3): 177–184.

Sveykovsky, J., and S. Muskat. 2006. Real-time detection of oil slick thickness patterns with a portable multispectral sensor. Technical report, July 31, 2006.

Svejkovsky, Jan, Judd Muskat, and Joseph Mullin. 2009. Adding a multispectral aerial system to the oil spill response arsenal. *Sea Technology* 50(8): 17–+.

Todd, Martin C., Richard Washington, Jose Vanderlei Martins, Oleg Dubovik, Gil Lizcano, Samuel M'Bainayel, and Sebastian Engelstaedter. 2007. Mineral dust emission from the Bodele depression, Northern Chad, during BodEx 2005. *Journal of Geophysical Research-Atmospheres* 112(D6).

Tränkle, Eberhard and Bernd Mielke. 1994. Simulation and analysis of pollen coronas. *Applied Optics* 33(21): 4552–4562.

Vapnik, Vladimir Naumovich. 1982. *Estimation of dependences based on empirical data*. In *Springer Series in Statistics*. New York: Springer.

Vapnik, Vladimir Naumovich. 1995. *The nature of statistical learning theory*. New York: Springer.

Vapnik, Vladimir Naumovich. 2000. *The nature of statistical learning theory*. In *Statistics for Engineering and Information Science*, 2nd edn. New York: Springer.

Vapnik, Vladimir Naumovich. 2006. *Estimation of dependences based on empirical data; empirical inference science: Afterword of 2006*. In *Information Science and Statistics*, 2nd ed. New York: Springer.

Voukantsis, Dimitris, Harri Niska, Kostas Karatzas, Marina Riga, Athanasios Damialis, and Despoina Vokou. 2010. Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. *Atmospheric Environment* 44(39): 5101–5111

Washington, R., and M.C. Todd. 2005. Atmospheric controls on mineral dust emission from the Bodele depression, chad: The role of the low level jet. *Geophysical Research Letters* 32 (17).

Washington, R., M.C. Todd, G. Lizcano, I. Tegen, C. Flamant, I. Koren, P. Ginoux, S. Engelstaedter, C.S. Bristow, C.S. Zender, A.S. Goudie, A. Warren, and J.M. Prospero. 2006a. Links between topography, wind, deflation, lakes and dust: The case of the Bodele depression, chad. *Geophysical Research Letters* 33(9).

Washington, R., M.C. Todd, S. Engelstaedter, S. Mbainayel, and F. Mitchell. 2006b. Dust and the low-level circulation over the Bodele depression, Chad: Observations from BodEx 2005. *Journal of Geophysical Research-Atmospheres* 111(D3).

Wayne, Peter, Susannah Foster, John Connolly, Fakhri Bazzaz, and Paul Epstein. 2002. Production of allergenic pollen by ragweed (Ambrosia artemisiifolia l.) is increased in CO2-enriched atmospheres. *Annals of Allergy, Asthma and Immunology* 88(3): 279–282.

Weinhold, Bob. 2013. *New primary standard set for fine particulate matter*.

WHO. *7 million premature deaths annually linked to air pollution*. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Accessed: 2016-08-29.

Wu, Daji, Gebreab K. Zewdie, Xun Liu, Melanie Kneed, and David J. Lary. 2017. Insights into the morphology of the East Asia pm2.5 annual cycle provided by machine learning. *Environmental Health Insights* 11: 1–7.

Wu, Daji, David J. Lary, Gebreab K Zewdie, and Xun Liu. 2019. Using machine learning to understand the temporal morphology of the pm2.5 annual cycle in East Asia. *Environmental Monitoring and Assessment* 191(272), 1–14.

Zewdie Gebreab, and David J. Lary. 2018. Applying machine learning to estimate allergic pollen using environmental, land surface and NEXRAD radar parameters. In *AGU Fall Meeting Abstracts*.

Zewdie, Gebreab K., David J. Lary, Estelle Levetin, and Gemechu F. Garuma. 2019a. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *International Journal of Environmental Research and Public Health* 16(11): 1992.

Zewdie, Gebreab K., David J. Lary, Xun Liu, Daji Wu, and Estelle Levetin. 2019b. Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. *Environmental Monitoring and Assessment* 191(7): 418.

Zhao, Feng, Amr Elkelish, Jörg Durner, Christian Lindermayr, J. Barbro Winkler, Franziska Ruff, Heidrun Behrendt, Claudia Traidl-Hoffmann, Andreas Holzinger, Werner Kofler, et al. 2016. Common ragweed (Ambrosia artemisiifolia L.): allergenicity and molecular characterization of pollen after plant exposure to elevated NO2. *Plant, Cell and Environment* 39(1): 147–164.

Advancement in Airborne Particulate Estimation Using Machine Learning

Lakitha Omal Harindha Wijeratne, Gebreab K. Zewdie, Daniel Kiv, Adam Aker, David J. Lary, Shawhin Talebi, Xiaohe Yu, and Estelle Levetin

Introduction

The air we breathe is vital and largely invisible (except when the pollution levels are very high). Every single minute, an average human being breathes around 10 liters of air. However, we often do not think about the composition of the air that we breathe and the impact it may be having on our health. Often, the air we breathe contains pollutant particles. Although it is apparent that air pollution results in increased hospital visits, missed school days, as well as missed work days (due to respiratory diseases), it is harder to localize exactly where unhealthy air resides. The World Health Organization (WHO) reports that nine out of ten people worldwide breathe polluted air which results in an estimated 7 million deaths per year (Nada Osseiran 2018).

Air Pollution Episodes in History

Historically, we have seen that air pollution episodes can result in significant loss of life. A few of example episodes include:

- Great smog of London (1952): In December of 1952, a severe smog covered many parts of the British Isles (Wilkins 1954). The episode lasted 5 days (December 5–December 9, 1952), and more than 4000 deaths occurred before the end of the year. Within the next 10 weeks, a further 8000 people lost their lives (Black 2003). The primary cause of the episode was extensive burning of high-sulfur coal (Polivka 2018). Following the incident, the British Parliament passed the Clean Air Act of 1956 which restricted burning of coal in urban areas.
- New York City smog (1966): On the Thanksgiving weekend of 1966, smog containing damaging levels of toxic pollution (comprised of carbon monoxide and sulfur dioxide) covered New York City. Carlson (2009) reports that Thanksgiving weekend of 1966 was the smoggiest day in the city's history. Although regional leaders announced a first-stage alert, it is believed that more than 200 people lost their lives due to the air pollution episode. In fact, Glasser et al. (1967) estimate that 24 excess deaths per day occurred in New York City during the air pollution episode (November 23–November 29, 1966). In the wake of such environmental pollution events, as a means of limiting and eradicating environmental pollution, the US EPA (United States Environmental Protection Agency) was established on December of 1970.
- Eastern China smog (2013): On December 7, 2013, a hazardous smog stretched a distance of about 2 km within China (Levy 2014). The episode lasted for 8 days between the 2nd and 9th of December 2013. Huang et al. (2016) state that within the duration of the smog episode, the average $PM_{2.5}$ was 212 $\mu g/m^3$, which was three times higher than the usual $PM_{2.5}$ concentration (76 $\mu g/m^3$) within the same area. In China, coal still remains to be the main energy source, and it's regarded to be the primary cause of fine PM pollution in China. The Chinese cities of Baoding, Shijiazhuang, and Handan reported more than 30,000 deaths in 2013 per city, which can be linked to pollution (Solomon 2016).

The original version of this chapter was revised: Updated chapter has been uploaded to Springerlink. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-71377-5_21.

L. O. H. Wijeratne · G. K. Zewdie · D. Kiv · A. Aker · D. J. Lary (✉)
S. Talebi · X. Yu · E. Levetin
Hanson Center for Space Sciences, The University of Texas at Dallas, Richardson, TX, USA
e-mail: lhw150030@utdallas.edu; david.lary@utdallas.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_13

- Great smog of New Delhi (2016): WHO announced New Delhi as the most polluted city in the world in 2014 (Saravanan et al. 2017). On November of 2017, air pollution levels in New Delhi went up to 999 on the AQI (Air Quality Index) scale. This is an air pollution level equivalent to smoking 50 cigarettes per day (Basu 2019). The cities' visibility level reduced to more than 50 m during the episode (Terry et al. 2018). It is assumed that the major causes of air pollution in New Delhi are burning coal, petrol, diesel, gas, biomass, and waste, along with industries, power plants, and firecrackers (Saravanan et al. 2017).

Such episodes remain a constant reminder of the devastation that air pollution can cause. The first step in fighting air pollution is to quantify the problem.

Making the Invisible, Visible

Conventional air quality management systems generally rely on a small number of regulatory-grade sensors across an urban area, for example, across the Dallas-Fort Worth Metroplex with a population of over seven million, there are just three airborne particulate sensors. Due to the substantial cost of these regulatory-grade sensing systems, they fail to provide adequate spatial and temporal resolution for characterizing air quality on a neighborhood scale. As such, these sensor systems do not adequately inform us about the situation within our neighborhoods, where people live, work, and play. Recent studies have demonstrated that air quality varies on very fine spatial and temporal scales (Harrison 2015; Harrison et al. 2015). As such, it is apparent that the world needs air quality sensing systems at the neighborhood scale (i.e., with a spatial resolution of less than a km). For example, a study that made near daily fine scale measurements at least every meter over a 100 km^2 in north Texas (Harrison 2015; Harrison et al. 2015) used variograms to characterize the spatial scale of airborne particulates. These studies revealed that the spatial scales over the study period depended on the synoptic situation and varied between 0.5 and 7.5 km.

The initiation of air quality sensing system requires an understanding of what parameters to measure, where the sensors are to be located, and the budgetary constraints that the system is to be bounded by. Such an understanding can be gained via the knowledge of the mix of pollutants that may be residing within the study and the general mindset of the people at stake. Another key aspect in getting started in a project is to look into available technologies that can abide by the requirements defined.

Airborne Particulates

Airborne atmospheric aerosols are an assortment of solid or liquid particles suspended in air (Boucher 2015). Aerosols, also referred to as particulate matter (PM), are associated with a suite of issues relevant to the global environment (Charlson et al. 1992; Ramanathan et al. 2001; Dubovik et al. 2002; Guenther et al. 2006; Hallquist et al. 2009; Kanakidou et al. 2005; Allen et al. 2014), atmospheric photolysis, and a range of adverse health effects (Dockery et al. 1993a; Oberdörster et al. 2005; Pope III et al. 2002; Pope et al. 2006; Cheng and Liu 2009; Chin 2009; Lim et al. 2012). Atmospheric aerosols are usually formed either by direct emission from a specific source (e.g., combustion) or from gaseous precursors (Stocker 2014). Although individual aerosols are typically invisible to the naked eye, due to their small size, their presence in the atmosphere in substantial quantities means that their presence is usually visible, e.g., as fog, mist, haze, smoke, dust plumes, etc. (Seinfeld 1986). Airborne aerosols vary in size, composition, and origin as well as in spatial and temporal distributions (Chin 2009; Pöschl 2005). As a result, the study of atmospheric aerosols has numerous challenges. The following aerosol classifications provide some useful insights.

Aerosol Classification

Characterization of atmospheric aerosols can be based on their origin, concentration, size, chemical composition, phase, and morphology (Seinfeld 1986). However, one of the main forms of aerosol classification is via their sources.

Source-Based Classification

The formation of atmospheric aerosols can be complex. As a result, the determination of global aerosol sources is approximate (Kondratyev et al. 2006). Three main terrestrial sources are typically quoted (Kokhanovsky 2008).

- Cosmic aerosols: Particles migrating through space are usually considered cosmic particles (Carslaw et al. 2002; Kirkby et al. 2011).
- Primary aerosols: Particles directly emanating from the earth's surface are usually termed primary aerosols (Holben et al. 2001; Streets et al. 2003; Bond et al. 2004; Kanakidou et al. 2005), for example, aerosols formed due to the agitation of oceanic or terrestrial surfaces by wind.
- Secondary aerosols: Secondary particles occur from condensation of gaseous species (aerosol precursors) (Atkinson 2000; Kanakidou et al. 2005; Hallquist et al. 2009; Jimenez et al. 2009). These may endure one or many chemical transformations prior their formation.

Primary and secondary aerosols are further subdivided depending on their origin into natural and anthropogenic (man-made) aerosols (Schauer et al. 1996; Andreae and Crutzen 1997; Yunker et al. 2002; Pöschl 2005; Kondratyev et al. 2006; Boucher 2015; Colbeck and Lazaridis 2010).

Most emissions from the oceans, vegetation, forest fires, and volcanoes are considered natural in origin. Anthropogenic sources are dominated by the emissions from the combustion of fossil and biofuels.

Shape-Based Classification

Colbeck (2014) describes the three main distinctions based on the shapes of atmospheric aerosols.

- Isometric particulates: The three dimensions of isometric particles are defined to be similar. Spherical particulates belong to this category (Wachs 2009). This study is done under the assumption of sphericity (isometric particulates).
- Platelets: Platelets have two longer dimensions compared to the third one. Disk-like particles fall under this classification.
- Fibers: Fibers are particulates with two smaller dimensions and one longer dimension. Asbestos is one well-known fiber.

Classification Based on Chemical Composition

Ambient PM is usually comprised of a mixture of one or more of the following chemical compounds: geological material (oxides of aluminum, silicon, calcium, titanium, iron, and other metal oxides), sulfates, nitrates, ammonium, sodium chloride, organic carbon, elementary carbon, and liquid water (Chow et al. 1998). A more generalized classification is derived from considering the chemical purity of PM by distinguishing between internal and external mixtures.

- External mixture: In an external mixture, individual particles within are chemically pure.
- Internal mixture: In an internal mixture, individual particulates are a mix of chemical species. A perfect internal mixture is said to have the same mix of chemical species for all particulates.

Typical atmospheric particulates would be in the middle ground between perfect internal and external mixtures (Boucher 2015). The optical properties of atmospheric aerosols, and in turn the radiative forcing due to atmospheric aerosols, are partly determined by the state of mixing (externally or internally) of the chemical species involved (Lesins et al. 2002).

Spatial Classification

Aerosols are also categorized with respect to their localized regions. The classification gives rise to these categories: urban aerosols, marine aerosols, rural continental aerosols, free troposphere aerosols, stratospheric aerosols, polar aerosols, and desert aerosols. In some cases, a geospatial classification might be inexact due to the possibility of long-range aerosol transportation. However, the regional aerosol classification is useful when local effects eclipse the more generic effects of aerosols (Boucher 2015).

Size-Based Classification

Aerosol size distribution and chemical composition play a role in their atmospheric transportation (Colbeck and Lazaridis 2010). Most atmospheric particles are not spherical. However, in atmospheric sciences, particles with equivalent settling velocities are considered to be of equal size irrespective of their actual size or composition. The microscopic properties of aerosols differ significantly depending on the type of aerosol. Nevertheless, generic models are defined to describe the main microscopic properties of a given aerosol with its appropriately assumed diameter (Kokhanovsky 2008). The two most generic definitions of such assumed diameters are as follows:

- Aerodynamic diameter: The diameter of a unit density sphere which has similar aerodynamic properties as the particle considered.
- Stokes diameter: The diameter of a sphere which has similar density as the particle considered.

These definitions are introduced to avoid ambiguities of size measurements that may occur due to using different types of instrumentation (Colbeck 2014). This study uses the aerodynamic diameter for size-based distinctions. There are two distinct means of aerosol classification with respect to size:

- Modal distributions: The size-based classification of aerosols is mainly devised on five modes (Boucher 2015; Alfarra 2004; Stier et al. 2005; Sỳkorová et al. 2016):
  1. The nucleation mode or ultrafine mode with a diameter of less than 0.01 μm.
  2. The Aitken mode with a diameter in the range 0.01 μm $-$ 0.1 μm.
  3. The accumulation mode with a diameter in the range 0.1 μm $-$ 1 μm.
  4. The coarse mode with a diameter in the range 1 μm $-$ 10 μm.
  5. The super-coarse mode with a diameter of greater than 10 μm.

  Each of these modes corresponds to the relative maximums of number, surface, and volume distributions of atmospheric aerosols.
- Variables related to human exposure: The term "fine" (or ultrafine) particulates usually refers to particulates less than 1 μm in aerodynamic diameter (PM_1) and particulates less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$). For air pollution control, particulates up to 10 μm in diameter (PM_{10}) are also considered (Pöschl 2005). Cur-

rently, the US EPA (United States Environmental Protection Agency) regulates $PM_{2.5}$ and PM_{10} due to the human health effects associated with $PM_{2.5}$ and PM_{10} (US EPA 2004). Some air quality monitors also measure the total suspended particle (TSP) size fraction which includes particulates up to 40 μm (Chow et al. 1998). Another division of occupational health-based size-selective sampling is defined by assessing the subset of particles that can reach a selective region of the respiratory system. On this basis, three main fractions are defined: inhalable, thoracic, and respirable (Hinds 2012). The current study focuses on measurements of the six variables PM_1, $PM_{2.5}$, PM_{10}, respirable (alveolic), thoracic, and inhalable size fractions.

Health Context

The effects on human health due to air pollution may be the most controversial (Seinfeld 1986). Nevertheless, it is by far the most important. Studies have shown that exposure of excess particulate matter has alarming negative health effects (Mannucci 2017). The smallest size ranges of (less than 2.5 μm) PM is capable of penetrating through to the lungs or even to one's bloodstream. As such, the highest mortality is associated with $PM_{2.5}$ (Chen et al. 2011). HEI (2017) reports that more than 90% of the world's population lived with unhealthy air in 2015. The American Thoracic Society (ATS) has a slightly higher guideline of 11 $\mu g/m^3$ annual mean concentrations as compared with the WHO's 10 $\mu g/m^3$ for $PM_{2.5}$. However, it is reported that 14% of countries with valid design values for atmospheric pollution exceed the said recommendation by the ATS (Cromar et al. 2016). The results of the Aphekom project conducted in 25 European cities reveal that complying with the WHOs PM guidelines for $PM_{2.5}$ would increase life expectancy by 22 months while also giving financial savings of €31 billion annually (Pascal et al. 2013). The health hazard created by excess airborne PM also creates critical expenditures within developing countries. The estimated economic cost due to $PM_{2.5}$ pollution for the city of Delhi was estimated to be $6394.74 million in 2015, up from $2714.10 million in 2005 (Maji et al. 2017). Due to these reasons, considerable amounts on research are done on the health hazards caused by PM. Table 1 provides an overview of research done on specific health concerns with respect to PM_{10}, $PM_{2.5}$, and UFPs (ultrafine particles).

Aerosol concentration, size, structure, and chemical composition are key factors in driving the health outcomes caused. However, these parameters are highly irregular in temporal and spatial schemes (Pöschl 2005). As such, even though the effect of PM exposure can be substantial, predicting a link between PM and human health can be challenging. Most studies rely on obtaining the level of morbidity and mortality for a given disease which can be attributed to the exposure to PM. Some studies also employ questionnaires in collecting health-related data.

Long-term exposure to $PM_{2.5}$ increases the risk of total and cardiovascular disease (CVD) mortality. The study by Thurston et al. (2016) concludes that $PM_{2.5}$ exposure has a substantial association with both total mortality and CVD mortality, with CVD having the highest hazard ratio of 1.10 for the study set of participants between 50 and 71 years. Pope et al. (2004) state that a 10 $\mu g/m^3$ increase in fine PM results in an 8%–18% increase in the mortality risk. A study conducted in six cities across the United States with a total of 8111 participating adults found that fine particulate air pollution was linked with excess mortality (Dockery et al. 1993b).

The ATS report (Cromar et al. 2016) for 2011–2013 found that 26% out of 21,400 excess morbidities and 26% out of 9320 excess deaths were associated with elevated $PM_{2.5}$ in the United States per year. A European study (Boldo et al. 2006) estimated that a reduction of the $PM_{2.5}$ abundance by 15 $\mu g/m^3$ of $PM_{2.5}$ would prevent 16,926 deaths annually within a subset of 23 European cities and that such a reduction would likely increase the life expectancy between 1 month to more than 2 years. The study included major cities, London, Paris, Athens, Barcelona, Madrid, and Valencia. In the study, excess exposure to $PM_{2.5}$ was viewed as a modifiable factor which causes cardiovascular morbidity and mortality. Maji et al. (2017) found that the mortality in Mumbai and Delhi during 2015 was associated with PM_{10} and lead to 32,014 and 48,651 deaths, respectively.

Cerebrovascular accidents are a prominent cause of morbidity throughout the world. It was estimated that an increase of 10 $\mu g/m^3$ of $PM_{2.5}$ accounts for 1.29% (95% CI 0.552%–2.03%) increase in the risk of emergency hospital admissions (Santibañez et al. 2013). Sulfate aerosols are known to cause respiratory throat and fever symptoms (Onishi et al. 2018).

In some cases, the maternal exposure to excess particulate matter has resulted in lower birth weights (LBW). A multicountry evaluation of LBW reveals that a 10 $\mu g/m^3$ increase in PM_{10} (odds ratio (OR) = 1.03; 95% confidence interval (CI), 1.01–1.05) and $PM_{2.5}$ (OR = 1.10; 95% CI 1.03–1.18) exposure during the entire pregnancy is positively correlated with LBWs (Dadvand et al. 2013).

Excess PM exposure can also be behind excess stress among individuals. Evidence has been found that mitochondrially encoded TRNA phenylalanine (MT-TF) and mitochondrially encoded 12S RNA (MT-RNR1) is linked with metal-rich PM_1 (Byun et al. 2013). Both mitochondrial MT-TF and MT-RNR1 DNA methylation are sources of oxidative stress which responds to foreign environments. Short-term exposure to $PM_{2.5}$ also prompts a mechanism involving pulmonary oxidative stress which in turn induces vascular insulin resistance and inflammation (Haberzettl et al. 2016).

Table 1 Health Concerns due to PM 10, PM 2.5, and ultrafine particles (UFPs). Table adapted from Ruckerl et al. (2006)

Health outcomes	Short-term studies			Long-term studies		
	PM_{10}	$PM_{2.5}$	UFP	PM_{10}	$PM_{2.5}$	UFP
Mortality						
All causes	xxx	xxx	x	xx	xx	x
Cardiovascular	xxx	xxx	x	xx	xx	x
Pulmonary	xxx	xxx	x	xx	xx	x
Pulmonary effects						
Lung function, e.g., PEF	xxx	xxx	xx	xxx	xxx	
Lung function growth				xxx	xxx	x
Asthma and COPD exacerbation						
Acute respiratory symptoms		xx	x	xxx	xxx	
Medication use			x			
Hospital admission	xx	xxx		x	x	
Lung cancer						
Cohort				xx	xx	x
Hospital admission				xx	xx	x
Cardiovascular effects						
Autonomic nervous system	xxx	xxx		x	x	
ECG-related endpoints						
Autonomic nervous system	xxx	xxx	xx			
Myocardial substrate and vulnerability		xx	x			
Vascular function						
Blood pressure	xx	xxx	x			
Endothelial function	x	xx	x			
Blood markers						
Pro-inflammatory mediators	xx	xx	xx			
Coagulation blood markers	xx	xx	xx			
Diabetes	x	xx	x			
Endothelial function	x	x	xx			
Reproduction						
Premature birth	x	x				
Birth weight	xx	x				
IUR/SGA	x	x				
Fetal growth						
Premature birth	x					
Infant mortality	xx	x				
Sperm quality	x	x				
Neurotoxic effects						
Central nervous system		x	xx			

Notes: X, few studies (6 or less); XX, many studies (7–10); XXX, large number of studies (>10).
Abbreviations: UFP, ultrafine particle; PEF, peak expiratory flow; COPD, chronic obstructive pulmonary disease; IUG, intrauterine growth restriction; SGA, small for gestational age

Environmental pollution is a potential cause of lung cancer. Tandem repeats are DNA sequences which lie adjacent to each other in the same orientation (direct tandem repeats) or in the opposite direction to each other. These DNA sequences are generally hypomethylated in cancer patients. A case study done on two contrasting groups on air pollution exposure of truck drivers and office workers reveals that PM is linked with hypomethylation of some tandem repeats (SATα, NBL2) (Guo et al. 2014).

The most likely candidates to be affected by unhealthy air are the elderly and infants. Pun et al. (2017) concluded that $PM_{2.5}$ is linked to both depressive and anxiety symptoms within older adults with the strongest association to individuals with lower socioeconomic measures. Shy et al. (1973) confirm that school children between the age of 9 and

13 exposed to elevated air pollution experience ventilatory problems. A study conducted with a collection of 40 fifth grade school children revealed that the "soot" fraction of $PM_{2.5}$ is strongly linked with pollution-related asthma attacks affecting children residing beside roadways (Spira-Cohen et al. 2011).

Using a business-as-usual emission scenario model, (Lelieveld et al. 2015) estimate that premature mortality due to outdoor air pollution could double by 2050. As such, it is of utmost importance to conduct in-depth research on PM and other air pollutant sources in order to enforce proper air pollution policies (Kelly and Fussell 2016).

Difficulty in Estimating Airborne Particulates

Conventional regulatory-grade instrumentation is accurate, but expensive. This makes it challenging to provide neighborhood-scale measurements due to the substantial costs involved. So in this study, we present two different case studies where we use machine learning to utilize different sensor types. First, we use low-cost optical particle counters that can be deployed at scale across neighborhoods. Second, we use remotely sensed observations made using weather RADARs.

What Is Machine Learning?

Machine learning has already proved useful in a wide variety of applications in science, business, healthcare, and engineering. Machine learning allows us to *learn by example* and to *give our data a voice*. It is particularly useful for those applications for which we do *not* have a complete theory, yet which are of significance. Machine learning is an automated implementation of the scientific method (Domingos 2015), following the same process of generating, testing, and discarding or refining hypotheses. While a scientist or engineer may spend his entire career coming up with and testing a few hundred hypotheses, a machine learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is therefore not surprising that machine learning is currently revolutionizing many areas of science, technology, business, and medicine (Lary et al. 2016, 2018).

Machine learning is now being routinely used to work with large volumes of data in a variety of formats such as image, video, sensor, health records, etc. Machine learning can be used in understanding this data and creating predictive and classification tools. When machine learning is used for regression, empirical models are built to predict continuous data, facilitating the prediction of future data points, e.g., algorithmic trading and electricity load forecasting. When machine learning is used for classification, empirical models are built to classify the data into different categories, aiding in the more accurate analysis and visualization of the data. Applications of classification include facial recognition, credit scoring, and cancer detection. When machine learning is used for clustering, or unsupervised classification, it aids in finding the natural groupings and patterns in data. Applications of clustering include medical imaging, object recognition, and pattern mining. Object recognition is a process for identifying a specific object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. These technologies are being used for applications such as driver-less cars, automated skin cancer detection, etc.

Machine learning is an automated approach to building empirical models from the data *alone*. A key advantage of this is that we make *no* a priori assumptions about the data, its functional form, or probability distributions. It is an empirical approach. However, it also means that for machine learning to provide the best performance, we do need a *comprehensive representative set of examples*, which spans as much of the parameter space as possible. This comprehensive set of examples is referred to as the *training data*.

So, for a successful application of machine learning, we have *two* key ingredients, both of which are essential, a machine learning algorithm and a comprehensive training dataset. Then, once the training has been performed, we should test its efficacy using an independent validation dataset to see how well it performs when presented with data that the algorithm has *not* previously seen, i.e., test its *generalization*. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.

It should be noted that with a given machine learning algorithm, the performance can go from poor to outstanding with the provision of a progressively more complete training dataset. Machine learning really is learning by example, so it is critical to provide as complete a training dataset as possible. At times, this can be a labor-intensive endeavor.

A key part of machine learning studies is an independent validation to objectively test the "generalization" of the empirical models. This is often done by randomly splitting the available data into two portions. One portion, the training dataset, is used to train the empirical machine learning model. The other portion, the independent validation dataset, is used to objectively test the empirical model by using data not seen in the training process.

We have used machine learning in many previous studies (Brown et al. 2008; Lary et al. 2009a; Lary and Aulov 2008; Lary et al. 2004; Malakar et al. 2013; Lary 2010; Malakar et al. 2012a; Lary 2013, 2007; Albayrak et al. 2011; Lary et al. 2003; Malakar et al. 2012b; Lary 2014; Lary et al.

2015b; Kneen et al. 2016; Lary et al. 2010; Medvedev et al. 2016; Lary et al. 2016; O et al. 2017; Wu et al. 2017; Nathan and Lary 2019; Lary et al. 2019, 2018; Wu et al. 2019; Alavi et al. 2016; Ahmad et al. 2016; Zewdie and Lary 2018; Malakar et al. 2018; Zewdie et al. 2019a,b; Chang et al. 2019; Choi et al. 2019). In this study, we have used machine learning for multivariate nonlinear non-parametric regression. Some of the commonly used regression algorithms include neural networks (McCulloch and Pitts 1943; Haykin 2001, 2007, 1994, 1999; Demuth et al. 2014; Bishop 1995), support vector machines (Vapnik 1982, 1995; Cortes and Vapnik 1995; Vapnik 2000, 2006), decision trees (Safavian and Landgrebe 1991), and ensembles of trees such as random forests (Ho 1998; Breiman 1984, 2001). Previously, we have used a similar approach to cross-calibrate satellite instruments (Lary and Aulov 2008; Brown et al. 2008; Lary et al. 2009a, 2016, 2018). Recently, other studies have also used machine learning to calibrate low-cost sensors (Li et al. 2014; Dong et al. 2015).

Case study: Using Machine Learning for the Calibration of Airborne Particulate Sensors

Low-cost sensors that can also be accurately calibrated are of particular value. For the last two decades, we have pioneered the use of machine learning to cross-calibrate sensors of all kinds. This was initially done for very expensive orbital instruments onboard satellites (awarded an IEEE paper prize and specially commended by the NASA MODIS team) (Lary et al. 2009a). We are now using this approach operationally for low-cost sensors distributed at scale across dense urban environments as part of our smart city sentinels. The approach can be used for very diverse sensors, but as a useful illustrative example that has operational utility, we describe here a use case for accurately calibrated low-cost sensors measuring the abundance and size distribution of airborne particulates, with the implicit understanding that many other sensor types could easily be substituted. These sensors can be readily deployed at scale at fixed locations, mobile on various robotic platforms (walking, flying, etc.) or vehicles, carried, or deployed autonomously as a mesh network, either by operatives or by robots (walking, flying, etc.).

Building-in calibration will enable consistent data to be retrieved from all the low-cost sensors. Otherwise, the data will always be under some suspicion as the inter-sensor variability among low-cost nodes can be substantial. While much effort has been recently placed on providing the connectivity of large disbursed low-cost networks, little to no effort has been spent on the automated calibration, bias-detection, and uncertainty estimation necessary to make sure the information collected is sound. A case study of providing this critical calibration using machine learning is the focus of this paper.

Any sensor system benefits from calibration, but low-cost sensors are typically in particular need of calibration. The inter-sensor variability among low-cost nodes can be substantial. In addition to the pre-deployment calibration, once the sensors have been deployed, the paradigm we first developed for satellite validation of constructing probability distribution functions of each sensor's observation streams can be used to both monitor the real-time calibration of each sensor in the network by comparing its readings to those of its neighbors and also answer the question "how representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?"

Using Probability Distribution Functions to Monitor Calibration and Representativeness in Real Time

It is useful to be able to answer the question "how representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?" We can answer this question by considering a probability distribution function (PDF) of all the observations made by a sensor over some temporal and spatial window. The width of this probability distribution is termed the representativeness uncertainty for that temporal and spatial window. The PDFs of all observations made by each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors can include measurements from primary reference sensors that may be available. This approach is used to estimate the measurement uncertainty and inter-instrument bias for the last hour, day, etc. We continuously accumulate the PDF for each sensor over a variety of time scales and compare it to its nearest neighbors within a neighborhood radius. Any calibration drift in a sensor will be quickly identified as part of the fully automated real-time workflow where we will automatically be comparing each sensor's PDFs to its neighbor's PDFs and to the reference instruments PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different.

Characterizing the Temporal and Spatial Scales of Urban Air Pollution

This study focused on the calibration of low-cost sensors as part of a larger endeavor with the goal of characterizing the temporal and spatial scales of urban pollution. The temporal

and spatial scales of each atmospheric component are intimately connected. The resolution used in atmospheric chemistry modeling tools is often driven by the computational resources available. The spatial resolution of observational networks is often determined by the fiscal resources available. It is worth taking a step back and characterizing what the actual spatial scales are for each chemical component of urban atmospheric chemistry. Based on our street-level surveys providing data at less than a meter resolution, it is clear that the spatial scales are dependent on several factors such as the synoptic situation, the distribution of sources, the terrain, etc. In the larger study, we characterize the spatial scales of multi-species urban pollution by using a hierarchy of measurement capabilities that include (1) a zero emission electric survey vehicle with comprehensive gas, particulate, irradiance, and ionizing radiation sensing and (2) an ensemble of more than 100 street-level sensors making measurements every few seconds of a variety of gases, particulates, light levels, temperature, pressure, and humidity. Each sensor is accurately calibrated against a reference standard using machine learning. This paper documents an example of low-cost sensor calibration for airborne particulate observations.

Societal Relevance

What are the characteristic spatial scales of each chemical species, and how does this depend on issues such as the synoptic situation? These are basic questions that are helpful to quantify when considering atmospheric chemistry, when looking forward to the next generation of modeling tools and observing system (whether from space or ground-based networks), and when evaluating mitigation strategies, especially with regard to co-benefits for air pollution and greenhouse gas reduction and investigating the evolution of urban air composition in a warming climate. To be able to quantify these spatial and temporal scales, we need a comprehensive observing system; being able to use low-cost sensors is of great assistance in achieving this goal.

The Dallas Fort Worth (DFW) Metroplex (where our study was conducted) is the largest inland urban area in the United States and the nation's fourth largest metropolitan area. Nearly a third of Texans, more than seven million inhabitants, live in the DFW area. A population which is growing by a thousand people every day. DFW is an area with an interesting variety of specific pollution sources with unique signatures that can provide a useful testbed for generalizing a measurement strategy for dense urban environments. For more than two decades, the DFW area has been in continuous violation of the Clean Air Act. DFW will be one of only ten non-California metropolitan areas still in violation of the Clean Air Act in 2025 unless major changes take place. This has already had a detrimental health impact, e.g., even though the Texas average childhood asthma rate is 7%, and the national average is 9%, the DFW childhood asthma rate is 20–25%. Second only to the Northeast, DFW ranks second in the number of annual deaths due to smog. Further, a leading factor in poor learning outcomes in high schools is absenteeism, a leading cause of absenteeism is asthma, and key trigger for asthma is airborne pollution (Lary et al. 2019). Physical exertion in the presence of high pollution levels is more likely to lead to an asthma event. The sensors calibrated in this study are being provided to high schools and high school coaches so that simple practical decisions can be made to reduce adverse health outcomes, e.g., given the levels of pollen/pollution today, should physical education/practice be outside or inside?

The Datasets Used

All of the measurements were made at our own field calibration station in the ambient environment. The calibration of the low-cost AlphaSense OPC occurs prior to their deployment across the dense urban environment of DFW. In this study, we use machine learning to bring together two distinct types of data. First, we use accurate in situ observations made by a research-grade particulate spectrometer. Second, we use observations from inexpensive optical particle counters. The inexpensive sensors are particularly useful as they can be readily deployed at scale.

Research-Grade Optical Particle Counter

The particulate spectrometer is a laser-based optical particle counter (OPC). In this study we used a GRIMM Laser Aerosol Spectrometer and Dust Monitor Model 1.109. The sensor has the capability of measuring particulates of diameters between 0.25 μm and 32 μm distributed within 32 size channels. Such a wide range of diameter space is made possible due to intensity modulation of the laser source. Particulates pumped into the sensor are detected through scattering a laser beam of 655 nm into a light trap. The laser beam is aimed at particulates coming through a sensing chamber at a flow rate of 1.21 l/min. The device classifies particulates into specific size classes subject to its intensity (Broich et al. 2012). The optical arrangement of the sensor is staged such that a curved optical mirror placed at an average scattering angle of 90° collects and redirects the scattered light toward a photo sensor. The wide angle of the optical mirror (120°) is meant to increase the light intensity redirected toward the photo sensor within the Rayleigh scattering domain which decreases the minimum detectable particle size. Furthermore, it compensates for Mie scattering undulations caused by monochromatic illumination. The sensing period

of the GRIMM sensor was set to 6 s and for each time window provides three standardized mass fractions, namely, based on occupational health (repairable, thoracic, and alveolic) according to EN 481 as well as PM_1, $PM_{2.5}$, and PM_{10}.

Low-Cost Optical Particle Counters

There are several readily available optical particle counters (OPC) which are useful, but much less accurate compared to research grade sensors. In this study, we focus on using such sensors, together with machine learning, to get as close as possible to the accuracy of research-grade PM sensors. After the application of the machine learning calibration, these lower-cost sensors perform admirably. In order for low-cost sensors to provide an improved picture of PM levels, a careful calibration is required. The current study uses an Alpha Sense OPCN3 (http://www.alphasense.com/) together with a cheaper environmental sensor (Bosch BME280) as data collectors. The OPC-N3 is compact, 75 mm × 60 mm × 65 mm in size, and weighs under 105 g which uses similar technology to the conventional OPCs where particle size is determined via a calibration based on Mie scattering. Unlike most OPCs, the OPC-N3 doesn't include a pump and a replaceable particle filter in order to pump aerosol samples through a narrow inlet tube, hence avoiding the need for regular maintenance. Sufficient airflow through the sensor is made possible with a low-powered micro-fan producing a sample flow rate of 280 mL/min. The OPC-N3 is capable of onboard data logging as well as measuring particulates of diameters up to 40 μm. This enables the OPC-N3 to measure pollen and other biological particulates. The onboard data is saved within an SD card which can be accessed through a micro-USB cable connected to the OPC. Furthermore, the OPC-N3's lower sensing diameter is reduced to 0.35 μm as opposed to its predecessor's (OPC-N2) lower limit of 0.38 μm. The wider range of sensing is made possible via the OPC switching between high and low gain modes automatically. The OPC-N3 calculates its PM values using the method defined by the European Standard EN 481 (Alphasense 2018).

Caveat: Particulate Refractive Index

The observations made by optical particle counters are sensitive to the refractive index of the particulates and their light absorbing properties. The retrieved size distributions and the mass concentrations can be biased, depending on the nature of the particulates. The current study does not explore the accuracy implications of this. A future study is underway which includes direct measurements of black carbon that will allow us to begin to explore these aspects. The machine learning paradigm is readily extensible to include these aspects, even though not explicitly addressed in this study.

Machine learning is an ideal approach for the calibration of lower-cost optical particle counters.

Ensemble Machine Learning

Multiple approaches for nonlinear non-parametric machine learning were tried including neural networks, support vector regression, and ensembles of decision trees. The best performance was found using an ensemble of decision trees with hyperparameter optimization (Safavian and Landgrebe 1991; Ho 1998; Breiman 1984, 2001). Ensemble methods use multiple learners to obtain better predictive performance that could be obtained from any of the individual learners alone. A good example of an ensemble of learners is a random forest, which uses an ensemble of decision trees. In this study, the specific implementation used was that provided by the MathWorks in the fitrensemble function which is part of the MATLAB Statistics and Machine Learning Toolbox. Hyperparameter optimization was used so that the optimal choice was made for the following attributes: learning method (bagging or boosting), maximum number of learning cycles, learning rate, minimum leaf size, maximum number of splits, and the number of variables to sample. During hyperparameter optimization, we use an optimization approach (e.g., Bayesian optimization) to choose a set of optimal hyperparameters for our learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

In this study, there were 72 inputs to our multivariate nonlinear non-parametric machine learning regression; these include the particle counts for each of the 24 size bins measured by the OPC-N3; the OPC-N3 estimates of PM_1, $PM_{2.5}$, and PM_{10}; a suite of OPC performance variables including the reject ratio; and particularly important, the ambient atmospheric pressure, temperature, and humidity. The OPC-N3 sensor includes two photodiodes that record voltages which are eventually translated into particle count data. However, particles which are not entirely in the OPC-N3 laser beam, or are passing down the edge, are rejected, and this is recorded in the "reject ratio" parameter. This leads to better sizing of particles and hence plays an important role within the machine learning calibration.

Each of the six outputs we wished to estimate had its own empirical model. The performance of these six models in their independent validation is shown in Figs. 1 and 2. The outputs we estimated were the six variables measured by the reference instrument, the research-grade optical particle counter, namely, PM_1, $PM_{2.5}$, and PM_{10}, and the standardized occupational health respirable, thoracic, and alveolic mass fractions. The alveolic fraction is the mass fraction of in-

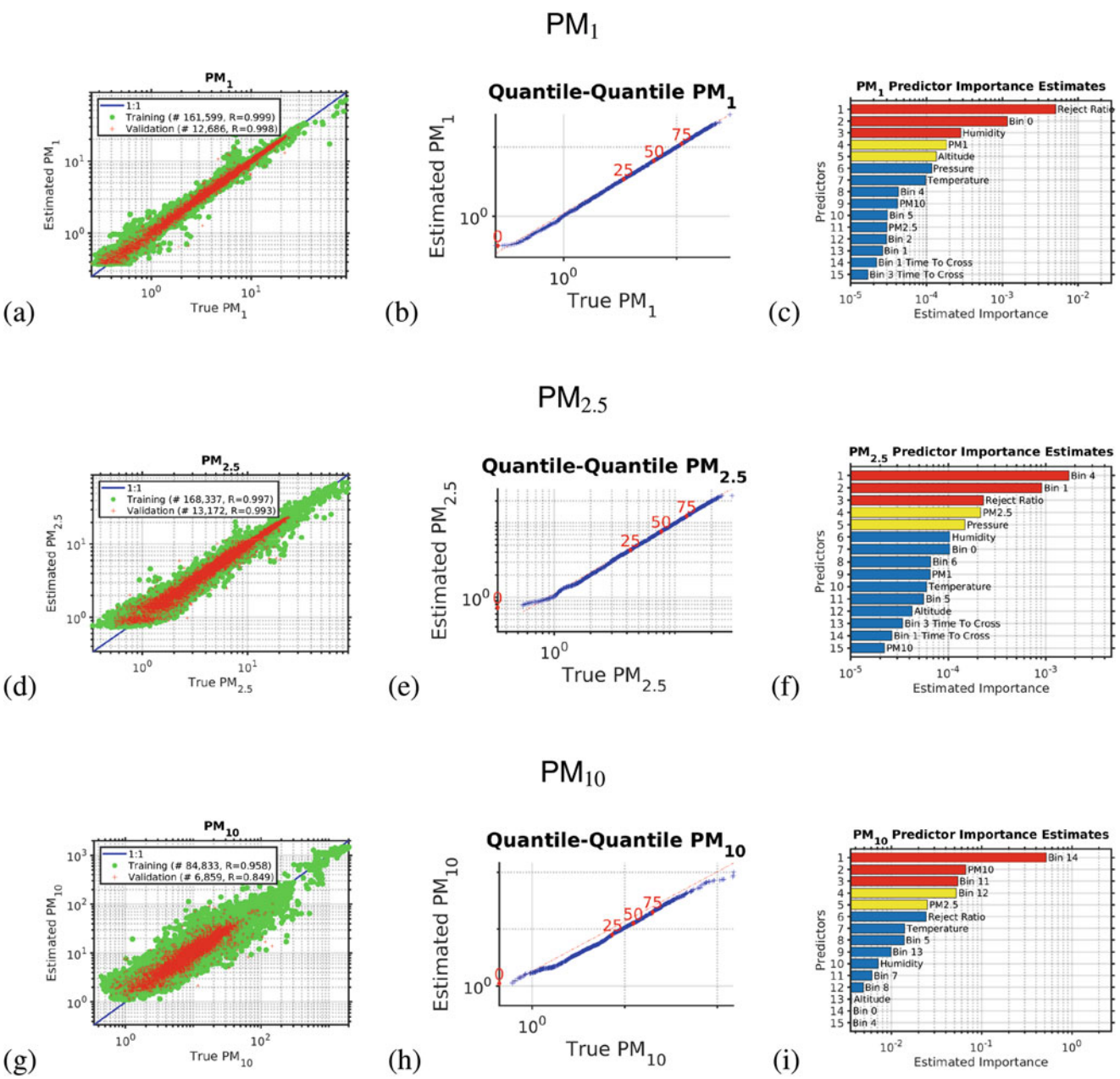


Fig. 1 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression for PM_1 (panels **a**–**c**), $PM_{2.5}$ (panels **d**–**f**), and PM_{10} (panels **g**–**i**). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data, and the red pluses are the independent validation data. The blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

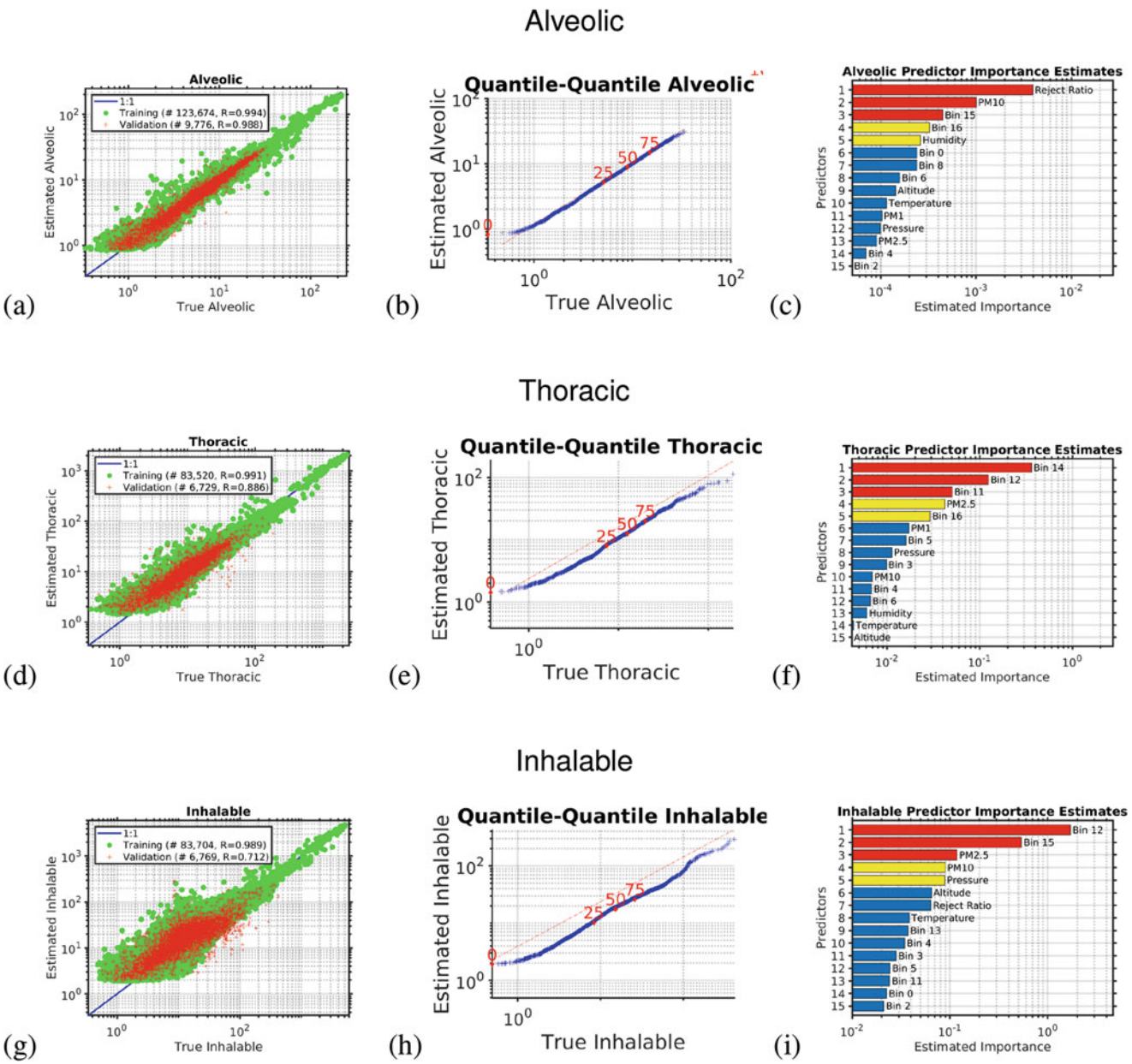


Fig. 2 This figure shows the results of the multivariate nonlinear nonparametric machine learning regression for the alveolic (panels **a**–**c**), thoracic (panels **d**–**f**), and inhalable size fractions (panels **g**–**i**). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data, and the red pluses are the independent validation dataset. The blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

haled particles penetrating to the alveolar region (maximum deposition of particles with a size $\approx 2\,\mu m$). The thoracic fraction is the mass fraction of inhaled particles penetrating beyond the larynx ($<10\,\mu m$). The respirable fraction is the mass fraction of inhaled particles penetrating to the unciliated airways ($<4\,\mu m$). The inhalable fraction is the mass fraction of total airborne particles which is inhaled through the nose and mouth ($<20\,\mu m$). For each of these six parameters, we created an empirical multivariate nonlinear non-parametric machine learning regression model with hyperparameter optimization.

Calibrating the Low-Cost Optical Particle Counters Using Machine Learning

Figure 1 shows the results of the multivariate nonlinear non-parametric machine learning regression for PM_1 (panels a to c), $PM_{2.5}$ (panels d to f), and PM_{10} (panels g to i). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning.

For the left-hand column of plots in Fig. 1 (the scatter diagrams), for a perfect calibration, the scatter plot would be a straight line with a slope of 1 and a y-axis intercept of 0; the blue line shows the ideal response. We can see that multivariate nonlinear non-parametric machine learning regression that we have used in this study employing an ensemble of decision trees with hyperparameter optimization has performed very well (panels a, d, and g). In each scatter diagram, the green circles are the data used to train the ensemble of decision trees, and the red pluses are the independent validation data used to test the generalization of the machine learning model.

We can see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red points), we see that $r_{PM_1} > r_{PM_{2.5}} > r_{PM_{10}}$.

For the middle column of plots in Fig. 1 (the quantile-quantile plots), we are comparing the *shape* of the probability distribution (PDF) of all the PM abundance data collected by the expensive reference instrument to that of the the PM abundance provided by calibrating the low-cost instrument using machine learning. A $\log_{10}$ scale is used with a tick mark every decade. The dotted red line in each quantile-quantile plot shows the ideal response. The red numbers indicate the percentiles (0, 25, 50, 75, 100). If the quantile-quantile plot is a straight line, that means both PDFs have *exactly* the same shape as we are plotting the percentiles of one PDF against the percentiles of the other PDF. Usually we would like to see a straight line at least between the 25th and 75th percentiles; in this case, we have a straight line over the entire PDF, which demonstrates that the machine learning calibration has performed well.

The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning. The relative importance metric is a measure of the error that results if that input variable is omitted. In the right-hand column of bar plots, we have sorted the importance metric into descending order, so the variable represented by the uppermost bar in each case was the most important variable for performing the calibration, the second bar is the second most important, etc. We note that along with the number of particles counted in each size bin, it is important to measure the temperature, pressure, and humidity to be able to accurately calibrate the low-cost OPC against the reference instrument. The data also suggests that the parameter "reject ratio" carries a higher deal of importance with respect to the calibration. OPC-N3 comprises two photodiodes which record voltages eventually translated into particle count data. However, particles which are not entirely in the beam or are passing down the edge are rejected and reflected on the parameter "reject ratio." This leads to better sizing of particles and hence plays a vital role within the ML calibration.

Another division of occupational health based size-selective sampling is defined by assessing the subset of particles that can reach a selective region of the respiratory system. On this basis three main fractions are defined: inhalable, thoracic, and respirable (Bickis 1998; Hinds 2012; Brown et al. 2013). Studies have shown that exposure of excess particulate matter has alarming negative health effects (Mannucci 2017). The smallest size ranges of particulate matter are capable of penetrating through to the lungs or even to one's bloodstream.

Figure 2 is similar to Fig. 1 and shows the results of the multivariate nonlinear non-parametric machine learning regression for the alveolic, thoracic, and inhalable size fractions. As would be expected, we see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red points), we see that $r_{Alveolic} > r_{Thoracic} > r_{Inhalable}$.

Operational Use of the Calibration and Periodic Validation Updates

The calibration just described occurs pre-deployment of the sensors into the dense urban environment. Once these initial field calibration measurements are made over a period of several months, in the manner described above, the multi-

variate nonlinear non-parametric empirical machine learning model is applied in real time to the live stream of observations coming from each of our air quality sensors deployed across the dense urban environment of the Dallas Fort Worth Metroplex. These corrected measurements are then made publicly available as Open Data as well as depicted on a live map and dashboard.

Building-in continual calibration to a network of sensors will enable long-term, consistent, and reliable data. While much effort has been recently placed on the connectivity of large disbursed IoT networks, little to no effort has been spent on the automated calibration, bias detection, and uncertainty estimation necessary to make sure the information collected is sound. This is one of our primary goals. This is based on extensive previous work funded by NASA for satellite validation.

After deployment, a zero emission electric car carrying our reference is used, to routinely drive past all the deployed sensors to provide ongoing routine calibration and validation. An electric vehicle does not contribute any ambient emissions and so is an ideal mobile platform for our reference instruments.

For optimal performance, the implementation combines edge and cloud computing. Each sensor node takes a measurement at least every 10 s. The observations are continually time-stamped at the nodes and streamed to our cloud server, the central server aggregating all the data from the nodes and managing them. To prevent data loss, the sensor nodes store any values that have not been transmitted to the cloud server for reasons, including communication interruptions, in a persistent buffer. The local buffer is emptied to the cloud server at the next available opportunity.

Data from all sensors are archived and serve as an open dataset that can be publicly accessed. The observed probability distribution functions (PDFs) from each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors include measurements from the electric car/mobile validation sensors. This comparison is used to estimate the size-resolved measurement uncertainty and size-resolved inter-instrument bias for the last hour, day, week, month, and year. We continuously accumulate the PDF for each sensor over a variety of time scales (an hour, day, week, month, and year) and compare it to its nearest neighbors within a neighborhood radius.

Any calibration drift in a sensor will be quickly identified as part of a fully automated real-time workflow, where we will automatically be comparing each sensor's PDFs to its neighbor's PDFs and to the reference instruments PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different. We have previously shown that machine learning can be used to effectively correct these inter-sensor biases (Lary et al. 2009b). As a result, the overall distributed sensing system will not just be better characterized in terms of its uncertainty and bias but also provide improved measurement stability over time.

Case Study: Using Weather Radars and Machine Learning to Estimate Airborne Particulates

The application of radar for atmospheric meteorology started soon after the end of the Second World War. It was during the Second World War in the 1930s that radar technology was first used to locate and track war planes. The interference from scatterers such as rainfall prompted the notion that radar can also be applied to measure atmospheric precipitation. Subsequently, the construction of radar networks for meteorologic purposes commenced. The first radar network for meteorologic purposes was the Weather Surveillance Radar-1957 (WSR-57) in the United States. Currently, the WSR radar network has been upgraded to WSR-88D (Weather Surveillance Radar, 1988). WSR-88D has about 160 Doppler radars all over the United States. Technically WSR-88D radar is known as the Next-Generation Radar (NEXRAD). The following sections present the measurements of the NEXRAD radar and its application to identify aerosols.

Weather radars are mainly designed for determining and forecasting atmospheric phenomena such as precipitation, cloud coverage, wind direction and magnitude, and other associated meteorological events. In addition to these daily atmospheric conditions, radar can detect other objects and particles of small size such as dust, sand, insects, bird migrations, ground clutter, etc. The weather radar can also detect variations in the refractive index of the atmosphere caused by variations in the ambient temperature.

Atmospheric radars employed for meteorologic purposes transmit electromagnetic pulses of various frequencies. The frequency range used in the design of the radar determines the purpose and observation capability of the radar. For example, radars designed for observing the amount, type, and motion of precipitation have frequencies from 3–10 GHz (in terms of wavelength, 10–3 cm, respectively). Radars having this frequency range are very convenient for meteorological purposes. Radars having higher frequencies are useful to observe small-size droplets and particles. Small-size cloud particles, light snow, fog, and light rainfall are observed by high-frequency meteorological radars. At relatively low frequencies (in the range of less than 100–1000 MHz), the radar can detect fluctuations in the refractive index of the clear atmosphere. Low-frequency radars are best suited for profiling wind speed and direction.

The NEXRAD radar is in general operated in two different modes based on atmospheric weather conditions. These two

modes are the precipitation mode and the clear air mode. In the precipitation mode, the NEXRAD radar is operated at fast rotations at various elevations up to about 19.5°. In precipitation mode, a high emphasis is given for measurements at several elevations in order to see vertical storm profiles. In clear air mode, the radar is operated slowly, and it is sensitive to observe scattering from small objects such as pollen, other particulate matter, dust, smoke, insects, and birds (Gali 2010). The approximate time for a volume scan is 6 and 10 min. for precipitation and clear air modes, respectively.

Direct measurements of pollen and other particulates are rarely done using the NEXRAD radar. However, a few exceptional research projects have been reported showing observation of large aerosols using the NEXRAD weather radar (Madonna et al. 2010). Consequently radar scattering from large aerosols such as pollen is hard to identify. But NEXRAD measurements of Doppler velocity, direction, and speed of wind which are meteorological variables controlling the distribution and dispersal of pollen and large particulate matter. Other meteorological variables such as cloud coverage, precipitation, and rainfall are also pollen-controlling variables associated with the radar base reflectivity. For example, Eq. (1) shows the rainfall estimation techniques based on NEXRAD reflectivity.

$$Z = aR^b \tag{1}$$

where Z and R, respectively, represent reflectivity and rainfall and a and b are experimentally determined constants. a and b are determined experimentally comparing radar reflectivity and rain gauge measurements. The National Weather Service default value of a and b are 300 and 1.4, respectively.

The lack of a complete functional relationship between NEXRAD measurements and airborne particulates motivates us to seek other options. The machine learning approach of "learning" by example from large datasets is the perfect candidate for this problem. In machine learning, we estimate a variable based on a large number of input variables (data), and the method is becoming popular in a wide variety of fields.

In this study, the inputs to our multivariate nonlinear non-parametric machine learning regression were the remotely sensed parameters provided by the weather radar. The outputs we wished to estimate were the variables measured by the in situ optical particle counter.

Estimating Aerosol Size Distribution

Figure 3 shows the results of the multivariate nonlinear non-parametric machine learning regression as a function of particle size. The x-axis shows the particle size on a log scale. The y-axis shows the quality of fit using the correlation coefficient of the scatter diagram for each particle size fraction; a perfect fit would have a correlation coefficient of 1. The blue line shows the results for the training data. The red line shows the results for the independent validation data. We can see that the machine learning can effectively use the NEXRAD data for the particles with a size of less than 7 μm. We can see a little more detail in Fig. 4 which shows the results of using an ensemble of regression trees for multivariate nonlinear non-parametric machine learning for three size fractions, 0.25 μm (panels a–c), 2.5 μm (panels d–f), and 25 μm (g–i).

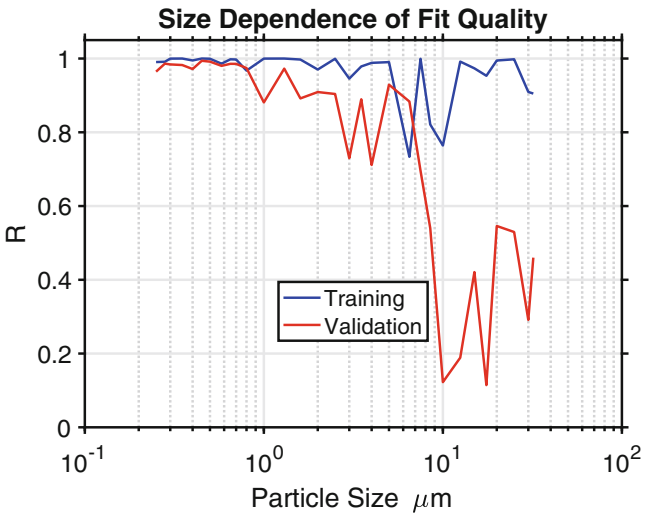


Fig. 3 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression as a function of particle size. The x-axis shows the particle size on a log scale. The y-axis shows the quality of fit using the correlation coefficient of the scatter diagram for each particle size fraction; a perfect fit would have a correlation coefficient of 1. The blue line shows the results for the training data. The red line shows the results for the independent validation data. We can see that the machine learning can effectively use the NEXRAD data for the particles with a size of less than 7 μm

The left-hand column of plots in Fig. 4 shows the scatter diagrams with the x-axis showing the actual number of particles observed by the in situ optical particle counter and the y-axis showing the number of particles estimated from the NEXRAD data using machine learning. The green circles are the training data, the red pluses are the independent validation dataset, and the blue line shows the ideal response. We can see that for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, e.g., those with a size of 0.25 μm (Fig. 4a), we have a very good scatter diagram and that the training and independent validation data have almost the same correlation coefficient. The same is true for particles with a diameter of 2.5 μm (Fig. 4d). However, for the larger particles that sediment rapidly, e.g., those with a diameter of 25 μm (Fig. 4g), the independent validation does not do well.

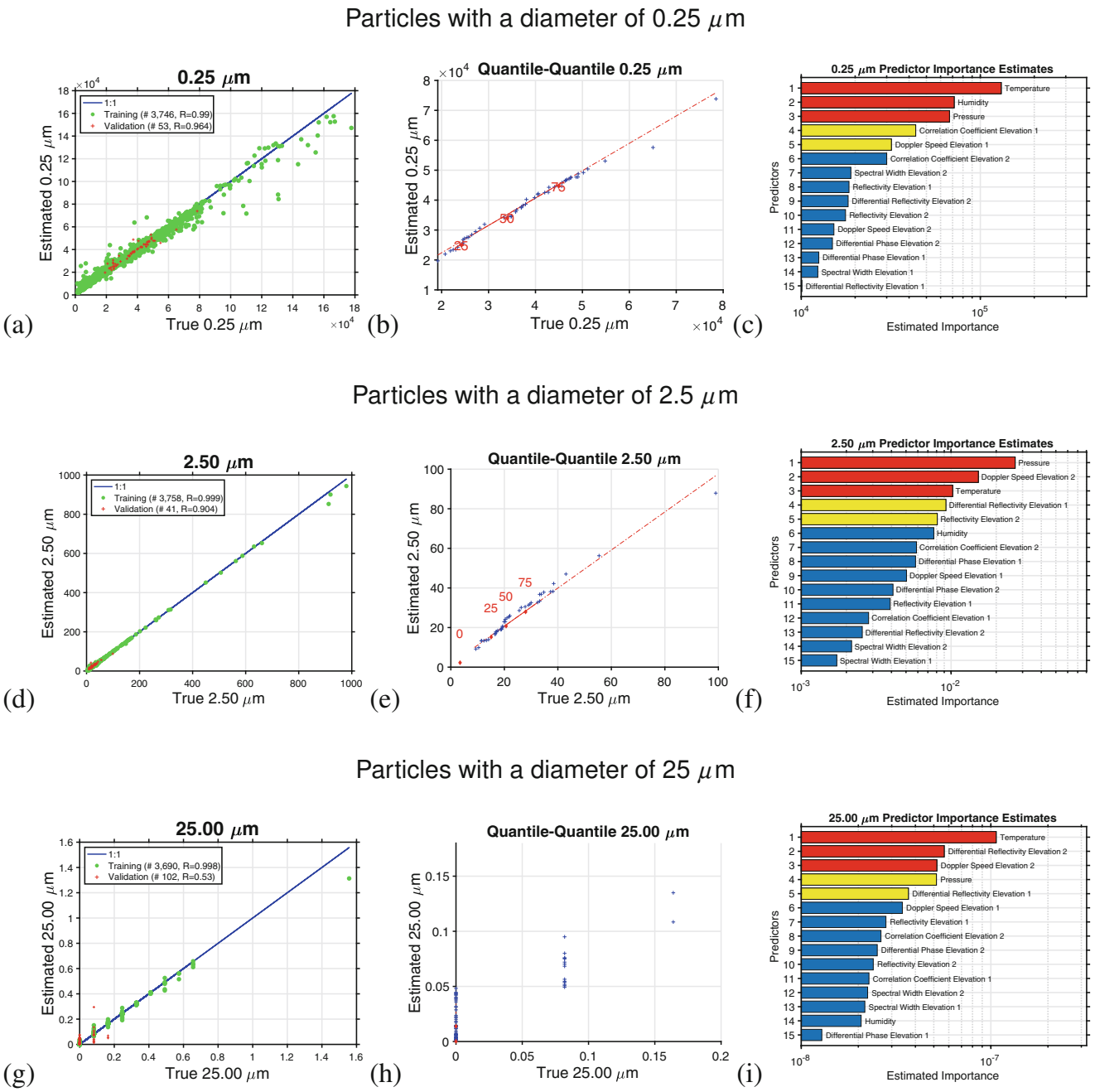


Fig. 4 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression for three size fractions, 0.25 μm (panels **a**–**c**), 2.5 μm (panels **d**–**f**), and 25 μm (**g**–**i**). The left-hand column of plots shows the scatter diagrams with the x-axis showing the actual number of particles observed by the in situ optical particle counter and the y-axis showing the number of particles estimated from the NEXRAD data using machine learning. The green circles are the training data, the red pluses are the independent validation dataset, and the blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the observed number of particles measured by the in situ optical particle counter and the y-axis showing the percentiles from the probability distribution function of the estimated number of particles. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for estimating the number of particles using machine learning

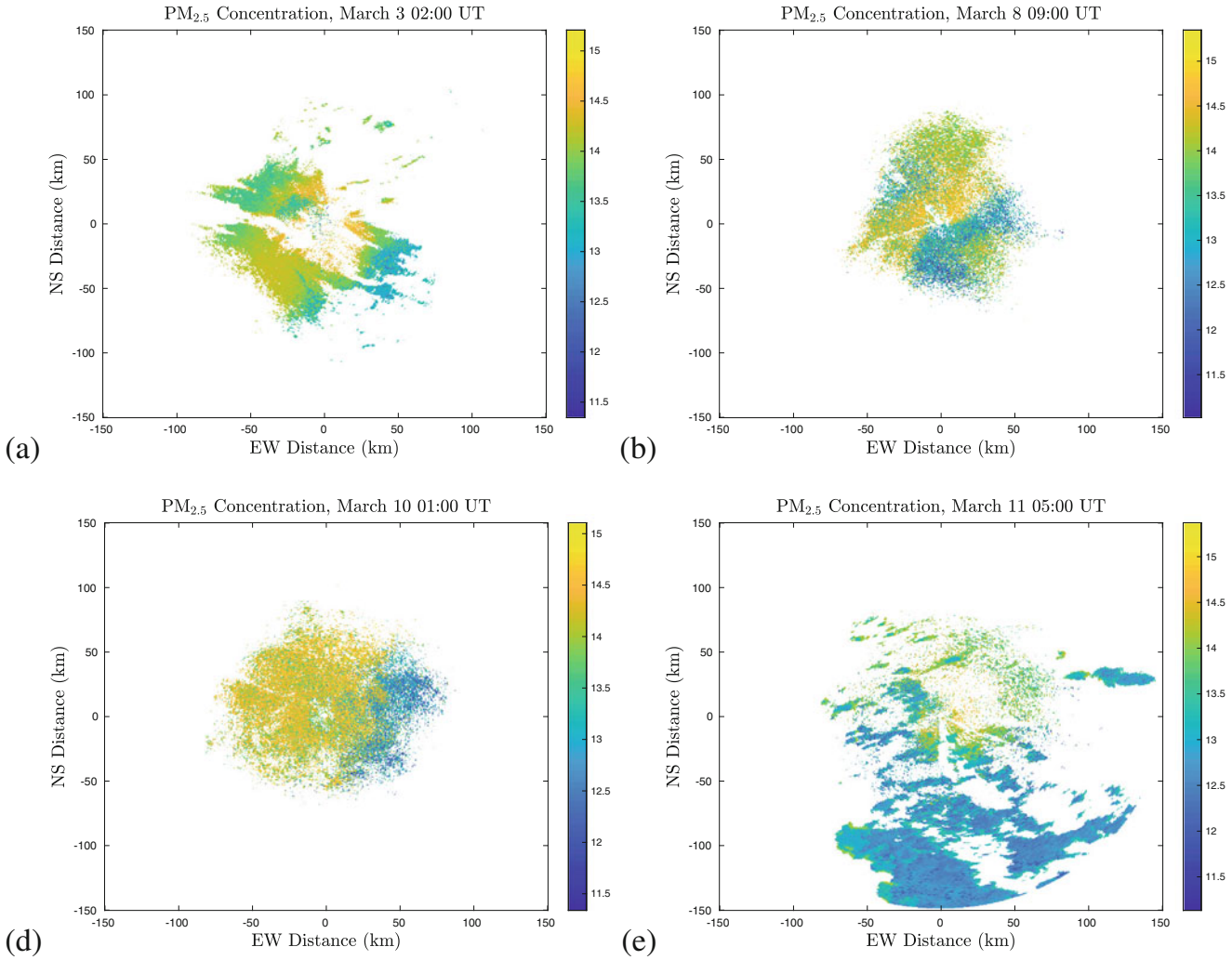


Fig. 5 Showing examples of the distribution of $PM_{2.5}$ over a large spatial area centered at the location of the NEXRAD radar. In this case, the NEXRAD radar measurements over a 0.5 km × 0.5 km are used to estimate the $PM_{2.5}$ concentrations

The middle column of plots in Fig. 4 shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the observed number of particles measured by the in situ optical particle counter and the y-axis showing the percentiles from the probability distribution function (PDF) of the estimated number of particles. The dotted red line shows the ideal response. We can see that for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, e.g., those with a size of 0.25 μm (Fig. 4b), the shape of the observed and estimated PDFs are almost the same; note that we have a straight line between the 25th and 75th percentiles. The same is true for particles withPlease provide explanation for part labels (a) to (e) in the caption of Fig. 5. a diameter of 2.5 μm (Fig. 4e). However, for the larger particles that sediment rapidly, e.g., those with a diameter of 25 μm (Fig. 4h), the independent validation does not do well.

The right-hand column of plots in Fig. 4 shows the relative importance of the input variables for estimating the number of particles using machine learning. We note that in each case, the temperature and pressure and sometimes the humidity are key factors. For the small particles with a diameter of 0.25 μm, the NEXRAD variables providing the most information are the correlation coefficient and Doppler speed at elevation 1. For the particles with a diameter of 2.5 μm, the NEXRAD variables providing the most information are the Doppler speed at elevation 2 and the differential reflectivity at elevation 1.

Figure 5 shows the spatial distribution of $PM_{2.5}$ particulates estimated over a large spatial area at 0.5 km × 0.5 km resolution. In this case, the machine learning model was developed at 10 km × 10 km pixel, and the model was applied to each pixel using the NEXRAD and atmospheric weather measurements as input.

Summary

Airborne particulates are of particular significance for their human health impacts and their roles in both atmospheric radiative transfer and atmospheric chemistry. Observations of airborne particulates are typically made by environment agencies using rather expensive instruments. Due to the expense of the instruments usually used by environment agencies, the number of sensors that can be deployed is limited. In this study, we have shown two different case studies illustrating the utility of using machine learning for studying airborne particulates.

We have shown that machine learning can be used to effectively calibrate lower-cost optical particle counters. For this calibration, it is critical that measurements of the atmospheric pressure, humidity, and temperature are included. Once the machine learning calibration has been applied to the low-cost sensors, independent validation using scatter diagrams and quantile-quantile plots shows that not only is the calibration effective, but the shape of the resulting probability distribution of observations is very well preserved.

These low-cost sensors are being deployed at scale across the dense urban environment of the Dallas Fort Worth Metroplex for both characterizing the temporal and spatial scales of urban air pollution and providing high schools and high school coaches a tool to assist in making better decisions to reduce adverse health outcomes, e.g., given the levels of pollen/pollution today, should physical education/practice be outside or inside?

In this study, we have also shown that observations made by NEXRAD weather radars can be used with machine learning to effectively estimate the abundance of airborne particulates with a diameter in the size range 0.1–7 μm. For this estimation, it is critical that measurements of the atmospheric pressure, humidity, and temperature are also made. Once machine learning has been applied, scatter diagrams and quantile-quantile plots show that not only is the approach effective, but the shape of the resulting probability distribution of observations is preserved.

References

Ahmad, Z., W. Choi, N. Sharma, J. Zhang, Q. Zhong, D.-Y. Kim, Z. Chen, Y. Zhang, R. Han, D. Shim, et al. 2016. Devices and circuits in CMOS for thz applications. In *Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM)*, pp. 29–8. New York: IEEE.

Alavi, Amir H., Amir H. Gandomi, and David J. Lary. 2016. *Progress of machine learning in geosciences*.

Albayrak, Arif, J.C. Wei, Maksym Petrenko, D.J. Lary, and G.G. Leptoukh. 2011. Modis aerosol optical depth bias adjustment using machine learning algorithms. In *AGU Fall Meeting Abstracts*.

Alfarra, Mohammedrami. 2004. *Insights into atmospheric organic aerosols using an aerosol mass spectrometer*. PhD thesis, Manchester: University of Manchester.

Allen, Myles R., Vicente R. Barros, John Broome, Wolfgang Cramer, Renate Christ, John A. Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, Navroz K. Dubash, et al. 2014. *IPCC fifth assessment synthesis report-climate change 2014 synthesis report*.

Alphasense. 2018. *Alphasense user manual opc-n3 optical particle counter*.

Andreae, Meinrat O., and Paul J. Crutzen. 1997. Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry. *Science* 276(5315): 1052–1058.

Atkinson, Roger. 2000. Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment* 34(12–14): 2063–2101.

Basu, Mausumi. 2019. The great smog of Delhi. *Lung India: Official Organ of Indian Chest Society* 36(3): 239.

Bickis, Ugis. 1998. Hazard prevention and control in the work environment: airborne dust. *World Health* 13: 16.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press. 95040465 Christopher M. Bishop. ill.; 24 cm. Includes bibliographical references (p. [457]-475) and index.

Black, J. 2003. Intussusception and the great smog of London, December 1952. *Archives of Disease in Childhood* 88(12): 1040–1042.

Boldo, Elena, Sylvia Medina, Alain Le Tertre, Fintan Hurley, Hans-Guido Mücke, Ferrán Ballester, Inmaculada Aguilera, et al. 2006. Apheis: Health impact assessment of long-term exposure to pm2. 5 in 23 European cities. *European Journal of Epidemiology* 21(6): 449–458.

Bond, Tami C., David G. Streets, Kristen F. Yarber, Sibyl M. Nelson, Jung-Hun Woo, and Zbigniew Klimont. 2004. A technology-based global inventory of black and organic carbon emissions from combustion. *Journal of Geophysical Research: Atmospheres* 109(D14).

Boucher. O. 2015. *Atmospheric Aerosols: Properties and Climate Impacts*. Netherlands: Springer. ISBN 978-9-40-179648-4. https://books.google.co.in/books?id=RXDCoQEACAAJ&redir_esc=y.

Breiman, Leo. 1984. *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series. Belmont: Wadsworth International Group.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.

Broich, Anna V., Lydia E. Gerharz, and Otto Klemm. 2012. Personal monitoring of exposure to particulate matter with a high temporal resolution. *Environmental Science and Pollution Research* 19(7): 2959–2972.

Brown, Molly E., David J. Lary, Anton Vrieling, Demetris Stathakis, and Hamse Mussa. 2008. Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. *International Journal of Remote Sensing* 29(24): 7141–7158.

Brown, James S., Terry Gordon, Owen Price, and Bahman Asgharian. 2013. Thoracic and respirable particle definitions for human health risk assessment. *Particle and Fibre Toxicology* 10(1): 12.

Byun, Hyang-Min, Tommaso Panni, Valeria Motta, Lifang Hou, Francesco Nordio, Pietro Apostoli, Pier Alberto Bertazzi, and Andrea A Baccarelli. 2013. Effects of airborne pollutants on mitochondrial dna methylation. *Particle and Fibre Toxicology* 10(1): 18.

Carlson, Jen. 2009. Flashback: The city's killer SMOG. https://gothamist.com/news/flashback-the-citys-killer-smog#photo-1.

Carslaw, K.S., R.G. Harrison, and J. Kirkby. 2002. Cosmic rays, clouds, and climate. *Science* 298(5599): 1732–1737.

Chang, Howard H., Anqi Pan, David J. Lary, Lance A. Waller, Lei Zhang, Bruce T. Brackin, Richard W. Finley, and Fazlay S. Faruque. 2019. Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS. *Environmental Monitoring and Assessment* 191(280).

Charlson, Robert J., S.E. Schwartz, J.M. Hales, Ro D. Cess, Jr J.A. Coakley, J.E. Hansen, and D.J. Hofmann. 1992. Climate forcing by anthropogenic aerosols. *Science* 255(5043): 423–430.

Chen, Renjie, Yi Li, Yanjun Ma, Guowei Pan, Guang Zeng, Xiaohui Xu, Bingheng Chen, and Haidong Kan. 2011. Coarse particles and mortality in three Chinese cities: the china air pollution and health effects study (capes). *Science of the Total Environment* 409(23): 4934–4938.

Cheng, M., and W. Liu. 2009. *Airborne Particulates*. New York: Nova Science Publishers. ISBN 978-1-60-692907-0. https://books.google.co.in/books?id=3H5wPgAACAAJ&redir_esc=y.

Chin, M. 2009. *Atmospheric Aerosol Properties and Climate Impacts*. Collingdale: DIANE Publishing Company. ISBN 978-1-43-791261-6. https://books.google.co.in/books?id=IgJZXXgtHmQC&redir_esc=y.

Choi, Wooyeol, Qian Zhong, Navneet Sharma, Yaming Zhang, Ruonan Han, Z. Ahmad, Dae-Yeon Kim, Sandeep Kshattry, Ivan R. Medvedev, David J. Lary, et al. 2019. Opening terahertz for everyday applications. *IEEE Communications Magazine* 57(8): 70–76.

Chow, Judith C., John G. Watson, et al. 1998. Guideline on speciated particulate monitoring. In *Report prepared for US Environmental Protection Agency, Research Triangle Park, NC*. Reno: Desert Research Institute.

Colbeck, I. 2014. *Aerosol Science: Technology and Applications*. New York: Wiley. ISBN 978-1-11-997792-6. https://books.google.co.in/books?id=eKUTAgAAQBAJ&redir_esc=y.

Colbeck, Ian, and Mihalis Lazaridis. 2010. Aerosols and environmental pollution. *Naturwissenschaften* 97(2): 117–131.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3): 273–297. Times Cited: 3429.

Cromar, Kevin R., Laura A. Gladson, Lars D. Perlmutt, Marya Ghazipura, and Gary W. Ewart. 2016. American thoracic society and marron institute report. Estimated excess morbidity and mortality caused by air pollution above American thoracic society–recommended standards, 2011–2013. *Annals of the American Thoracic Society* 13(8): 1195–1201.

Dadvand, Payam, Jennifer Parker, Michelle L. Bell, Matteo Bonzini, Michael Brauer, Lyndsey A. Darrow, Ulrike Gehring, Svetlana V. Glinianaia, Nelson Gouveia, Eun-hee Ha, et al. 2013. Maternal exposure to particulate air pollution and term birth weight: A multi-country evaluation of effect and heterogeneity. *Environmental Health Perspectives* 121(3): 267.

Demuth, Howard B., Mark H. Beale, Orlando De Jess, and Martin T. Hagan. 2014. *Neural Network Design*. Martin Hagan, USA, 2nd edn. ISBN 0-9717-3211-6, 978-0-97-173211-7.

Dockery, D.W., C.A. Pope, X.P. Xu, J.D. Spengler, J.H. Ware, M.E. Fay, et al. 1993a. An association between air-pollution and mortality in 6 United-States cities. *New England Journal of Medicine* 329(24): 1753–1759. *Find this article online*.

Dockery, Douglas W., C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris Jr, and Frank E. Speizer. 1993b. An association between air pollution and mortality in six US cities. *New England Journal of Medicine* 329(24): 1753–1759.

Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. London: Basic Books.

Dong, W., G. Guan, Y. Chen, K. Guo, and Y. Gao. 2015. Mosaic: Towards city scale sensing with mobile sensor networks. In *Proceedings of the 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 29–36. https://doi.org/10.1109/ICPADS.2015.12.

Dubovik, Oleg, Brent Holben, Thomas F. Eck, Alexander Smirnov, Yoram J. Kaufman, Michael D. King, Didier Tanré, and Ilya Slutsker. 2002. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *Journal of the Atmospheric Sciences* 59(3): 590–608.

Gali, Rohith Kumar. 2010. *Assessment of NEXRAD P3 data on streamflow simulation using SWAT for North Fork Ninnescah Watershed, Kansas*. PhD thesis, Manhattan: Kansas State University.

Glasser, Marvin, Leonard Greenburg, and Franklyn Field. 1967. Mortality and morbidity during a period of high levels of air pollution: New York, Nov 23–25, 1966. *Archives of Environmental Health: An International Journal* 15(6): 684–694.

Guenther, A., T. Karl, Pedro Harley, C. Wiedinmyer, P.I. Palmer, and C. Geron. 2006. Estimates of global terrestrial isoprene emissions using MEGAN (model of emissions of gases and aerosols from nature). *Atmospheric Chemistry and Physics* 6(11): 3181–3210.

Guo, Liqiong, Hyang-Min Byun, Jia Zhong, Valeria Motta, Jitendra Barupal, Yinan Zheng, Chang Dou, Feiruo Zhang, John P McCracken, Anaité Diaz, et al. 2014. Effects of short-term exposure to inhalable particulate matter on DNA methylation of tandem repeats. *Environmental and Molecular Mutagenesis* 55(4): 322–335.

Haberzettl, Petra, Timothy E. O'Toole, Aruni Bhatnagar, and Daniel J. Conklin. 2016. Exposure to fine particulate air pollution causes vascular insulin resistance by inducing pulmonary oxidative stress. *Environmental Health Perspectives* 124(12): 1830.

Hallquist, Mattias, John C. Wenger, Urs Baltensperger, Yinon Rudich, David Simpson, M. Claeys, J. Dommen, N.M. Donahue, C. George, A.H. Goldstein, et al. 2009. The formation, properties and impact of secondary organic aerosol: Current and emerging issues. *Atmospheric Chemistry and Physics* 9(14): 5155–5236.

Harrison, William Alan. 2015. *In-situ observation of atmospheric particulates*. Dallas: The University of Texas.

Harrison, William A., David Lary, Brian Nathan, and Alec G Moore. 2015. The neighborhood scale variability of airborne particulates. *Journal of Environmental Protection* 6(05): 464.

Haykin, Simon S. 1994. *Neural Networks: A Comprehensive Foundation*. New York: Macmillan. 93028092 Simon Haykin. ill.; 26 cm. Includes bibliographical references (p. 635–690) and index.

Haykin, Simon S. 1999. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River: Prentice Hall, 2nd edn., 98007011 Simon Haykin. ill.; 25 cm. Includes bibliographical references (p. 796–836) and index.

Haykin, Simon S. 2001. Kalman Filtering and Neural Networks. In *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley. 2001049240 edited by Simon Haykin. ill.; 24 cm. A Wiley Interscience publication. Includes bibliographical references and index.

Haykin, Simon S. 2007. *New Directions in Statistical Signal Processing: From Systems to Brain*. Neural Information Processing Series. Cambridge: MIT Press. 2005056210 GBA671791 013536699 (OCoLC)ocm62302576 (OCoLC)62302576 edited by Simon Haykin ... [et al.]. ill.; 26 cm. Includes bibliographical references (p. [465]-508) and index. Modeling the mind: from circuits to systems/Suzanna Becker—Empirical statistics and stochastic models for visual signals/David Mumford—The machine cocktail party problem/Simon Haykin, Zhe Chen—Sensor adaptive signal processing of biological nanotubes (ion channels) at macroscopic and nano scales/Vikram Krishnamurthy—Spin diffusion: a new perspective in magnetic resonance imaging/Timothy R. Field—What makes a dynamical system computationally powerful?/Robert Legenstein, Wolfgang Maass—A variational principle for graphical models/Martin J. Wainwright, Michael I. Jordan—Modeling large dynamical systems with dynamical consistent neural networks/Hans-Georg Zimmermann ... [et al.]—Diversity in communication: from source coding to wireless networks/Suhas N. Diggavi—Designing patterns for easy recognition: information transmission with low-density parity-check codes/Frank R. Kschischang, Masoud Ardakani—Turbo processing/Claude Berrou, Charlotte Langlais, Fabrice Seguin—Blind

signal processing based on data geometric properties/Konstantinos Diamantaras—Game-theoretic learning / Geoffrey J. Gordon—Learning observable operator models via the efficient sharpening algorithm/Herbert Jaeger . . . [et al.].

Health Effects Institute HEI. 2017. *State of global air 2017*. Special.

Hinds, William C. 2012. *Aerosol technology: properties, behavior, and measurement of airborne particles*. New York: Wiley.

Ho, T.K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832–844.

Holben, B.N., D. Tanre, A. Smirnov, T.F. Eck, I. Slutsker, N. Abuhassan, W.W. Newcomb, J.S. Schafer, B. Chatenet, F. Lavenu, et al. 2001. An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET. *Journal of Geophysical Research: Atmospheres* 106(D11): 12067–12097.

Huang, Fang, Renjie Chen, Yuetian Shen, Haidong Kan, and Xingya Kuang. 2016. The impact of the 2013 eastern china smog on outpatient visits for coronary heart disease in shanghai, china. *International Journal of Environmental Research and Public Health* 13(7): 627.

Jimenez, Jose L. M.R. Canagaratna, N.M. Donahue, A.S.H. Prevot, Qi Zhang, Jesse H. Kroll, Peter F. DeCarlo, James D. Allan, H. Coe, N.L. Ng, et al. 2009. Evolution of organic aerosols in the atmosphere. *Science* 326(5959): 1525–1529.

Kanakidou, M. J.H. Seinfeld, S.N. Pandis, I. Barnes, F.J. Dentener, M.C. Facchini, R. Van Dingenen, B. Ervens, ANCJSE Nenes, C.J. Nielsen, et al. 2005. Organic aerosol and global climate modelling: a review. *Atmospheric Chemistry and Physics* 5(4): 1053–1123.

Kelly, Frank J., and Julia C Fussell. 2016. Health effects of airborne particles in relation to composition, size and source. *Airborne Particulate Matter*, 344–382.

Kirkby, Jasper, Joachim Curtius, João Almeida, Eimear Dunne, Jonathan Duplissy, Sebastian Ehrhart, Alessandro Franchin, Stéphanie Gagné, Luisa Ickes, Andreas Kürten, et al. 2011. Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* 476(7361): 429.

Kneen, Melanie A., David J. Lary, William A. Harrison, Harold J. Annegarn, and Tom H. Brikowski. 2016. Interpretation of satellite retrievals of pm2.5 over the Southern African interior. *Atmospheric Environment* 128: 53–64.

Kokhanovsky, Alexander A. 2008. *Aerosol optics: light absorption and scattering by particles in the atmosphere*. Berlin: Springer.

Kondratyev, Kirill Ya, Lev S. Ivlev, Vladimir F. Krapivin, and Costas A. Varostos. 2006. *Atmospheric aerosol properties: Formation, processes and impacts*. Berlin: Springer.

Lary, D. 2007. Using neural networks for instrument cross-calibration. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2010. *Artificial intelligence in geoscience and remote sensing*. London: INTECH Open Access Publisher.

Lary, David J. 2013. Using multiple big datasets and machine learning to produce a new global particulate dataset: A technology challenge case study. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2014. Bigdata and machine learning for public health. In *142nd APHA Annual Meeting and Exposition 2014*. Washington: APHA.

Lary, D.J. and O. Aulov. 2008. Space-based measurements of hcl: Intercomparison and historical context. *Journal of Geophysical Research: Atmospheres* 113(D15).

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2003. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics Discussions* 3(6): 5711–5724.

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2004. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics* 4(1): 143–146.

Lary, David J., L.A. Remer, Devon MacNeill, Bryan Roscoe, and Susan Paradise. 2009a. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* 6(4): 694–698.

Lary, D.J., L.A. Remer, D. MacNeill, B. Roscoe, and S. Paradise. 2009b. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* 6(4): 694–698.

Lary, D.J., A. Nikitkov, D. Stone, and Alexey Nikitkov. 2010. Which machine-learning models best predict online auction seller deception risk. *American Accounting Association AAA Strategic and Emerging Technologies*.

Lary, David J., Fazlay S. Faruque, Nabin Malakar, Alex Moore, Bryan Roscoe, Zachary L. Adams, and York Eggelston. 2014. Estimating the global abundance of ground level presence of particulate matter (pm2. 5). *Geospatial Health* 8(3): 611–630.

Lary, D.J., T. Lary, and B. Sattler. 2015a. Using machine learning to estimate global pm2.5 for environmental health studies. *Environmental Health Insights* 9: EHI–S15664.

Lary, D.J., T. Lary, and B. Sattler. 2015b. Using machine learning to estimate global pm2. 5 for environmental health studies. *Environmental Health Insights* 9(Suppl 1): 41.

Lary, David J., Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7(1): 3–10.

Lary, David J., Gebreab K. Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J. Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, et al. 2018. Machine learning applications for earth observation. In *Earth Observation Open Science and Innovation. ISSI Scientific Report Series* vol. 15, pp. 165–218. Berlin: Springer.

Lary, Maria-Anna, Leslie Allsop, and David John Lary. 2019. Using machine learning to examine the relationship between asthma and absenteeism. *Environmental Modeling and Assessment* 191(332): 1–9.

Lelieveld, Jos, John S. Evans, M. Fnais, Despina Giannadaki, and Andrea Pozzer. 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* 525(7569): 367.

Lesins, Glen, Petr Chylek, and Ulrike Lohmann. 2002. A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing. *Journal of Geophysical Research: Atmospheres* 107(D10): AAC–5.

Levy, Robert. 2014. Smog shrouds Eastern China. https://earthobservatory.nasa.gov/images/82535/smog-shrouds-eastern-china.

Li, Linglong, Yixin Zheng, and Lin Zhang. 2014. Demonstration abstract: Pimi air box: a cost-effective sensor for participatory indoor quality monitoring. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, pp. 327–328. New York: IEEE Press.

Lim, Stephen S., Theo Vos, Abraham D. Flaxman, Goodarz Danaei, Kenji Shibuya, Heather Adair-Rohani, Mohammad A. AlMazroa, Markus Amann, H. Ross Anderson, Kathryn G. Andrews, et al. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet* 380(9859): 2224–2260.

Madonna, F., A. Amodeo, G. D'Amico, L. Mona, and G. Pappalardo. 2010. Observation of non-spherical ultragiant aerosol using a microwave radar. *Geophysical Research Letters* 37(21).

Maji, Kamal Jyoti, Anil Kumar Dikshit, and Ashok Deshpande. 2017. Disability-adjusted life years and economic cost assessment of the health effects related to pm2. 5 and pm10 pollution in Mumbai and Delhi, in India from 1991 to 2015. *Environmental Science and Pollution Research* 24(5): 4709–4730.

Malakar, Nabin K., David J. Lary, A. Moore, D. Gencaga, Bryan Roscoe, Arif Albayrak, and Jennifer Wei. 2012a. Estimation and bias correc-

tion of aerosol abundance using data-driven machine learning and remote sensing. In *Proceedings of the2012 Conference on Intelligent Data Understanding*, pp. 24–30. New York: IEEE.

Malakar, N.K., D.J. Lary, R. Allee, R. Gould, and D. Ko. 2012b. Towards automated ecosystem-based management: A case study of northern Gulf of Mexico water. In *AGU Fall Meeting Abstracts*.

Malakar, N.K., D.J. Lary, D. Gencaga, A. Albayrak, and J. Wei. 2013. Towards identification of relevant variables in the observed aerosol optical depth bias between MODIS and Aeronet observations. In *AIP Conference Proceedings*, vol. 1553, pp. 69–76. College Park: AIP.

Malakar, Nabin K., D.J. Lary, and B. Gross. 2018. Case studies of applying machine learning to physical observation. In *AGU Fall Meeting Abstracts*.

Mannucci, Pier Mannuccio. 2017. *Air pollution levels and cardiovascular health: Low is not enough*.

McCulloch, W.S., and W. Pitts. 1943. *Bulletin of Mathematical Biophysics* 5: 115. https://doi.org/10.1007/BF02478259.

Medvedev, Ivan R., Robert Schueler, Jessica Thomas, O. Kenneth, Hyun-Joo Nam, Navneet Sharma, Qian Zhong, David J. Lary, and Philip Raskin. 2016. Analysis of exhaled human breath via terahertz molecular spectroscopy. In *Proceedings of the 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz)*, pp. 1–2. New York: IEEE.

Nada Osseiran, Lindmeier, Christian. 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.

Nathan, Brian J., and David J. Lary. 2019. Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale. *Environmental Modeling and Assessment* 191(337).

O, K.K., Q. Zhong, N. Sharma, W. Choi, R. Schueler, I.R. Medvedev, H.-J. Nam, P. Raskin, F.C. De Lucia, J.P. McMillan, et al. 2017. Demonstration of breath analyses using CMOS integrated circuits for rotational spectroscopy. In *International Workshop on Nanodevice Technologies, Hiroshima, Japan*.

Oberdörster, Günter, Eva Oberdörster, and Jan Oberdörster. 2005. Nanotoxicology: An emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives* 113(7): 823–839.

Onishi, Kazunari, Tsuyoshi Thomas Sekiyama, Masanori Nojima, Yasunori Kurosaki, Yusuke Fujitani, Shinji Otani, Takashi Maki, Masato Shinoda, Youichi Kurozawa, and Zentaro Yamagata. 2018. Prediction of health effects of cross-border atmospheric pollutants using an aerosol forecast model. *Environment International* 117: 48–56.

Pascal, Mathilde, Magali Corso, Olivier Chanel, Christophe Declercq, Chiara Badaloni, Giulia Cesaroni, Susann Henschel, Kadri Meister, Daniela Haluza, Piedad Martin-Olmedo, et al. 2013. Assessing the public health impacts of urban air pollution in 25 european cities: results of the Aphekom project. *Science of the Total Environment* 449: 390–400.

Polivka, Barbara J. The great London smog of 1952. *AJN The American Journal of Nursing* 118(4): 57–61 (2018).

Pope, C. Arden, Richard T. Burnett, George D. Thurston, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, and John J. Godleski. 2004. Cardiovascular mortality and long-term exposure to particulate air pollution. *Circulation* 109(1): 71–77.

Pope, C., Arden Dockery, and Douglas W. 2006. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air and Waste Management Association* 56(6): 709–742.

Pope C. ArdenIII, Richard T. Burnett, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, Kazuhiko Ito, and George D. Thurston. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama* 287(9): 1132–1141.

Pöschl. Ulrich. 2005. Atmospheric aerosols: composition, transformation, climate and health effects. *Angewandte Chemie International Edition* 44(46): 7520–7540.

Pun, Vivian C., Justin Manjourides, and Helen Suh. 2017. Association of ambient air pollution with depressive and anxiety symptoms in older adults: results from the NSHAP study. *Environmental Health Perspectives* 125(3): 342.

Ramanathan, V.C.P.J., P.J. Crutzen, J.T. Kiehl, and Dm Rosenfeld. 2001. Aerosols, climate, and the hydrological cycle. *Science* 294(5549): 2119–2124.

Ruckerl, R., A. Ibald-Mulli, W. Koenig, A. Schneider, G. Woelke, J. Cyrys, and A. Peters. 2006. Air pollution and markers of inflammation and coagulation in patients with coronary heart disease. *American Journal of Respiratory and Critical Care Medicine* 173(4): 432–441.

Safavian, S.R., and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 21(3): 660–674. ISSN 0018–9472. https://doi.org/10.1109/21.97458.

Santibañez, Daniela A., Sergio Ibarra, Patricia Matus, Rodrigo Seguel, et al. 2013. A five-year study of particulate matter (pm2. 5) and cerebrovascular diseases. *Environmental Pollution* 181: 1–6.

Saravanan, J., M. Jayadurgalakshmi, and R. Karthickraja. 2017. China's Nanjing vs India's Delhi–a perspective for vertical forest. *International Journal of Civil Engineering and Technology* 8: 12.

Schauer, James J., Wolfgang F. Rogge, Lynn M. Hildemann, Monica A. Mazurek, Glen R. Cass, and Bernd R.T. Simoneit. 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment* 30(22): 3837–3855.

Seinfeld, J.H. 1986. *Atmospheric chemistry and physics of air pollution*. A Wiley-Interscience publication. New York: Wiley. ISBN 978-0-47-182857-0. https://books.google.co.in/books?id=NAhSAAAAMAAJ&redir_esc=y.

Shy, Carl M., Victor Hasselblad, Robert M. Burt, Cornelius J. Nelson, and Arlan A. Cohen. 1973. Air pollution effects on ventilatory function of us schoolchildren: Results of studies Cincinnati, Chattanooga, and New York. *Archives of Environmental Health: An International Journal* 27(3): 124–128.

Solomon, Feliz. 2016. China's SMOG is as deadly as smoking, new research claims. https://time.com/4617295/china-smog-smoking-environment-air-pollution/.

Spira-Cohen, Ariel, Lung Chi Chen, Michaela Kendall, Ramona Lall, and George D. Thurston. 2011. Personal exposures to traffic-related air pollution and acute respiratory health among Bronx schoolchildren with asthma. *Environmental Health Perspectives* 119(4): 559.

Stier, P., J. Feichter, S. Kinne, S. Kloster, E. Vignati, J. Wilson, L. Ganzeveld, I. Tegen, Martin Werner, Y. Balkanski, et al. 2005. The aerosol-climate model echam5-ham. *Atmospheric Chemistry and Physics* 5(4): 1125–1156.

Stocker, Thomas. 2014. *Climate change 2013: The physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Streets, D. Ga, T.C. Bond, G.R. Carmichael, S.D. Fernandes, Q. Fu, D. He, Z. Klimont, S.M. Nelson, N.Y. Tsai, M. Qm Wang, et al. 2003. An inventory of gaseous and primary aerosol emissions in asia in the year 2000. *Journal of Geophysical Research: Atmospheres* 108(D21).

Sỳkorová, Barbora, Marek Kucbel, and Konstantin Raclavskỳ. 2016. Composition of airborne particulate matter in the industrial area versus mountain area. *Perspectives in Science* 7: 369–372.

Terry, James P., Gensuo Jia, Robert Boldi, and Sarah Khan. 2018. The Delhi 'gas chamber': smog, air pollution and the health emergency of november 2017. *Weather* 73(11): 348–352.

Thurston, George D., Jiyoung Ahn, Kevin R Cromar, Yongzhao Shao, Harmony R. Reynolds, Michael Jerrett, Chris C. Lim, Ryan Shanley, Yikyung Park, and Richard B. Hayes. 2016. Ambient particulate matter air pollution exposure and mortality in the nih-aarp diet and health cohort. *Environmental Health Perspectives* 124(4): 484.

US EPA. 2004. *Air quality criteria for particulate matter*, vol. 2. US Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment.

Vapnik, Vladimir Naumovich. 1982. *Estimation of Dependences Based on Empirical Data*. In *Springer Series in Statistics*. New York: Springer.

Vapnik, Vladimir Naumovich. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, Vladimir Naumovich. 2000. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science, 2nd edn. Springer, New York.

Vapnik, Vladimir Naumovich. 2006. *Estimation of Dependences Based on Empirical Data; Empirical Inference Science: Afterword of 2006*. In *Information Science and Statistics*, 2nd edn. New York: Springer.

Wachs, Anthony. 2009. A dem-dlm/fd method for direct numerical simulation of particulate flows: Sedimentation of polygonal isometric particles in a Newtonian fluid with collisions. *Computers and Fluids* 38(8): 1608–1628.

Wilkins, E.T. 1954. Air pollution and the London fog of December 1952. *Journal of the Royal Sanitary Institute* 74(1): 1–21.

Wu, Daji, Gebreab K. Zewdie, Xun Liu, Melanie Anne Kneen, and David John Lary. 2017. Insights into the morphology of the East Asia pm2. 5 annual cycle provided by machine learning. *Environmental Health Insights* 11: 1178630217699611.

Wu, Daji, David J. Lary, Gebreab K. Zewdie, and Xun Liu. 2019. Using machine learning to understand the temporal morphology of the pm2.5 annual cycle in East Asia. *Environmental Monitoring and Assessment* 191(272): 1–14.

Yunker, Mark B., Robie W. Macdonald, Roxanne Vingarzan, Reginald H. Mitchell, Darcy Goyette, and Stephanie Sylvestre. 2002. PAHs in the Fraser river basin: a critical appraisal of PAH ratios as indicators of PAH source and composition. *Organic Geochemistry* 33(4): 489–515.

Zewdie, Gebreab, and David J. Lary. 2018. Applying machine learning to estimate allergic pollen using environmental, land surface and NEXRAD radar parameters. In *AGU Fall Meeting Abstracts*.

Zewdie, Gebreab K., David J. Lary, Estelle Levetin, and Gemechu F. Garuma. 2019a. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *International Journal of Environmental Research and Public Health* 16(11): 1992.

Zewdie, Gebreab K., David J. Lary, Xun Liu, Daji Wu, and Estelle Levetin. 2019b. Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. *Environmental Monitoring and Assessment* 191(7): 418.

Linking Disease Outcomes to Environmental Risks: The Effects of Changing Spatial Scale

Chetan Tiwari, David Sterling, and Leslie Allsopp

Introduction

Geographic information systems (GIS) enable the assessment of environmental risks and their potential impacts on human health outcomes by providing a mechanism to overlay maps of exposure and disease outcomes. While the ability to overlay various geographic datasets is a common function provided by most GIS software packages, the ability to derive meaningful associations between these layers is limited by data quality, issues pertaining to the accuracy of exposure assessment, and problems of representing the intensity of disease outcomes over space and time. In a recent review of the role of geographic information science (GISc) in the analysis of health and place, Mennis and Yoo (2018) identify major challenges and opportunities including problems associated with scale of analysis in health research. In this context, they argue that most GIS-based health research in this area has focused on problems of sparse or missing data and emphasizes the need for more research to understand the implications of resolution and spatial and temporal sampling frameworks on assessments of individual-level environmental exposures (Mennis and Yoo 2018). The problem of deriving associations between layers of geographic data with inconsistent scales is of particular concern in the context of big data where personalized health information through electronic health records and individual measures of exposure via low-cost sensors are becoming more common. Limitations on the use of such data due to privacy concerns or inconsistent quality often lead to the production and dissemination of datasets aggregated to different levels of spatial resolution. While GIS software can be used to combine such layers of geographic data, it is critical to note that inconsistent scales and misaligned boundaries resulting from the use of disparate spatial units will likely result in incorrect and/or misleading conclusions.

The problems of changing spatial scales and misaligned geographic boundaries are well documented in discussions of spatial uncertainty. Spatial uncertainty is broadly defined as the problem of identifying and quantifying error in the geographic location of objects. Such error may result in biased interpretations of the true relationships between the location of objects in space and surrounding contextual or environmental factors. There are two issues associated with spatial uncertainty – the change of support problem (CoSP) and the uncertain geographic context problem (UGCoP). CoSP is concerned with the problem of drawing inferences about observations at a spatial scale that is different from the scale at which those observations have occurred. Kwan (2012) defines the uncertain geographic context problem (UGCoP) as the problem of identifying the effects of spatial displacement between the geographic definitions of contextual units and the "true causally relevant" context. The problem presented by aggregation may be considered as a subset of the CoSP and is similar to the well-known modifiable areal unit problem (MAUP) which states the patterns observed on

C. Tiwari (✉)
Departments of Geosciences & Computer Science, Georgia State University, Atlanta, GA, USA

Center for Disaster Informatics & Computational Epidemiology, Georgia State University, Atlanta, GA, USA
e-mail: ctiwari@gsu.edu

D. Sterling
Department of Biostatistics and Epidemiology, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA

School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

L. Allsopp
Department of Biostatistics and Epidemiology, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA
e-mail: Leslie.allsopp@unthsc.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_14

a map; therefore, any inferences drawn are likely to change if the shape and scale of the map unit change. In many population-based studies, the UGCoP arises from the need to use area-based measures of statistical observations, such as census tracts, to reconstruct a fundamentally vague definition of space, such as the risk of disease that arises from some kind of exposure to environmental risk. See Kwan (2018), Fisher et al. (2018), and Griffith (2018) for an overview of spatial analytic approaches to identify, measure, and address this problem.

The utility and pitfalls of GIS for exposure assessment are well described in Nuckols et al. (2004). They describe several studies that use GIS in various ways to estimate exposures to a variety of environmental risks. Some examples of studies cited in their work include an assessment of associations between residential proximity to landfill sites and adverse birth outcomes (Elliott et al. 2001), examination of possible neurobehavioral effects of exposure to trichloroethylene using a simulation model (MODFLOW) (Reif et al. 2003), and a population-based study to evaluate lung cancer outcomes to urban air pollution using a combination of dispersion modeling and geostatistical techniques (Bellander et al. 2001; Nyberg et al. 2000). Other studies that use GIS to produce fine-scale assessments of environmental risk include an assessment of the health risks posed by urban heat islands detected using remote sensing methods including airborne or satellite platforms (Tomlinson et al. 2011), an overview of techniques to produce fine-scale estimates of fine particulate matter using remotely sensed data and geostatistical approaches (Al-Hamdan et al. 2009, 2014), estimation of spatio-temporal variations in hot weather conditions of Hong Kong using statistical techniques (Shi et al. 2019), and mapping of exposures to particulate matter using remotely sensed data and geostatistical modeling techniques (Leelasakultum and Kim Oanh 2017). In all cases, assessments of the risk of environmental exposure are estimated for various levels of spatial resolution that may not necessarily conform to the scale at which disease data are commonly made available.

Under current practice, makers of disease maps select census or other administrative units for which both disease and demographic data are available. GIS software are then used to compute and visualize disease *rates* among the populations contained within those administrative units. The choice of administrative unit influences the resolution and statistical reliability of observed disease rates. Patterns of disease rates displayed on maps produced using such spatial units represent a tradeoff between spatial resolution and statistical reliability. Maps with high degrees of spatial resolution generally exhibit poor statistical reliability as the population support – numbers of persons at risk – used in calculating each rate is often small. As more sources of georeferenced health and demographic data become available, so does the opportunity to control the numbers of people at risk and the geographic size of the areas mapped. In geographical circles, the spatial resolution of a map refers to the size or area used to measure the spatial variation of a disease rate. If the areas mapped were of equal size, the map would be said to have the same geographic resolution across the map. Since most maps use administrative areas as the spatial units to map, the common spatial resolution of a map is the average size of the administrative areas used. A second meaning of spatial resolution is when the minimum size mapped is the smallest size for which the common geography between disease data and demographic data realizes a fixed level of statistical reliability. The goal of the actual spatial resolution achieved by the map is not, therefore, a common spatial size, but, instead, a minimum sized spatial unit at any location on the map that realizes the statistical reliability desired by the mapmaker. A third meaning of spatial resolution has arisen more recently in the era of digital maps when the smallest spatial unit on the map is a pixel. If the map is constructed so that pixel values change according to some function of relative location, then the earth size corresponding to one pixel is the geographic resolution of the map in question. In this chapter, we advocate for a disease mapping approach that focuses on a deliberate choice of geographic resolution and statistical reliability. To this end, we demonstrate how the two characteristics can be controlled using a simulated dataset on disease outcomes that are influenced by four randomly selected locations of environmental exposure.

Relevance of Disease Mapping for Assessing Public Health Impacts

Disease mapping refers to the process of constructing a map that shows the spatial distribution of disease within a specific geographic region. Disease maps improve public health decision-making by providing a mechanism to identify geographic areas that are in most need of interventions or resources (Bertollini and Martuzzi 1999; Moore and Carpenter 1999; Ricketts 2003). They can help answer such as the following questions: What populations are at risk? Where they are located? What are the underlying conditions in those areas? The common spatial context enables researchers and public health practitioners to link various geographic layers of data to explore associations between a multitude of complex processes that include various combinations of social, cultural, and environmental determinants. In 1850, John Snow created the first disease map of cholera distribution in London and initially showed the importance of cartographic representation of disease in serving public health (Koch 2004). Snow's point map shown in Fig. 1 describes the spatial patterns of cholera deaths and its geographical association with other features on the landscape, including the broad street pump, which was subsequently identified as serving the

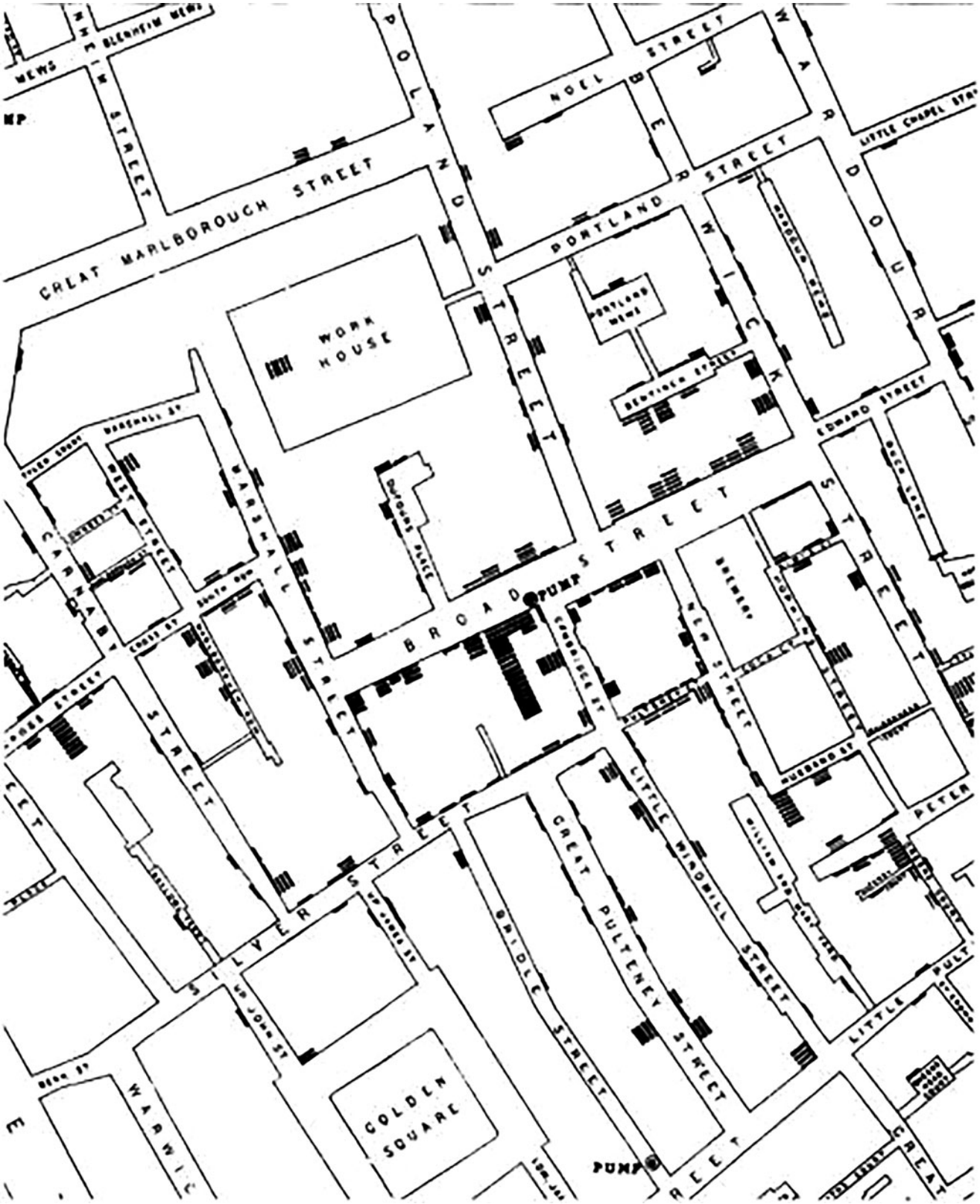


Fig. 1 John Snow's map of cholera death in London (McLeod 2000)

population with sewage tainted water (McLeod 2000; Shiode et al. 2015). John Snow's map was among the first studies to utilize disease maps for understanding public health issues.

In the modern context, disease maps are commonly used to identify spatial relationships between disease outcomes, risk factors in the environment, and population characteristics (Croner et al. 1996; English et al. 1999; Gatrell et al. 2003; Glass et al. 1995; Goodchild et al. 1992). The rise of computerized mapping software and easy access to aggregated health data have enabled the production and delivery of maps via interactive websites or as applications on mobile devices like cell phones. The Centers for Disease Control and Prevention (CDC) publishes mortality and other environmental datasets for the United States via a web-based portal called CDC WONDER. Users of this website are also able to create online maps of various health outcomes. The purpose of this web portal is twofold – (1) it enables researchers and practitioners to create their own maps to aid research and/or public health intervention and (2) it allows the public to produce maps for their own interest. Other examples of publicly available mapping portals for health data include AIDSVu, National Cancer Institute's GIS Portal, among others. While emerging GIS technology has led to the democratization of mapping, thus enabling better public participation in understanding the social and environmental determinants of health, it may also lead to misleading or biased perceptions when maps are not interpreted or used correctly. The use of choropleth maps as the default map type for representing disease burdens is of particular concern for various reasons discussed in the section titled "Methods: Linking Maps of Disease Outcomes to Environmental Risks" (see Fig. 5 for an example). Maps that represent unreliable information may lead to biased and/or incorrect perceptions about the complex relationships between environmental risks, disease burdens, and population characteristics. In addition to careful selection of map type, it is imperative that mapmakers communicate information about the intended purpose of the map, the process by which it was produced, and other information considered important for interpreting the observed patterns.

Data

The synthetic data generation process consists of four stages as described in Fig. 2. In stage 1, block-group-level population data for Denton County in Texas was used to create a point distribution representing individuals. Population values were divided by 10 for computational efficiency. The resulting dataset consists of 66,092 points where each point represents an individual. Note that the spatial distribution of these points is proportional to the block-group population distribution in Denton County. In stage 2, four sites of simulated environmental risk were selected in Denton County. These sites were selected to cover urban and rural contexts. We assumed a 1-mile radius of "exposure" around each of these four sites. We will refer to these buffers as "high-risk" areas. In stage 3, case data were created using two levels of simulated disease risk: (1) a 1% risk of disease among the population overall and (2) a 5% risk of disease among populations within high-risk areas. Finally, in stage 4, the simulated datasets were converted into GIS layers for use in subsequent analysis. The final synthetic dataset consisted of 89 simulated cases and 1508 individuals in the high-risk area (rate = 0.0097) compared to 625 simulated cases and 63,870 individuals overall (rate = 0.059). Block-group-level population data were obtained from the US Census. Alteryx software was used to create the synthetic datasets.

Methods: Linking Maps of Disease Outcomes to Environmental Risks

A dot density map is the simplest way to represent disease patterns over space. Such maps are typically produced by randomly placing dots or other point symbols within the spatial extent of each geographic unit such that the total number of dots within that unit is equal or proportional to the observed number of disease cases. When producing such maps, the mapmaker chooses a numerical value that each dot represents on the map; for example, the mapmaker may decide that one dot represents five disease cases. Areas containing many dots indicate high concentrations of disease cases, whereas areas with fewer dots represent lower concentrations. As illustrated in Fig. 3, the dot value along with dot size can result in maps with vastly different presentations of disease concentration and spread. Larger spatial units such as the census tracts located in the northern and western parts of the county are more likely to be distorted as the dots are not placed in accordance with population density – instead they are randomly disbursed across the entire spatial extent of each tract. Further, such maps do not take population into consideration and are generally inadequate for measuring the intensity of a disease within a population.

Choropleth maps are a commonly used alternative (Diggle 2000). They are constructed by grouping areas (typically representing administrative units) into categories and are assigned a color based on the value of the variable being mapped. Choropleth maps are commonly used for many reasons – they are easy to produce and interpret; they rely on existing spatial units that typically represent administrative boundaries for which other demographic and secondary data are collected; and the process of aggregating data to some administrative unit often addresses privacy and confidentiality concerns. The process of constructing a choropleth maps typically requires the following three major decisions:

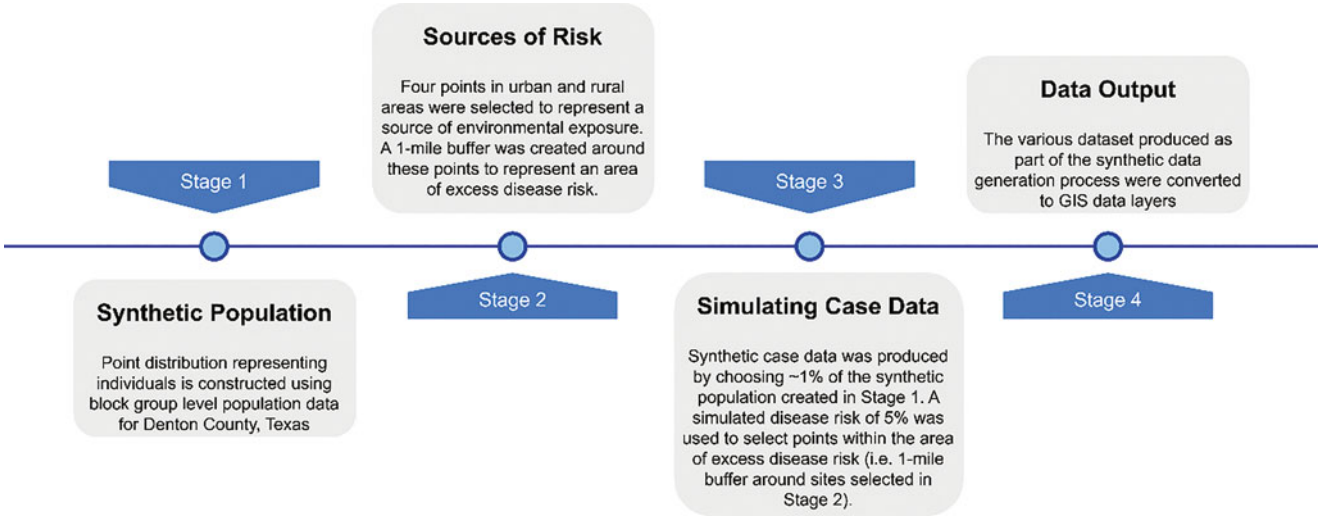


Fig. 2 Synthetic data generation process

1. Choice of Map Unit

   Choropleth maps are based on an existing system of boundaries which form the basic spatial units using which the map is produced. These units represent the geography at which data are collected and/or made available for mapping. In the United States, census entities such as tracts, block groups, or zipcode tabulation areas (ZCTAs) form the basis of many choropleth maps. The choice of map unit influences the patterns that are observed on the map and present several challenges that are discussed in detail below. Figure 4 shows various census entities that are commonly used in the United States. Census tracts are represented by the dark black borders. Census block groups represent finer spatial units and are represented by the yellow lines. Note that census block groups are perfectly contained within census tracts. Zipcode tabulation areas or ZCTAs are a census unit that approximates area representations of zipcode service areas that are created by the US Postal Service for the purposes of mail delivery. Zipcodes are a dynamic entity that do not conform to traditional census statistical data units such as block groups or tracts. This presents a problem **wherein** demographic and/or socioeconomic data collected by the census cannot be linked to zipcodes, which are commonly used descriptions of residential addresses. Although ZCTAs provide a mechanism to link census data to residential zipcodes, it is important to note that they are approximations of zipcodes. The error between the "true zipcode boundary" and ZCTAs is not consistent over space with some areas presenting a greater magnitude of misclassification compared to others. On a related note, one must also be careful when comparing choropleth maps constructed from different spatial units as the underlying geography supporting the statistic being visualized may be different across maps.

2. Choice of Classification Method

   The process of classification takes a large number of observations and groups them into categories or classes. Creating maps from fewer, well-defined classes makes them easier to read and understand when compared to a map produced from raw data values. The mapmaker typically specifies the number of classes and classification method. Generally, a map must not have more than seven classes. Although more classes result in less data generalization, they may clutter the map with too much detail, thus rendering it ineffective. Commonly employed classification methods include equal intervals, quantiles, and natural breaks. The equal interval method divides the data into equal-sized classes (Fig. 5a). It works best when data values are spread across the entire range. This method must not be applied on a skewed dataset as it may result in a washed-out map where one color (class) dominates. The quantile method places an equal number of observations within each class (Fig. 5b). This method generally results in attractive maps as every color (class) has approximately equal representation. A drawback of this method is that it may result in classes that have varying numerical ranges. The natural breaks method examines the data to identify natural groupings of data that aim to group similar values while maximizing difference between classes (Fig. 5c). The Jenks Natural Breaks algorithm (Jenks 1963) is used in most common GIS software.

3. Choice of Color and Map Context

   To produce an effective map, the mapmaker must think about the aesthetic qualities of the final map. Considerations include choice of map colors, inclusion of map elements such as a north arrow and scalebar, use of data layers to provide context, labeling styles, among others. Qualitative data are represented using differences

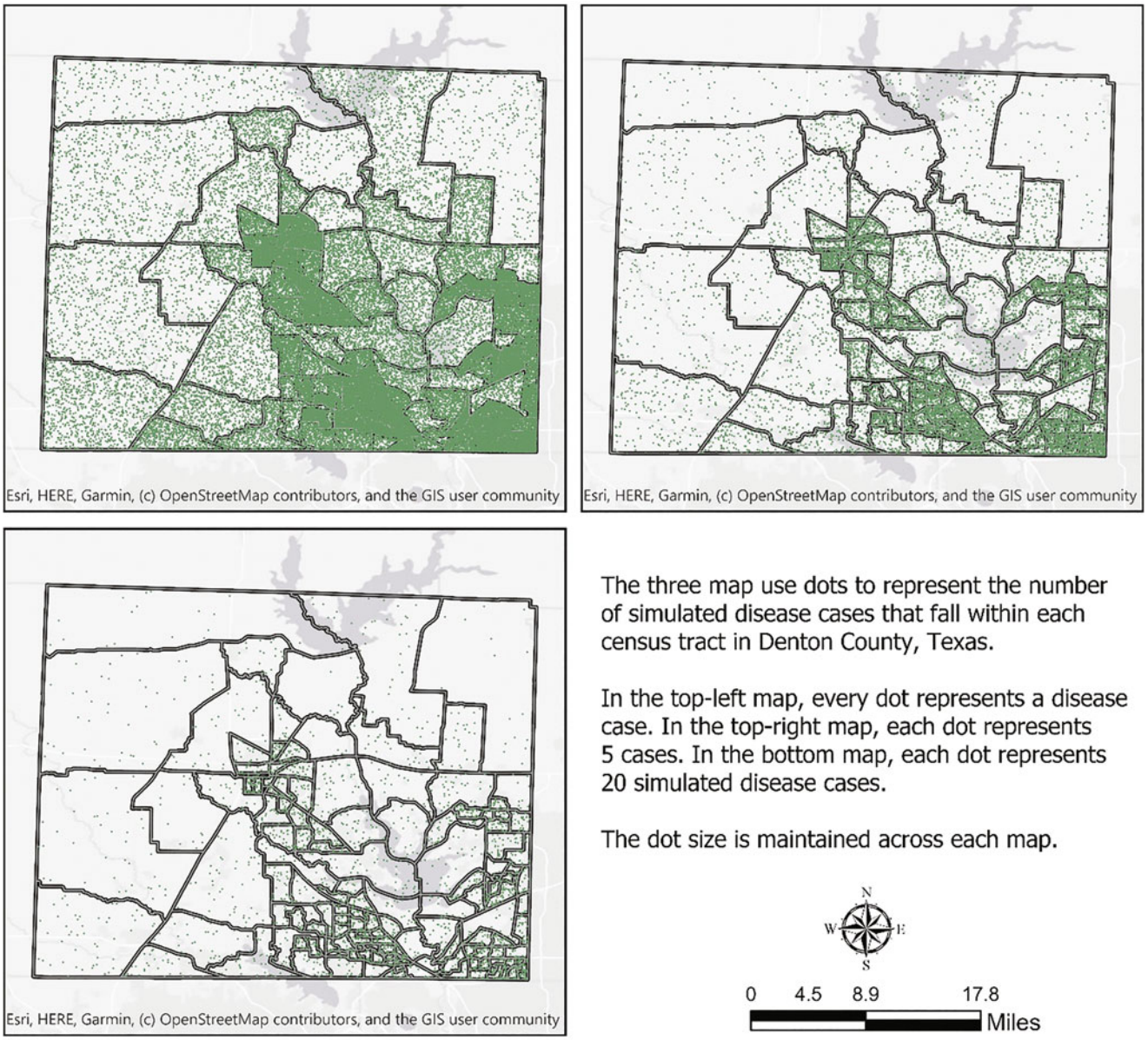


Fig. 3 Using dot density maps to display disease outcomes

in hue, while quantitative data that contain a progression of low to high values are represented by varying the levels of saturation or lightness of a particular color. Brewer et al. (1997) provide several guidelines on selecting color schemes for mortality maps. The colorbrewer2.org website is an excellent resource for mapmakers looking for recommendations of color schemes based on the type of data and map use (Brewer 2003; Harrower and Brewer 2003). Most GIS software allow mapmakers to selectively include various map elements such as north arrows, neatlines, and scalebars. GIS software including ArcGIS and QGIS typically include various options for each map element, thus allowing for high levels of customization in the production of the final map. Secondary data layers such as road networks, satellite imagery, or topographic maps can be used to provide background or contextual information that can aid the map reader. Examples of how such data can be used in disease maps can be found in Beyer et al. (2012).

While choropleth maps are easy to produce and interpret, they also present several problems, particularly for portraying rates of disease in a population. Such maps are subject to the modifiable areal unit problem (MAUP) which states that any change in the scale (level of aggregation) or shape of map units (such as administrative boundaries) will result in

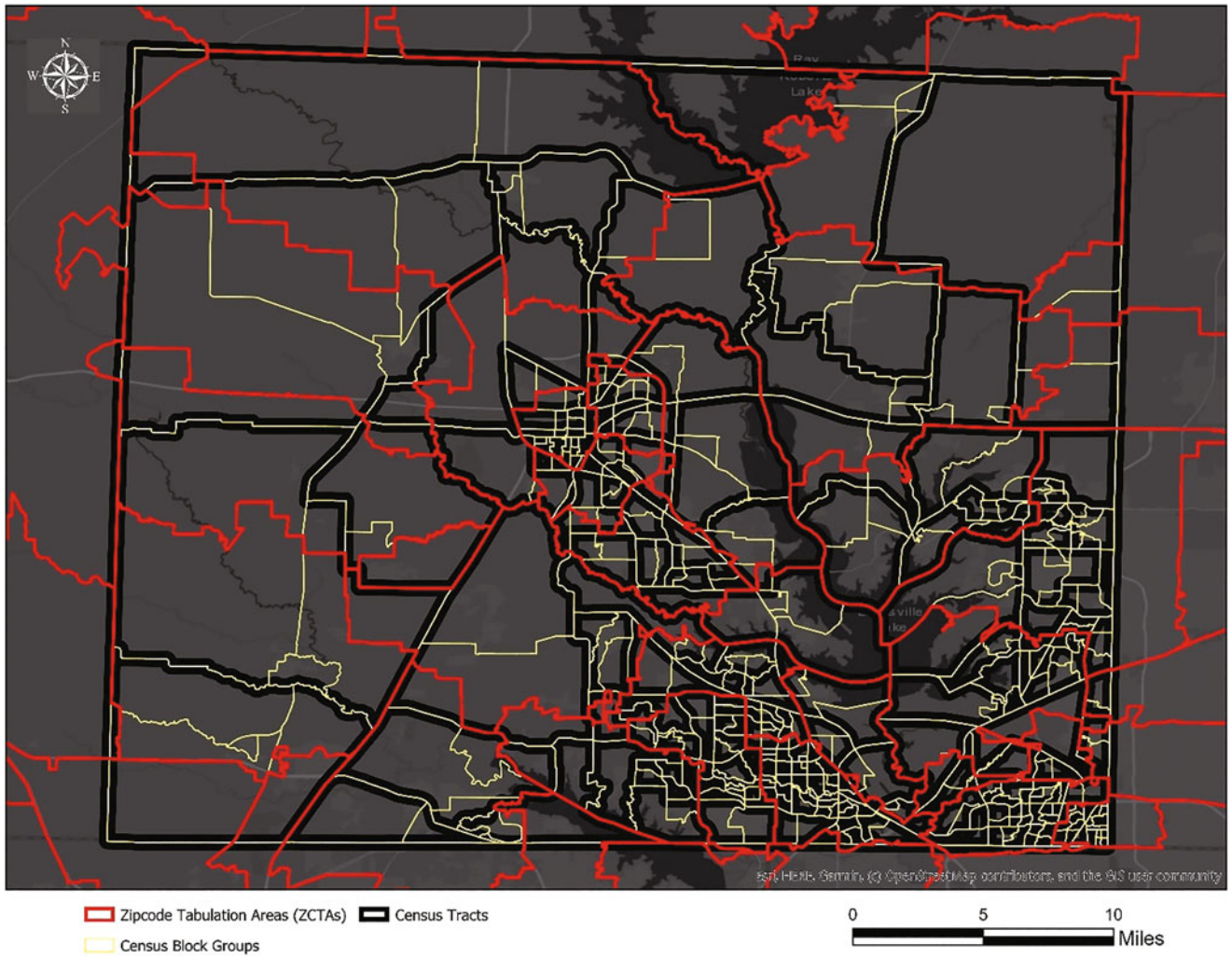


Fig. 4 Commonly used census boundaries in the United States

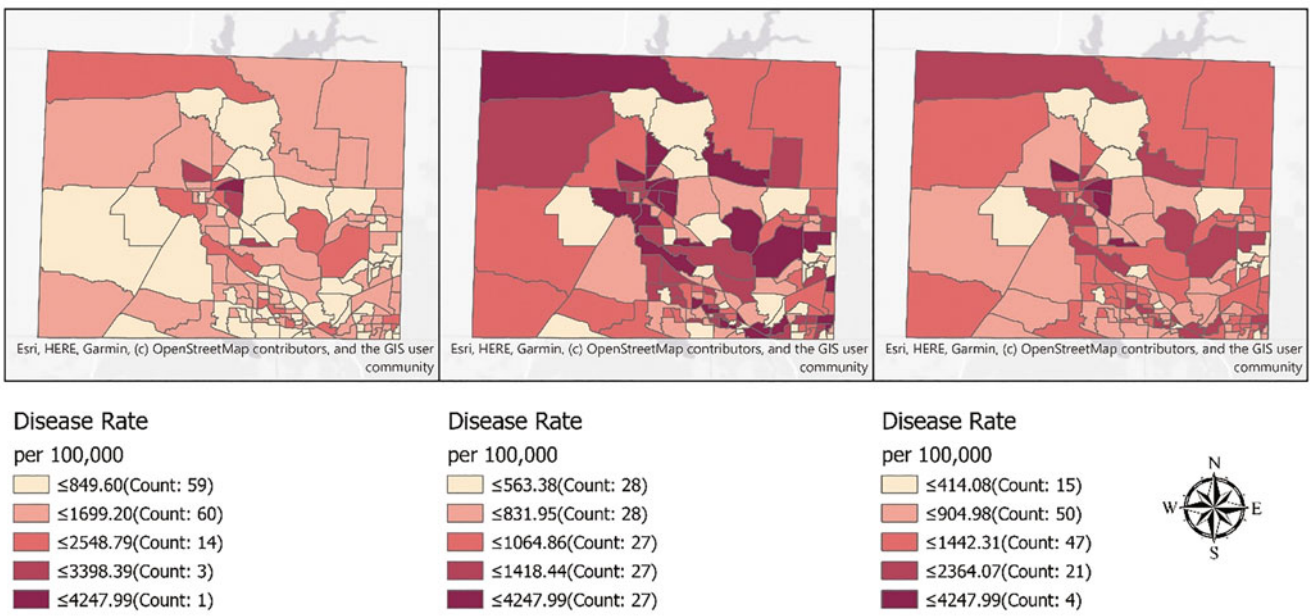


Fig. 5 Commonly used map classification methods

changing map patterns. Simply put, any change in the shape or size of the unit being mapped will result in maps with different spatial patterns of disease burdens. Cressie (1993) showed that administrative boundaries tend to change based on socioeconomic, demographic, and environmental criteria for which health event data are collected and can influence the observed rates and patterns of disease distribution. Bell et al. (2006) added that health data are aggregated on predefined spatial scales and any change in boundary does not represent true numerical information about the region. In other words, the aggregation of data into arbitrary administrative units can lead to loss of information about how diseases are distributed within those units themselves. Further, choropleth maps of disease rates are subject to statistical variability due to small numbers problem. In other words, areas with sparse population counts are likely to yield estimates of disease rates that are highly unstable and may dramatically change with the addition or deletion of a few cases.

Methods to address the small numbers problem aim to increase the population basis of support by aggregating data over space and/or time to create collections of larger, contiguous spatial units known as *spatial supports (*Beyer et al. 2012*;* Hansen 1991*;* Mungiole et al. 1999*;* Rushton et al. 2000*)*. Other methods rely on the use of geostatistical modeling approaches (Berke 2005; Goovaerts 2005; Goovaerts 2006) or other types of statistical techniques (Clayton and Kaldor 1987; Devine et al. 1994; Lawson et al. 2000; Marshall 1991; Mollie and Richardson 1991). A third category of disease maps represents disease risk as a continuous function over geographical space. Kernel density estimation methods are commonly used to produce such maps (Talbot et al. 2000; Tiwari and Rushton 2005). Maps produced using these methods use a kernel or spatial filter characterized by a particular shape, size, and density function (Carlos et al. 2010; Shi 2010) to compute the intensity of a disease along a set of sampling locations overlaid across the study area. Disease rates are computed at each sample location by dividing the number of cases that fall within a kernel placed at that point by the population contained within it. The size of the kernel is determined using one of two strategies: (1) a fixed size is used at each sample point, thus ensuring consistent spatial support but variable population support and (2) kernel sizes expand or contract to meet a minimum population threshold, thereby ensuring consistent population support but variable spatial support (Talbot et al. 2000; Tiwari and Rushton 2005; Tiwari 2013). Variable sized kernels or adaptive spatial filters are preferred over fixed-size filters as they address problems of undersmoothing or oversmoothing. Undersmoothing results when the kernel size is not large enough and continues to compute disease rates using sparse population counts. This may occur in rural areas where population densities tend to be low. Oversmoothing occurs when the kernel size is larger than what would be needed to compute a stable disease rate. Oversmoothing occurs in densely populated urban areas and results in loss of resolution on a map. Variable sized kernels contract and expand in size such that each kernel contains some minimum, user-defined population threshold. Resulting maps provide consistent levels of statistical reliability across all areas and high levels of geographic detail in areas where such detail is expected (e.g., urban contexts). In the work discussed in this chapter, we used the Web-Based Disease Mapping and Analysis Program (WebDMAP) to produce such maps. Following are the three major steps involved.

1. Create Data Files

   WebDMAP requires three data files to compute disease rates using the kernel density estimation method. The grid file provides point locations on which kernels or spatial filters will be constructed. The other two files provide the locations of disease cases and populations, respectively. If individual-level data are available, each location represents an individual. Alternatively, each location can also represent aggregated counts of case/population data for some spatial unit such as a census block group or ZCTA. Location data must be provided in unprojected coordinates (i.e. latitude and longitude). Simulated disease and population data used in this chapter can be downloaded from http://webdmap.com/kdedata.
2. Define Minimum Population Threshold

   Recall that the size of the kernel/spatial filter that is placed at each grid point is determined by some user-defined minimum population size value. Note that the size of the spatial filters is determined by this user-specified parameter. Large population thresholds in areas with sparse populations will result in the largest filter sizes. Conversely, small population thresholds in areas with dense populations will result in the smallest filter sizes. In the work discussed in this chapter, we used a population threshold of 1000 individuals. The study area, Denton County, comprises dense urban areas (central and south-eastern portions) as well as sparsely populated rural areas (northwestern portions). Correspondingly, we see a combination of small and large filter sizes across the study region (Fig. 6).
3. Compute Rates and Produce Maps

   The algorithm for computing rates using this method is described below:

   (a). Compute distance strings for the case and population data. Distance strings are a kind of data structure that were originally designed for efficiently storing information about travel costs between nodes

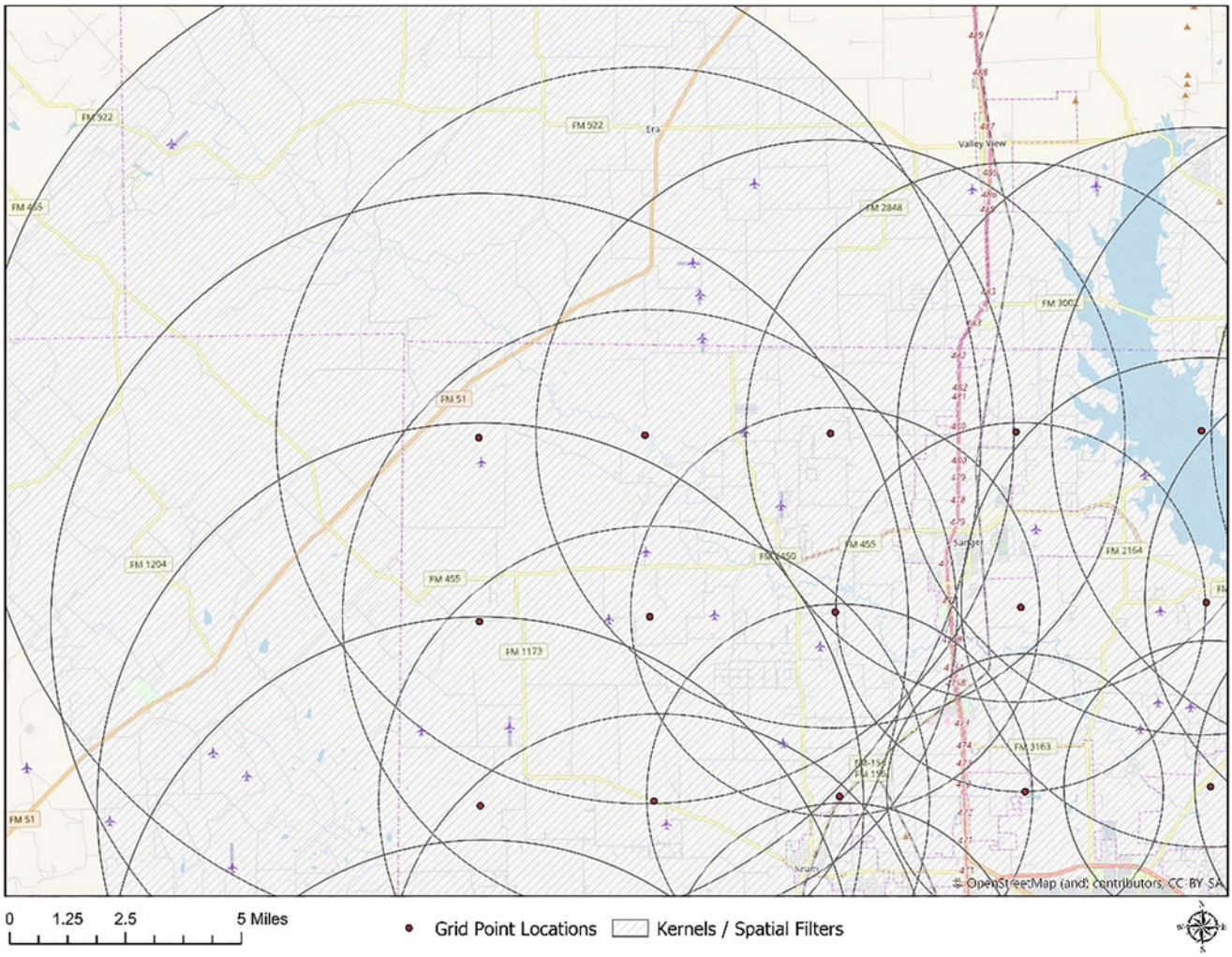


Fig. 6 Grid points and spatial filters

and were used frequently in location allocation algorithms (Densham and Rushton 1992; Hillsman 1984; Sorensen and Church 1995). The basic idea behind a distance string is that it stores information about travel costs (i.e., distance) between a base node (i.e., every grid point) and all other nodes (i.e., cases and population locations) in an increasing order of distance. The procedure is implemented in a PostgreSQL database, thus enabling the calculation of distance strings for large datasets. To improve computational efficiency, distance strings are truncated at a user-defined cutoff value. By doing so, we assume that spatial filters will never be larger than a certain size and therefore terminate distance string calculations at the user-defined cutoff value. In this analysis, the distance string cutoff was set to 100 miles – i.e., we assumed that spatial filters will never exceed a 100-mile radius. To further improve computational efficiency, spatial indexes were created for all data tables, thereby resulting in substantially faster database query processing times. See Nguyen (2009) for details on how spatial indexes work within the PostgreSQL/PostGIS relational database. Distance strings are computed for the case and population data.

(b). For each grid point, use the population distance strings table to identify the distance associated with the user-defined population threshold value. This is implemented using database functions that query the population distance strings table to identify the distance value that corresponds with the row where the cumulative population weight exceeds the user-defined population threshold. This is the size of the spatial filter. For example, in Fig. 7, if the user-defined population threshold is set to 200, the algorithm will

gridpoint	populationlocation	distance	population	cumulativepopulation
1	a	1.1 miles	30	30
1	b	1.4 miles	100	130
1	e	1.9 miles	75	205
1	g	3.4 miles	20	225
...		...		
2	t	4.1 miles	10	10
2	f	4.8 miles	5	15
2	b	7.4 miles	15	30
2	a	12.3 miles	25	55
...		...		

Fig. 7 Population distance strings example

gridpoint	caselocation	distance	cases	cumulativecases
1	a	1.1 miles	5	5
1	b	1.4 miles	2	7
1	e	1.9 miles	7	14
1	g	3.4 miles	3	17
...		...		
2	t	4.1 miles	4	4
2	f	4.8 miles	1	5
2	b	7.4 miles	5	10
2	a	12.3 miles	8	18
...		...		

Fig. 8 Case distance strings example

select the record highlighted in orange to define the size of the spatial filter (i.e., 1.9 miles). Note that the actual population contained within this spatial filter is 205. This may occur when aggregated data are used. This process is repeated for every grid point.

(c). For each grid point, query the case distance strings table and note the cumulative weight value that is associated with the distance noted in step b above. This is the number of cases that fall within the spatial filter size at that grid point. For example, if the distance value at grid point 1 is 1.9 miles (step b), then the number of disease cases that are contained within the spatial filter constructed at that grid point is 14.

(d). Compute a rate at every grid point by dividing the cumulative number of cases (step c) (Fig. 8) by the cumulative population (step b) (Fig. 7).

(e). Repeat steps b through d to compute a rate for all the grid points.

(f). A continuous surface map can be created from the grid points using the inverse distance weighted (IDW) interpolation method in any standard GIS software. The IDW method with 8 neighbors and a power of at least 2 is recommended to avoid any "double" smoothing that may occur in addition to what has already been performed by the spatially adaptive filter method.

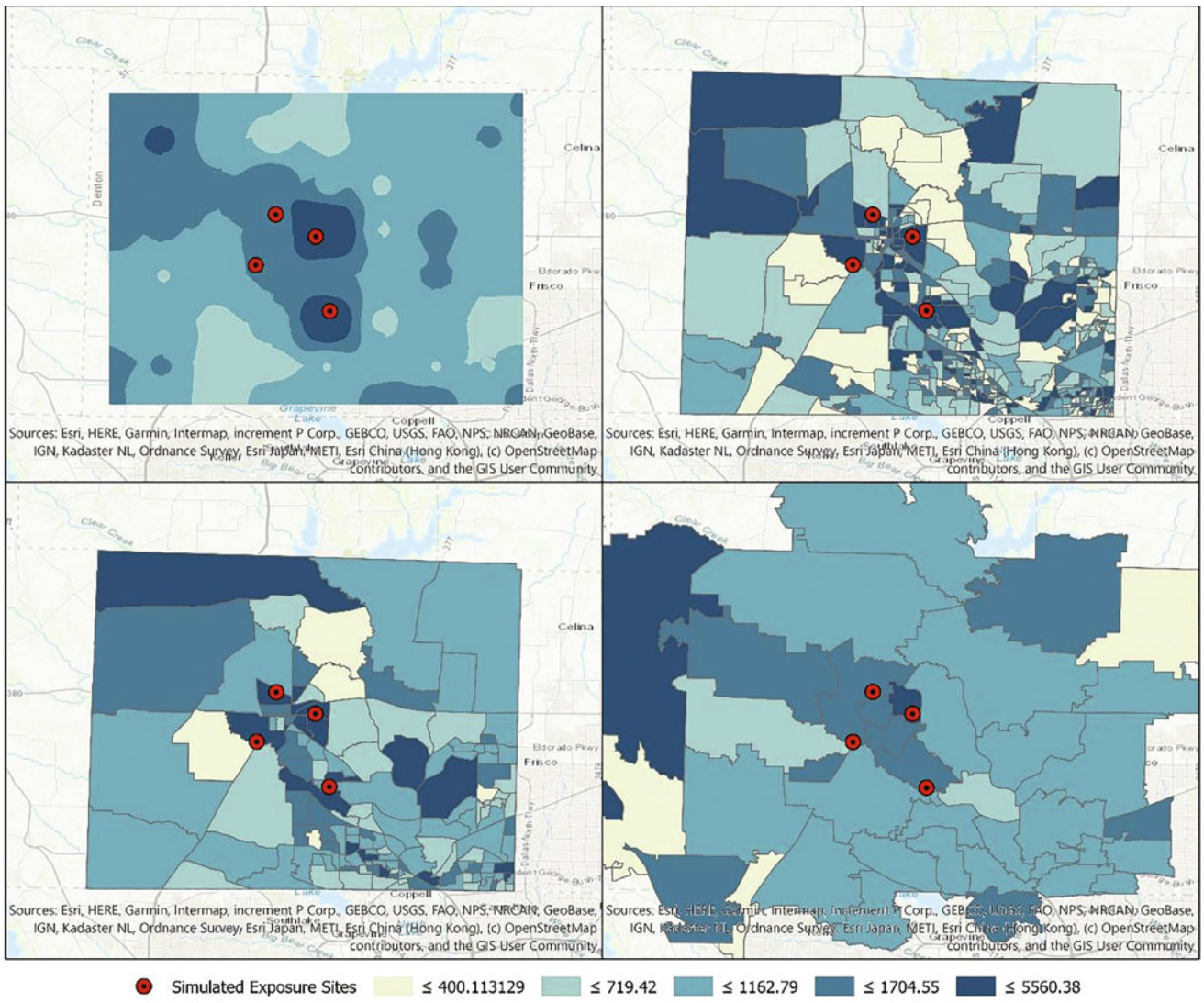


Fig. 9 Disease maps produced by us

The maps shown in Fig. 9 were constructed using kernel density estimation (Fig. 9a) and as choropleth maps (Fig. 9b–d). The four red dots indicate the sites of some environmental exposure. These sites were randomly selected to include urban and rural contexts. As described earlier, areas within a 1-mile buffer around each of the four points represent an area of elevated disease risk (five times the overall rate). To construct the map in Fig. 9a, a grid of points, placed 4 miles apart, was overlaid on top of the study area (Denton County, Texas). At each point of this grid, variable sized kernels were constructed such that each kernel or spatial filter contained exactly 1000 individuals. If aggregated data were used instead of individual-level point data, each kernel would contain a collection of spatial units with a minimum population size of 1000. Kernel size (radius) ranged from a minimum of 0.72 miles to a maximum of 10.834 miles. The average kernel size was 4.163 miles. The number of cases falling within each kernel were assigned to each grid point. These ranged from a minimum of 4 cases to a maximum of 33 with an average of 10.736 cases. Rates were computed at each grid point by dividing the case count by population. Disease rates computed at each grid point ranged from a minimum value of 400 per 100,000 to a maximum value of 3300 per 100,000. The average rate was 1073 cases per 100,000 population. Rate values at each grid point were converted into a continuous surface of disease risk using the inverse distance weighted (IDW) interpolation technique. Disease rates were computed using the Web-Based Disease Analysis and Mapping Program (Web-DMAP). Final map output was created using ArcGIS Pro.

The maps in Fig. 9b–d were created by aggregating case and population counts to three sets of administrative boundaries with varying levels of spatial resolution. The maps used census block groups, census tracts, and zip code tabulation areas (ZCTAs), respectively. For each spatial unit (i.e., each block group, tract, or ZCTA), a rate was calculated by dividing the total number of cases within that unit by its population. For all four maps, rates were classified into five groups using the quantile classification method.

The spatial patterns of disease rates in each of the four maps present slight variations when compared to each other. Note that the underlying data used in each map are identical. Differences in observed patterns are a result of the different levels of aggregation and the method used to construct the map. Among the four maps, Fig. 9b presents the most geographic detail. However, due to the relatively small size of census block groups, they also contain the most variability in populations ranging from a minimum of 6 persons within a block group to a maximum of 646. Due to the variable population sizes, block groups also portrayed the most variability in disease rates with an average rate and standard deviation of 1045.67 and 939.26 per 100,000 population, respectively. The highest rates of disease were observed around the four exposure sites along with pockets in the northwest and eastern sections of the county. In contrast, the map in Fig. 9d contains the least amount of geographic detail. The use of large ZCTA boundaries tends to wash away any fine-scale variations in disease rates. Only one of the four exposure sites is located in an area of highest disease risk. Pockets of high rates are observed in the northwestern parts of the county. While the level of geographic detail presented in this map is low, it also contains the most stable estimates of disease rates. The average rate in ZCTAs in Denton County was found to be 1066.53 cases per 100,000 population with a standard deviation of 574.17. The map in Fig. 9c represents a balance between the maps presented in 9b and 9d. It uses census tracts, which are slightly larger in size (and population) compared to block groups and considerably smaller in size compared to ZCTAs. Areas surrounding the four exposure sites are classified as areas of highest disease risk in addition to pockets of high rates in the northwestern and eastern parts of the county. The average disease rate is estimated at 1040.8 per 100,000 population along with a standard deviation of 642.97. Finally, the map in Fig. 9a identifies areas of highest disease risk surrounding two of the four exposure sites. Unlike a choropleth map, this map does not use discrete spatial units to represent rates across Denton County. Instead, a continuous surface of disease risk is used to identify areas of highest and lowest rates. As discussed earlier in this chapter, the average rate is found to be 1073 cases per 100,000 population with a standard deviation of 462.99. Among the four maps, the one produced using kernel density estimation presents the best balance between resolution and reliability. The ability to control the population basis of support ensures that consistent sample sizes are used in the calculation of every disease rate across the map. This map presents a desirable tradeoff between geographic resolution and reliability – maintaining high levels of geographic detail in urban areas while preserving high levels of statistical reliability in rural areas. While the map produced using census block groups presents high levels of visual detail, such maps must be used with caution due to the problem of unstable rates caused by small population counts. Conversely, a map that uses coarse spatial units such as ZCTAs not only maintains statistical reliability but also leads to severe loss in geographic detail across the entire map.

The choice of disease map and/or the spatial resolution at which disease data are available influence the ability to detect associations between disease burdens and environmental exposures. The maps in Fig. 10a–d represent the spatial patterns of population exposure to four point locations of simulated environmental risk. These four locations are denoted by red dots in Figs. 9a–d and 10a–d. Exposure is measured as the Euclidean distance between each individual, represented as a point in the synthetic dataset, and the closest point location representing a site of environmental risk. Map 10a was produced by interpolating distances computed for each individual point in Denton County. Lighter colors represent closer distances compared to darker colors. As expected, areas close to the four red dots on the map show lower distance values. The maps in Fig. 10b–d represent the average exposure distance for populations aggregated to census block groups, tracts, and ZCTAs, respectively. As expected, areas within close proximity of the four red dots portray lower exposure distances. Note that the block-group-level map shows better geographic resolution when compared to the other maps. This is generally a desirable property in maps of environmental exposure when compared to disease maps, where high levels of geographic detail typically represent poor statistical reliability. However, it is critical to note that valid map comparisons can only be made when the underlying geographic or spatial basis of support is consistent across all maps that are being compared. Inconsistent spatial supports can result from differences in resolution, scale, or boundary definitions. For example, one cannot directly compare a disease map constructed using the KDE method with a block-group-level map of environmental exposure. Dasymetric mapping (REF) or other geostatistical modeling techniques including interpolation (REF PYCNO) may be used to reconcile maps that do not have consistent spatial supports.

The scatter plots in Fig. 11a–d show the directionality and strength of the relationships between exposures and disease outcomes. Figs. 11a, b represent the relationship between disease and exposure data measured at fine geographic scales (individual- and block-group levels), whereas plots in Fig.

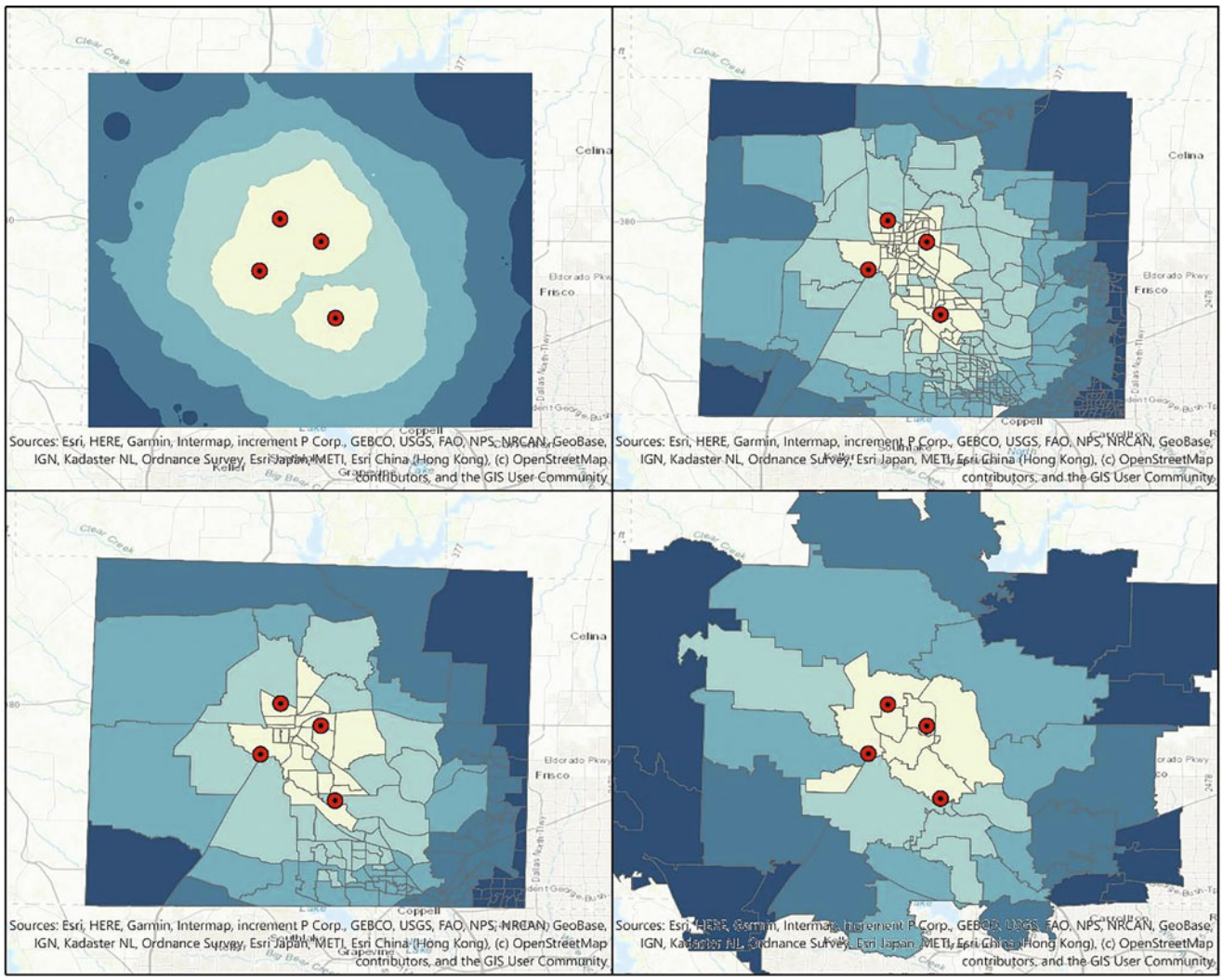


Fig. 10 Exposure maps

11c, d represent this relationship at coarser geographic scales (tract and ZCTA levels, respectively). It is interesting to note that scatter plots produced using data at finer geographic scales correctly describe a negative relationship between distance and disease intensity. Note that the synthetic dataset was produced using a five times greater risk of disease in individuals within 1 mile of a simulated site of environmental exposure. Conversely, the scatter plots produced using coarser data represent an inverse relationship, suggesting that disease risk increases as one moves away from these sites of environmental concern. The synthetic data produced do not support this conclusion. It is also important to note that, although the relationship between distance and disease rate is correctly represented in Fig. 11b, the variability in disease rates as indicated by the boxplot beside the y-axis is likely to bias the strength of the relationships between distance and disease rates.

Conclusions

The type of mapping method used to produce maps of disease outcomes or environmental risks as well as their parameters influences the observed patterns of disease distribution and consequently our interpretations of associated risk factors. It is important to remember that a map merely represents one abstraction of complex underlying processes that control how diseases and environmental risks manifest themselves across space and time. The construction of an "honest map" requires full disclosure of the methods used, scale of analysis, quality of data, and other parameters used in the final construction of the map. The objective of this chapter is not to identify the "best" mapping method but to demonstrate that each method comes with advantages and disadvantages and, importantly, have an impact on the patterns and relationships that are observed.

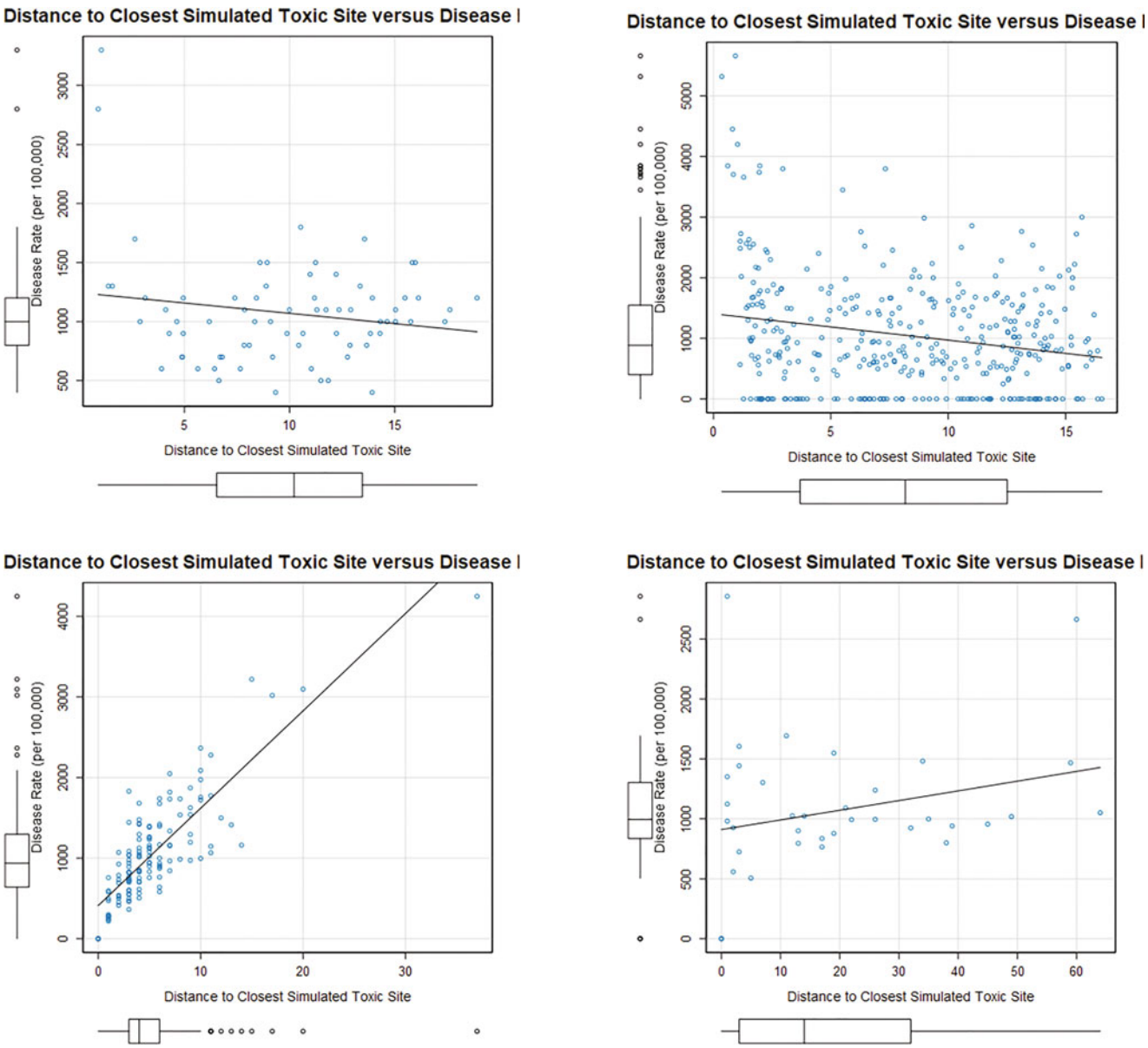


Fig. 11 Relationships between exposures and disease outcomes

References

Al-Hamdan, M.Z., W.L. Crosson, S.A. Economou, M.G. Estes Jr., S.M. Estes, S.N. Hemmings, et al. 2014. Environmental public health applications using remotely sensed data. *Geocarto International* 29 (1): 85–98.

Al-Hamdan, M.Z., W.L. Crosson, A.S. Limaye, D.L. Rickman, D.A. Quattrochi, M.G. Estes Jr., et al. 2009. Methods for characterizing fine particulate matter using ground observations and remotely sensed data: Potential use for environmental public health surveillance. *Journal of the Air & Waste Management Association* 59 (7): 865–881.

Bell, B.S., R.E. Hoskins, L.W. Pickle, and D. Wartenberg. 2006. Current practices in spatial analysis of cancer data: Mapping health statistics to inform policymakers and the public. *International Journal of Health Geographics* 5 (1): 49.

Bellander, T., N. Berglind, P. Gustavsson, T. Jonson, F. Nyberg, G. Pershagen, and L. Järup. 2001. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environmental Health Perspectives* 109 (6): 633–639.

Berke, O. 2005. Exploratory spatial relative risk mapping. *Preventive Veterinary Medicine* 71 (3–4): 173–182.

Bertollini, R., and M. Martuzzi. 1999. Disease mapping and public health decision-making: Report of a WHO meeting. *American Journal of Public Health* 89 (5): 780.

Beyer, K.M., C. Tiwari, and G. Rushton. 2012. Five essential properties of disease maps. *Annals of the Association of American Geographers* 102 (5): 1067–1075.

Brewer, C.A. 2003. A transition in improving maps: The ColorBrewer example. *Cartography and Geographic Information Science* 30 (2): 159–162.

Brewer, C.A., A.M. MacEachren, L.W. Pickle, and D. Herrmann. 1997. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers* 87 (3): 411–438.

Carlos, H.A., X. Shi, J. Sargent, S. Tanski, and E.M. Berke. 2010. Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics* 9 (1): 39.

Clayton, D., & Kaldor, J. 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*: 671–681.

Cressie, N.A. 1993. *Statistics for spatial data*. New York: John Wiley and Sons. Inc.

Croner, C.M., J. Sperling, and F.R. Broome. 1996. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15 (18): 1961–1977.

Densham, P.J., and G. Rushton. 1992. Strategies for solving large location-allocation problems by heuristic methods. *Environment and Planning A* 24 (2): 289–304.

Devine, O.J., Louis, T.A., & Halloran, M.E. 1994. Empirical bayes methods for stabilizing incidence rates before mapping. *Epidemiology*: 622–630.

Diggle, P.J. 2000. Overview of statistical methods for disease mapping and its relationship to cluster detection. *Spatial Epidemiology: Methods and Applications* 87: 103.

Elliott, P., D. Briggs, S. Morris, C. de Hoogh, C. Hurt, T.K. Jensen, et al. 2001. Risk of adverse birth outcomes in populations living near landfill sites. *BMJ* 323 (7309): 363–368.

English, P., R. Neutra, R. Scalf, M. Sullivan, L. Waller, and L. Zhu. 1999. Examining associations between childhood asthma and traffic flow using a geographic information system. *Environmental Health Perspectives* 107 (9): 761–767.

Fisher, R., T. Walshe, P. Bessell-Browne, and R. Jones. 2018. Accounting for environmental uncertainty in the management of dredging impacts using probabilistic dose–response relationships and thresholds. *Journal of Applied Ecology* 55 (1): 415–425.

Gatrell, A.C., J.C. Harman, B.J. Francis, C. Thomas, S.M. Morris, and M. McIllmurray. 2003. Place of death: Analysis of cancer deaths in part of north West England. *Journal of Public Health* 25 (1): 53–58.

Glass, G.E., B.S. Schwartz, J.M. Morgan III, D.T. Johnson, P.M. Noy, and E. Israel. 1995. Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health* 85 (7): 944–948.

Goodchild, M., R. Haining, and S. Wise. 1992. Integrating GIS and spatial data analysis: Problems and possibilities. *International Journal of Geographical Information Systems* 6 (5): 407–423.

Goovaerts, P. 2005. Geostatistical analysis of disease data: Estimation of cancer mortality risk from empirical frequencies using poisson kriging. *International Journal of Health Geographics* 4 (1): 31.

———. 2006. Geostatistical analysis of disease data: Accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. *International Journal of Health Geographics* 5 (1): 52.

Griffith, D.A. 2018. Uncertainty and context in geography and GI-Science: Reflections on spatial autocorrelation, spatial sampling, and health data. *Annals of the American Association of Geographers* 108 (6): 1499–1505.

Hansen, K.M. 1991. Head-banging: Robust smoothing in the plane. *IEEE Transactions on Geoscience and Remote Sensing* 29 (3): 369–378.

Harrower, M., and C.A. Brewer. 2003. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40 (1): 27–37.

Hillsman, E.L. 1984. The p-median structure as a unified linear model for location—Allocation analysis. *Environment and Planning A* 16 (3): 305–318.

Jenks, G.F. 1963. Class intervals for statistical maps. *International Yearbook Cartography* 3: 119–134.

Koch, T. 2004. The map as intent: variations on the theme of John Snow. *Cartographica: The International Journal for Geographic Information and Geovisualization* 39(4): 1–14.

Kwan, M. 2012. The uncertain geographic context problem. *Annals of the Association of American Geographers* 102 (5): 958–968.

———. 2018. The limits of the neighborhood effect: Contextual uncertainties in geographic, environmental health, and social science research. *Annals of the American Association of Geographers* 108 (6): 1482–1490.

Lawson, A.B., A.B. Biggeri, D. Böhning, E. Lesaffre, J.F. Viel, A. Clark, et al. 2000. Disease mapping models: An empirical evaluation. Disease mapping collaborative group. *Statistics in Medicine* 19 (17): 2217–2241.

Leelasakultum, K., and N.T. Kim Oanh. 2017. Mapping exposure to particulate pollution during severe haze episode using improved MODIS AOT-PM10 regression model with synoptic meteorology classification. *GeoHealth* 1 (4): 165–179.

Marshall, R.J. 1991. Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 40 (2): 283–294.

McLeod, K.S. 2000. Our sense of snow: The myth of john snow in medical geography. *Social Science & Medicine* 50 (7–8): 923–935.

Mennis, J., and E.E. Yoo. 2018. Geographic information science and the analysis of place and health. *Transactions in GIS* 22 (3): 842–854.

Mollie, A., and S. Richardson. 1991. Empirical bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine* 10 (1): 95–112.

Moore, D.A., and T.E. Carpenter. 1999. Spatial analytical methods and geographic information systems: Use in health research and epidemiology. *Epidemiologic Reviews* 21 (2): 143–161.

Mungiole, M., L.W. Pickle, and K.H. Simonson. 1999. Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine* 18 (23): 3201–3209.

Nguyen, T.T. 2009. Indexing PostGIS databases and spatial query performance evaluations. *International Journal of Geoinformatics* 5 (3): 1.

Nuckols, J.R., M.H. Ward, and L. Jarup. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* 112 (9): 1007–1015.

Nyberg, F., Gustavsson, P., Järup, L., Bellander, T., Berglind, N., Jakobsson, R., & Pershagen, G. 2000. Urban air pollution and lung cancer in Stockholm. *Epidemiology*: 487–495.

Reif, J.S., J.B. Burch, J.R. Nuckols, L. Metzger, D. Ellington, and W.K. Anger. 2003. Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply. *Environmental Research* 93 (3): 248–258.

Ricketts, T.C. 2003. Geographic information systems and public health. *Annual Review of Public Health* 24 (1): 1–6.

Rushton, G., G. Elmes, and R. McMaster. 2000. Considerations for improving geographic information system research in public health. *Urisa-Washington DC* 12 (2): 31–50.

Shi, X. 2010. Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science* 24 (5): 643–660.

Shi, Y., C. Ren, M. Cai, K.K. Lau, T. Lee, and W. Wong. 2019. Assessing spatial variability of extreme hot weather conditions in Hong Kong: A land use regression approach. *Environmental Research* 171: 403–415.

Shiode, N., S. Shiode, E. Rod-Thatcher, S. Rana, and P. Vinten-Johansen. 2015. The mortality rates and the space-time patterns of john snow's cholera epidemic map. *International Journal of Health Geographics* 14 (1): 21.

Sorensen, P.A., & Church, R.L. 1995. A comparison of strategies for data storage reduction in location-allocation problems (95-4).

Talbot, T.O., M. Kulldorff, S.P. Forand, and V.B. Haley. 2000. Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine* 19 (17–18): 2399–2408.

Tiwari, C. 2013. Methods for creating smoothed maps of disease burdens. In *Geographic health data: Fundamental techniques for analysis*, ed. F. Boscoe. Wallingford: CABI.

Tiwari, C., & Rushton, G. 2005. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. *Developments in spatial data handling* (pp. 665–676). Springer, Berlin, Heidelberg.

Tomlinson, C.J., L. Chapman, J.E. Thornes, and C. Baker. 2011. Remote sensing land surface temperature for meteorology and climatology: A review. *Meteorological Applications* 18 (3): 296–306.

The Influence of MATUP on Identifying Spatiotemporal Emerging Hot Clusters on Public Health Issues: Cases of Dengue Fever and Lung Cancer

Huiyu Lin and Jay Lee

Spatiotemporal analysis has become a hot topic in geographic research, which is now widely applied in the fields of public health, climate change, earthquake analysis, criminology, and many others; see for examples Nakaya and Yano (2010); Hsueh et al. (2012); Zhang et al. (2013); or Gong et al. (2019). Spatiotemporal analytic is particularly useful for exploring if public health data exhibit any spatiotemporal trends, such as how incidents of a certain disease distribute spatiotemporally and how their spatiotemporal patterns evolve. But to use the analytics appropriately, it is essential to understand those patterns and how the spreading of such diseases progresses so that suitable strategies or intervention programs can be designed and implemented.

Additional examples include many recent studies on dengue fever (DF) that had explored the spatiotemporal patterns of DF incidents based on the nature of the disease. DF is a vector-borne disease that is spread when certain conditions are met. Such conditions involve both spatial and temporal aspects. For example, locations with environments that are ideal for mosquitoes' reproduction may see elevated DF cases (Yu et al. 2011; Casas et al. 2017; Gong et al. 2019). Also, weather conditions, such as higher humidity and temperature, would contribute to more activities of mosquitoes, thereby increasing DF cases (Chien and Yu 2014; Delmelle et al. 2016). Consequently, temporal patterns of DF incidents are critical in the spread of DF cases as temperatures fluctuate across seasons and even daily. In addition, different demographics, as well as the socioeconomic levels of different studied areas, were found to be linked to the uneven spatial patterns and varying temporal trends of DF cases (Hu et al. 2012; Koyadun et al. 2012). As a result, it is crucial to explore the spatiotemporal patterns of DF further.

On the other hand, non-infectious chronic diseases such as cancers also have heterogeneity spatial and temporal patterns over space and time (Guo et al. 2016). For example, due to the increasing urbanization level which often causes elevated environmental pollution and stress to urban dwellers, especially in developing countries such as China, more and more scholars have turned their interests to analyzing the spatiotemporal pattern of lung cancer (LC)[1] and the relationships between LC cases and pollutants (Jerrett et al. 2013; Guo et al. 2016, 2017; Chen et al. 2016). However, in order to protect the privacy of cancer patients, and the data availability limitation, the spatiotemporal patterns of cancer studies were often done at a granular geographic level such as reporting results from analyzing data that are aggregated to counties or states. Such results, however, are often not precise enough for use as references when developing public health programs for localized programs.

It is interesting to note that, due to the availability of data and different research purposes, inconsistent spatial and temporal units were used in different research (Hsueh et al. 2012; Jerrett et al. 2013; Guo et al. 2016). For example, transmitted diseases which require immediate actions to block the transmission, such as dengue fever, were often investigated at a finer level such as point level due to the highly contagious nature, whereas chronic diseases such as cancers were often studied at more aggregated levels such as county and states. Nevertheless, the use of different spatiotemporal units may cause the research to yield different results due to the exis-

H. Lin
College of Environment and Planning, Henan University, Kaifeng, Henan, China
e-mail: hlin25@kent.edu

J. Lee (✉)
Department of Geography, Kent State University, Kent, OH, USA
e-mail: jlee@kent.edu

[1] Cancer data and results from analyzing the data are presented without base maps to respect patient privacy and the confidentiality of the data.

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_15

tence of an issue that we refer to as the modifiable areal and temporal units' problem (MATUP).

In this chapter, two case studies, DF cases from 2003 to 2008 in Kaohsiung City, Taiwan Province, China, and a dataset of LC cases from 2000 to 2013 in a US county, are analyzed and compared using different spatiotemporal units to present the effects of how MATUP impacts on analytical results. In addition, this chapter demonstrates the use of the *Emerging Hot Spot Analysis* tool available in ArcGIS (version 10.5 or higher, ESRI, Inc., Redlands, California). The tool was used to identify spatiotemporal hot clusters.

Modifiable Areal and Temporal Unit Problem (MATUP)

The spatiotemporal dimension of our data consists of a two-dimensional plane and a temporal axis. Necessarily, time and space are continuous. Hence, it is all artificial when it comes to dividing time or space. As a result, strong bias may be introduced to the interpretations and conclusions of the statistical results when such divisions are implemented differently.

Dividing a continuous space into smaller units introduces the modifiable areal unit problem (MAUP). MAUP was first introduced by Openshaw and Taylor (1979). When conducting spatial analysis to the same geographical area, the variation of results can be caused by either the different scale or resolution used, which is called the *scale effect*, or by the different regions divided, which is called the *zone effect* (Openshaw 1984; Wong 2009). For example, even at the same analytical scale, if the zones or boundary delimitations are different for the study area, the results may be different as well.

Time is also a continuous variable which is not discrete in nature. Meentemeyer (1989) considered that every geographical event has its own timely resolution. Harrower et al. (2000) pointed out that, when visualizing time series data, some patterns can only be observed at certain time granularity. Hence, it is essential to pick the appropriate time unit for a specific event based on its characteristics, because it will affect the result like those affected by MAUP (Gibson et al. 2000; Hornsby and Egenhofer 2002). Coltekin et al. (2011) called this phenomenon the modifiable temporal unit problem (MTUP) and proposed that researchers should pay attention to choosing the appropriate time unit when analyzing geographic processes.

Coltekin et al. (2011) mentioned three aspects of the temporal units, including the duration (how long), the temporal resolution (how often), and the point in time (when). The duration refers to the time span of the data, which is similar to the boundary that defines the enclosure of the events into a spatial unit. For time, a time span includes the start and end time of all events recorded in the dataset. Data for a short period may not form any pattern yet, or new patterns may be newly formed but have not yet been recorded into the analyzed dataset. This is similar to the boundary effects in space.

Temporal resolution is the time granularity or scale, which can be seconds, minutes, hours, days, weeks, months, or years. Different temporal resolutions would include different numbers of incidents within that period, which, in turn, would further cause different statistical results. Similar to the spatial scale effects, smaller temporal units may accumulate too few events to be considered significant, while longer temporal units may be too generalized to detect patterns.

Furthermore, the point in time of an event is the time when the event occurred. This is similar to the zone effects in MAUP. Cheng and Adepeju (2014) discussed how to divide time and how that affects the analytical result. Under the same time granularity, there are different ways to divide time. For example, a week can be divided into Monday to Sunday with Monday as the starting date. At the same time, a week can also be divided into Sunday to the next Saturday, and so on. Although both divisions of days into a week all include 7 days, the starting and ending time of the analyzed data would be different. Some events are closely related to time. For example, transportation clearly shows different temporal patterns between those during weekdays and those over weekends.

Similar to discussions on the effects of the MAUP and MTUP, the effects MATUP on analytical results would be classified into three categories, including the scale, the division of units, and the boundary of spatiotemporal events/data.

Spatiotemporal Scale Effect

The spatiotemporal scale corresponds to the spatial resolution and temporal granularity. They are essentially the artificially dividing units over the continuous space and time. There are different ways for dividing a space, such as artificially defined political boundaries, areas divided based on population or grids according to certain distances (feet, mile, kilometer), or spatial units determined by natural barriers (rivers, mountains, etc.) Time is often divided into standard units such as second, minute, hour, day, week, month, season, and year. As a result, the units of time and space are not unified. Hence, time and space are described when conducting the unit division individually, such as the geographical events happen within a year and within a 1 square kilometer radius from a certain center point in space.

The data of LC and DF were analyzed in this chapter to demonstrated how different spatiotemporal scales affect

the identification of hot cluster[2] results. Three spatial units, including 0.5 km^2, 1 km^2, and 1.5 km^2, and 5 temporal periods, including 1 week, 1 month, 3 months, 6 months, and 1 year, were used to form a total of 15 spatiotemporal units. Although it is expected that larger spatiotemporal units may be generalized too much such that detail information could be lost, the purpose of the comparative studies discussed in this chapter is to verify if the results are better if finer spatiotemporal units are used.

Spatiotemporal Zone Effect

As it is mentioned earlier, geographical space and time are continuous; all spatial and temporal divisions are therefore artificial. For example, geographical space is used to being divided into political boundaries, such as states, counties, and cities, or census units. As a result, the same area can be divided into different spatial schemes. The sequencing of spatial units may also be differently and artificially defined. For example, a place may be divided into certain zonal structure, and the zones can be sequenced from east to west or west to east, or the sequence of zones can start from the center and move outward.

Days are conventionally aggregated into weeks by grouping 7 consecutive days into a week, but different countries have different customs for using different days as the first day of a week. In some cases, moreover, people only consider a week as 7 consecutive days rather than defining precisely when each week starts. As a result, for both space and time, there are many ways for spatiotemporal aggregation, even on the same spatiotemporal scale. Hence, if the characteristics of a phenomenon can only be found following a specified spatiotemporal pattern, the analytical results would be affected by the way that space and time divided.

In this chapter, the zone effect was tested using the same resolution but different spatiotemporal division scheme. The spatial division of fishnet grid and hexagon grid and the time step alignment of the end time and start time were compared.

Spatiotemporal Boundary Effect

The spatiotemporal boundaries of a set of events include the boundaries of both the geographical space and time, which describe the spatiotemporal range of the events. For some events, the probability of them occurring is related to not only the environment of their locations but also their surrounding environment, which may include economy, demography, and climate, among other environmental conditions. For example, to identify whether a region is a crime hot spot over a period of time, apart from considering the crime events that happened in this location during that time period, the incidents that happened in its adjacent locations may also be considered. As a result, the boundaries chosen to delineate the events as to form a dataset may affect the outcome. Hsueh et al. (2012) investigated the spatiotemporal patterns of DF using the same dataset. They analyzed the patterns for each year and detected different patterns. Therefore, it is apparent that results of analyzing spatiotemporal patterns of a set of events may be affected by the different spatiotemporal boundaries applied in the analysis.

For the discussion in this chapter, the spatial extents of the two datasets are fixed by the data. So, the spatial boundary effects are not tested here. Only the temporal boundary effects are examined.

Method

Spatiotemporal Pattern Detection

The spatiotemporal patterns of the examined data were identified using the *Space Time Pattern Mining Tools* toolbox, which is available in ArcGIS 10.5 or higher.

Different spatiotemporal units were used to explore the scale effects of the MATUP on results of analyzing spatiotemporal patterns. Three spatial units, including fishnet grids with the width of 0.5 km, 1 km, and 1.5 km, and 5 temporal periods, including 1 week, 1 month, 3 months, 6 months, and 1 year, were used to form a total of 15 spatiotemporal units used in detecting the spatiotemporal hot clusters. The temporal alignment was set such that the last date of the dataset was set as the END_TIME and the rest were defined by counting backward from that time.

In addition, the zone effect was tested on the DF dataset using the same spatiotemporal unit, but the zonings of the spatial units are different. This comparative analysis used a hexagonal spatial grid to partition the study area and started counting temporal units at the *START_TIME* and forward.

Two tools inside the ArcGIS 10.5 toolbox were used for the identification of spatiotemporal hot clusters. First, the *Create Space Time Cube* tool was used to create space-time bins from the original data. Defined by the selected spatiotemporal units, space-time bins are the aggregated three-dimensional spatiotemporal units for the analysis. All bins belonging to the same time range are called a time slice. Bins that belong to the same location across time are called Bin Time Series, as it is shown in Fig. 1a. Moreover, then, the *Emerging Hot Spot Analysis* tool used the bins created earlier to construct spatiotemporal patterns.

The Create Space Time Cube tool summarizes a set of points into a NetCDF output file by aggregating them

[2] *Hot Clusters* is used for spatiotemporal data as *Hot Spots* are for spatial data.

into space-time bins. NetCDF is a structure which supports the storage of multi-dimensional scientific data, such as geographical coordinates, time, temperature, humidity, etc. Within each bin, the number of points (space-time events) is counted, and their associated attribute values are aggregated. For all bin locations, the trend for counts and summary field values is evaluated. The block of space-time cubes is the only acceptable input data format for the *Emerging Hot Spot Analysis* tool.

Details on the tool's parameters, their meanings, and requirements of the Create Space-Time Bin tool can be found in the Help document of ArcGIS (http://desktop.arcgis.com/en/arcmap/10.3/tools/space-time-pattern-mining-toolbox/create-space-time-cube.htm). Three parameters below have the most effects on the division of the study spatiotemporal extent into spatiotemporal units:

1. Time Step Interval: The number of seconds, minutes, hours, days, weeks, or years that represent a single time step. Examples for valid entries for this parameter are 1 day, 2 weeks, 3 months, and so on.

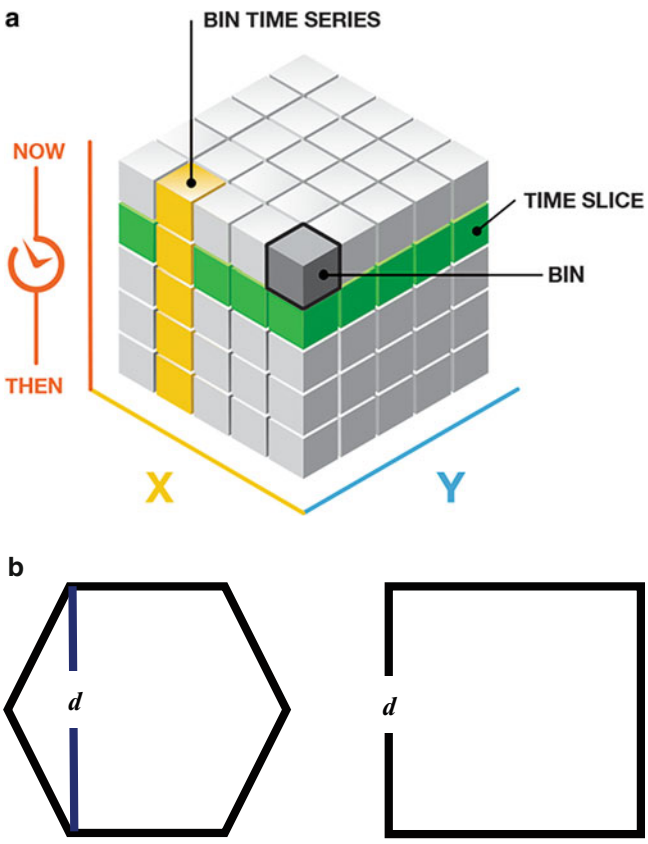


Fig. 1 (**a**) Space-time cube. (From ArcGIS 10.5 Help document, retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/space-time-pattern-mining-toolbox/create-space-time-cube.htm). (**b**) Distance interval (d)

2. Time Step Alignment: Defines how aggregation will occur based on a given time step interval.
3. Distance Interval (d): The spatial extent of the bins used to aggregate the input features. For a fishnet grid, the distance interval, d, is the side length of the square. If the analysis were to aggregate data into hexagons, the distance interval is the height of each hexagon, or d, as it is shown in Fig. 1b.

In this research, several time step intervals were set to 1 week, 1 month, 3 months, 6 months, and 1 year. Distance intervals include 0.5 km, 1 km, and 1.5 km. The default value of END_TIME was used for the time step alignment, which aligned time steps to the last time event of the data and aggregates time backward. Also, time step alignment was set to START_TIME for the spatiotemporal unit of 1 month by 1 km^2 and 3 months by 1 km^2 so as to support a comparison of the identified hot clusters to see whether they cause any differences for the results.

The output file created by the Create Space-Time Bin tool were the input file for the Emerging Hot Spot Analysis tool. This tool needs to be given a search neighborhood range, including the Neighborhood Distance and the Neighborhood Time Step. The Neighborhood Distance was set to be two times the unit's distance interval correspondingly while the Neighborhood Time Step was set to be one.

The *Emerging Hot Spot Tool* combines two statistical measures to identify trends within the three-dimensional space. The tool uses the Getis-Ord G_i^* statistics to evaluate every bin within the search over the adjacent neighborhoods of each time slice. As a result, every bin was determined whether it was a hot or cold spot based on the calculated z-score and p-value.

The formula calculating the Getis-Ord G_i^* can be written as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \overline{X}\sum_{j=1}^{n} W_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}}$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features, and

$$\overline{X} = \frac{\sum_{j=1}^{n} x_j}{n},$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\overline{X}\right)^2}.$$

Next, based on the evaluation result of each bin with respect to other locations across time series, the Mann-Kendall

trend test was used to evaluate the temporal trend across the temporal dimension (Mann 1945; Kendall and Gibbons 1990; Hamed 2009). With the Mann-Kendall test, the bin value for the first period was compared to the bin value for the second period. If the first is smaller than the second, the result is a $+1$, indicating an increasing trend. Otherwise, the result is -1, representing a decreasing trend. If the two values are tied, the result is 0, meaning no trends are detected. The results are summed after comparing with each neighbor. Therefore, a random trend has the expected sum of 0, indicating no trend over time. Based on the variance for the values in the time series of bins, the number of ties, and the number of periods, the observed sum was compared to the expected sum (0) to determine if the difference was statistically significant or not. The trend for each bin's time series was recorded as a z-score and a p-value. A small p-value indicates that the trend was statistically significant. The sign associated with the z-score determines if the trend was an increase in bin values (positive z-score) or a decrease in bin values (negative z-score).

With the ArcGIS software, the tool categorizes every bin as one of the following statuses: new hot spot, consecutive hot spot, intensifying hot spot, persistent hot spot, diminishing hot spot, sporadic hot spot, oscillating hot spot, historical hot spot, new cold spot, consecutive cold spot, intensifying cold spot, persistent cold spot, diminishing cold spot, sporadic cold spot, oscillating cold spot, historical cold spot, or no pattern detected. The entire definition of all hot spot types can be found on the ArcGIS website (http://desktop.arcgis.com/en/arcmap/10.3/tools/space-time-pattern-mining-toolbox/learnmoreemerging.htm).

However, because different datasets are merely a sample of the population, some types of hot spots may not be detected. For example, if the spatiotemporal units were too small for the data, they may result in a small number of cases aggregated into one unit. With that, the variance between bins might be too small to be detected. Therefore, different spatiotemporal units were used and tested in this research to showcase how they can affect the analysis results.

Datasets

The dataset of the DF cases is from the Health Bureau of Kaohsiung City in Taiwan. The dataset includes all daily reported DF cases from 2003 to 2008 for a total of 1408 records. The reported DF cases were geocoded to latitude/longitude coordinates under the transverse Mercator projection. The temporal attribute of the data includes the date when the cases were reported and the date of confirmation. Given that all reported DF cases shown in the dataset were later confirmed to be DF, the analysis only used the reported dates as the temporal information.

The LC dataset was for an un-named county in the USA and was artificially distorted by randomly shifting each location a ± 200 feet in all x, y directions to protect patient privacy. The resulting dataset contains over 6000 cases whose space-time attribute values were slightly modified from their original values. Given that the shifts had a mean of 0, the space-time variance should remain the same. With the random shifts, each LC case had a set of latitude and longitude coordinates and a date that ranged from 1996 to 2009. Although the dataset does not contain original locations of cancer cases, the dataset generally simulated the spatiotemporal distribution of LC in an average-sized US county. For the validity of the study reported here, as long as the dataset is the same when testing different spatiotemporal analytical units, the different impacts on the results can still stand.

Furthermore, this dataset has a much higher number of cases than the DF dataset, as well as a more extended time period. Therefore, it can be an excellent comparison to the DF analysis as well as showcasing the effects of the MATUP on results of chronic diseases research. However, the result of this dataset cannot be used to interpret real-life cancer cases. It merely contributes to discussing the effects of MATUP.

Results and Discussion

Impacts of Different Spatiotemporal Scales

There is a total of 11 types of hot/cold spots detected from the spatiotemporal distribution of the LC cases. The results are shown in Fig. 2. Only the fishnet grids that had been identified as a hot/cold cluster were plotted. In Fig. 2, all windows have the same geographic scale and have the same spatial extent.

As it is shown in Fig. 2, there are many varying results. Each row of figures represents results using the same temporal unit, and each column of figures contains results using the same spatial unit. The spatial unit is marked as the width of the fishnet grid. For the results at the finer temporal units such as 1 week and 1 month, only sporadic hot clusters were detected. When the spatiotemporal units or either the spatial or temporal unit becomes larger, more variety of patterns could be detected. Overall, the figures indicate that, potentially, more variance can be identified if the spatiotemporal unit is substantial enough for data to accumulate.

However, the same conclusion cannot be drawn on the DF cases. As it is shown in Fig. 3, only three types of hot clusters were identified for the DF spatiotemporal patterns. Due to the data limitation (i.e., scarcity), only 1 month, 3 months, and 6 months were used as the temporal units. It is worth to mention that no matter at which spatial level was used, using 1 week as the temporal extent failed to reveal any pattern.

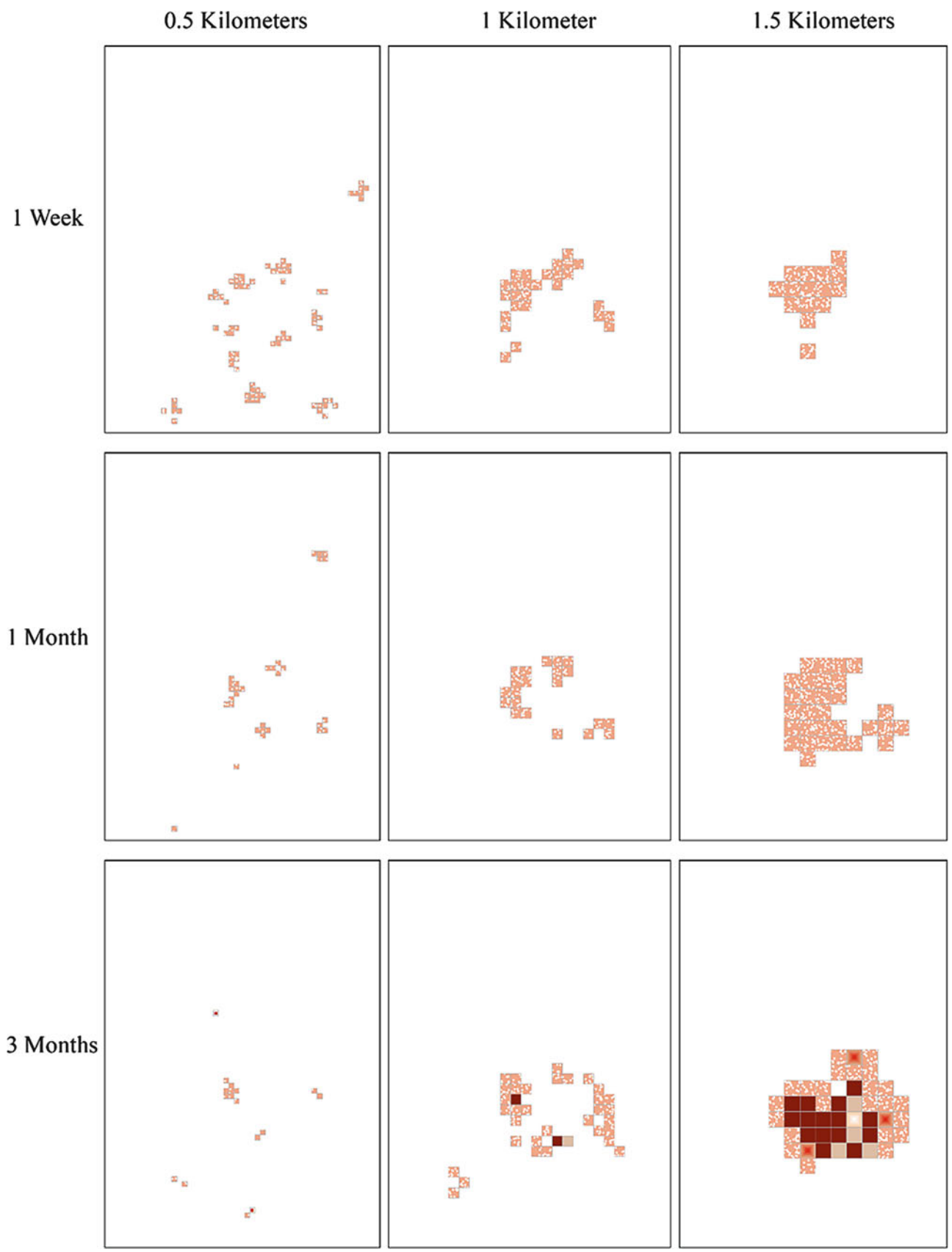


Fig. 2 Emerging hot spots of LC under different spatiotemporal scales

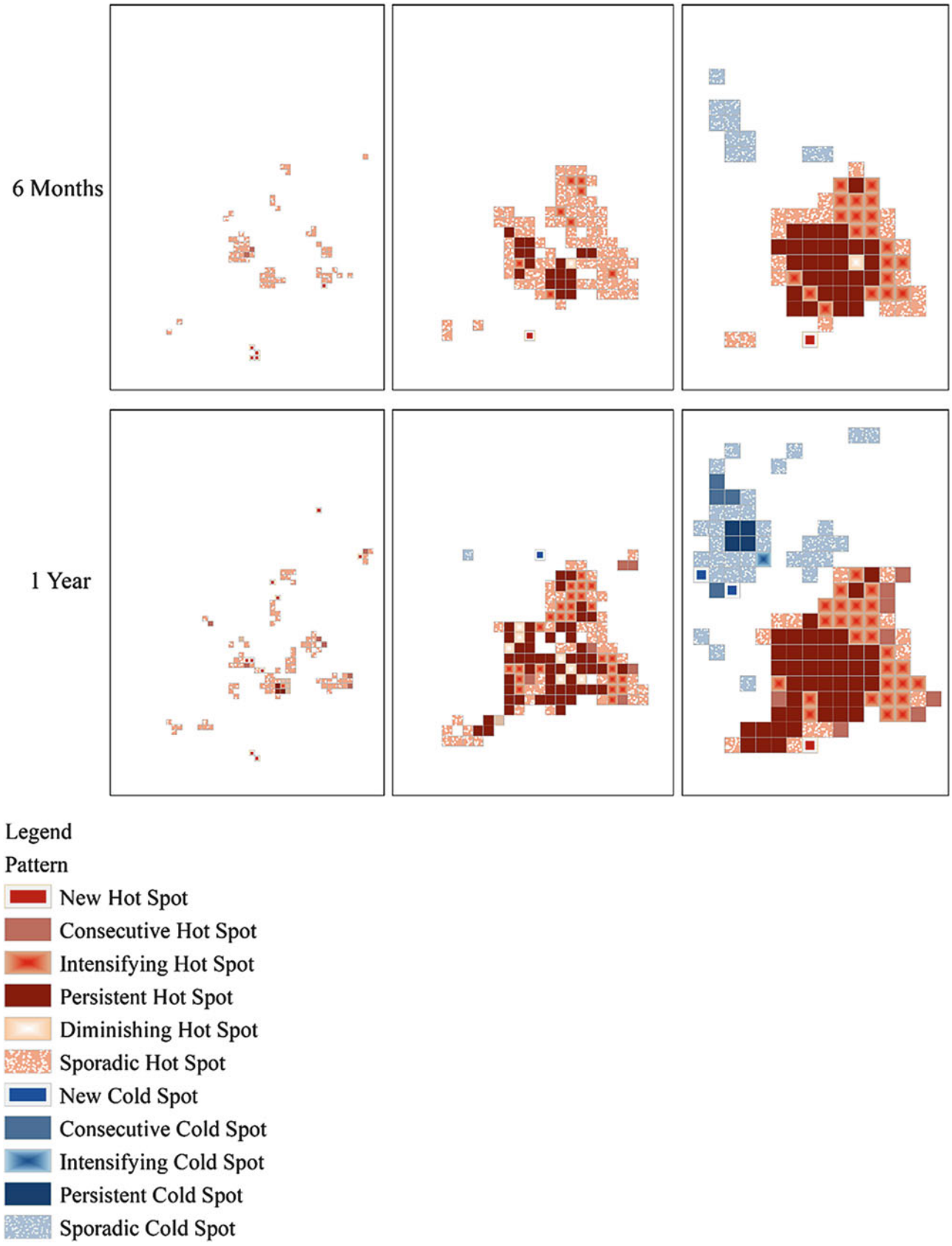


Fig. 2 (continued)

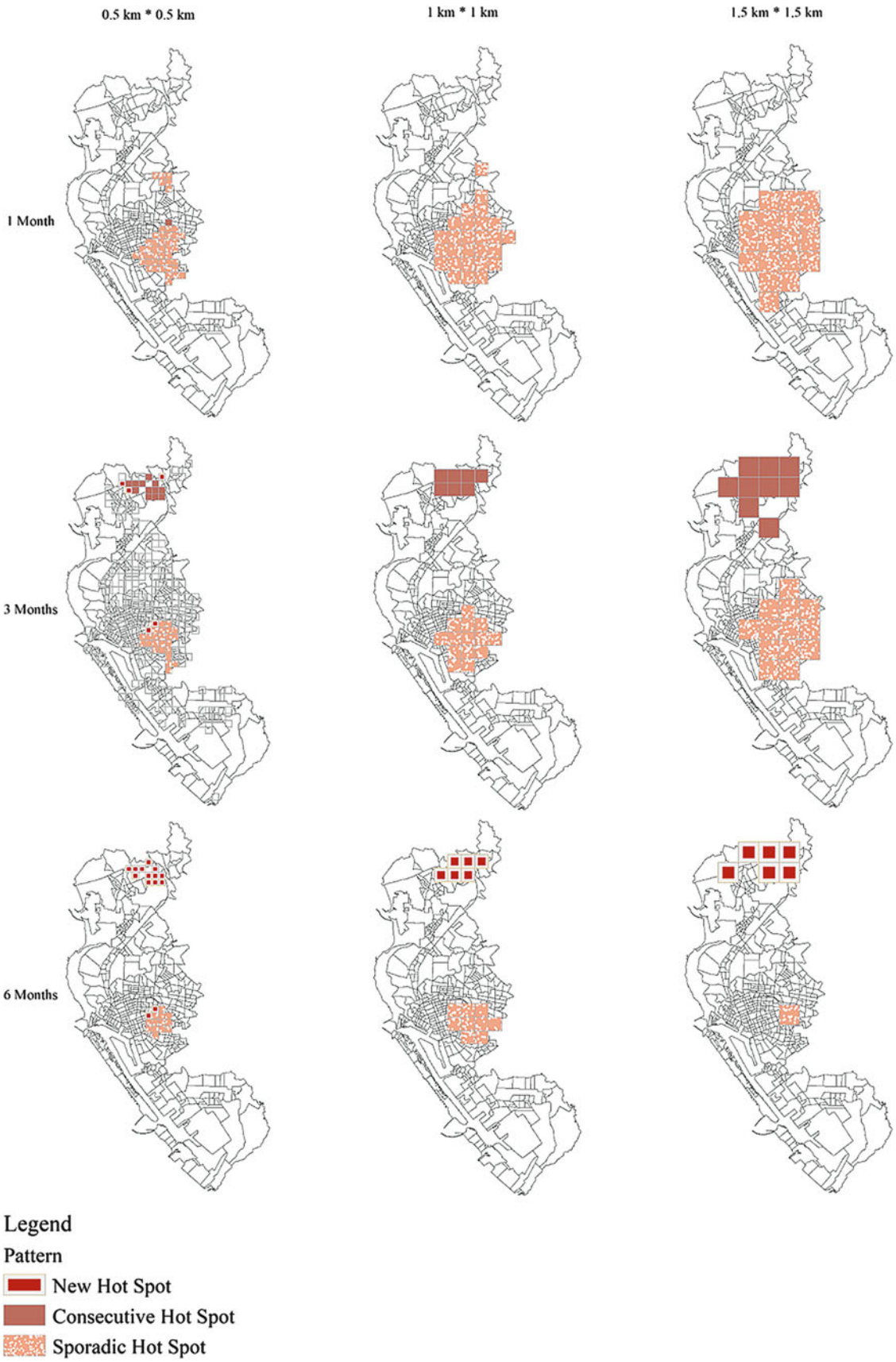


Fig. 3 Emerging hot spots of DF under different spatiotemporal scales

This might be caused by having too few cases accumulated within one unit (spatiotemporal bin), so the algorithm was not able to detect any variance among bins.

In both cases, if the spatial unit stayed the same when the temporal unit was lengthened, the tool could not always detect more patterns than when the temporal unit was short. This situation is evident for DF cases. The patterns detected from the DF cases show some regional variance. For the northern part of Kaohsiung, patterns can only be identified for larger spatiotemporal units, and the types of hot clusters are different. However, for the central regions of Kaohsiung, when the spatial extent was controlled to be the same, there were fewer patterns detected if the temporal units are increasing – an outcome that is contrary to the assumption that the longer time period used for a bin, the more cases would be included in a bin.

Furthermore, the result from using the largest spatiotemporal unit (6 months by 1.5 km^2) in the detection of hot clusters was showing the least variance (result showed the least number of units detected as hot spots compared to results from other spatiotemporal units), whereas the results using the temporal units of 3 months, regardless of which spatial units used, seem to show the most varying patterns (results showed both the most number of hot spots detected, and more types of hot spots identified – especially for the unit of 3 months by 0.5 km^2). These results might be caused by the nature of the disease as DF often has intense cyclic occurrences that generally start in mid or late summer and end in late fall (Hsueh et al. 2012). Therefore, temporal units that were too small failed to form any pattern, while temporal units that were too large tended to conceal detail information. Hence, it is essential to select the best spatiotemporal units for health research. Ideally, the units can be defined based on the nature of the disease. However, if the nature of the disease is unknown, the researcher should try more combinations for the best results, given the increasing computation power that we now have with modern computers. In the DF cases, it appears that using the temporal unit of 3 months might be the right choice for future studies.

However, in the LC cases, when the spatial unit was 0.5 km^2, the number of locations identified as hot clusters decreased as the temporal units were increased from 1 week to 3 months. Nevertheless, the phenomenon was later reverse as more locations and types of hot clusters were identified when the temporal unit was increased.

With larger spatial units such as the 1 km^2 grids and 1.5 km^2 grids, when the temporal unit was increased, more varying patterns were detected (except for the pattern identified at the 1 month by 1 km^2 which was less statistically significant than that of the 1 week by 1 km^2). Furthermore, the most patterns were detected when the temporal unit was set to be 1 year. The result indicates that for research on chronic diseases, using a spatiotemporal unit that is too small may not reveal comprehensive spatiotemporal patterns of the disease.

Impacts of Spatiotemporal Zone

The analysis used different division schemes when the spatiotemporal resolution remains the same to demonstrate the impacts of the zone effect. For different spatiotemporal zones, in the *Create Space Time Cube* tool, the time step alignment parameter was set to either END_TIME or START_TIME, and the aggregation shape type was set to be either FISHNET_GRID or HEXAGON_GRID. For the DF cases, the comparison was made with the spatiotemporal resolution of 90 days by 1 km^2.

The temporal unit was set to 90 days to approximate the temporal units of 3 months (i.e., a season). Based on the previous result (see Fig. 3), using 3-month time periods may produce a better result compared to results from using other temporal units. However, the original dataset had 72 months, which could be divided evenly by 3 months. Therefore, if the analysis were to use 3-month periods as the temporal unit, there might be much detectable difference – 3 is such a small numeric value. Hence, the spatiotemporal resolution was set to be 90 days by 1 km^2. The results were shown in Fig. 4. The maps of each row have the same spatial zone, while each column has the same temporal division.

As it is shown in Fig. 4, the change of spatiotemporal zoning scheme did impact the results. When using the same spatial zone, changing the time step alignment affected the types of clusters detected, especially in the northern parts of the study area. However, it seems like when using the same temporal alignment, the types of clusters identified showed little differences, while the location of the clusters did show to be changed.

Furthermore, the effects on the temporal trends can be further verified by mapping the trends as they are shown in Fig. 5. The numbers of different trends by confidence levels under different time step alignments are shown in Tables 1 and 2.

Overall, a more significant change occurred when the time step alignment changed. More locations were found to have increasing trends if the alignment was set to START_TIME. Also, although some locations were found to have an uptrend, the level of confidence did change.

Furthermore, the impact of zone effects on the analysis of LC cases was shown on the maps in Figs. 6 and 7. The spatiotemporal unit for the analysis LC cases was 365 days by 1.5 km^2 *. The unit of 365 days was comparable to the unit of 1 year. The numbers of different trends by confidence levels under different time step alignments are listed in Tables 3 and 4. The zoning effect has an impact on the results when either the spatial or temporal divisions changes. Compared to the DF cases, the effect was more evident for the different spatial division schemes.

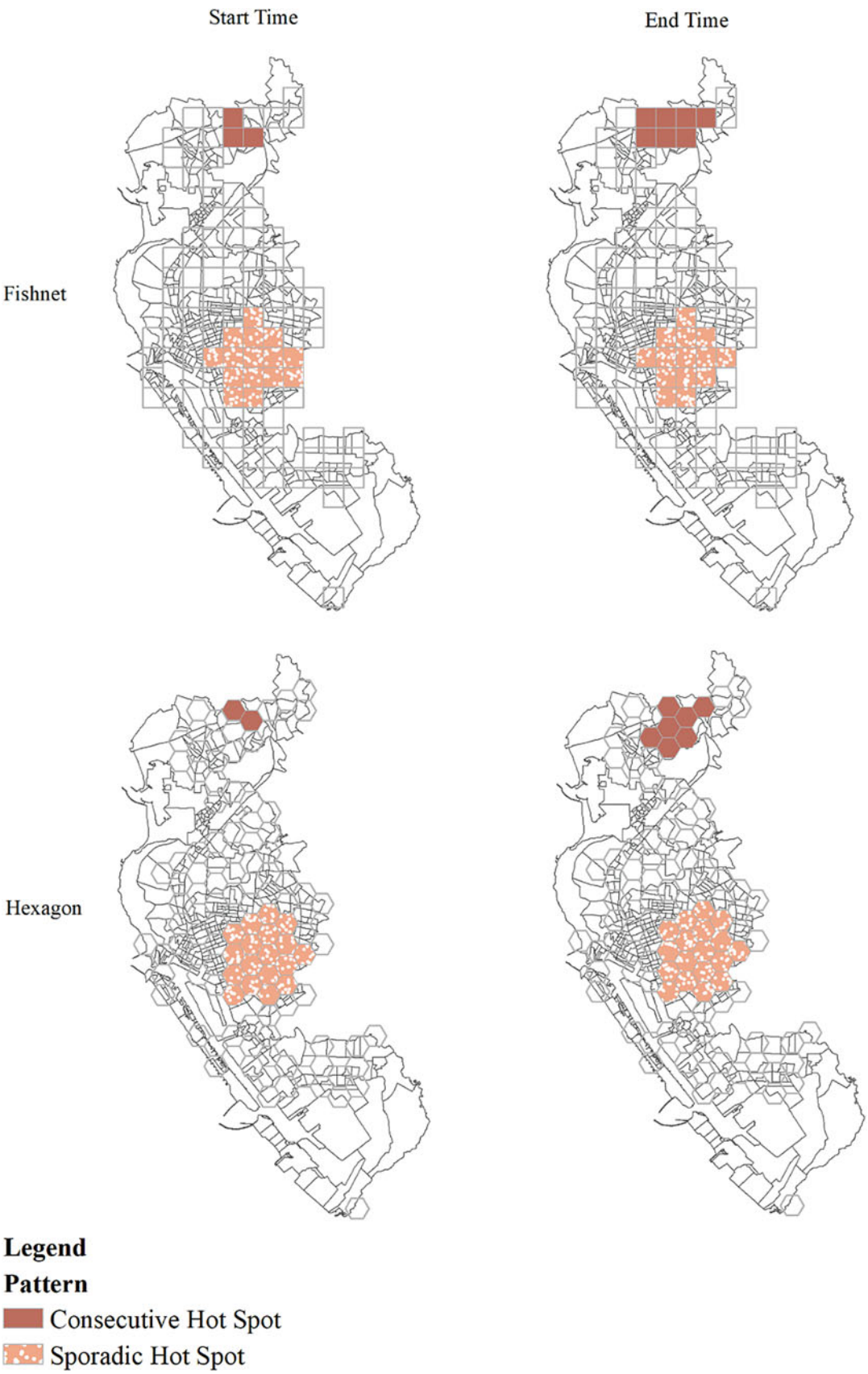


Fig. 4 Emerging hot spots under spatiotemporal zone effects on DF analysis

Fig. 5 Temporal trend under spatiotemporal zone effects on DF analysis

Table 1 DF analysis: the number of different trends by confidence under different time step alignments – fishnet grid

Time step alignment	Uptrend 99% confidence	Uptrend 95% confidence	Uptrend 90% confidence	No significant trend
END_TIME	4	21	16	63
START_TIME	13	21	21	49

Table 2 DF analysis: the number of different trends by confidence under different time step alignments – hexagon grid

Time step alignment	Uptrend 99% confidence	Uptrend 95% confidence	Uptrend 90% confidence	No significant trend
END_TIME	5	16	18	75
START_TIME	12	24	22	56

However, changing the temporal alignment scheme has caused a significant impact on the result than by changing the spatial division scheme. The most exciting result is that for most of the locations identified as historical hot clusters, they were detected as oscillating or consecutive hot clusters when the time step alignment changed from START_TIME to END_TIME. Besides, more locations were identified to have an increasing trend as it showed in Fig. 7. According to the Toolbox's description, it is better to use the END_TIME as the temporal alignment if the study concerns more on the most recent trends (http://desktop.arcgis.com/en/arcmap/10.3/tools/space-time-pattern-mining-toolbox/learnmorecreatecube.htm). Therefore, the benefit of using END_TIME showed promising results on LC cases.

Impacts by Different Spatiotemporal Boundary Schemes

Due to the data limitation, only temporal boundary effects were explored in this chapter. In order to demonstrate the boundary effect, the data of the last 3 months from the original DF dataset and the data of the first 1 year from the LC dataset were removed. For the DF cases, the temporal boundary changed from January 1, 2003, to December 31, 2008, to January 1, 2003, to September 30, 2008. The spatiotemporal unit used for the DF cases is 3 months by 1 km^2 *. For the LC cases, the temporal boundary was changed from January 1, 1996–December 31, 2009, to January 1, 1997–December 31, 2009. For both cases, the time step alignment was END_TIME, and the aggregation shape type was fishnet grid.

Due to the limitation of the software, if the data cannot be broken up evenly into the temporal units (e.g., 3-month or 1-year intervals), there would be a time step that does not have data over its entire span. This will produce bias to the result because the algorithm would assume that the biased time step has significantly fewer events than other time steps. Therefore, data of the most recent 3 months and the oldest year correspondingly in the DF and LC dataset were removed. In this way, the boundary effects of the MATUP can be tested.

The results are shown in Figs. 8 and 9. The spatial scale for all maps is 1:300,000. As it is shown in Fig. 8, the temporal boundary does have an impact on the results. Also, the effects were more evident in the DF cases as no pattern was detected in central Kaohsiung and the types of clusters completely changed in the northern part of the city. However, for the LC cases, the changes seem less obvious. The reason for this phenomenon might be that for the LC dataset, it was the oldest data being removed. Therefore, when the time step alignment was END_TIME, the most recent pattern remained.

To further test the boundary effect, data of the most recent year was removed from the LC dataset, while other conditions remained the same. The result is shown in Fig. 10. Compared to the result shown in Fig. 9, the changes in the pattern seem more substantial, especially on the east, northeast, and northwest portion of the study region. This result indicates that not only the boundary difference affects the result but there is a combination of effects from both the temporal boundary and the temporal alignment.

Strength and Weakness

In this chapter, two sets of diseases cases, including a transmitted disease and a chronic disease, were used to demonstrate the effects of the MATUP. The newly developed Space Time Pattern Mining Tools toolbox in ArcGIS was used for the analysis, including two functions of *Create Space Time Cube* and *Emerging Hot Spot Analysis*. However, there were a strength and weakness of the two tools.

The advantage of the *Create Space Time Cube* is that it aggregates raw point data into a cube of three-dimensional space-time bins, which is the foundation for the *Emerging Hot Spot Analysis*. The tool requires the setting of two parameters to define the spatiotemporal unit, including the *time step interval* and the *distance interval*. Also, the user can define the *time step alignment* rules which have found to have effects on the results in this chapter. If the user is more interesting in recent patterns, using END_TIME as the temporal alignment produced better results.

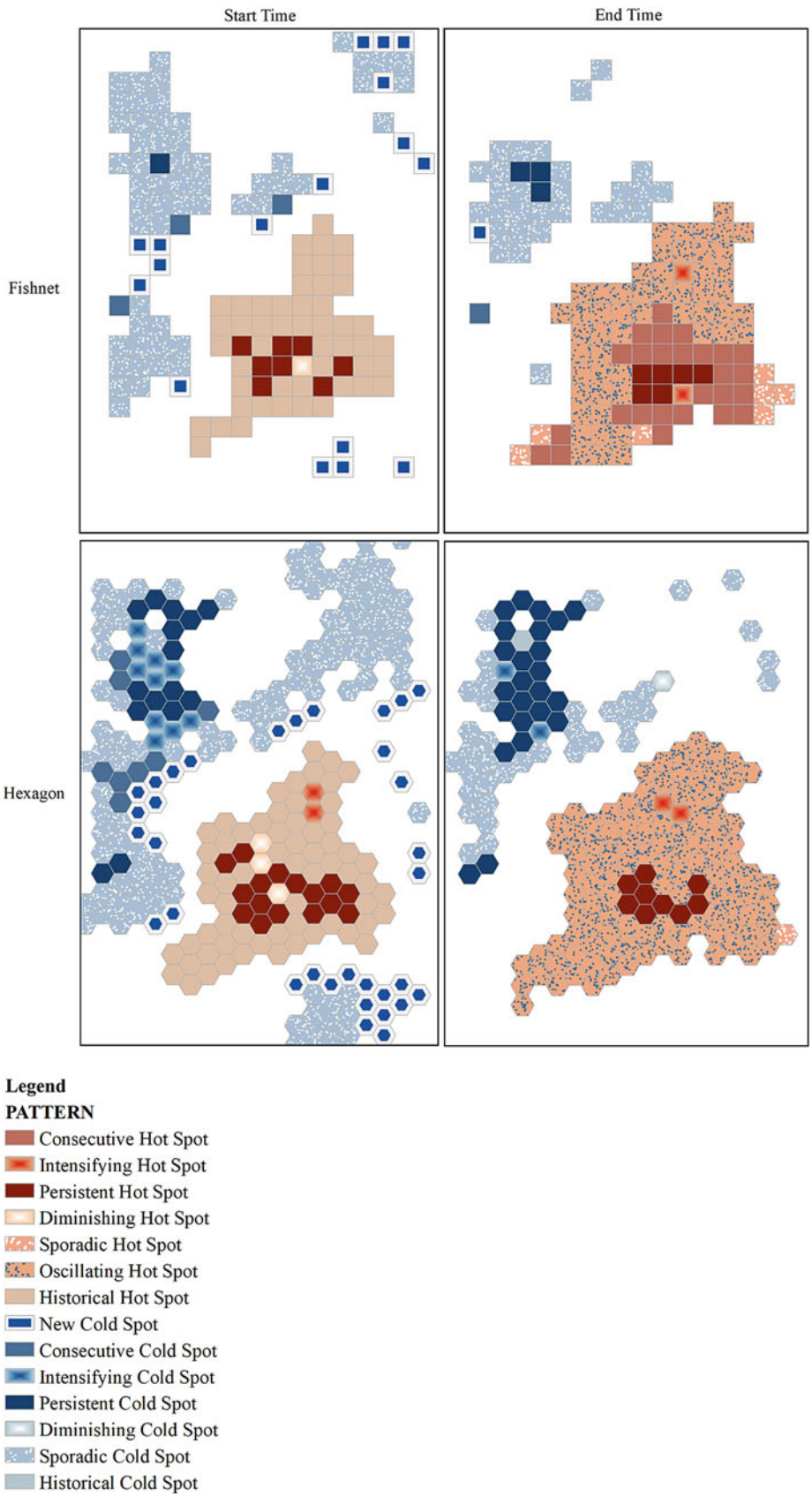


Fig. 6 Emerging hot spots under spatiotemporal zone effects on LC analysis

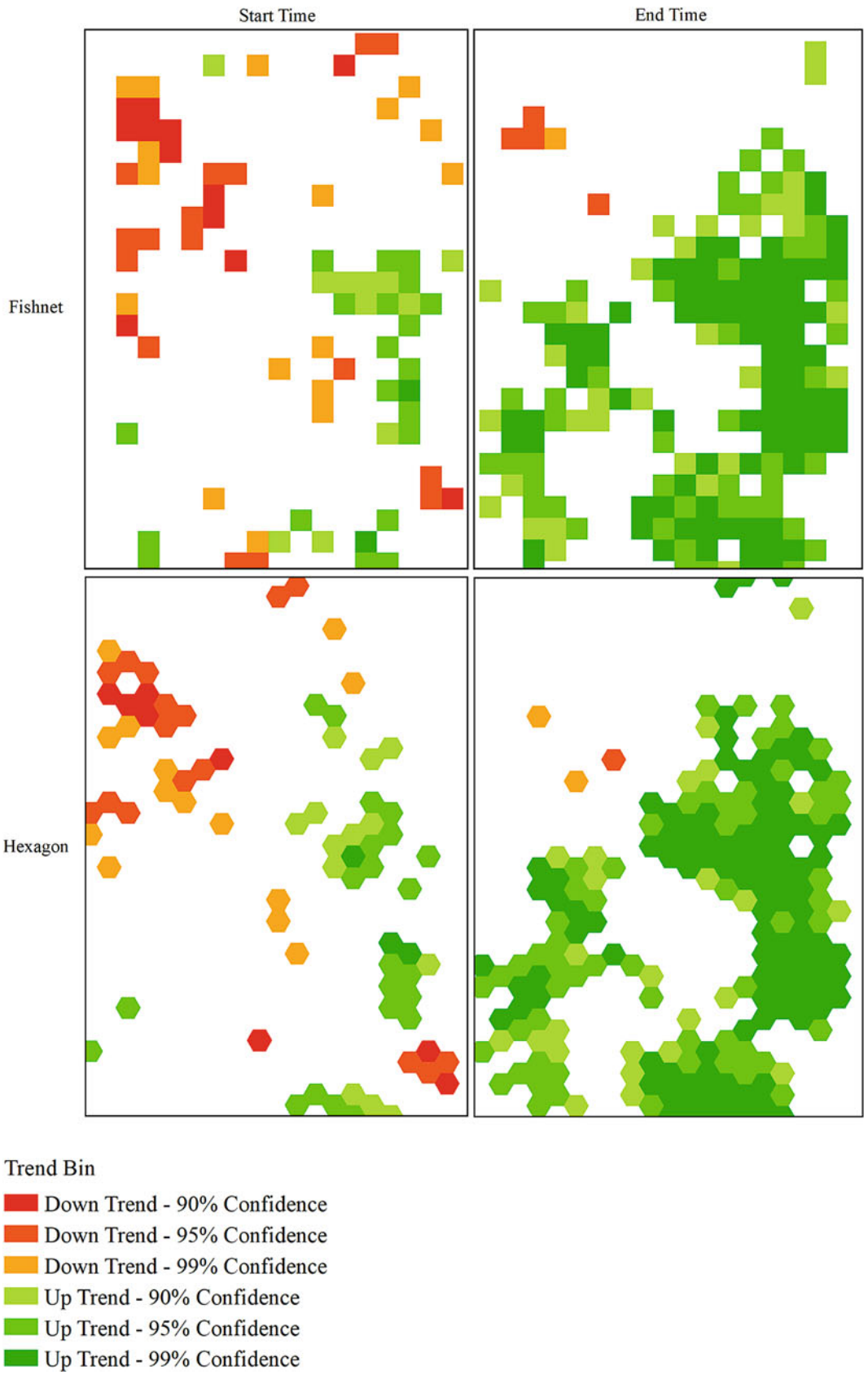


Fig. 7 Temporal trend under spatiotemporal zone effects on LC analysis

Table 3 LC analysis: the number of different trends by confidence under different time step alignments – fishnet grid

Time step alignment	Uptrend 99% confidence	Uptrend 95% confidence	Uptrend 90% confidence	No significant trend	Downtrend 90% confidence	Downtrend 95% confidence	Downtrend 99% confidence
END_TIME	0	5	2	216	37	60	104
START_TIME	14	21	20	329	11	26	3

Table 4 LC analysis: the number of different trends by confidence under different time step alignments – hexagon grid

Time step alignment	Uptrend 99% confidence	Uptrend 95% confidence	Uptrend 90% confidence	No significant trend	Downtrend 90% confidence	Downtrend 95% confidence	Downtrend 99% confidence
END_TIME	0	1	3	245	32	69	131
START_TIME	13	22	16	376	21	29	4

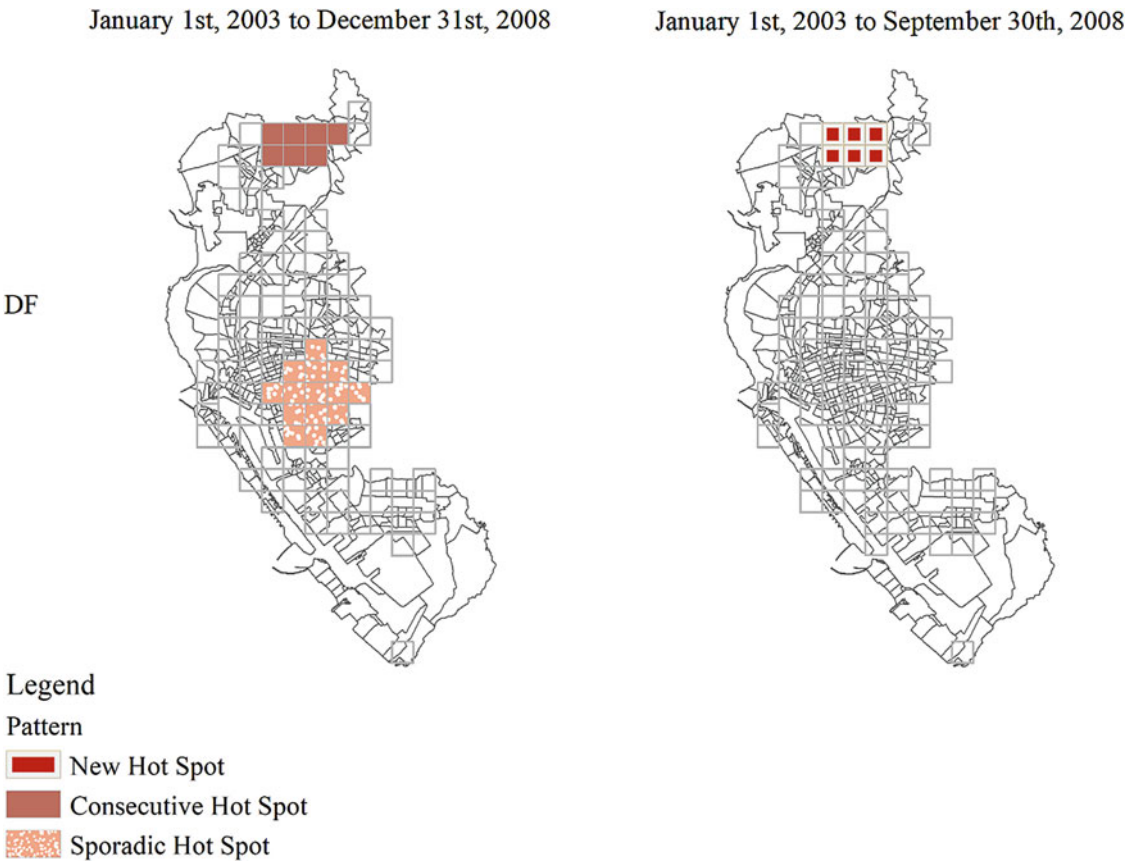


Fig. 8 Emerging hot spots under the spatiotemporal boundary effect on DF analysis. (January 1, 2003, to December 31, 2008, and January 1, 2003, to September 30, 2008)

However, the spatial feature has some limitations. Although users can define whether to create a fishnet grid or a hexagon grid, both aggregation shapes are often not aligned with real-life political or police jurisdiction boundaries. As a result, this brings more difficulty in the interpretation of the result. To this issue, a possible workaround is to have the spatial units as small as possible and then re-assemble the spatial units to match political/artificial boundaries as close as possible after deriving the results.

Furthermore, the *Emerging Hot Spot Analysis* combined Getis-Ord G_i^* statistics for the spatial analysis and the Mann-Kendall test for the temporal trend detection. Also, the tool can identify eight types of hot clusters and eight types of cold clusters based on both the spatial and temporal patterns of the data. Overall, the tool presented a substantial improvement from the hot spot analysis and trend analysis that were based only on spatial dimensions.

Nevertheless, there is still some weakness in this method. First, the *Emerging Hot Spot Analysis* merely combines two well-established statistics rather than proposing a new spatiotemporal statistical method that could have merged

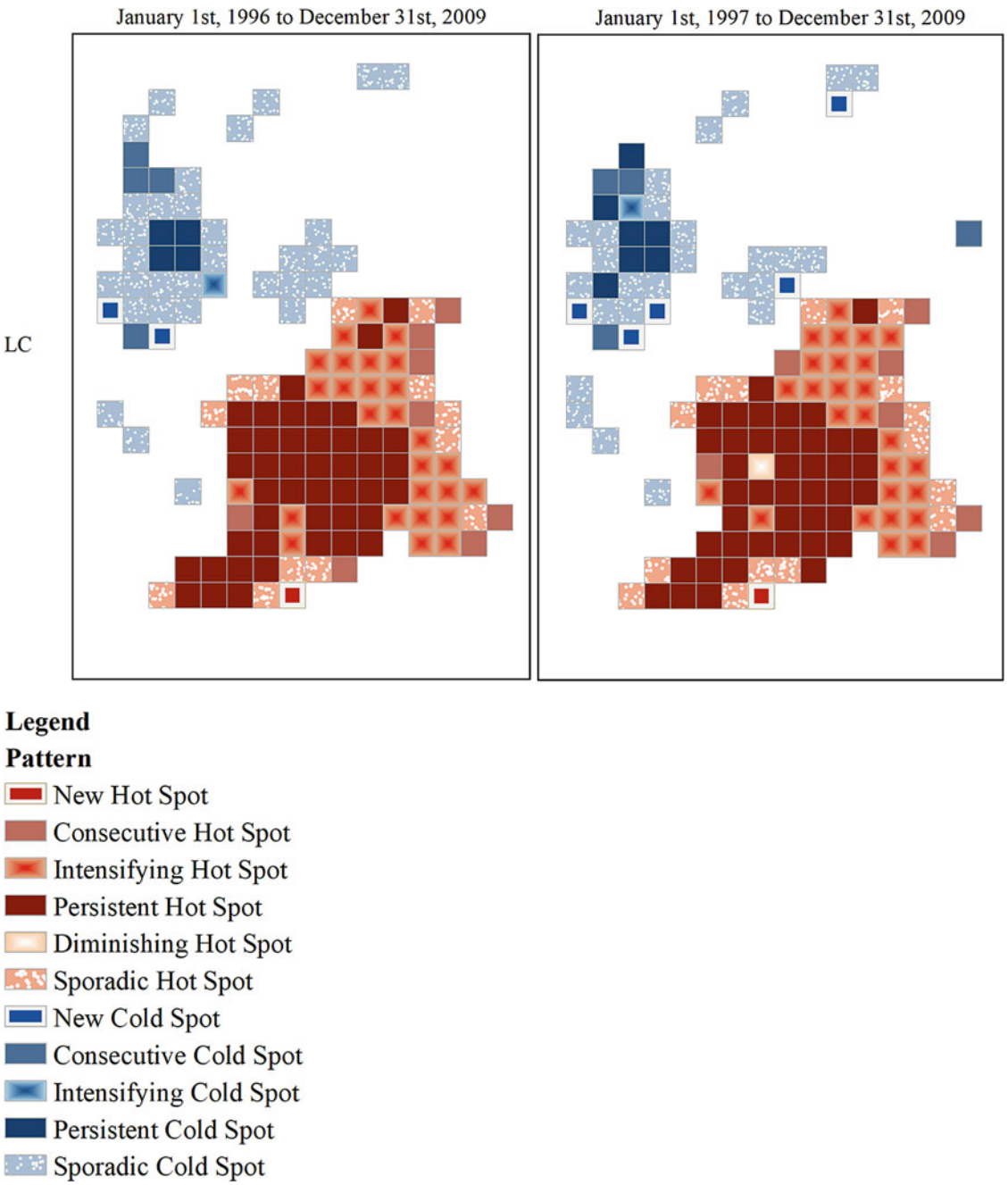


Fig. 9 Emerging hot spots under the spatiotemporal boundary effect on LC analysis. (January 1, 1996, to December 31, 2009, and January 1, 1997, to December 31, 2009)

the spatial and temporal measures organically. Second, the definition of each hot/cold clusters is arbitrary in most cases so that it may not be suitable for some studies that have to use artificially defined boundaries. For example, the definition of the new hot spot describes it as "A location that is a statistically significant hot spot for the final time step and has never been a statistically significant hot spot before." What if a location only becomes a hot spot for the second to last time step while having never been a hot spot before? In real-life practice, we may consider it also to be a new hot spot, while the algorithm might reach a different result. Although 16 types of hot spots are an improvement from the previous methods, whether they stand the test of time needs more experiments.

Furthermore, as the boundary and zone effect of the MATUP shown in this chapter, the result of the analysis is subject to a substantial change if new data enter the dataset. After all, one of the most intriguing applications of this method is to find clusters that are newly developed or have

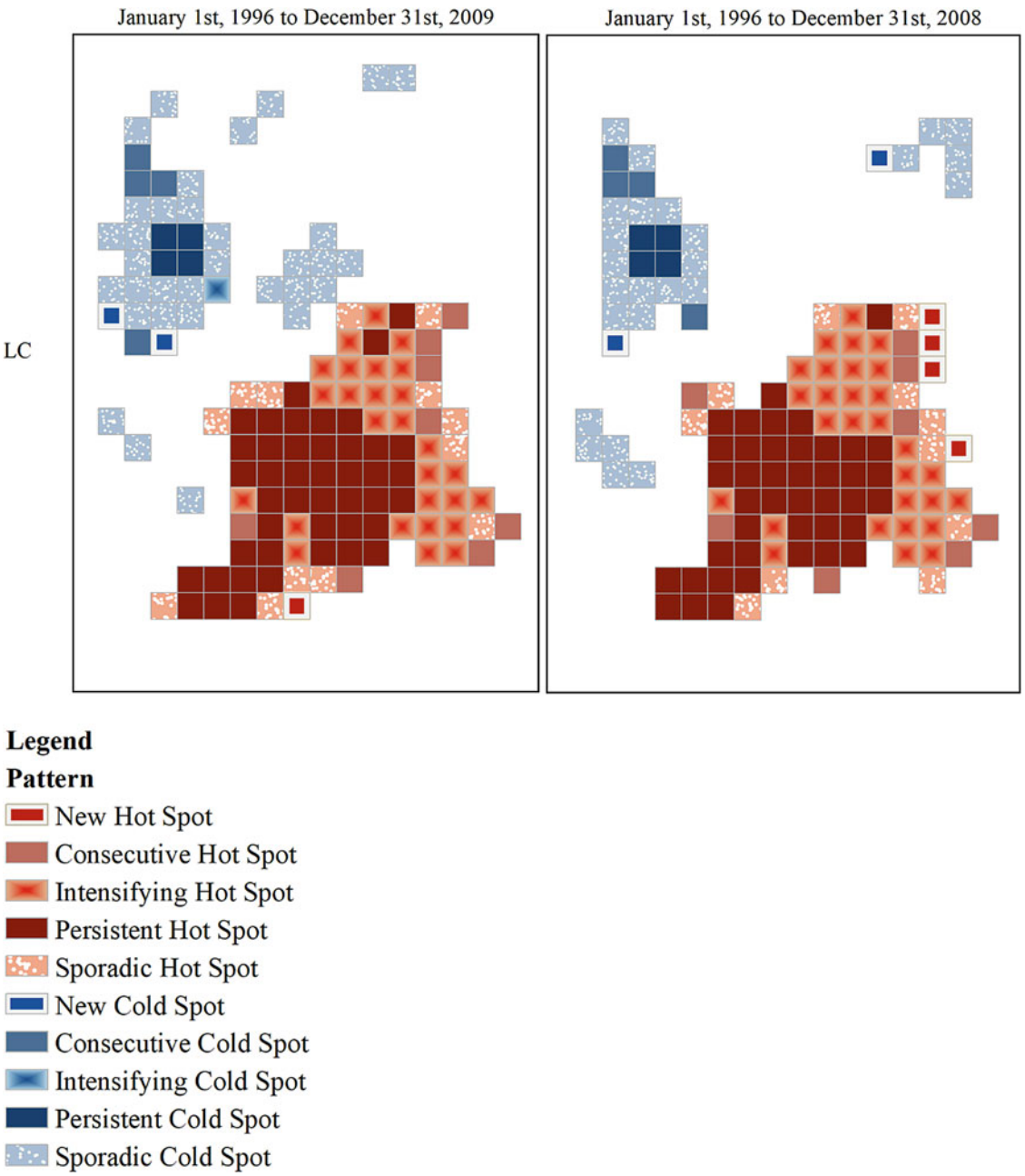


Fig. 10 Emerging hot spots under the spatiotemporal boundary effect on LC analysis. (January 1, 1996, to December 31, 2009, and January 1, 1996, to December 31, 2008)

been intensifying, so that timely response can be enforced. Therefore, it might be more useful to develop a dynamic computer program which has the ability to record new data and produce timely results for the user.

Overall, the method is an improvement from the previous spatial and temporal analysis, which has the potential to be applied in different research projects regarding spatiotemporal patterns of events. However, due to MATUP effects and how the clusters are designed, users should be cautious in interpreting results to avoid overlooking important information.

Conclusion

This chapter discussed the effects of the modifiable areal and temporal unit problem (MATUP) on identifying spatiotemporal hot clusters. There are three categories of the MATUP effects, including the spatiotemporal scale, zone, and boundary. The cases of dengue fever and lung cancer cases were examined using the *Space Time Pattern Mining* toolbox available in ArcGIS 10.5 or later versions. The

results were compared and investigated so that the effects of MATUP can be further understood.

Based on the results, all three types of effects do impact the results of hot spot detection. Using different scales indicates different levels of data accumulation in each unit. Judging from the results, spatiotemporal units that are too small cause fewer incidents accumulation in each unit, so that the statistical results are less significant because less variance has shown between units. However, larger spatiotemporal units may not always show more varying patterns than small units, such as the example of the DF cases. For any given dataset, the most suitable temporal scale should be based on the nature of the events.

Moreover, different results were produced when using different temporal alignments, such as START_TIME and END_TIME, for analysis. This indicates the changes in how space and time are divided because that is directly related to the variance among results. If the priority of the project is to identify recent patterns, using END_TIME is better as the alignment rule.

Furthermore, when different temporal boundaries were used for the analysis of the same dataset, detected patterns would change. However, the effects of changing boundaries may be magnified when combined with different temporal zoning schemes, such as in the case of analyzing the LC cases.

In summary, the analysis in this paper shows the effects of MATUP on spatiotemporal cluster patterns. The analysis reveals the existence of MATUP so that the structure of the spatiotemporal analysis is not entirely independent of the spatiotemporal units used in the research. Since the results of the spatiotemporal analysis rely on the spatiotemporal units selected, systematic analysis using different spatiotemporal units can be experimented when conducting analysis, so that the effects on the variance of results can be minimized.

References

Casas, Irene, Eric Delmelle, and Elizabeth C. Delmelle. 2017. Potential versus revealed access to care during a dengue fever outbreak. *Journal of Transport and Health* 4: 18–29. https://doi.org/10.1016/j.jth.2016.08.001.

Chen, Xi, et al. 2016. Long-term exposure to urban air pollution and lung cancer mortality: A 12-year cohort study in northern China. *Science of the Total Environment* 571 (22): 855–861. https://doi.org/10.1016/j.scitotenv.2016.07.064.

Cheng, T., and M. Adepeju. 2014. Modifiable temporal unit problem (mtup) and its effect on space-time cluster detection. *PLoS One* 9 (6): e100465.

Chien, L.C., and H.L. Yu. 2014. Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. *Environment International.* 73: 46–56.

Coltekin, A., S.D. Sabbata, C. Willi, I. Vontobel, S. Pfister, M. Kuhn, et al. 2011. Modifiable temporal unit problem. ICC2011 Workshop.

Delmelle, Eric, Michael Hagenlocher, Stefan Kienberger, and Irene Casas. 2016. A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta Tropica* 164: 169–176. https://doi.org/10.1016/j.actatropica.2016.08.028.

Gibson, C.C., E. Ostrom, and T.K. Ahn. 2000. The concept of scale and the human dimensions of global change: A survey. *Ecological Economics* 32 (2): 217–239.

Gong, Junfang, Shengwen Li, and Jay Lee. 2019. Space, time, and disease on social media: A case study of dengue fever in China. *Geomatica.* 72. https://doi.org/10.1139/geomat-2018-0016.

Guo, Yuming, et al. 2016. The association between lung cancer incidence and ambient air pollution in China: A spatiotemporal analysis. *Environmental Research* 144: 60–65. https://doi.org/10.1016/j.envres.2015.11.004.

———. 2017. The burden of lung cancer mortality attributable to fine particles in China. *Science of the Total Environment* 579: 1460–1466. https://doi.org/10.1016/j.scitotenv.2016.11.147.

Hamed, K.H. 2009. Exact distribution of the Mann-Kendall trend test statistic for persistent data. *Journal of Hydrology* 365: 86–94.

Harrower, M., A. MacEachren, and A.L. Griffin. 2000. Developing a geographic visualization tool to support earth science learning. *American Cartographer* 27 (4): 279–293.

Hornsby, K., and M.J. Egenhofer. 2002. Modeling moving objects over multiple granularities. *Annals of Mathematics & Artificial Intelligence* 36 (1–2): 177–194.

Hsueh, Ya-hui, Jay Lee, and Lisa Beltz. 2012. Spatio-temporal patterns of dengue fever cases in Kaohsiung City, Taiwan, 2003 e 2008. *Applied Geography* 34: 587–594.

Hu, W., et al. 2012. Spatial patterns and socioecological drivers of dengue fever transmission in Queensland, Australia. *Environmental Health Perspectives* 120 (2): 260–266.

Jerrett, Michael, et al. 2013. Spatial analysis of air pollution and mortality in California. *American Journal of Respiratory and Critical Care Medicine* 188 (5): 593–599.

Kendall, M.G., and J.D. Gibbons. 1990. *Rank correlation methods*. 5th ed. London: Griffin.

Koyadun, Surachart, Piyarat Butraporn, and Pattamaporn Kittayapong. 2012. Ecologic and sociodemographic risk determinants for dengue transmission in urban areas in Thailand. *Interdisciplinary Perspectives on Infectious Diseases* 2012: 907494.

Mann, H.B. 1945. Nonparametric tests against trend. *Econometrica* 13: 245–259.

Meentemeyer, V. 1989. Geographical perspectives of space, time, and scale. *Landscape Ecology* 3 (3–4): 163–173.

Nakaya, T., and K. Yano. 2010. Visualising crime clusters in a space–time cube: An exploratory data-analysis approach using space–time kernel density estimation and scan statistics. *Transactions in GIS* 14 (3): 223–239. https://doi.org/10.1111/j.1467-9671.2010.01194.x.

Openshaw, S. 1984. *The modifiable areal unit problem*, Concepts and techniques in modern geography No. 38. Geo Books, Norwich, England.

Openshaw, S., and P.J. Taylor. 1979. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical applications in the spatial sciences*, ed. N. Wrigley, 127–144. London: Pion.

Wong, D. 2009. The modifiable areal unit problem (MAUP). In A. S. Fotheringham and P. A. Rogerson, editors. In *The SAGE handbook of spatial analysis*. 105–123, SAGE Publications, Los Angeles, California, USA.

Yu, Hwa-Lung, Shang-Jen Yang, Hsin-Ju Yen, and George Christakos. 2011. A spatio-temporal climate-based model of early dengue fever warning in Southern Taiwan. *Stochastic Environmental Research and Risk Assessment* 25: 485–494.

Zhang, Hao, et al. 2013. Analysis of land use/land cover change, population shift, and their effects on spatiotemporal patterns of urban heat islands in metropolitan Shanghai, China. *Applied Geography* 44: 121–133. https://doi.org/10.1016/j.apgeog.2013.07.021.

The Spatial Non-stationarity in Modeling Crime and Health: A Case Study of Akron, Ohio

Huiyu Lin, Jay Lee, and Gregory Fruits

Introduction

The overall purpose of this study is to examine if violent crime rates are a good predictor for community health. Specifically, this study used local obesity rates as a proxy to community health. From an ecological standpoint, researchers often study obesity through investigating their associations with environmental characteristics (Sandy et al. 2013; Burdette and Whitaker 2004; Shahid and Bertazzon 2015; Ruijsbroek et al. 2015) and/or social structural characteristics, such as poverty (Halleröd and Larsson 2008; Chen and Truong 2012; Salois 2012; Rybarczyk et al. 2015; Huang et al. 2018) and race (Fan and Jin 2014). Furthermore, sociologists often argued that the fear of crime and the lack of appropriate infrastructure led to less physical activities, which resulted in obesity among residents. To this end, however, there have not been direct causal associations or universal relationships found between crime and obesity in the literature.

The selection of analytical methods may contribute to the inconsistent results found in previous research. Current research mainly used global regression models which failed to consider the spatial non-stationarity within the relationships between variables (Sandy et al. 2013; Brown Barbara et al. 2014. Therefore, spatially weighted analytics such as the geographically weighted regression (GWR)/geographically weighted Poisson regression (GWPR) have become increasingly recognized and used in public health research (Gilbert and Chakraborty 2011; Nakaya et al. 2005; Yang and Matthews 2012; Comber et al. 2011). This study examines whether violent crime can be a good predictor to local obesity prevalence and if such association is spatially varying using the GWR.

This paper demonstrates the use of GWR in modeling crime, health, demographic, and environmental data. The discussions are to answer the questions: Is it necessary to use GWR in crime and health modeling? How to organize data? How to select the appropriate variables for model building? How to interpret the results? Is the GWR result better than the ordinary least square (OLS) regression? What should we pay attention when mapping the results?

Geographically Weighted Regression and Ordinary Least Square Regression

Spatial Non-stationarity

The concept of spatial non-stationarity was first introduced by Fotheringham, Charlton, and Brunsdon (Fotheringham et al. 1996). In their paper, they pointed out that even though researchers had recognized the spatial component in data, global models were still widely used in studies. However, according to the *First Law of Geography* (Tobler 1970), global parameters could not be able to capture the spatial variances which existed in the relationships between the explanatory variables and the dependent variable (Fotheringham et al. 1996; Brunsdon et al. 1996).

In an effort to address the issue of spatial non-stationarity, researchers proposed localized spatial statistics such as the G statistics (Getis and Ord 1992), local indicators of spatial association (LISA) statistics (Anselin 1995), local ordinary least square regression (OLS) and local Bi-square (Fotheringham et al. 1996), and geographically weighted regression

H. Lin · J. Lee (✉) · G. Fruits
Department of Geography, Kent State University, Kent, OH, USA
e-mail: jlee@kent.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_16

(Brunsdon et al. 1996). Among the aforementioned methods, the first two measure the levels of spatial clustering in geographical events and capture the spatial heterogeneity among them. The latter extended traditional regression models by adding components that measured the strength of spatial associations.

Since the introduction of the geographically weighted regression (GWR), there have been an increasing number of studies that used the method in research related to public health (e.g., Chen and Truong 2012; Chi et al. 2013; Wen et al. 2010; Chalkias et al. 2013). However, with concerns that building GWR models may yield a higher correlation between model variables than those from OLS models (Cahill and Mulligan 2003; Cahill and Mulligan 2007; Troy et al. 2012; Deng 2015; Rybarczyk et al. 2015), many previous and current research still applied global linear regression such as OLS for model building (Vandewater et al. 2004; Singh et al. 2008; Carroll-scott et al. 2020). Given this, it is crucial to weight the strength and weakness of the two methods so that future researchers can choose the most appropriate one.

The GWR model can be expressed as:

$$y_i = \beta_0 (u_i, v_i) + \sum_{j=1}^{k} \beta_j (u_i, v_i) x_{ij} + \varepsilon_i$$

where y_i is the estimated value of the dependent variable at the location i and (u_i, v_i) describe the coordinates of i, β_0 is the intercept value, and β_j is a set of parameters at point i. The value of β_j will vary for different space-time locations. It is assumed that the observed data close to point i have a greater influence in the estimation of β_j than others. Detail explanation regarding the GWR method can be found in Fotheringham, Charlton, and Brunsdon (Fotheringham et al. 2002).

In addition to GWR, the spatial matrix was introduced to other regression models to expand the spatial statistics such as the GWPR mentioned earlier (Nakaya et al. 2005) and Gaussian semi-parametric GWR (SGWR, Villarraga et al. 2014).

Although GWR can be used to explore the spatial non-stationarity among variables, it might not be suitable for use for all datasets. Therefore, it needs to be justified that using the spatially weighted method is indeed superior to global modeling before applying it to a specific dataset in the research.

Justification for Using GWR

There are a few criteria that we can use for assessing if GWR is suitable for use in a study. A global regression model may be built for an initial assessment. After applying the OLS regression, Jarque-Bera test (Jarque and Bera 1980; Thadewald and Büning 2007; Barbu 2012) and Koenker (BP) statistics (Wallace 2011; Avila-flores et al. 2010) can be applied to evaluate the model residuals for assessment. Combining both tests assesses whether the relationships shown by the model have any bias or are consistent over the study region. If both statistics are statistically significant, we would be confident that the global model is biased, and another method, such as GWR, should be used (Ortolano et al. 2018; Avila-flores et al. 2010).

However, the Jarque-Bera test and the Koenker (BP) statistics do not necessarily reflect whether the spatial non-stationarity causes the bias. Therefore, other criteria should be included to justify whether using GWR produces a better model. Overall, we may compare the residual squares, the Akaike information criterion (AIC) (Akaike 1974) which serves as a goodness-of-fit indicator (Yang and Matthews 2012), and the R^2 and/or adjusted R^2 between the GWR and OLS model to determine whether it is necessary to use GWR and whether GWR models perform better than the OLS model.

The residual squares are the sum of the squared residuals in the model. Models with smaller residual squares have a closer fit between estimated values to the observed data. Furthermore, the model with lower corrected AIC (AICc) value reflects a better fit to the observed data. If the AICc value of the GWR model is lower than that of the OLS model with the difference larger than 3, it can be asserted that using the GWR model is beneficial.

Also, R^2 represents the goodness of fit, which shows the proportion of the dependent variable variance accounted for by the independent variables. The value of R^2 and adjusted R^2 shows the strength of the association between the dependent variable and independent variables. The larger the value, the better the fit of the model. However, adding any collinearity among explanatory variables might inflate the value of R^2. Therefore, the adjusted R^2 should also be evaluated.

Bandwidth Selection

Different from the ordinary least square (OLS) regression model, which treats a study area as having the same association between dependent and independent variables everywhere, the GWR uses a moving kernel with a fixed or adaptive bandwidth for defining the different spatial weight that a given local unit should be weighted in the analysis of the association between dependent and independent variables. Such a model produces localized regression parameters, i.e., a local R^2 and local regression coefficients for each spatial unit and each independent variable in the model. Therefore, GWR can be used to search for locations that exhibit significantly strong (or weak) associations between

the independent and dependent variables or to detect "hot spots" (Fotheringham et al. 2002).

The spatial weights matrix may be structured to be based on distances between local spatial units. Based on the *First Law of Geography* (Tobler 1970), the selection of a bandwidth may significantly affect the model outcome (Cahill and Mulligan 2007). The bandwidth that is too small may include fewer data points for estimation, which may result in an instability of the parameter estimates, while a bandwidth that is too large may smooth the spatial variation at the estimation point.

Many GIS software allows users to determine whether they want to use a pre-defined fixed bandwidth or an adaptive bandwidth. The value of a fixed bandwidth may come from previous experience or literature. However, in this research, adaptive bandwidth was selected. There are two popular parameters usually provided in GIS software to calculate the optimal bandwidth – the corrected Akaike information criterion (AICc) and the cross-validation (CV). The two parameters produce similar results, while the AICc (Akaike 1974) also can serve as a goodness-of-fit indicator (Yang and Matthews 2012), which makes it more popular among researchers (Cahill and Mulligan 2007; Fotheringham et al. 2002).

Attribute Selection Process

It is also crucial to select a proper set of independent variables that are correlated with the dependent variable while avoiding multicollinearity among one another so that the model is reliable. The first step of choosing variables is to always refer to the literature and previous studies so that the model makes sense. This study referred to ecological studies including opportunity theory (Cohen and Felson 1979) and environmental criminology theory (Brantingham and Brantingham 1975) that suggested the included variables for representing the racial component. Selected explanatory variables include Black population percentage; economic components such as housing occupancy, income, and employment rate; and physical/built environment characteristics. However, not all variables were suitable for building the health-crime regression model.

As introducing variables that are highly correlated may affect the model outcome by falsely inflating the R^2, there are a few criteria that we should use to select the most appropriate set of independent variables for the model building. A correlation analysis may be performed to both dependent and independent variables before building the model. Independent variables that are not statistically significantly correlated with the dependent variable can be discarded. Among independent variables, the ones that were not highly correlated with each other can be retained in the final GWR model. Also, the model's variance inflation factor (VIF) can be calculated to refer to whether there was any redundancy among explanatory variables. If the VIF was less than 5, the variable could be included in both the OLS and the GWR model.

When building an OLS model, it often excludes the not significant variables, for example, at the 95% confidence (t-value >1.96 or <−1.96). However, since spatial non-stationarity may exist in the relationship between the dependent variable and independent variables, when an independent variable is sometimes not significantly correlated with the dependent variable in an OSL model, it may be acceptable in the GWR model. This is because it may be significantly related to the dependent variable at certain places.

In this research, the variable selection process for building the GWR model was done by using GWR4 software. In the model set, the function of "Geographical variability test" was selected, which reported a statistic for each variable. This statistic is called the "difference of criterion" (DIFF of Criterion). It indicates whether the variable presents any spatial variability. For variables that have the values of the DIFF of Criterion that are greater than 2, it was suggested that these variables should be assumed as global variables but not local ones. These variables were best removed manually if they also were not significantly related to the dependent variable at the global term.

Study Area and Data

The city of Akron in the Summit County of Ohio was the study area which is located in the center of Summit County of Northeast Ohio. Akron was one of the fastest-growing cities in America during the 1920s with a population peak of over 300,000 people. However, the city's population has been continuously declining since then. The 2010 census showed that the city had 199,110 people. Figure 1 shows the locations of the downtown Akron neighborhoods and the population density by block groups in 2010 (population data retrieved from US Census Bureau). As it is shown in Fig. 1, the population of Akron is concentrated in neighborhoods surrounding downtown.

The spatial units for analysis are census block groups, which are the smallest unit to have an extensive selection of census variables available. Therefore, data retrieved in other spatial units such as point data of crime and BMI (for obesity) were aggregated into block groups for analysis.

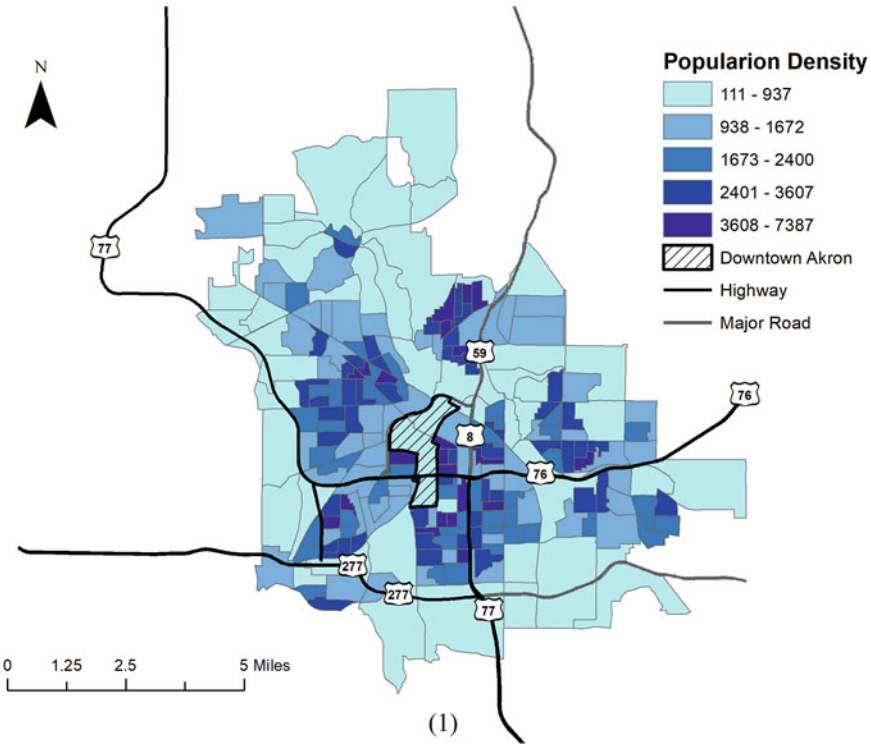


Fig. 1 Population density in the city of Akron, 2010

Obesity and Crime Data

Self-reported data were used to calculate the BMI for individuals. These data came from Summit County's Department of Motor Vehicles. The dataset contained a complete listing of self-reported heights and weights for all residents who held drivers' licenses. The data cover all holders of drivers' licenses between 2009 and 2014, which is a 5-year spectrum, corresponding to the time duration that each license has to be renewed.

To further reduce potential biases, only data from the file that recorded heights and weights for adults aged between 16 and 21 were considered for this study. This is based on the assumption that the first-time reported heights and weights are more accurate than the ones from the licenses that were renewed later because many renewals probably were not given updated heights and weights. The data file of all license holders has approximately 440,000 records.

Crime data comes from the Akron Police Department. Each year's data file contains the time and date of the crime events. Collectively, there are crime data from 2009 to 2012. Each record in these files contains location information as geocoded by the Akron Police Department in the form of latitude and longitude.

Crime types of violent crimes were selected for analysis. Violent crime in this dataset includes assault, battery, murder, homicide, and manslaughter. It is noted that crimes of sexual assault were not included because of the complex nature of the crime.

The Uneven Distribution of Obesity Cases, Crimes, and Neighborhood Characteristics

Figure 2a, b shows the distribution of obesity rates and violent crime rates (i.e., the number of crime incidents per 100,000 populations in each spatial unit). The distribution of obesity, as shown in Fig. 2a, is similar to the distribution of population densities. However, the distribution pattern of the violent crime rates, as it is shown in Fig. 2b, is different from the distribution of that of obesity rates which the downtown center often observes the highest crime rate

In addition, neighborhood characteristics, including socioeconomic and environmental variables, were included in the analysis. The socioeconomic data were obtained from the US Census Bureau, and the 5-year estimates of 2014 were selected. The variables include population density, Black percentages, renter-occupied housing percentage, median household income (MHHI, US dollars), and unemployment rates. These variables represent several socioeconomic components of Akron, including racial, housing occupancy, and economic status.

The built environment data were retrieved from the County of Summit GIS Hub and Open Data (http://data-summitgis.opendata. arcgis.com/), including the impervious surface percentage and tree cover percentage per block group and road density.

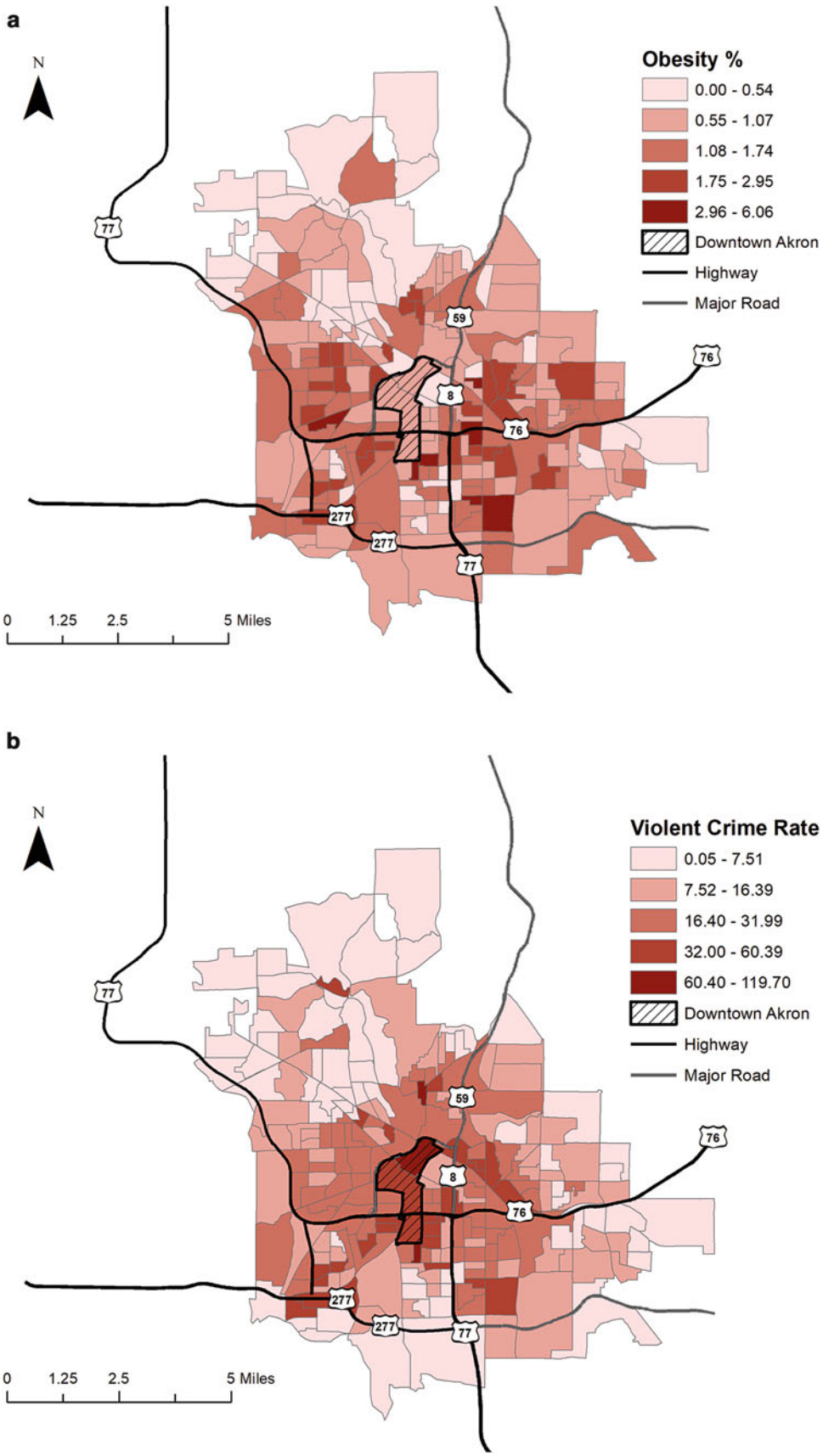


Fig. 2 (**a**) Distribution of obesity rates; (**b**) Distribution of violent crime rates

Data Organization and Software

The obesity percentage of each block group was the dependent variable in the model. It was calculated by first calculating the individual's BMI from the sample. Those individuals who had BMI larger than 30 were considered to be obese and were included in this study. The BMI >30 criterion is according to the definition by the United States Centers for Disease Control and Prevention (https://www.cdc.gov/obesity/adult/defining.html). The number of the obese populations was aggregated to block groups and standardized by dividing the block groups' population counts to derive the obesity percentages (i.e., obesity rates). Independent variables in the model include violent crime rates and SES and environmental attributes of Akron, including Black population percentages, renter-occupied housing percentages, median household income, unemployment rates, impervious surface percentages, tree cover percentages, and road densities.

GWR was performed using GWR4* software (Nakaya 2014; software available at https://gwrtools.github.io/) to build models between violent crime rates and obesity rates. An adaptive Gaussian kernel was selected for building the models, and AICc were used to assess the model's goodness of fit. ArcGIS 10.5 (ESRI, Redlands, CA) was used for mapping the results.

Obesity, Violence, and Neighborhood Characteristics: Regression Analysis Results

OLS Regression Result

Table 1 shows the regression coefficients in the OLS model. Table 1 shows the best models with the subsets of the independent variables. The R^2 and adjusted R^2 of the OLS model are 43.43% and 41.31%, and the minimum AICc value is 421.58. The VIF values for the impervious surface percentages and tree cover rates are higher than 5 and lower than 7.5, which indicate the model being moderately problematic. Other VIF values indicate that there were no problematic levels of multicollinearity. All coefficients other than that of the environmental variables, including impervious surface percentages, road densities, and tree cover rates, were statistically significant at the 5% level.

The OLS model shows that violent crime rates and the Black population percentages are positively related to the obesity rates. Although it cannot be asserted that there exists a causal relationship between crime, race, and obesity, elevated crime rates and Black percentages in a neighborhood may be related to increasing obesity rates as indicated by the model outcome.

Also, median household income, unemployment rates, and renter-occupied housing rates are negatively related to the obesity rates. These results show an inconsistent relationship between the economic status of the neighborhoods and obesity rates as an increase of median household income and an increase of the unemployment rate are related to lower the obesity rate in the neighborhood. Higher unemployment rates and renter-occupied housing rates are typically considered to be reflecting lower SES of a neighborhood, while a higher median household income often indicates a higher SES.

Overall, inconsistent results were found in the OLS model. Both the Koenker (BP) and Jarque-Bera tests are statistically significant, indicating that the OLS model is bias and not reliable. This may be explained as that the OLS model has failed to consider the spatial variations within the relationships (Fotheringham et al. 2002). Therefore, building a GWR model is necessary to analyze the local variations.

Geographically Weighted Regression Results

Four variables entered the GWR model, including the road densities, violent crime rates, Black population percentages, and unemployment rates. The model's AICc value is 402.40,

Table 1 OLS model result

Variable	Coefficient	Std. error	*t*-Statistics	Significance	VIF
Intercept	2.097489	0.465264	4.508174	0.000***	
Impervious surface %	−0.004801	0.007132	−0.673158	0.502	6.691355
Road density	−0.000166	0.000105	−1.577789	0.116	1.767722
Tree cover %	−0.006307	0.007292	−0.864936	0.388	5.301768
Violent crime %	0.027152	0.003438	7.897950	0.000***	1.712260
Black %	0.007354	0.001814	4.054199	0.000***	1.522596
MHHI	−0.000011	0.000004	−3.141825	0.002***	2.750956
Unemployment %	−0.007841	0.003737	−2.098080	0.037**	1.189063
Renter %	−0.008773	0.002532	−3.464068	0.001***	2.388320

Number of observations, 222; AICc, 421.58
R^2, 0.4343; adjusted R^2, 0.4131
**significant at the 99% level; **significant at the 95% level

which is reduced by 19.18 from that of the previous OLS model. The R^2 and adjusted R^2 are 50.00% and 45.78%, which are higher than those of the OLS model. The optimal bandwidth is 50.86.

Figure 3a–d show maps of individual variables' *t*-values, which are presented using graduated colors. A positive *t*-value represents a positive association between the variable and obesity, while a negative value shows otherwise. Locations that have absolute *t*-values higher than 1.96 or 2.58 (or lower than −1.96 or − 2.58), which correspond to the 95% or 99% significance levels, respectively, reveal statistically significant relationships. Mapping is done in ArcMap 10.5. For variables of road densities and unemployment rates, values are divided into five categories based on the level of significance. Due to all locations showing positive associations between violent crime rates/Black population percentages and obesity rates, the *t*-values for these two variables are divided using Natural Break in ArcMap 10.5

As it is shown in Fig. 3a, road densities are negatively related to obesity rates in Akron neighborhoods. Block groups located on the south and east of Akron observed the most robust relationships. However, most block groups in Akron experience no significant relationships between road densities and obesity rates.

Figure 3b shows that the violent crime rates are positively related to obesity rates across the whole study area. Violent crimes in block groups of the south, southwest, and southeast have stronger associations with obesity rates. The race variable is also observed to have a positive association with obesity in the study area, as it is shown in Fig. 3c. Stronger such relationships are mostly concentrated in downtown and north Akron

In addition, the unemployment rates are negatively related to obesity rates. This is shown in Fig. 3d. Significant relationships are found in the south and east sides of Akron, as well as in downtown neighborhoods.

Discussion

Overall, the results of our analysis are consistent with previous research in that locations that experienced higher violent crime rates also experienced higher obesity rates. The results from the GWR models showed better results than those of OLS models. As shown in Fig. 3b, violent crimes showed overall positive associations with obesity rates in the study area. These results are consistent with findings from previous research (Taylor 1995; Stafford et al. 2007; Sandy et al. 2013). The results of the GWR models showed that there were spatial non-stationarity in the associations between obesity, crimes, racial, socioeconomic, and environment variables. Some locations were more vulnerable to violent crime and obesity than others, especially locations in the urban center and the south side of Akron.

In general, it is expected that increasing crime rates, especially in neighborhoods located in the urban areas, may be associated with elevated obesity rates. However, different neighborhoods reported having different degrees of the associations between crime and obesity. Therefore, it is worth to mention that locations that have higher regression coefficients between crime and obesity are not necessarily the locations that have high crime rates and high obesity rates as it is shown in Fig. 2a, b. Locations with higher crime rates and obesity rates are located in the urban areas, in and around neighborhoods adjacent to the urban center. Also, areas that have the highest coefficients are located in the southern part of Akron. These neighborhoods are older urban residential communities where housing values and household incomes are low and renter-occupied rates are relatively lower than national average (around 42% compared to the national average of 36%), according to the US Census Bureau (https://www.census.gov/quickfacts/fact/table/US/PST045218). Nevertheless, renter-occupied percentage was found to not have a significant contribution to obesity rate in the final GWR model. However, to use the variable as the sole explanatory variable, local significant relationships were found. As a result, more detailed research should be done locally to investigate whether more reasons may contribute to poor safety or health situation.

Furthermore, no strong or significant associations are found between environmental variables and obesity, except road densities. However, only some block groups reported significant *t*-values between the variables and obesity. These neighborhoods are located in the south side of Akron, which has low housing values and household income. The result also shows that there are spatial non-stationarity in associations between the variable and obesity. More research should be done to explore the effects and extent of that association.

The racial variable showed to be significantly associated with obesity. As shown in Fig. 3c, block groups located in downtown and the northern neighborhoods show stronger relationships with obesity. These neighborhoods have high housing rental rates and low household income according to the Akron Neighborhood Profiles.

Also, economic variables of unemployment rates are found to be related to obesity in some Black groups. Locations in the eastern part of Akron observe negative associations between unemployment and obesity. Such a result is not consistent with findings from previous research in that low economic status contributes to higher obesity rates (Laitinen et al. 2002). Therefore, solely trying to increase local income levels may not be the best strategy for improving local health status. Instead, more surveys or research should be done on the levels and ways of food consumption and nutrition levels of the residents.

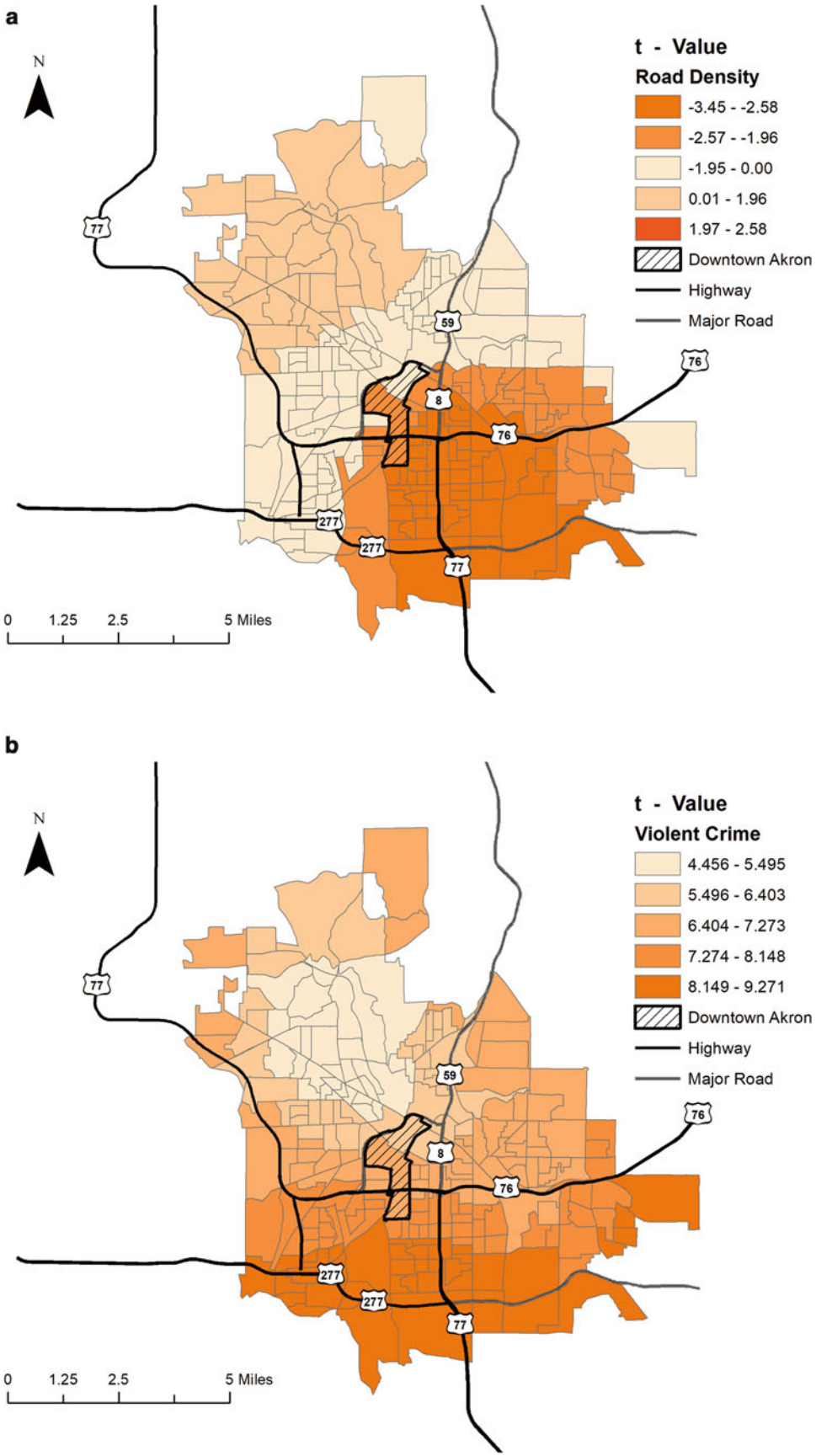
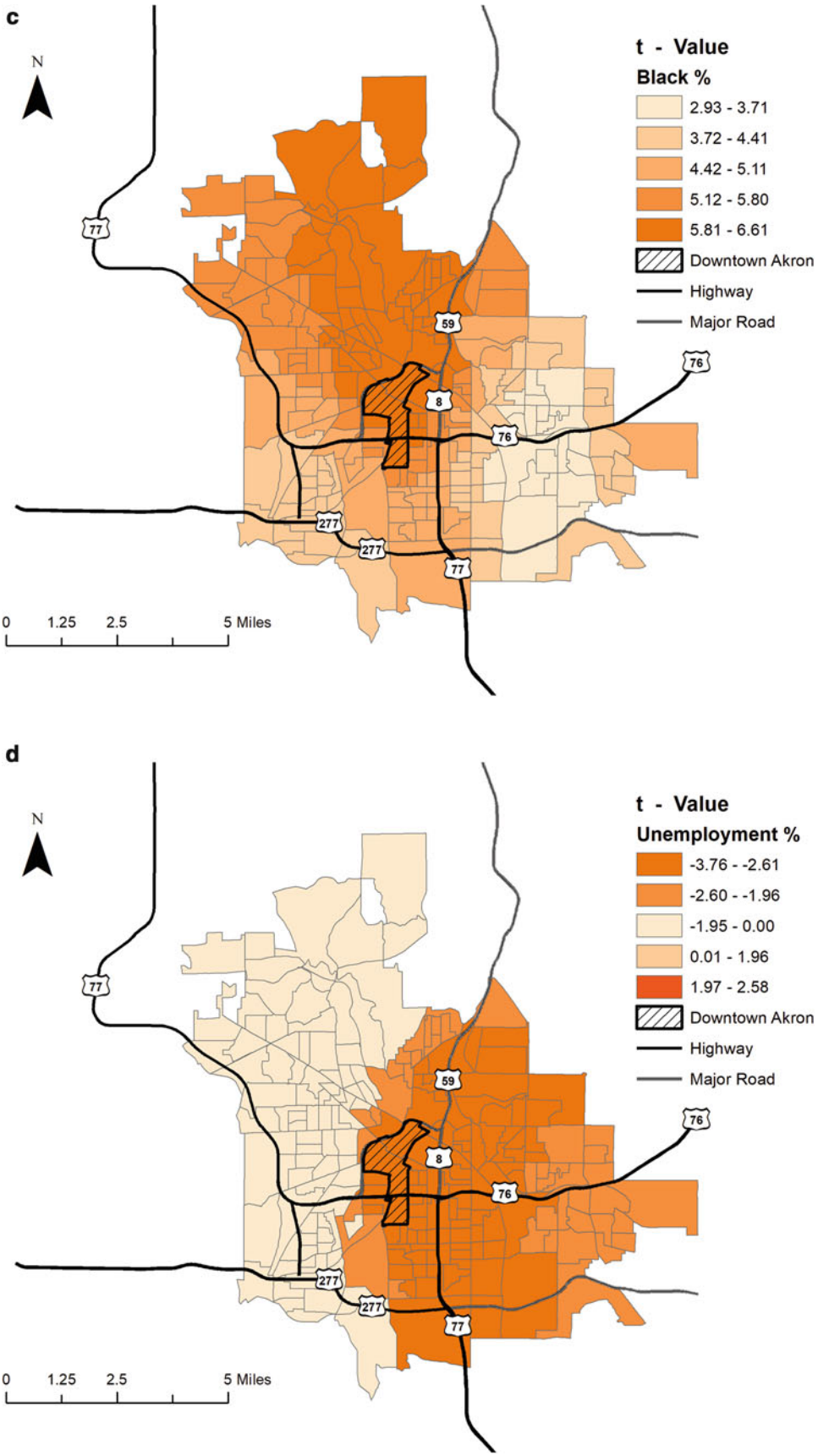


Fig. 3 (**a**) Road density *t*-values; (**b**) violent crime rate *t*-values; (**c**) Black population percentage *t*-values; (**d**) unemployment rate *t*-values

Fig. 3 (continued)



Strength and Weakness

Comparing the results from the OLS model and those from the GWR model, we demonstrated that the GWR model produces statistically significant results that seem to be more practical than those from the OLS model. However, GWR is not a panacea for all model-building problems. Based on what it has been discussed above, this section summarizes some weakness of the method and what users should pay attention to when addressing the issue of spatial non-stationarity.

First, it is necessary to justify whether GWR is needed. The procedure presented in this research shows a combination of the Jarque-Bera test and Koenker (BP) statistics to indicate that the OLS model is bias and unreliable so that applying GWR is preferable to achieve potentially better results. However, although the GWR method produced a significantly better model than the OLS model in this paper, the increase of the overall R^2 is not substantial. Therefore, further investigation and other measures may be taken to improve the explanatory power of the independent variables in the model.

Second, the selection of the bandwidth is crucial in model building. The most common measures are the AICc and CV. Researchers can also assign a bandwidth based on existing literature and their particular research needs. Keep in mind, however, that both fixed and adaptive bandwidths may introduce a certain degree of generalization over the spatial non-stationarity. Ideally, the selected bandwidth not only would have the best statistical results but also is the most meaningful.

Third, the selection of model variables is also critically important. In this research, variables in the final model were selected based on the following criteria. First, the population density was not suitable to be a local predictor as its DIFF of Criterion is greater than 2. Also, impervious surface and tree cover percentages were not statistically significant to explaining the variations in the dependent variable since their local absolute *t*-values were less than 1.96 or 2.58. After removing these three variables, the variable MHHI, renter percentages contribute less to the model as compared to other variables. However, if removing a variable caused little to no impact on the overall R^2, the user should eliminate such variable. Therefore, both variables were not included in the final GWR model.

In summary, as a quantitative method, a user of the GWR model should be careful with the model parameters in order to produce an optimal model and appropriately interpreted the results.

Concluding Remarks

To study into why existing studies found inconsistent results, this article applied OLS and GWR methods to explore the associations between obesity and crime, SES, and environment. In the OLS model, both the Koenker (BP) and Jarque-Bera tests reported significant, which indicate that the global model was not sufficient nor appropriate to explore the associations between crime and obesity. Overall, GWR models revealed spatial non-stationarity in the associations being investigated and produced better results than OLS models.

Violent crimes were found to be generally positively related to obesity. It is worth to note that locations that have already experience higher crime and obesity rates may not have the same effects as other places. Therefore, it is worth to look into specific neighborhoods as of why even though it currently does not have a high crime and obesity rate, it is still vulnerable to a change. Policies such as revitalize or gentrify downtown Akron may attract new investments into downtown, so as to improve local economies. Accordingly, the police department should increase patrol in urban areas to help ensure the safety of neighborhoods.

Both the environmental and SES variables showed local variances in the GWR model. These variances confirm the existence of the spatial non-stationarity among their relationships with local health status. According to the result, even for a single city, the same strategies might not work for all neighborhoods. Policies must be adjusted to target on local situations. Given the spatial non-stationarity concluded in this study, more detailed investigations should be conducted locally so that appropriate measures can be taken to reduce the problems of neighborhood crime and health issues.

Finally, the study reported here is one of the relatively few that look at the associations between violent crime and public health from a quantitative perspective. Results reported here should contribute to our understanding of, spatially, how violence, socioeconomic, and environmental conditions may influence local health.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6): 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Anselin, L. 1995. Local indicators of spatial association' LISA. Geographical Analysis 27 (2):93–115.https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Avila-flores, Diana, Marin Pompa-Garcia, and Xanat Antonio-amiga. 2010. Driving factors for forest fire occurrence in the Durango State of Mexico: A geospatial perspective. 20 (2008): 491–497. https://doi.org/10.1007/s11769-010-0437-x.

Barbu, Ionel. 2012. Econometric study over the arrivals in agrotouristic pensions in the Crişana region. *WSEAS Transactions on Business and Economics*: 290–295.

Brantingham, Patricia, and Paul Brantingham. 1975. Residential burglary and urban form. *Urban Studies* 12 (3): 273–284. https://doi.org/10.1080/00420987520080531.

Brown Barbara, B., Carol M. Werner, Ken R. Smith, Calvin P. Tribby, and Harvey J. Miller. 2014. Physical activity mediates the relationship between perceived crime safety and obesity. *Preventive Medicine* 66: 140–144. https://doi.org/10.1016/j.ypmed.2014.06.021.

Brunsdon, Chris, A. Stewart Fotheringham, and Martin E. Charlton. 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geographical Analysis. 28: 281298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Burdette, Hillary L., and Robert C. Whitaker. 2004. Neighborhood playgrounds, fast food restaurants, and crime: Relationships to overweight in low-income preschool children. *Preventive Medicine* 38 (1): 57–63. https://doi.org/10.1016/j.ypmed.2003.09.029.

Cahill, M., and G. Mulligan. 2003. The determinants of crime in Tucson, Arizona. *Urban Geography* 24 (7): 582–610. https://doi.org/10.2747/0272-3638.24.7.582.

———. 2007. Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review* 25 (2): 174–193.

Carroll-Scott, Amy, Kathryn Gilstad-Hayden, Lisa Rosenthal, Susan M. Peters, Catherine Mccaslin, Rebecca Joyce, and Jeannette R. Ickovics. 2020. Social Science & Medicine Disentangling Neighborhood Contextual Associations with child body mass index, diet, and physical activity : The role of built, socioeconomic, and social environments. *Social Science & Medicine* 95 (2013): 106–114. https://doi.org/10.1016/j.socscimed.2013.04.003.

Chalkias, Christos, Apostolos G. Papadopoulos, Kleomenis Kalogeropoulos, Kostas Tambalis, Glykeria Psarra, and Labros Sidossis. 2013. Geographical heterogeneity of the relationship between childhood obesity and socio-environmental status: Empirical evidence from Athens, Greece. *Applied Geography* 37: 34–43. https://doi.org/10.1016/j.apgeog.2012.10.007.

Chen, Duan-Rung, and Khoa Truong. 2012. Using multilevel modeling and geographically weighted regression to identify spatial variations in the relationship between place-level disadvantages and obesity in Taiwan. *Applied Geography* 32 (2): 737–745. https://doi.org/10.1016/j.apgeog.2011.07.018.

Chi, Sang-Hyun, Diana S. Grigsby-Toussaint, Natalie Bradford, and Jinmu Choi. 2013. Can geographically weighted regression improve our contextual understanding of obesity in the US ? Findings from the USDA Food Atlas. *Applied Geography* 44: 134–142. https://doi.org/10.1016/j.apgeog.2013.07.017.

Cohen, E. Lawrence, and Marcus Felson. 1979. Social change and crime rate trends: A routine activity american sociological review. *American Sociological Review* 44 (4): 588–608.

Comber, Alexis J., Chris Brunsdon, and Robert Radburn. 2011. A spatial analysis of variations in health access: Linking geography, socio-economic status, and access perceptions. *International Journal of Health Geographics* 10 (1): 1–11.

Deng, Chengbin. 2015. Integrating multi-source remotely sensed datasets to examine the impact of tree height and pattern information on crimes in Milwaukee, Wisconsin. *Applied Geography* 65: 38–48. https://doi.org/10.1016/j.apgeog.2015.10.005.

Fan, Maoyong, and Yanhong Jin. 2014. Do neighborhood parks and playgrounds reduce childhood obesity? *American Journal of Agricultural Economics* 96 (1): 26–42. https://doi.org/10.1093/ajae/aat047.

Fotheringham, A.S., M. Charlton, and C. Brunsdon. 1996. The geography of parameter space: An investigation of spatial non-stationarity. *International Journal of Geographical Information Systems* 10 (5): 605–627. https://doi.org/10.1080/026937996089021 00.

Fotheringham, A.S., C. Brunsdon, and M. Charlton. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. West Sussex: Wiley.

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24(3):189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Gilbert, Angela, and Jayajit Chakraborty. 2011. Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. *Social Science Research* 40 (1): 273–286. https://doi.org/10.1016/j.ssresearch.2010.08.006.

Halleröd, Björn, and Daniel Larsson. 2008. Poverty, welfare problems, and social exclusion. *International Journal of Social Welfare* 17 (1): 15–25. https://doi.org/10.1111/j.1468-2397.2007.00503.x.

Huang, Xi, Christian King, and Jennifer McAtee. 2018. Exposure to violence, neighborhood context, and health-related outcomes in low-income urban mothers. *Health and Place* 54: 138–148. https://doi.org/10.1016/j.healthplace.2018.09.008.

Jarque, C.M., and A.K. Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economic Letters* 6: 255–259.

Laitinen, J., C. Power, E. Ek, U. Sovio, and M.R. Järvelin. 2002. Unemployment and obesity among young adults in a northern Finland 1966 birth cohort. *International Journal of Obesity* 26 (10): 1329–1338. https://doi.org/10.1038/sj.ijo.0802134.

Nakaya, T. 2014. GWR4 User Manual. Available online: https://www.st-andrews.ac.uk/geoinformatics/wpcontent/uploads/GWR4manual_201311.pdf

Nakaya, T., A.S. Fotheringham, C. Brunsdon, and M. Charlton. 2005. Geographically weighted poisson regression for disease association mapping. 2004: 2695–2717. https://doi.org/10.1002/sim.2129.

Ortolano, Gaetano, Roberto Visalli, Gaston Godard, and Rosolino Cirrincione. 2018. Quantitative X-Ray Map Analyser (Q-XRMA): A new GIS-based statistical approach to mineral image analysis. *Computers & Geosciences* 115: 56–65. https://doi.org/10.1016/j.cageo.2018.03.001.

Ruijsbroek, Annemarie, Alet H. Wijga, Ulrike Gehring, Marjan Kerkhof, and Mariël Droomers. 2015. School performance: A matter of health or socio-economic background? Findings from the PIAMA birth cohort study. *PLoS One* 10 (8): e0134780. https://doi.org/10.1371/journal.pone.0134780.

Rybarczyk, Greg, Alex Maguffee, and Daniel Kruger. 2015. Linking public health, social capital, and environmental stress to crime using a spatially dependent model. *City* 17 (1): 17–33.

Salois, Matthew J. 2012. The built environment and obesity among low-income preschool children. *Health and Place* 18 (3): 520–527.https://doi.org/10.1016/j.healthplace.2012.02.002.

Sandy, Robert, Rusty Tchernis, Jeffrey Wilson, Gilbert Liu, and Xilin Zhou. 2013. Effects of the built environment on childhood obesity: The case of urban recreational trails and crime. *Economics and Human Biology* 11 (1): 18–29. https://doi.org/10.1016/j.ehb.2012.02.005.

Shahid, Rizwan, and Stefania Bertazzon. 2015. Local spatial analysis and dynamic simulation of childhood obesity and neighbourhood walkability in a Major Canadian City. *AIMS Public Health* 2 (4): 616–637. https://doi.org/10.3934/publichealth.2015.4.616.

Singh, Gopal K., Michael D. Kogan, and Peter C. Van Dyck. 2008. A multilevel analysis of state and regional disparities in childhood and adolescent obesity in the United States: 90–102. https://doi.org/10.1007/s10900-007-9071-7.

Stafford M., Chandola, T., and Marmot, M. 2007. Association between fear of crime and mental health and physical functioning. Am J Public Health, 97 (11):2076–2081. https://doi.org/10.2105/AJPH.2006.097154

Thadewald, Thorsten, and Herbert Büning. 2007. Jarque–Bera test and its competitors for testing normality – A power comparison. *Journal of Applied Statistics* 34 (1): 87–105. https://doi.org/10.1080/02664760600994539.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46 (Supplement): 234–240.

Troy, A. J., Grove, M, and O'Neil-Dunne, J. 2012. The relationship between tree canopy and crime rates across an urban-rural gradient in the greater baltimore region. Landscape and urban planning, 106 (3): 262–70. https://doi.org/10.1016/j.landurbplan.2012.03.010

Taylor, R. B. 1995. The impact of crime on communities. The annals of the American academy of political and social science, 539, 28–45.

Vandewater, Elizabeth A., Mi-suk Shim, and Allison G. Caplovitz. 2004. Linking obesity and activity level with children' s television and video game use. 27: 71–85. https://doi.org/10.1016/j.adolescence.2003.10.003.

Villarraga, Hernán G., Albert Sabater, and Juan A. Módenes. 2014. Modelling the spatial nature of household residential mobility within municipalities in Colombia. *Applied Spatial Analysis and Policy* 7 (3): 203–233. https://doi.org/10.1007/s12061-014-9101-7.

Wallace, Brian. 2011. Geographic information systems correlation modeling as a management tool in the study effects of environmental variables' effects on cultural resources.

Wen, Tzai-hung, Duan-rung Chen, and Meng-Ju Tsai. 2010. Identifying geographical variations in poverty-obesity relationships : Empirical evidence from Taiwan. *Geospatial Health* 4 (2): 257–265.

Yang, Tse-Chuan, A. Stephen, and Matthews. 2012. Health & place understanding the non-stationary associations between distrust of the health care system, health conditions, and self-rated health in the elderly : A geographically weighted regression approach. *Health & Place* 18 (3): 576–585. https://doi.org/10.1016/j.healthplace.2012.01.007.

Challenges of Assessing Spatiotemporal Patterns of Environmentally Driven Infectious Diseases in Resource-Poor Settings

Alina M. McIntyre, Karen C. Kosinski, and Elena N. Naumova

Introduction

Modern geographic information systems (GIS), global remote sensing data repositories, and spatial analytic tools are enabling better understanding of complex patterns of environmentally driven, climate-sensitive, and often preventable infections. Waterborne and water-related infections are among the most common infectious diseases in low-income countries, yet the path for their prevention is achievable by meeting the SDGs. Diarrheal infections are diseases of poverty aggravated by lack of proper infrastructure such as sanitation facilities, sanitary landfills, and improved water sources. These diseases typically exhibit strong associations with climate, meteorological, and environmental parameters. Climate-sensitive diseases tend to show pronounced seasonal patterns that repeat annually. Some agents, like rotavirus and *Cryptosporidium*, are well known for their universal fluctuations over the course of a year, associated with the changes in temperature and precipitation (Jagai et al. 2009, 2012a). Successful intervention programs aiming to improve water treatment, sanitation, and hygiene and to prevent diseases with vaccination could result in changes of temporal patterns that are manifested by the reduced intensity of seasonal spikes. Many water-related diseases exhibit strong seasonal patterns that are distinct for each pathogen in a given population and locality. Disturbances in human-environment interactions due to emerging novel pathogens, viral mutations, drug resistance, and severe weather episodes might affect the timing and intensity of infectious outbreaks. Therefore, maps of climate-sensitive diseases with strong seasonality can be dramatically different in different seasons.

In this chapter, we focus on two preventable diseases, with both high burden and potentially debilitating consequences, to illustrate the challenges of mapping disease incidence and their risk factors using nationally representative records. Urogenital schistosomiasis (UGS), caused by *S. haematobium*, is a notoriously focal disease with substantial spatial heterogeneity (Ekpo et al. 2008); it is a water-related disease that is highly endemic in Ghana (Adenowo et al. 2015). Diarrheal disease prevalence and severity are highest in infants and often associated with poverty, overcrowding, and inadequate sanitation.

In order to understand the spatiotemporal patterns of environmentally driven diseases, at least three types of information are needed: (1) disease records, (2) demographic data, and (3) environmental data. Disease records corrected for population density enable the comparison of disease incidence across geographical regions, patient profiles, and time periods. Compositions of sociodemographic and environmental indicators help to identify risk factors to mark locations, seasons, and groups that require further public health focus to minimize health risks and reduce disease burden.

In the data-rich settings of high-income countries, studies that used nationally representative data from multiple data streams to examine health outcomes and their environmental risks are common and continue to improve with respect to methods and rigor (Ayanian et al. 1993; Hajak 2001; Cummins et al. 2005; Joy et al. 2008; Sallis et al. 2009). This type of analysis has been enabled by the wide use of GIS-based platforms and systems. One of the most widely used GIS platforms in the USA originates from software created by

A. M. McIntyre · K. C. Kosinski
Department of Community Health, School of Arts and Sciences, Tufts University, Medford, MA, USA
e-mail: amcintyr@bu.edu; Karen.Kosinski@tufts.edu

E. N. Naumova (✉)
Division of Nutrition Epidemiology and Data Science, Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, USA
e-mail: elena.naumova@tufts.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_17

Environmental Systems Research Institute (ESRI) at Harvard Labs in 1969, which then expanded to develop commercial GIS products for wider use and spatial analysis applications (ESRI 2019). GIS was developed to "store, visualize, analyze, and interpret geographic data," and it has become routine for many applications to study health issues spatially. In the health fields, GIS allows researchers to examine how location relates to factors such as sociodemographic status, health infrastructure, and health outcomes data (Centers for Disease Control and Prevention (CDC) 2016). In addition to mapping the size and scope of various health outcomes, GIS allows the user to conduct risk assessments, disease impact simulations, and pinpoint areas for health interventions and policy (Fradelos et al. 2014). With the growing supply of data available for spatial analysis, GIS technology is also growing in the quality and the breadth of applications, and researchers must stay current on platforms and analysis techniques updates. GIS-based platforms have become an integral part of nationally representative data repositories worldwide.

Census records are invaluable repositories of social, demographic, economic, and environmental records. The United Nations (UN) defines a census as "the total process of planning, collecting, compiling, evaluating, disseminating and analyzing demographic, economic and social data at the smallest geographical level pertaining, at a specified time, to all persons in a country or in a well-defined part of a country" (Mrkić n.d.). In addition, the UN emphasizes the general importance of conducting a population census: the data provides reliable, official statistics for government and the public to use to eventually make conclusions on the distribution of wealth, government services, and political representation in a country or region (Mrkić n.d.). Data are usually collected through a questionnaire distributed in person, in the mail, or electronically. Many countries, including the USA, collect census data every 10 years; some countries (e.g., Japan, New Zealand, Canada) conduct a census every 5 years (Mrkić n.d.). Census data is invaluable in assessing population characteristics. However, even in resource-rich settings, census information may not capture groups with low literacy rates, groups living in remote areas, and groups that are nomadic or experiencing homelessness (Bartley 2001).

In the data-sparse settings of low-income countries, census records are sometimes the only data available for a nationwide analysis (Arku et al. 2016; Barcus et al. 2007; Varenne et al. 2004; Timæus and Jasseh 2004; Ferraz et al. 2017). A small number of studies have matched census records with health data to assess relationships among diseases and various environmental and sociodemographic variables. For example, Barcus et al. (2007) showed, through the use of sociodemographic census data and malaria diagnoses in hospital records, that there was a higher risk of higher-grade parasitemia and severe malaria with fatal outcomes among the urban residents of Papua New Guinea than among the rural residents. In sub-Saharan Africa as a whole, Timæus and Jasseh (2004) used sociodemographic census data in combination with HIV mortality data from 26 Demographic and Health Surveys to find that excess mortality occurs among women ages 25–39 and men ages 30–44. In Ghana, Arku et al. (2016) used random sampling from the 2010 Population and Housing Census to examine under-5 mortality in comparison to various sociodemographic variables using Bayesian spatial analysis and demonstrated that higher use of liquefied petroleum gas (LPG) as cooking fuel was associated with lower under-5 child mortality after adjusting for other mortality risk factors. This increased use in LPG and associated decrease in mortality could be due to a move away from solid fuels responsible for household air pollution. Findings also indicate that even though under-5 mortality has decreased, the cross-district inequality in mortality has increased (Arku et al. 2016).

Data-Rich Environments: Challenges and Lessons

Large-scale studies have demonstrated the power of verified, uniformly, and routinely collected data in conducting epidemiological investigations at national and global levels. International organizations, like the World Health Organization (WHO), collect and disseminate health records with global coverage. In the USA, several data repositories, hosted by the US Centers for Disease Control and Prevention (CDC) and the Centers for Medicare and Medicaid Services (CMS), provide exhaustive national coverage. For several decades, the CMS has maintained the Medicare Provider Analysis and Review (MEDPAR) files with extensive individual information on patient age, sex, race, residence ZIP code, health provider, diagnostic codes, dates of admission, discharge and follow-up procedures, and total hospitalization charges. The CDC regularly publishes records of mandatory reported infections based on patient age, state, and week of confirmed or provisional infections. The value of national repositories increases dramatically when researchers amend location-specific information with GIS-based tools.

Using CMS records (15 years of ZIP code-based daily demographic, environmental, and 600 M individual entries) allowed us to develop models and visualization tools for tracking the spread in vulnerable populations of enteric infections (Chui et al. 2009, 2011a; Cohen et al. 2008; Jagai and Naumova 2009; Jagai et al. 2010, 2012b; Naumova et al. 2007; Mor et al. 2009), pneumonia and influenza (P&I) (Chui et al. 2011b; Cohen et al. 2010, 2011; Lofgren et al. 2007, 2010; Moorthy et al. 2012; Mor et al. 2011; Naumova et al. 2009), and health conditions resulting from exposure to extreme weather (Liss et al. 2017). These studies show

the importance of understanding the uncertainties related to place of residence (PoR) and place of health care (PoH) access in developing maps of disease incidence. In our study of hospitalizations due to P&I in patients with cognitive impairment, we hypothesized that access to care is limited by financial constraints, insurance status, individual preferences and perceptions, as well as geographic proximity, travel time, and assistance with transportation. Elderly people with cognitive impairment are at an elevated risk for infection-associated complications due to limited ability to communicate health problems, which may increase delay in special care delivery. Time between onset of symptoms that causes hospitalization and initiation of specialized medical care is an important factor for preventing severe outcomes from infections. Reliable estimates of such measures are difficult at a national and even at a regional scale. However, proxies like geographic distance, average travel time, or a minimum distance to a healthcare facility can now be derived using novel GIS technologies. We linked hospitalization records using the PoR and the PoH and estimated linear and network distances in rural and urban settings. Rural and poor communities had the highest rates of hospitalizations due to P&I, and moreover, P&I patients with dementia had a death rate 1.5 times higher than national averages (Naumova et al. 2009). These results suggest strong disparities in healthcare practices in rural locations and among vulnerable populations. Thus, spatial and temporal data were able to demonstrate that infrastructure, proximity, and access to proper care are significant predictors of P&I morbidity and mortality.

There is a need to better understand how traditional maps of seasonal flu could depend on human migration patterns with respect to seasonally changing point of care. Little has been done to address this issue yet. Specifically, we have shown that spatiotemporal hospitalization patterns of P&I fluctuate due to seasonal population migration in older adults (Chui et al. 2011b). This group experiences the most severe morbidity from influenza and also has the highest rates of seasonal migration within the USA. When we classified hospitalizations by state of residence, provider state, and date of admissions and compared the hospitalization profile data of Florida residents with that of out-of-state residents by state of primary residence and time of year (in season or out of season), we observed distinct seasonal patterns of nonresident P&I hospitalizations. This pattern was especially evident when comparing typical winter destination states, such as California, Arizona, Texas, and Florida, to other states. Although most other states generally experienced a higher proportion of non-resident P&I during the summer months (April–September), southern states had higher non-resident P&I during the traditional peak influenza season (October–March) (Chui et al. 2011b).

When researchers integrate information from various sources, new questions and hypotheses can be postulated and answered. We recently outlined a framework to evaluate state foodborne and waterborne surveillance systems using hospitalization records (Mor et al. 2014). Using a Bayesian modeling approach, we generated smoothed standardized morbidity ratios (SMR) and surveillance-to-hospitalization ratios (SHR) and compared predicted values to the observed surveillance counts and the number of hospitalized cases, respectively. We then identified municipalities that deviated from the norm and flagged them for potential uncertainties (Fig. 1). For each studied infection (*Campylobacter, Cryptosporidium*, *Giardia*, hepatitis A, non-typhoid *Salmonella*, *Salmonella typhi*, and/or *Shigella*), we examined and related the spatial distribution of SHR to the mean for the entire state adjusted for population age-structure. Our study confirmed that the spatial "signal" depicted by surveillance was influenced by inconsistent testing and reporting practices, since municipalities that reported fewer cases relative to the number of hospitalizations had a lower relative risk, as estimated by SMR. We made a first estimate of the degree of completeness and coherence of two major national data sources—CMS medical claims and CDC laboratory confirmed records—and we outlined the first step toward the harmonization and integration of related data streams derived from different sources (Mor et al. 2014).

With the growing availability of detailed records from surveillance systems and health providers, researchers are making commendable attempts to combine various data sources to improve primary data collection, validate findings, and better understand the burden and distribution of diseases. Alignment of the records in both temporal and spatial scales allows detailed examination of associations between health outcomes and environmental exposures, including the effects of air and water pollution, extreme weather, and natural disasters (Basu and Samet 2002; Curriero et al. 2002; Keatinge and Donaldson 2001; Conlon et al. 2011; Bhaskaran et al. 2010; Rogot and Padgett 1976; Liu et al. 2011; Michelozzi et al. 2007). The current work in spatiotemporal modeling reveals the gap between the desired data quality needed to build reliable maps and the existing infrastructure and the nature of complex health-host-agent-environment interactions. The key challenge relates to the uncertainties in places of exposure (PoE), PoR, and PoH and the way that researchers justify conceptually and analytically fundamental assumptions crucial for selecting and implementing appropriate study designs, data analysis techniques, visualization tools, and statistical inferences. A common task of detecting hot spots and linking them to exposures often suffers from the inability to decouple PoR from PoE. For example, in detecting hot spots of waterborne outbreaks, these challenges can manifest by a) missing or incomplete information about specific PoE; b) uncertainty in time and source of exposure due to intermittent and infrequent public water supply; and c) lack of geocoded maps

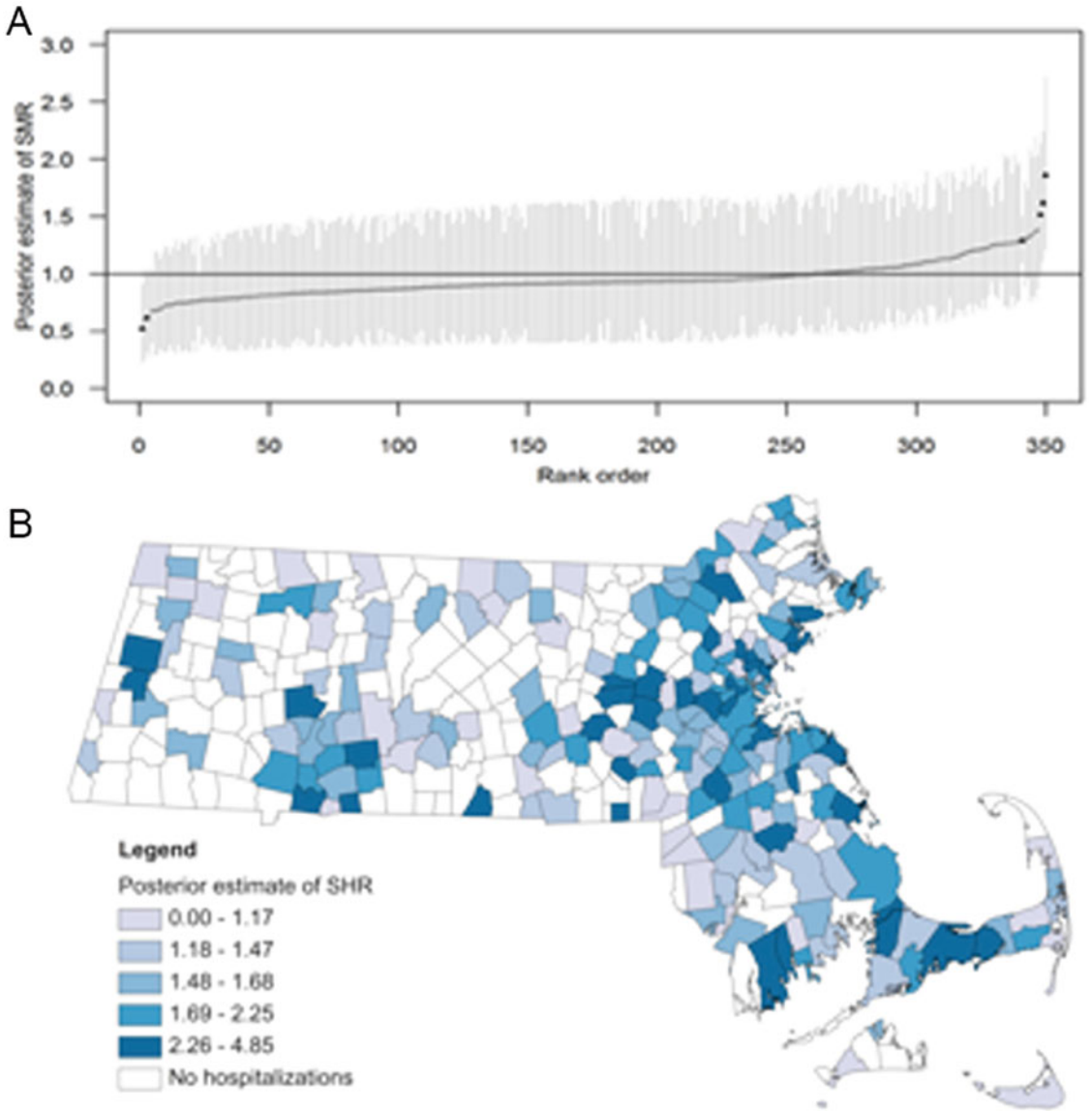


Fig. 1 Surveillance to hospitalization ratio (SHR) for salmonellosis [44]. (**a**) Point estimates and credible intervals for posterior estimates are shown for each municipality (grey) and for the entire state (horizontal line). Municipalities that had significantly lower SHR compared to the state mean are indicated with a black triangle. (**b**) Municipalities are categorized according to quintiles; with the lower quartile being suspect of underreporting

of water distribution networks and information on network integrity. The challenge of disentangling the PoE, PoR, and PoH is universal for both data-rich and data-sparse settings, although the magnitude of such a challenge is undoubtedly higher for low-income countries.

Data-Sparse Environments: Challenges and Opportunities

In data-sparse environments, national health monitoring systems are often expensive and sometimes cost-prohibitive, but the lack of data and poor data quality are major barriers to making population-level conclusions (Bhalla et al. 2010; Pandey et al. 2010; National Research Council 2012). Lack of data prevents targeted disease control programs, limits the assessment and effectiveness of health interventions, and limits long-term health-related investments. Yet many low-income counties are creating and supporting infrastructure to collect and process vital information. In many cases, these efforts are aligned with meeting the SDGs. Multiple SDGs focus on reducing diseases of poverty and diseases associated with poor water, sanitation, and hygiene.

Ghana serves as an inspiring example of building a robust data infrastructure using geospatial technologies. Our work in Ghana began in 2007 and focused on UGS and

water infrastructure. UGS is highly endemic in many parts of Ghana, but when we began our research, there were no updated data about which communities were at the greatest risk of serious pathology. Mass drug administration (MDA) had not yet been widely and regularly implemented due to the cost of praziquantel, and data about water and sanitation infrastructure in rural communities were outdated. Since 2007, MDA has broadly scaled up, and several studies have been done to estimate UGS prevalence in Ghana.

Over the last 12 years, we have conducted numerous field studies to gather detailed information on 75 communities in the Eastern Region of Ghana (Fig. 2). Because of the granularity of the data, in some cases with repeated measures, we have been able to look very carefully at both UGS and water infrastructure. We have studied incidence and prevalence of UGS by age and demographic group, key attributes of common diagnostic tests for UGS, the seasonality of UGS through time series analyses with remote sensing data, and the role that surface water and water infrastructure play in terms of UGS risk. Our research in Ghana has demonstrated the need for accurate maps to guide primary prevention programs for UGS (Kosinski et al. 2011a, b, 2012, 2016a, b; Kulinkina et al. 2017, 2019; Wrable et al. 2019). Disease control strategies still need to be supplemented with measures capable of interrupting transmission, such as water, sanitation, and hygiene (WASH) infrastructure (e.g., wells, swimming pools, washing stations); these infrastructure improvements must be cost-effective and sustainable, and they must also be correctly targeted to the communities that need, want and stand to benefit from them.

In order to expand this kind of research to a national scale, scientists, public health professionals, and practitioners have to compile health records based on a national disease monitoring system, census records, and environmental data. This step allows researchers to construct risk maps and predict potential hot spots of diseases at the national and regional levels. The GIS-based tools help to determine towns' networks, connectivity of and proximity to local health centers, sanitation, and water infrastructures. With data depicting land use, land cover, climate conditions, favorable conditions for vectors and hosts, and water infrastructure, researchers can examine spatial heterogeneity of climate-sensitive diseases, like UGS and diarrhea, and determine environmental predictors that govern disease variability. While not many low-income countries have a well-developed infrastructure to monitor disease, with the advancement of geospatial technologies, it is possible to accelerate technological leapfrogging, which usually occurs in three stages by (a) importing and absorbing highly modern technology; (b) replicating, producing, and improving the imported technology; and (c) moving on to innovations on one's own (Bhagavan n.d.). We conclude this chapter with suggestions about how to facilitate this process and accelerate the transition based on our experience in Ghana.

Disease Mapping: Data Sources and Supporting Platforms

In this section, we describe the national-level datasets that originate from different sources, including government agencies and academic institutions. We illustrate the use of two major national data repositories: the 2010 Population and Housing Census and the District Health Information Management System (DHIMS) maintained by Ghana Health Service (GHS). We also outline the role of the national platform and services that enables the development and investigation of spatiotemporal patterns in environmentally driven climate-sensitive infections.

District Health Information Management System (DHIMS) in Ghana

In Ghana, DHIMS offers a national repository of health records, supported by GHS. The data that feeds into DHIMS are acquired from individual clinics in Ghana that report monthly disease counts to GHS (President's Malaria Initiative 2014). After relying primarily on a paper-based system, GHS received support from the United States' President's Malaria Initiative (PMI) in 2012 to progress to a more comprehensive web-based system (President's Malaria Initiative 2014). Now, DHIMS is a free open-source health management data platform used by many organizations and governments (USAID 2016). DHIMS is supported and funded through various partnerships, including USAID, the Korean International Cooperation Agency, and Samsung Corporation (USAID 2016). Currently, these partnerships are attempting to implement a Community-based Health Planning and Services (CHPS) e-tracker to improve data collection quality and practices. Specifically, Samsung intended to provide digital tablets for CHPS needs, and USAID offers technical assistance for the e-tracker as well as training for CHPS health officers (USAID 2016). GHS executed DHIMS, which has substantially increased the quantity and quality of health data acquired in Ghana (President's Malaria Initiative 2014). DHIMS was our source for UGS and diarrheal disease monthly counts.

CERSGIS Data from Ghana

The Centre for Remote Sensing and Geographic Information Services (CERSGIS) provides GIS and remote sensing services for sustainable development and resource management in Ghana (Centre for Remote Sensing and Geographic Information Services (CERSGIS) 2019). CERSGIS is affiliated with the Department of Geography and Resource Development at the University of Ghana, Legon, and currently oper-

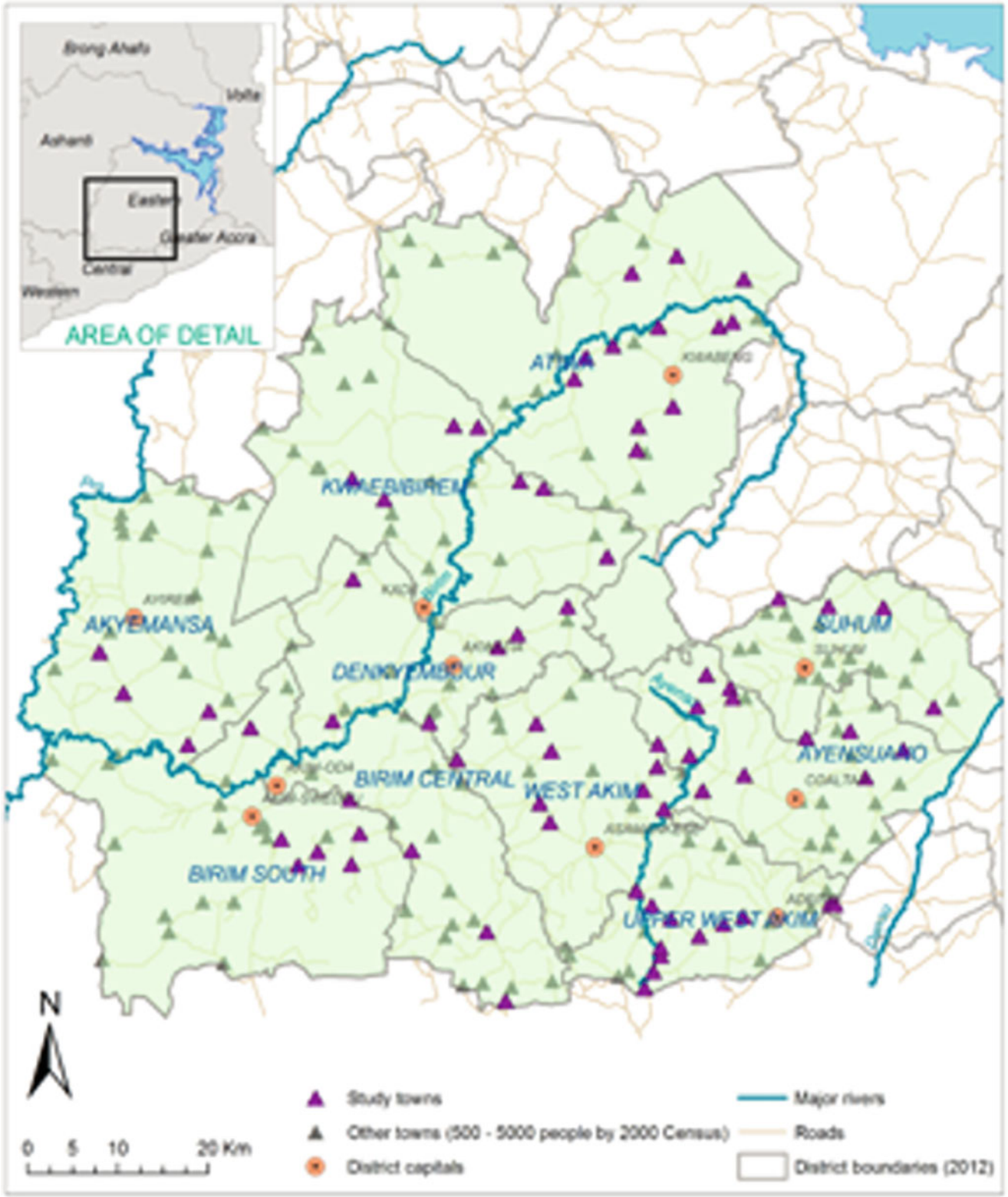


Fig. 2 Map of towns with pop. range 500 – 5,000; major rivers and roads are shown; 75 study towns are indicated with purple triangles (A. Kulinkina, personal communication)

ates as a non-profit organization. CERSGIS offers access to various spatial data, including point features (e.g., buildings, facilities), line features (e.g., roads, routes), and polygon features (e.g., districts, regions) (Centre for Remote Sensing and Geographic Information Services (CERSGIS) 2019). In addition to an outline of their available services (remote sensing, GIS and GPS, web application development, and training sessions), the CERSGIS website also has a published catalog of all projects conducted by its staff. Links to project information and spatial resources can be found, but most spatial data must be requested through contact information provided on the web. Data from CERSGIS enabled the analyses at the household, community, regional, and national levels.

Ghanaian Census Data from the Ghana Statistical Service

The Ghana Statistical Service (GSS) is a governmental organization responsible for censuses, surveys, and various sociodemographic data necessary for Ghana's development in both the public and private sectors (Ghana Statistical Service 2019a). GSS has 10 regional offices and >100 district offices throughout the country. At the time of our writing this chapter, GSS was in the process of redeveloping their website, where they publish raw data and collection methods along with their data acquisition goals. A recent, large-scale endeavor by GSS was the 2010 Population and Housing Census. All 216 Ghanaian districts were surveyed based on recommendations from the United Nations Principles and Recommendations Report for countries conducting censuses, with the objective "to provide information on the number, distribution and social, economic and demographic characteristics of the population of Ghana necessary to facilitate the socio-economic development of the country" (Ghana Statistical Service 2014). Key census topics adapted from UN recommendations include migration patterns, sociodemographic characteristics, infrastructure data, and education and economic status (Ghana Statistical Service 2014). GSS also added various topics they found important to Ghana such as religion, detailed housing and agricultural information, and cooking habits (Ghana Statistical Service 2014). For the 2010 Population and Housing Census, data were collected using a systematic sampling method of every tenth private dwelling in Ghana (Ghana Statistical Service 2013). GSS reports the total sample size as 2,466,289 households. GSS and the Ministry of Education distributed questionnaires, and enumerators were encouraged to use local languages (Ghana Statistical Service 2013).

We produced population density map(s) for Ghana and maps of UGS and diarrheal disease using census records and DHIMS data (Fig. 3). We also show population density by district for people classified as living in "rural" areas. As expected, the population varies by district (Fig. 3a), and percent of the population characterized as rural is greater in northern, northeastern, and southwestern areas of Ghana (Fig. 3b). Census data show that households average five persons, unemployment varied by district ranging between 1% and 7%, and most individuals are reported as literate (Table 1).

In order to map disease rates from January 2012 to December 2016, we normalized district-specific monthly counts through a multi-step process. The number of days in each month were recorded, accounting for leap years. Populations of the 216 districts were projected from the 2010 census data for each study year (2012–2016) using intercensal population growth rates (Codjoe et al. 2013). Disease counts were normalized by days (accounting for leap year) and population and then multiplied by 100,000 to show monthly disease rates. Diarrheal disease rates appear to be generally high throughout the country, especially in western areas of Ghana. UGS rates appear greater in some northern, central, and southern areas, most likely in rural regions of Ghana (Fig. 3c, d). Summary statistics for diarrheal disease and UGS show that district-level rates per 100,000 population over a 5-year period vary tremendously (12–130,613 cases for diarrheal disease; 0.5 to 5653 cases for UGS) (Table 1). The rates are also fluctuating over time exhibiting occasional spikes and potential for upward and downward trend for diarrheal disease and UGS, respectively (Fig. 4).

Spatiotemporal Predictors of Diarrheal Disease and UGS as Collected by Ghana Census

We focused our recent research on variables from the 2010 Population and Housing Census in Ghana that could be linked plausibly to environmental health: solid waste disposal, sanitation, drinking and domestic water, energy sources (lighting source and cooking fuel), literacy, and employment. These characteristics provide a general overview of sociodemographic factors and infrastructure in Ghana; they are also important predictors of diarrheal infections and UGS. We emphasize potential health risks for diarrheal disease and/or UGS based on the peer-reviewed literature and SDG expectations, and we highlight definitions, methods, and relevance of indicators while providing district-specific maps for selected variables.

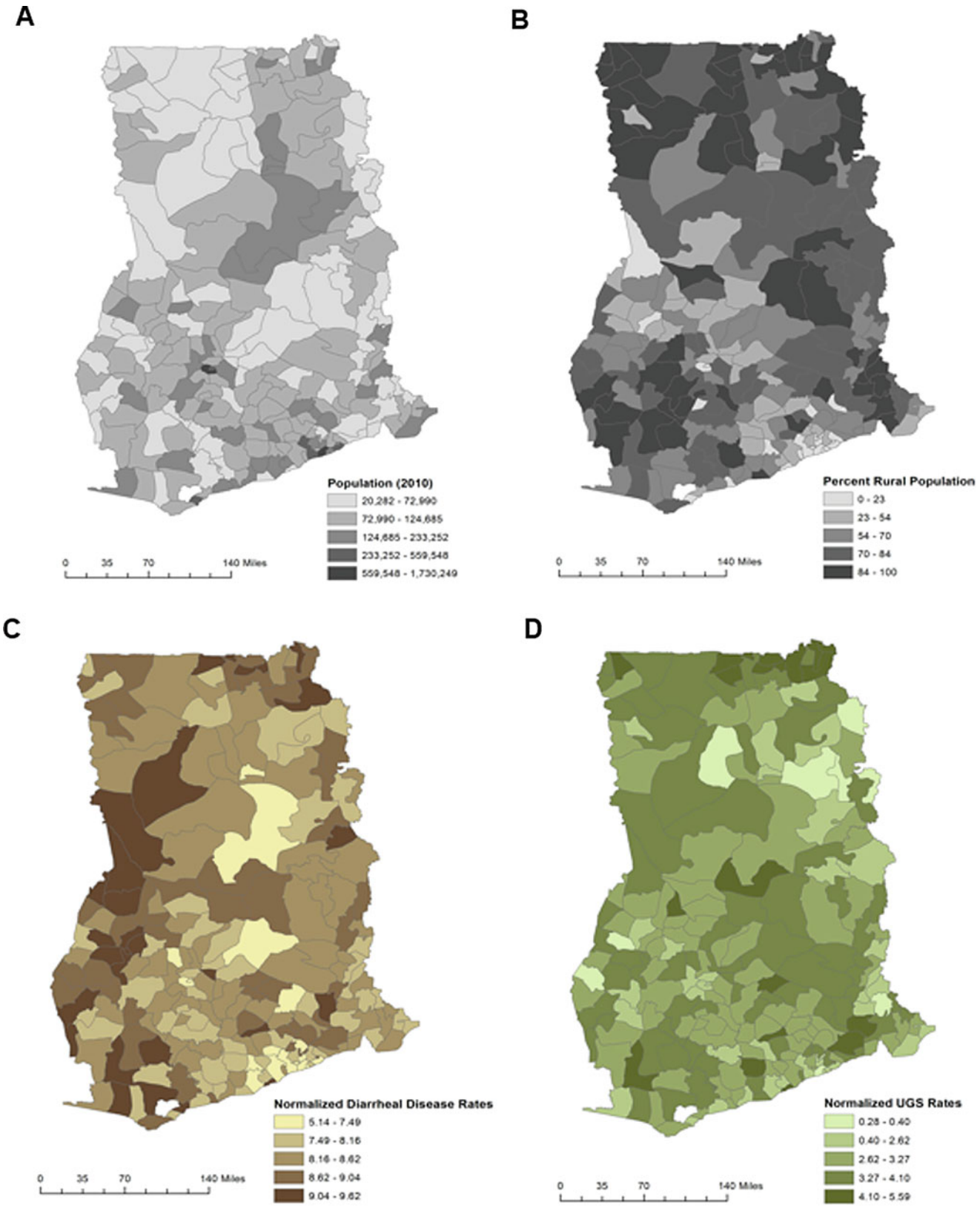


Fig. 3 Population density in Ghana and rates of diarrheal disease and urogenital schistosomiasis. (**a**) Population density for all 216 districts in Ghana according to data from the 2010 Population and Housing Census. (**b**) Population density by district for people classified as living in "rural" areas according to the 2010 Population and Housing Census. (**c**) Distribution of population-normalized, natural log-transformed diarrheal disease rates for January 2012–December 2016 for all 216 districts in Ghana (**d**) Distribution of population-normalized, natural log-transformed UGS rates for January 2012–December 2016 for all 216 districts in Ghana

Solid Waste Disposal

Solid waste disposal is generally measured and captured in census reports at the household (residential) or commercial/industrial level (World Bank 2018). Many methods are used to determine practices, but survey or self-reported data are often used for household-level disposal (World Bank 2018). The US Environmental Protection Agency (US EPA) defines waste as "any discarded, rejected, abandoned, unwanted or surplus matter whether or not intended for sale or for recycling, reprocessing, recovery or purification" (Samwine 2017). Solid waste consists of "all unwanted or discarded materials arising from both human and animal activities" (United States Environmental Protection Agency 2009). Solid waste is addressed by SDG 6.3, which states that it is necessary to "work to eliminate dumping and minimize release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increase recycling and safe reuse globally" (United Nations General Assembly 2015). Inappropriate solid waste disposal practices, such as open dumping of hazardous materials, can pose serious health threats. If dumping areas lack a proper covering material, they can become a breeding ground for scavenging animals, rodents, and disease vectors (Agyepong 2010). In addition, locations where solid waste is placed can produce gasses like methane, carbon dioxide, ammonia, and hydrogen sulfide, contributing to poor air quality (Mata-Alvarez 2002). Many commonly used disposal approaches create environmental hazards to both the natural ecosystem and the local community. Toxic and hazardous substance exposure is also a risk for people who scavenge in search of valuable items in the solid waste (Samwine 2017). Finally, dumping areas and/or landfills can produce leachate, which pollutes soil, groundwater, and surface water (Agyepong 2010).

Table 1 Descriptive statistics (mean, SD, range) for monthly rates of diarrheal disease and UGS per 100,000 population (natural log-transformed) between January 2012 and December 2016 in 216 districts in Ghana (data extracted from DHIMS) and for population and housing district characteristics (data extracted from Population and Housing Census 2010)

Characteristic	Mean	SD	Range
Monthly diarrheal disease rate (natural log-transformed)	8.40	0.74	2.52–11.78
Monthly UGS rate (natural log-transformed)	3.36	1.14	−0.7–8.64
Population	114,198	164,139	20,282–1,730,249
Household size	5	1.4	3.5–10
% unemployed	3	1.5	0.7–6.9
% literate	66.9	19.3	20.5–94

In the Ghanaian census, solid waste disposal practices were determined using a self-report survey at the household level: burn waste, use a public dump (sanitary landfill or open dumping), dump indiscriminately, bury waste, or "other" (Ghana Statistical Service 2014). In Ghana, specific solid waste management practices vary among districts and also between urban and rural settings (Samwine 2017) (Table 2). However, general practices include transport to landfills, especially in more urban areas, and open dumping. We outlined six common solid waste disposal practices and their accordance with SDG expectations (Table 2). Solid waste disposal practices that meet SDG expectations are practices that properly contain, maintain, and do not pollute surrounding areas (sanitary landfills, composting, and recycling). Practices that do not meet SDG expectations are often dangerous for the communities (open burning, incineration, and open dumping).

Waste reduction and waste recycling are considered best practices, followed by composting, converting waste to energy sources, and sanitary landfills (Table 3). In Ghana, unsanitary landfills and landfills that do not capture methane

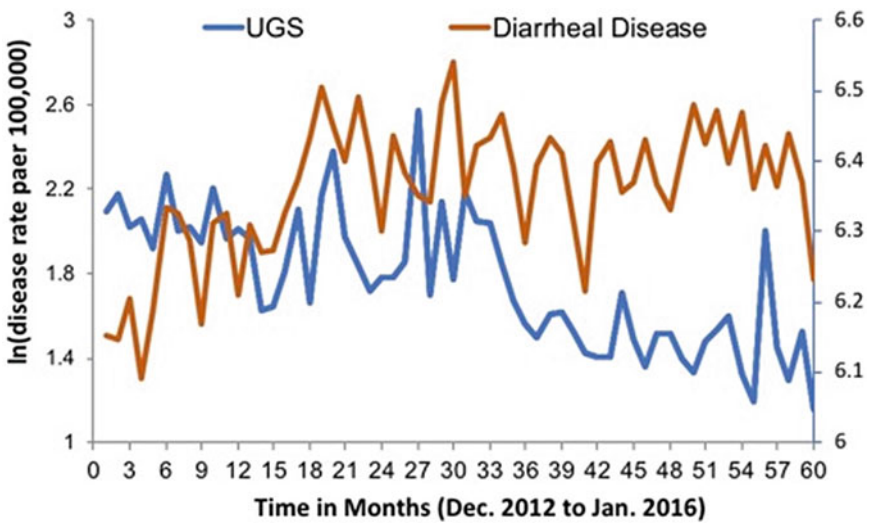


Fig. 4 Monthly time series of disease rates per 100,000 (log-transformed) for UGS (blue line, left axis) and diarrhea (red line, right axes) for 60-month period from December 2012 to January 2016 in 216 districts of Ghana

Table 2 Descriptions of solid waste disposal options and key information about whether practices meet Sustainable Development Goal 2030 expectations

Practices	Description (Mihelcic et al. 2009)	Meets SDGs	SDG explanation (United Nations General Assembly 2015)
Composting	Organic waste broken down by biogeochemical processes in the presence of oxygen	YES	Must be properly protected and maintained
Recycling	Reusing waste for other products or uses; recycling of paper and/or plastic if possible	YES	Must be properly protected and maintained
Sanitary landfill	Waste compacted, daily soil cover; groundwater not contaminated; waste separated from disease vectors	YES	Scavenging can occur; many landfills do not meet these standards
Incineration	Brick/block incinerator or metal drum incinerator used to burn waste	NO	Cannot guarantee separation of waste type (e.g., personal items, electronic, medical)
Open burning	Waste piled and burned outdoors, usually by community members	NO	Dangers include air pollution and attraction of disease vectors, especially near children
Open dumping	Waste thrown into yard, local area, or surface water	NO	Exposure to waste hazards and disease vectors, potential for water contamination

Table 3 Percentage (mean + SD) of households per district that use each type of solid waste disposal practice; data extracted from the Ghana Population and Housing Census (2010)

Solid waste disposal	Total	Rural	Urban
Public dump (open space)	46.2 ± 18.8	54.2 ± 18.3	37.9 ± 20.6
Public dump (container)	16.0 ± 13.1	5.5 ± 6.1	32.9 ± 20.1
Dumped indiscriminately	14.2 ± 11.2	18.2 ± 12.3	6.3 ± 6
Burned by household	11.4 ± 9.4	11.5 ± 9.6	11.5 ± 10.5
Collected	7.2 ± 11.5	4.9 ± 5.5	7.3 ± 12.4
Buried by household	3.9 ± 2.5	4.5 ± 2.9	3.2 ± 2.2
Other	1.1 ± 1.3	1.2 ± 1.4	0.9 ± 1.2

gas (CH_4) are common. Other challenges in Ghana include a lack of technology and tools to collect and break down waste, lack of planning capacity for waste quantity, and poor individual-level practices (Samwine 2017). Most households in Ghana do not use an improved practice for solid waste disposal (Table 3, Fig. 5a). The most common practice for solid waste disposal across all districts in Ghana is to use a public dump in an open space (open dumping) or without proper protection to separate the public dump from the rest of the environment (46.2 ± 18.8%) (Table 3, Fig. 5b). The second and third most common practices are public dumping into a container (16.0 ± 13.1%) and indiscriminate dumping (14.2 ± 11.2%) (Table 3, Fig. 5c, d). Unsanitary waste disposal practices pose a serious concern for public health, yet their specific impact on infection spread has not yet been examined in detail.

Sanitation

Census data that report the sanitation status of a population are captured by various methods; however, survey data and self-report based on a provided list of specific sanitation practices are often used (UNICEF 2017). Inadequate sanitation is estimated to cause 280,000 deaths annually due to diarrheal disease and is a major factor in several neglected tropical diseases: intestinal worms, schistosomiasis, and trachoma (World Health Organization 2018). Improving sanitation infrastructure and sanitation and hygiene practices can positively health impact (World Health Organization 2018). SDG 6 calls for countries to provide global "access to adequate and equitable sanitation and hygiene for all and end open defecation, paying special attention to the needs of women and girls and those in vulnerable situations" (United Nations General Assembly 2015). However, in 2015, only 39% of the global population used a safely managed sanitation service (World Health Organization 2018). A safely managed service is the use of a toilet or improved latrine that is not shared with other households, where excreta are treated or disposed of safely (World Health Organization 2018). Of all the World Bank regions, sanitation is most lacking in sub-Saharan Africa, where an estimated 695 million people still use unimproved facilities (WHO/UNICEF Joint Monitoring Programme (JMP) 2015).

Sanitation promotion, through different forms of education, can improve health outcomes (Al-Delaimy et al. 2014; Madon et al. 2018; Crocker et al. 2016). These studies have used various methods to improve sanitation including training local leaders to run community-led total sanitation (CLTS) (Madon et al. 2018; Crocker et al. 2016), implementing community health learning packages (Al-Delaimy et al. 2014) and creating school-based learning programs (Madon et al. 2018). The approach of community-led education interventions aims to help people control the determinants of health or underlying causes of poor health such as poverty and lack of health information or infrastructure in a sustainable, community-based style (Madon et al. 2018). When examining education levels before a health behavior intervention, studies have shown that a higher maternal level of education is associated with better child health, including a protective effect on diarrheal disease (Fink et al. 2011; Hobcraft et al. 1984; Rutstein 2000). Studies examining the

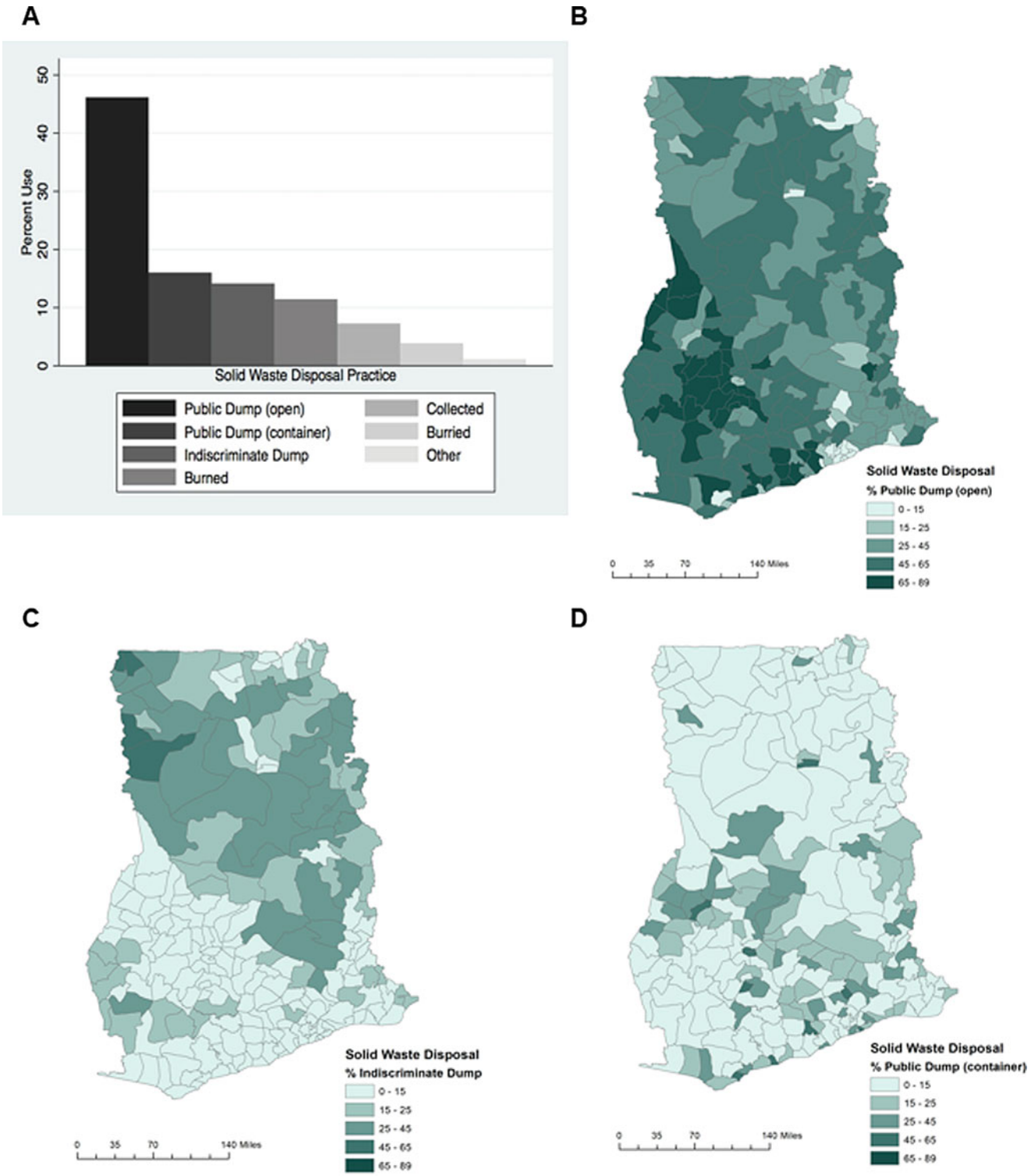


Fig. 5 Most common solid waste disposal practices in 216 districts in Ghana. (**a**) Bar chart of the most common methods of solid waste disposal for households in all districts. (**b**) Percentage of households per district that use open dumping in public spaces for solid waste disposal. (**c**) Percentage of households per district that use public dumping in containers for solid waste disposal. (**d**) Percentage of households per district that perform indiscriminate dumping for solid waste disposal

Table 4 Percentage (mean + SD) of households per district that use each type of sanitation practice; data extracted from the Ghana Population and Housing Census (2010)

Sanitation facility type	Total	Rural	Urban
No facilities (bush, beach, field)	31.2 ± 29.9	36.5 ± 32.3	20.4 ± 21.3
Public toilet (WC, KVIP, pit, pan, etc.)	30.9 ± 16.3	27.7 ± 17.9	37.9 ± 14.2
Pit latrine	21.1 ± 15.5	24.6 ± 17.6	16.6 ± 12.9
KVIP	9.1 ± 5.5	7.0 ± 4.6	13.2 ± 7.2
Water closet	7.0 ± 9.3	3.5 ± 5.6	11.0 ± 10.2
Bucket or pan	0.4 ± 0.5	0.2 ± 0.2	0.6 ± 0.8
Other	0.4 ± 0.3	0.4 ± 0.4	0.4 ± 0.4

relationship between sanitation practices and education at the level of the individual are lacking.

In the 2010 Ghanaian census, sanitation practices were determined using a self-report survey using the following categories: if a household used no facilities (bush/beach/field), water closet (WC), pit latrine, Kumasi Ventilated Improved Pit (KVIP) latrine, bucket/pan, public toilet, or other (Ghana Statistical Service 2014). In Ghana, only 15% of the population uses improved sanitation facilities (WHO/UNICEF Joint Monitoring Programme (JMP) 2015). After analyzing the 2010 census data, we found that the most commonly used sanitation practice in Ghana was "no facility" or practicing open defecation (31.2 ± 29.9%) (Table 4, Fig. 6). Using a public toilet was the second most common option (30.9 ± 16.3). The three most common sanitation practices varied by district: absence of sanitation facilities are common in the north, while public toilets and pit latrines are common in the south (Fig. 6b–d). Overall, northern Ghana has a greater percentage of open defecation than other areas.

Safe Drinking and Domestic Water

Census data on safe drinking and domestic water sources can be captured multiple ways, but survey data and self-report based on a provided list of specific drinking water sources are often used (UNICEF 2017). In the 2010 Ghanaian census, drinking water source used was collected in a self-report survey with the following categories: pipe-borne (inside the dwelling), pipe-borne (outside dwelling), a public tap/standpipe, borehole/tube well, protected well, rainwater, protected spring, bottled water, sachet water, from a taker supply/vendor, unprotected well, unprotected spring, river/stream, dugout/pond/lake/dam/canal, or other (Ghana Statistical Service 2014). Domestic water source was also collected in the census and used the same categories but did not include bottled water or sachet water as options (Ghana Statistical Service 2014).

The WHO/UNICEF (WHO/UNICEF Joint Monitoring Programme (JMP) 2015) recently published information stating that current access to safe drinking water in Ghana has reached 89% as of 2015 (93% in urban areas and 84% in rural areas). Safe drinking water is defined as "sources that, by nature of their construction or through active intervention, are protected from outside contamination, particularly fecal matter" (World Health Organization n.d.). With the high percentage of reported coverage in Ghana, the 2030 SDG 6 to achieve universal and equitable access to safe and affordable drinking water for all seems to be moving in a positive direction (United Nations General Assembly 2015). The SDG measurement for "access" to safe water is not the same thing as actual water use; the current census questions are not designed to capture the fact that many households use multiple drinking water sources; instead, the census captures the "main" drinking water source for a household. Thus, even though a high percentage of improved water "access" is reported, uncertainty remains about the details of daily drinking water use, and no nationally representative data are currently available to research this further with respect to source preferences or temporal/seasonal/spatial variation in source selection.

Adequate quantities of safe water for domestic use are necessary for human life and can reduce the spread of disease (Howard et al. 2003). The WHO defines domestic water use as "water used for all usual domestic purposes including consumption, bathing and food preparation" (World Health Organization 1993). Most importantly, domestic water supplies provide basic health protection in the forms of handwashing, cleaning, bathing, and laundry (Howard et al. 2003). Water scarcity may limit certain sanitary and hygiene habits that protect health (Howard et al. 2003). Poor hygiene has been known for decades to cause diarrheal disease; skin and eye diseases, in particular trachoma; and diseases related to infestations (Cairncross and Feachem 1993).

Use of safe drinking and domestic water sources is affected by many factors such as taste, smell, appearance, price, convenience, and proximity, among others, but there is no country-level data on these factors for most low-income countries. We used multiple methods to assess water use in Asamama, Eastern Region, Ghana (Kosinski et al. 2016a). Asamama faces challenges due to poor water quality and lack of sanitation, with associated risks of diseases including schistosomiasis and diarrheal disease. Data from the community were collected to assess water use by 247 households. Methods included UGS screening of school children, mapping borehole locations and river access points, assessing water quality, and holding focus group discussions with teen and adult participants. About 10.5% of participants reported using only borehole water, 35.2% reported using only river water, and 53.8% reported using both river and borehole. Focus group data revealed that water preferences vary due

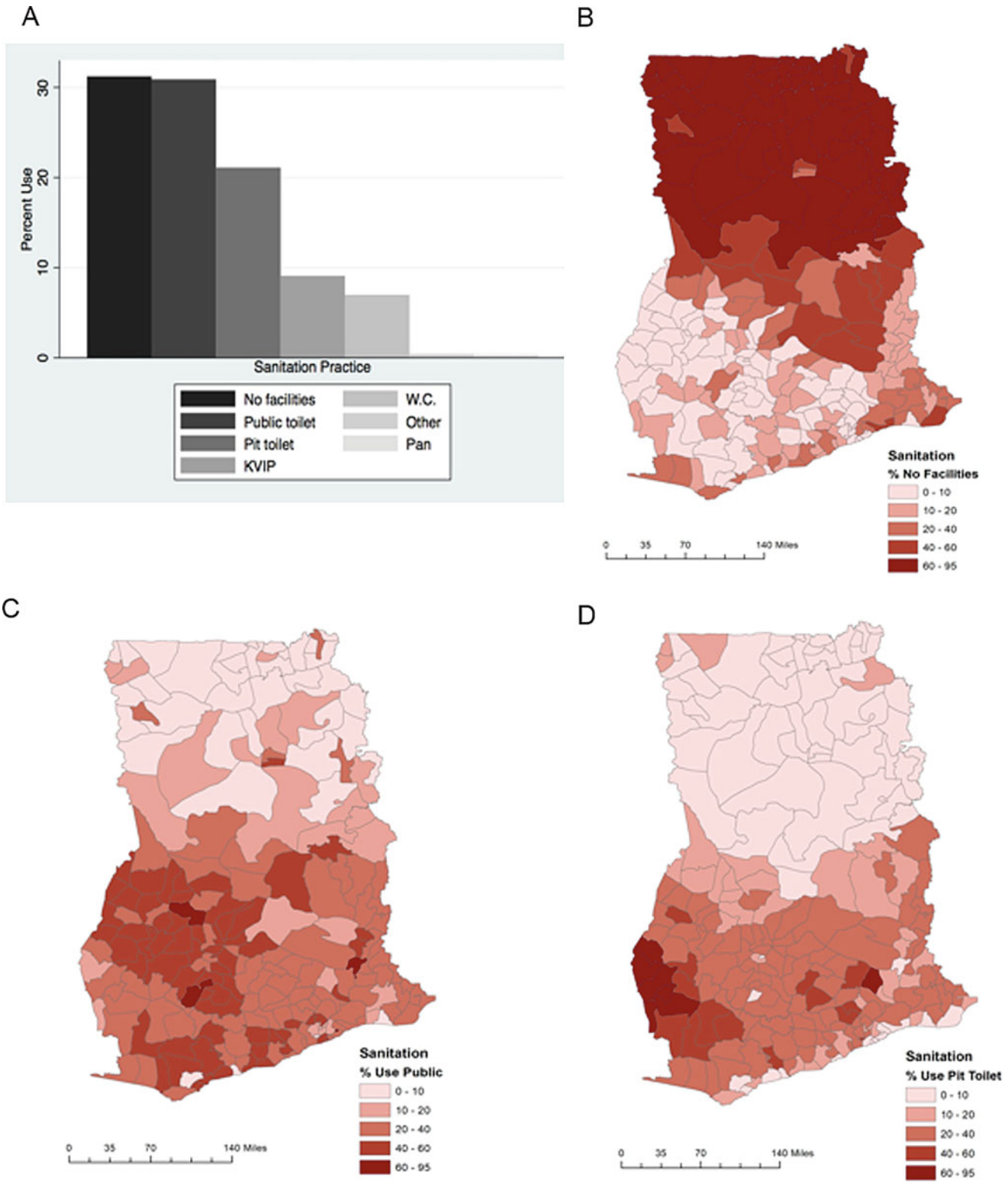


Fig. 6 Most common sanitation practices in 216 districts in Ghana. (**a**). Bar chart of the most popular methods of sanitation for households in all districts. (**b**) Percentage of households per district that use open defecation. (**c**) Percentage of households per district that use public toilets (WC, KVIP latrines, pit latrines, pans, etc.). (**d**). Percentage of households per district that use private pit latrines

Table 5 Drinking water sources with description (Cairncross and Feachem 1993) and accordance with SDG expectations (United Nations General Assembly 2015; Mihelcic et al. 2009)

Water source	Description	Meets SDG	SDG explanation
Borehole, pump, tube well	Well drilled into ground to reach groundwater, system for vessel to draw water or handle for manual pumping	YES	Groundwater needs to be free of contamination
Bottled water	Bottles, usually plastic, filled with safe drinking water, distributed to communities	YES	Can take advantage of communities; high selling price, cause of plastic waste
Pipe-borne	Water is piped from watershed supply, usually receives treatment, enters the primary distribution system, and finally enters a distribution system specific to the community	YES	Must be protected at source in community
Protected well	Well with protective cover: usually sealed concrete slab to protect from outside waste and water from contaminating safe water	YES	Protective cover needs to be functioning
Sachet water	Plastic containers filled with safe drinking water, distributed to communities	YES	Can take advantage of communities; high selling price, cause of plastic waste
Spring	Natural spring water, flows through a barrier (usually concrete or brick box) built around spring to protect against outside pollution	YES	Must be protected with a cover at the source
Standpipe or public tap	Pipe routed to water distribution system, usually placed in central location, faucet for community use	YES	Needs to be accessible to community, protected at the source
Tanker supply	Vehicle carries large amount of water	YES	Can take advantage of communities; high selling price
Dugout, pond, or lake	Natural body of water, without filtration system or protection from contamination	NO	Contamination occurs easily
River or unprotected spring	Natural water sources, no filter for outside contamination	NO	Contamination occurs easily
Unprotected well	Well without protection from outside contamination	NO	No barrier from outside contamination

to factors such as taste, appearance, perceived quality, proximity, and the physical condition of water sources. Improved infrastructure and individual water treatment options could improve water sources, but more research is needed to assess perceived health risks associated with certain water sources (Kosinski et al. 2016a).

We recently examined six drinking water sources that are commonly used in Ghana and their accordance with the SDGs (Table 5). Multiple sources meet SDG expectations, as long as these sources are protected, accessible, and affordable.

Based on our analysis of the 2010 Ghana Census Data, we found that boreholes/pumps/tube wells are the most commonly used drinking and domestic water sources ($32.5 \pm 20.8\%$) (Table 6, Fig. 7a). Pipe-borne water (pipes provide water to house compound outside of dwelling) and river water are the next most common sources for both drinking and domestic water (Table 6). For drinking water use specifically, borehole use is higher in the northern areas of Ghana, as well as along the western coast. There also appears to be high use in the southwest corner (Fig. 7b). Pipe-borne water (outside of the dwelling) use is higher in the southern areas of Ghana (Fig. 7c). River water use is higher in the eastern area of Ghana, as well as a corner of the southwest (Fig. 7d). This information, especially on a more granular scale, is essential for understanding the health impact of improved water sources. When the data were in their original form of 216 separate district reports, compiled in tables, it was not possible to view these interesting spatial patterns; visualizing the data geographically has enabled new perspectives on drinking water sources in the country.

When we examined domestic water use, we used the same techniques (maps not shown). We found that borehole use is highest in the northeast and southwest and river water for domestic use is higher in the eastern areas. For domestic water use, there is a more distinct pattern for outdoor piped water; higher percentages occur along the southern coast and some southern regions. Spatially, drinking water and domestic water have similar percent use patterns across Ghana; for both sources, borehole use is higher in the northwestern area, use of pipes is higher in the south, and river water use is higher in eastern areas. With enhanced GIS capacities, it would be possible to demonstrate the value of investment in water infrastructure at the community level and nationally.

Energy Sources, Lighting, and Cooking

According to SDG 7, all people should have "... access to affordable, reliable, sustainable and modern energy" (United Nations General Assembly 2015). Energy use is generally measured and captured in census reports at the household or commercial/industrial level (World Bank 2018). Survey or self-report data are common ways of measuring use (World Bank 2003). In the Ghanaian census, energy sources, such as lighting or cooking fuel used, are collected in a self-report survey with the following categories: electricity (main), electricity (private generator), kerosene lamp, gas lamp, solar

Table 6 Percentage (mean + SD) of households per district that use each type of water source practice; data extracted from the Ghana Population and Housing Census (2010)

	Drinking water			Domestic water		
Water source	Total	Rural	Urban	Total	Rural	Urban
Borehole, pump, tube well	32.5 ± 20.8	40.8 ± 21.8	18.4 ± 18.2	31.9 ± 19.5	39.1 ± 20.7	19.1 ± 17.2
Pipe-borne outside dwelling	14.6 ± 9.9	10.4 ± 8.3	21.7 ± 12.2	14.2 ± 10.3	10.1 ± 8.4	21.0 ± 12.1
River, stream	13.4 ± 12.5	18.2 ± 14.4	5.1 ± 11.5	15.6 ± 13.6	20.8 ± 15.1	6.6 ± 12.7
Public tap, standpipe	12.8 ± 10.1	10.4 ± 10.3	19.4 ± 16.8	12.1 ± 9.7	9.7 ± 9.8	18.4 ± 16.4
Pipe-borne inside dwelling	6.7 ± 9.0	2.8 ± 4.5	11.6 ± 11.6	7.3 ± 9.9	3.0 ± 5.1	12.4 ± 12.2
Protected well	6.7 ± 6.8	5.5 ± 5.6	9.8 ± 11.5	8.8 ± 8	6.7 ± 6.2	13.9 ± 14
Sachet water	5.1 ± 9.6	2.8 ± 7.5	7.7 ± 11.1	–	–	–
Unprotected well	3.1 ± 4.8	3.5 ± 5.0	2.1 ± 5.1	3.8 ± 5.5	4.0 ± 5.3	3.4 ± 6.6
Dugout, pond, lake, dam, canal	2.6 ± 5.2	3.3 ± 6.2	1.2 ± 6.0	3.2 ± 6.0	3.8 ± 6.6	1.6 ± 6.7
Rainwater	0.9 ± 2.2	0.8 ± 2.0	0.9 ± 3.0	0.9 ± 2.2	0.8 ± 1.6	1.0 ± 3.5
Tanker, vendor provided	0.9 ± 2.7	0.5 ± 1.7	1.1 ± 4.2	1.5 ± 5.3	0.9 ± 3.5	2.0 ± 7.0
Protected spring	0.4 ± 0.3	0.3 ± 0.3	0.4 ± 0.4	0.4 ± 0.3	0.3 ± 0.3	0.4 ± 0.5
Unprotected spring	0.3 ± 0.5	0.4 ± 0.6	0.1 ± 0.3	0.4 ± 0.5	0.5 ± 0.6	0.2 ± 0.4
Bottled water	0.2 ± 0.3	0.1 ± 0.3	0.3 ± 0.3	–	–	–
Other	0.1 ± 0.1	0.1 ± 0.2	0.1 ± 0.2	0.2 ± 0.2	0.3 ± 0.3	0.1 ± 0.2

energy, candle, flashlight/torch, firewood, crop residue, or other (Ghana Statistical Service 2014).

When we analyzed energy source use in Ghana based on 2010 census data, we found that the most common cooking fuel was wood (53.7 ± 28.9%) followed by charcoal (23.5 ± 16.2%) (Table 7). Wood as a principal cooking fuel source is high throughout Ghana. Use of charcoal is higher in the southern half of the country but is somewhat heterogeneous across districts. Use of gas as a principal cooking fuel is higher in some areas of the south and east. The most common lighting source was reported as "no energy source" (49.4 ± 21.5%) followed by gas (25.2 ± 16.8%) and burning crop residue (23.2 ± 17.3%); higher electricity use appears in the southern region of Ghana. In contrast, higher kerosene use appears in the northern region of Ghana, as well as eastern parts of the country. High rates of flashlight use as a principal lighting source occur in western Ghana. Little is known how improvements in the use of green energy sources may benefit health and well-being by technological leapfrogging.

Literacy and Employment

By 2030, SDG 4.6 states that all countries should "ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy" (UN General Assembly 2015). Lower literacy rates can be an independent risk factor for poor health outcomes (Baker et al. 1998; Dewalt et al. 2004; Wolf et al. 2005; Sudore et al. 2006). In a systematic review, DeWalt et al. (Dewalt et al. 2004) used 1980–2003 data from MEDLINE, the Cumulative Index to Nursing and Allied Health Literature, the Educational Resources Information Center, Public Affairs Information Service, *Industrial and Labor Relations Review*, PsycINFO, and AgeLine and found that limited literacy is associated with health variables such as knowledge about healthcare and various chronic diseases. Martel et al. (2019) recently showed that increased class year in school is associated with significantly higher knowledge of urogenital schistosomiasis attributes, which has the potential to result in improved protective behaviors. In a census, literacy is generally measured as the total number of literate persons—able to read and write—in a given age group, expressed as a percent of the total population in that age group (UNESCO 2008). Age groups usually include adult literacy rate (15+ years) and youth literacy rate (15–24 years), but may be further divided (UNESCO 2008). There is no international standard way to measure literacy, and many countries rely on self-reporting. During the 2010 Ghanaian census, respondents (aged 11+ years) were considered literate if they could read and write a simple statement with understanding (Ghana Statistical Service 2013).

Employment status collected for census data is generally measured by self-report in the form of a household survey reporting current or previous employment or unemployment administrative records (U.S. Census Bureau, Labor Force Statistics 2017). Methods for determining employment status vary and can include simply stating that a person is currently employed or not, listing a current place of employment, or using statistics describing persons using unemployment insurance (U.S. Census Bureau, Labor Force Statistics 2017). Limited research exists on the relationship between unemployment and health; however, unemployment causes increased mental health conditions, including anxiety and depression, as well as heart disease and its associated risk

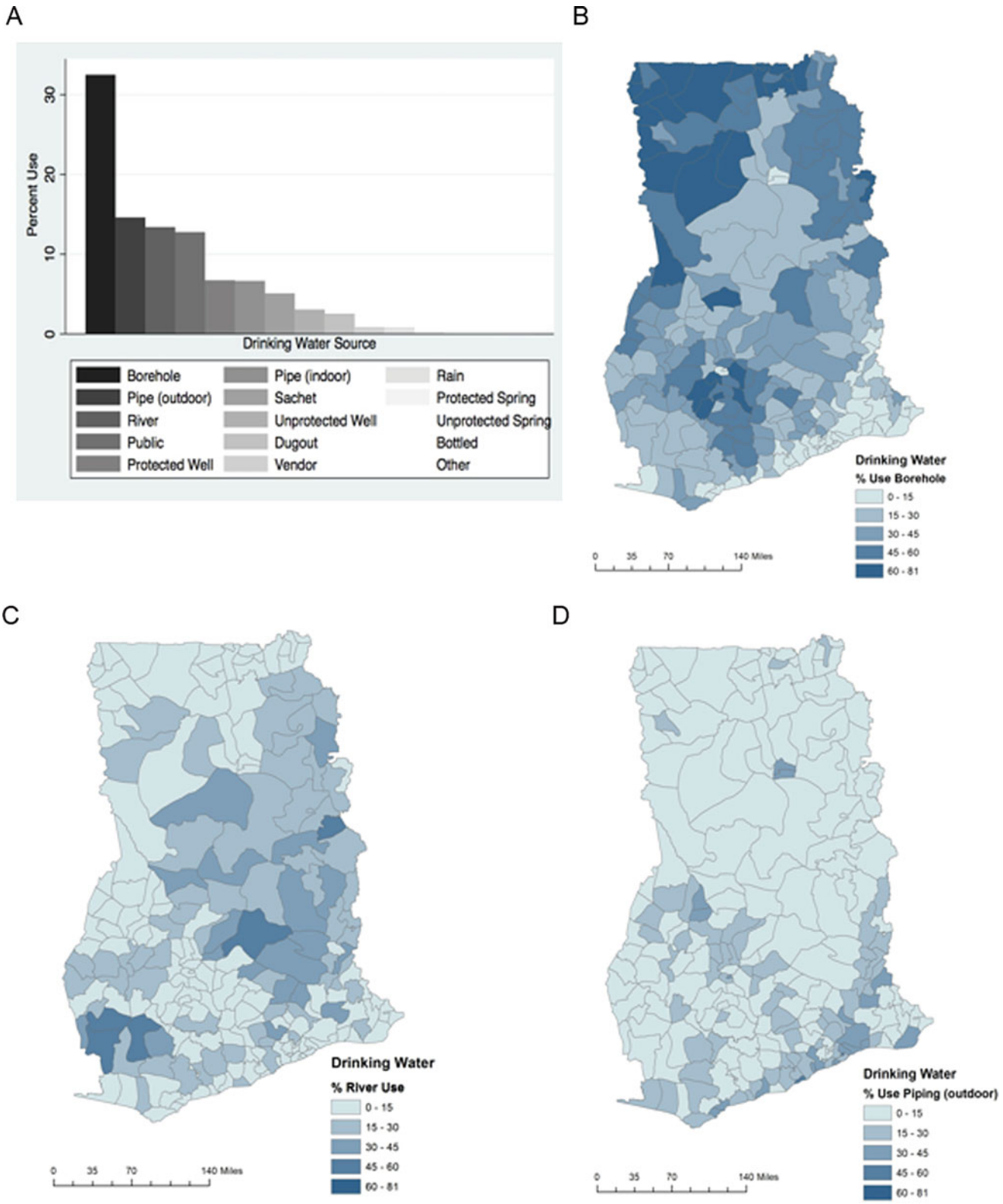


Fig. 7 Most common drinking water sources for 216 districts in Ghana. **A**. Bar chart of the most popular drinking water sources for households in all districts. **B**. Percentage of households per district that use borehole wells/pumps/tube wells for drinking. **C**. Percentage of households per district that use pipe-borne water that is outside the house for drinking. **D**. Percentage of households per district that use river/stream water for drinking

Table 7 Percentage (mean + SD) of households per district that use each type of fuel source for cooking and lighting; data extracted from the Ghana Population and Housing Census (2010)

	Cooking			Lighting		
Fuel source	Total	Rural	Urban	Total	Rural	Urban
Wood	53.7 ± 28.9	31.6 ± 22.5	67.6 ± 27.3	–	–	–
Charcoal	23.5 ± 16.2	40.8 ± 19.2	15.1 ± 13.6	–	–	–
Gas	11.2 ± 13.5	14.6 ± 11.2	7.9 ± 13.5	0.2 ± 0.1	0.2 ± 0.1	0.2 ± 0.2
None	8.0 ± 14.2	10.91 ± 18.4	5.5 ± 9.6	–	–	–
Electricity	0.4 ± 0.8	0.5 ± 2.0	0.3 ± 0.3	–	–	–
Kerosene	0.4 ± 0.3	0.5 ± 0.4	0.3 ± 0.2	25.2 ± 16.8	15.6 ± 11.0	30.6 ± 19.2
Crops	3.1 ± 9.1	1.1 ± 3.5	3.5 ± 10.1	0.1 ± 0.2	0.4 ± 4.9	0.2 ± 0.2
Sawdust	0.1 ± 0.2	0.1 ± 0.2	0.1 ± 0.2	–	–	–
Animal waste	0.0 ± 0.1	0.0 ± 0.1	0.0 ± 0.1	–	–	–
Other	0.1 ± 0.3	0.7 ± 7.3	0.1 ± 0.2	0.7 ± 6.2	0.2 ± 0.2	0.3 ± 0.4
Solar	–	–	–	0.3 ± 0.4	0.1 ± 0.2	0.4 ± 1.6
Candle	–	–	–	0.4 ± 0.6	0.5 ± 0.6	0.4 ± 0.6
Flashlight	–	–	–	23.2 ± 17.3	8.3 ± 6.7	28.4 ± 19.5
Fire	–	–	–	0.4 ± 0.4	0.2 ± 0.2	0.4 ± 0.5
Public electricity	–	–	–	49.4 ± 21.5	73.5 ± 13.4	38.0 ± 21.3
Private electricity	–	–	–	0.7 ± 0.6	0.8 ± 0.9	0.8 ± 0.4

factors (Wilkinson and Marmot 2003). Overall, as a social determinant of health, unemployment increases the risk of illness and premature death (Wilkinson and Marmot 2003).

Employment is addressed by SDG 8, which calls for countries to "promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all" (United Nations General Assembly 2015). In the 2010 Ghanaian census, employment status was determined with a self-report survey with eight categories for a person to mark as a best indicator of their status: employee, self-employed without employees, self-employed with employees, casual worker, contributing family worker, apprentice, domestic employee (house help), or other (Ghana Statistical Service 2014). In Ghana, between the years 2000 and 2010, the number of people employed increased from approximately 7.43 million to 10.24 million and has remained at a relatively stable increase of linear growth (Alagidede et al. 2013).

Summary of Ghanaian Case Study

Our case study in which we used data from multiple data sources shows that there were distinct spatial patterns when comparing environmental and health variables across districts. The analysis of the 2010 census data shows that there are spatial patterns of water use, sanitation infrastructure, and energy source use throughout Ghana, with more areas of unimproved and unsafe use appearing in the northern sections. There appear to be greater environmental health risks in northern areas in Ghana: low use of sanitation facilities and high use of kerosene and wood fuel. It is also highly plausible that there are distinct spatial patterns when considering rurality and urbanicity, but there are no available nationally representative data to study these important potential correlates of health outcomes. In the future, it would be incredibly useful to create shapefiles for all 216 districts in Ghana with rural and urban areas clearly delineated. Analyzing environmental health variables by urban and rural locality would be a fascinating next step. Findings from the WHO concerning rural disparities on water and sanitation state that approximately 55% of the world's rural population has access to safe water sources and 57% have access to improved sanitation. In contrast, approximately 85% of the world's urban population has access to safe water sources, and 92% has access to improved sanitation (WHO/UNICEF 2017). Additionally, rural communities have an increased reliance on kerosene and wood for fuel, contributing to increased risk for negative air quality and related health effects (Rehman et al. 2005). Moving forward, it will be invaluable for all countries to have access to high-quality, clearly georeferenced data based on locality in order to draw population-level conclusions on environmental health indicators and associated disease outcomes.

Ghana's Efforts to Improve Nationally Representative Dataset Availability and Accessibility

In this section, we highlight Ghana's efforts to improve nationally representative dataset availability and accessibility. Ghana has spearheaded efforts to develop and support online repositories and websites to track nationally representative

data. In 2018, the GSS posted a Central Data Catalog on the GSS website and outlined various datasets collected by a variety of governmental agencies and academic institutions. We explored GSS websites including the main GSS Website (Ghana Statistical Service 2019a), the GSS Data Catalog (Ghana Statistical Service 2019b), the Ghana Open Source Data Initiative (Ghana Open Source Data Initiative 2019), and the Ghanaian website on the SDGs (Ghana Statistical Service 2019c).

The GSS website (Ghana Statistical Service 2019b) provided a data catalog with information about data availability via download, including direct access; public access; licensed access, available through external repository; and data not available. Table 8 outlines the 59 datasets mentioned in the GSS Central Data Catalog, along with the organization that initiated or supported data collection. The GSS website has 26 datasets listed as "available," in some capacity, for download. However, many available datasets require a registration process in order to receive data. This process does not guarantee a response or delivery of data.

The most recent GSS website has a new "Open Data Initiative" that mirrors the format of the now-retired GSS Data Catalog and is visibly "in progress" to be finalized. The Open Data Initiative offers names of 133 datasets that will be eventually available. As of March, 2019, we were able to access 15 datasets. For the present study, SDG data were not used due to the limited amount of data published as of March 2019. Of the 59 datasets, only 4 datasets were available for immediate download. Remaining datasets required a registration process or were not available. Registration processes included creating an account on the GSS website; once registered, access to completing a data request form was granted. Data request forms included a description of proposed project or use for data, professional or academic affiliation information, and contact information (Table 9).

GSS also reports data organized by SDG on a separate, more recently updated website (Ghana Statistical Service 2019c). This innovative, in-progress effort by GSS reports the status of datasets that could be used to assess each SDG goal and sub-goal. Titled the "Ghana data for Sustainable Development Goal Indicators," GSS is seeking a total of 244 datasets to address all aspects of the SDGs; 33 datasets are currently available for download (Table 10).

Challenges of Data Compilation and Processing

As we were producing the maps and tables presented in this chapter, we scrupulously documented challenges and limitations. We created *a priori* categories to organize the encountered challenges in four domains: data collection and acquisition, data validation and verification, data analysis, and data presentation. These categories followed a common data life cycle and include data availability, access, completeness, quality, complexity, interpretability, delivery, and comprehension. We define these categories from a point of view of mapping diseases using census data, DHIMS records, and the CERSGIS platform and describe availability and access challenges as we download and compile records for district-level mapping. We documented challenges encountered due to incomplete, mismatched, or missing records or as we were forced to estimate some metrics based on the preprocessed data rather than the raw original data. We also paid attention to the ability to perform a complex analysis on a refined spatiotemporal scale and to assess long-term and seasonal variations. As many disease monitoring systems offer their clients various tabular or visual forms of reporting, we assessed their quality of delivery and comprehension.

2010 Population and Housing Census

For the present study, 2010 Population and Housing Census data were not available to download directly; instead, we extracted records from 216 District Analytical Reports as separate documents available in PDF format. We used an online PDF-to-Microsoft Excel table conversion tool for each variable for each of the 216 districts in Ghana. All variables in the reports were presented as percentages of the district population, and most were additionally broken down into percentages for urban and rural parts of districts. In addition to overall limited access to GSS data, the first step to acquiring census data was the most time-consuming; PDF-to-Excel spreadsheet data extraction proved tedious and prone to errors (Table 11). Data quality was negatively impacted because content sections in the 216 district reports were inconsistent and statistics did not always report on the same variables. Overall, the census data presented a wide variety of challenges that researchers face in working to use these data for health-related purposes with spatial applications.

District Health Information Management System (DHIMS)

DHIMS data were acquired by a direct request to GHS for case counts of UGS and diarrheal disease for 216 districts in Ghana. The data were provided at the district level and included monthly disease counts for diarrheal disease and UGS between January 2012 and December 2016. DHIMS data allowed us to create country-level maps of both UGS and diarrheal disease; however, there is additional information that could be collected in the future to further enhance data analysis options. The DHIMS database and GHS websites do not currently provide details about the quality or com-

Table 8 Catalog of data on the Ghana Statistical Service (GSS) website (Ghana Statistical Service 2019b), along with data availability as of March, 2019

Data source	Year	Producer	Governmental support	Access type
Academically based				
Afrint Household Level Data, Round 1 and 2	2001, 2002	Lund University	Swedish gov't	Direct
NetMark Insecticide-Treated Nets Survey, Baseline Household Evaluation	2004	Academy for Educational Development	USAID	External
INDEPTH Study on Global Ageing and Adult Health, Wave 1	2004	Navrongo Health Research Centre: GHS	US National Institute on Aging	External
Study on Global Ageing and Adult Health, Wave 1	2007–2008	Ghana Medical School	US National Institute on Aging	External
Governmentally based				
Fertility Survey	1979–1980	Central Bureau of Statistics	UNFPA, USAID, UK Overseas Development Administration	External
Population Census 1960	1960	Ghana Census Office	Not listed	External
Population Census 1970	1970	Ghana Census Office	Not listed	External
Population Census 1984	1984	GSS	Not listed	External
Population and Housing Census	2000	GSS—Office of the President	Gov't of Ghana and others	Public
Population and Housing Census	2010	GSS—Min. of Finance and Economic Planning	Gov't of Ghana, UNFPA, UNDP, and others	Public
Demographic and Health Survey	1988	GSS	UNFPA	External
Demographic and Health Survey	1993	GSS—Office of the President	USAID	Public
Ghana Demographic and Health Survey	1998	GSS—Office of the President	Gov't of Ghana, USAID	Public
Demographic and Health Survey	2003	GSS—Gov't of Ghana	Gov't of Ghana, USAID	Public
Demographic and Health Survey	2008	GSS, Min. of Health	USAID, UNFPA, UNICEF, Ghana AIDS Commission	Public
Demographic and Health Survey	2014	GSS—Office of the President	Not listed	External
Ghana Maternal Health Survey	2007	GSS	USAID	Public
Ghana Maternal Health Survey	2017	GSS	Not listed	
Emergency Obstetric And Newborn Care, Round 2	2011	GHS	Gov't of Ghana, UNICEF, UNFPA, WHO, USAID	Licensed
Annual Statistical Report on Births and Deaths, Round 1	2012	Births and Deaths Registry	Gov't of Ghana	Licensed
Annual Statistical Report on Births and Deaths, Round 2	2013	Births and Deaths Registry	Ministry of Local Gov't and Rural Development	External
Holistic Assessment of the Health Sector Programme of Work	2013	Ministry of Health	Gov't of Ghana	Licensed
Annual Schools Census—Basic Schools Info, Round 24	2012–2013	Ministry of Education	Gov't of Ghana	Direct
Annual Schools Census—Senior High School, Round 7	2012–2013	Ministry of Education	Gov't of Ghana	Direct
GLSS 1	1987–1988	GSS	Office of the President	Public
GLSS 2	1988–1989	GSS	Office of the President	Public
GLSS 3	1991–1992	GSS	Office of the President	Public
GLSS 4, with Labour Force Module	1998–1999	GSS	Office of the President	Public
GLSS 5, with Non-farm Household Enterprise Module	2005–2006	GSS	Office of the President	Public
GLSS 6, with Labour Force Module	2012–2013	GSS	Gov't of Ghana	Public
GLSS 7	2017	GSS		Public
Multiple Indicator Cluster Survey	1995	Ministry of Health	Gov't of Ghana	External
Multiple Indicator Cluster Survey	2006	GSS	Office of the President	Public
Multiple Indicator Cluster Survey	2008	GSS	Gov't of Ghana	Public
Multiple Indicator Cluster Survey	2011	GSS	Autonomous	Public
Ghana Child Labour Survey	2001	GSS	Ministry of Finance and Economic Planning	Public

(continued)

Table 8 (continued)

Data source	Year	Producer	Governmental support	Access type
Service Provision Assessment Survey	2002	GSS		External
Ghana Core Welfare Indicators Survey	1997, 2003	GSS	Office of the President	Public
National Industrial Census	2003	GSS	Autonomous	Public
Financial Service Survey	2006	GSS	Office of the President	Public
Job Tracking Survey	2006	GSS	Office of the President	Public
National Transport Household Survey	2007	GSS	Autonomous	Public
Public Expenditure Tracking Surveys	2007	GSS	Autonomous	External
Crime Victimization Survey	2009	GSS	Autonomous	External
Ghana Time Use Survey	2009	GSS		Public
Ghana User Satisfaction Survey	2012	GSS	Autonomous	Public
Transport Indicator Database Survey, Round 2	2012	GSS	Gov't of Ghana	Public
Social Accounting Matrix	2015	GSS	Office of the President	Public
Agricultural Production Survey—Minor Season, Round 2	2013	Ministry of Food and Agriculture	Gov't of Ghana	Licensed
Externally based				
World Health Survey	2002, 2003	WHO		External
Enterprise Survey	2007	World Bank		External
Global Financial Inclusion Database	2011	World Bank		External
People Security Survey	2002	International Labour Organization		External

GLSS Ghana Living Standards Survey, *GSS* Ghana Statistical Service, *GHS* Ghana Health Service, *WHO* World Health Organization

Table 9 Catalog of data listed on the Ghana Statistical Service (GSS) website (Ghana Statistical Service 2019b) and accessibility status out of 59 data files as of March, 2019

GSS data catalog	Number of data files	Immediate download	Registration required	Application required	Additional notes
Direct data access	4	Yes	No	No	N/A
Public use data files	31	No	Yes	Yes	N/A
Licensed data files	5	No	Yes	Yes	Extremely comprehensive application
Data available from external repository	2	No	Yes, on separate website	No	Link provided; log in to separate website
Data not available	17	No	No	No	Not available: only contextual documentation

pleteness of the data or the methods that individual clinics use to determine what constitutes a "case" or provide a case definition. Similar to many well-developed monitoring systems, the DHIMS website could provide, for example, annual estimates of disease incidence, treatment, and cure rates. This information will improve interpretability of maps and data analysis. Offering the data to researchers in a temporally and spatially disaggregated manner would allow for advanced analyses to occur. Both UGS and diarrheal disease may have a time lag between the onset of disease and reporting; for UGS, this could be years. A continuous monitoring system allows researchers and practitioners to track disease incidence over time and address this not yet well-understood time lag. Existing records do not delineate between urban and rural locality, and the ability to combine disease data with environmental characteristics relevant to urban-rural settings is limited. Overall, the spatial origin and incidence of infection cannot be accurately assessed with the currently available data (Table 12).

CERSGIS Data

Shapefiles and all country-level data layers for Ghana, including district and regional boundaries, water bodies, land use, and cities and towns, were purchased from CERSGIS by Tufts University. Information on updates to district boundary changes has not been effectively translated to existing GIS data layers. A shapefile of Ghana with all 216 districts was not available through CERSGIS, an academic repository, or central database. Because the available shapefiles

Table 10 Summary of datasets sought by the Ghana Statistical Services (GSS) to determine progress toward the 2030 Sustainable Development Goals (SDGs) (United Nations General Assembly 2015)

Sustainable development goal	Total datasets GSS desires	Datasets reported online	Datasets currently being sought	No data sources available yet; not currently being sought
Total	244	33	7	204
1. No poverty	14	2	0	12
2. Zero hunger	13	3	0	10
3. Good health and well-being	27	6	0	21
4. Quality education	11	1	2	8
5. Gender equality	14	2	0	12
6. Clean water and sanitation	11	2	1	8
7. Affordable and clean energy	6	1	0	5
8. Decent jobs and economic growth	17	2	0	15
9. Industry, innovation, and infrastructure	12	5	0	7
10. Reduced inequalities	11	0	0	11
11. Sustainable cities and communities	15	1	1	13
12. Responsible consumption and production	13	1	0	12
13. Climate action	8	0	0	8
14. Life below water	10	0	0	10
15. Life on land	14	0	1	13
16. Peace and justice—strong institutions	23	2	2	19
17. Partnerships for the goals	25	5	0	20

contained 137 districts rather than the currently existing 216, district-level data were digitized manually in ArcMap (Version 10.5.1). We also corrected district names to match corresponding district names in the census and health records (spelling, case, and spacing) and joined a district-level shapefile to spreadsheets containing both census and DHIMS data. Errors and misalignments in district names and boundary lines within shapefiles and across other spatial programs make the merging step time-consuming and did not warrant full data validation (Table 13).

Leapfrogging to the Modern Geospatial World

In order to build robust national systems for health monitoring, many high-income countries have invested in decades of innovation, iteration, and trial and error. With new technological solutions in low-income countries, particularly African ones, it is possible to skip the technological evolution process to "leapfrog" over now-obsolete technologies with a movement directly to modern infrastructure. Reliable and regularly collected data on health outcomes and potential risk factors are necessary for decision-making about interventions and health delivery infrastructure (Grimes et al. 2014; Nori-Sarma et al. 2017). Lack of data can severely limit analyses and may prevent public health officials and researchers from drawing valuable, nationally representative conclusions on the state of a population's health (National Research Council 2012). Ideally, a surveillance system should provide key information on infection transmission, vaccination, and treatment options and update the public on potential hot spots to minimize infection risk. Moving forward, monitoring systems for non-communicable diseases are expanding their role in developing and implementing preventive strategies. Similarly, geospatial systems for tracking environmental exposures are merging with health monitoring platforms to enrich our understanding of disease ecology, epidemiology, and ways to improve health and well-being.

There are both challenges and opportunities for low-income settings with respect to developing geospatial capacities in compiling health and environmental records. In the case study that focused on Ghana, some of the major challenges were that the 2010 Population and Housing Census data were not adequately available to researchers; infrastructural and environmental data from CERSGIS had inconsistent attribute labels and district names, which made merging datasets difficult; and health records from 2012 to 2016 from DHIMS were limited in both data completeness and granularity. Despite the challenges, there are many potential solutions, some of which may already be in process. These possible solutions address many of the data challenges that other studies have cited (Bhalla et al. 2010; Pandey et al. 2010; National Research Council 2012). We believe the following "low-hanging fruit" solutions would help in leapfrogging:

Table 11 Data challenges that remain to be addressed with the 2010 Population and Housing Census Data from Ghana

Availability	Half of the datasets listed on the GSS website are potentially available for download
	Data from some censuses not available for immediate download: registration on the GSS website is needed, followed by the completion of the "Application to Access to a Public Use Dataset"
	Long lag time (4–5+ years) following the census before the data can be requested results in delayed data analysis
Access	Raw data used to create figures and tables were not publicly available
	Extraction process to compile a basic set of environmental variables from all 216 district reports is time-consuming (took between 60 and 100 hours)
	Publicly available data were not presented in an analysis-ready format (e.g., CSV files); district-level reports with descriptive statistics and tables were downloaded individually from the GSS website as PDFs, and then data were extracted from reports manually, which is time-consuming and prone to errors. PDF-to-Excel converter (an open-source app) could only convert one PDF at a time; a researcher had to download each excel table and then clean/extract relevant data
	Manual scrolling through document was necessary (or using a "find" command) to find relevant charts/graphs
Completeness	Varying amounts and patterns of missing data found in 216 PDF reports; any process that may have been used to validate data completeness was not described and remains unknown
	In some instances, basic statistics in a tabular form were missing, and it forced users to estimate values of a variable of interest from graphs instead of using proper corresponding tables; occasionally variables of interest were not available in either graphs or tables
Quality	Raw or preprocessed data were not available for verification, so data quality is uncertain
	In some instances, tables and charts have errors (e.g., percentages do not sum to 100% or incorrect denominators used), limiting their usefulness and credibility
Complexity	Analyses at refined spatial scales are not currently possible (e.g., rural and urban data values cannot be spatially assigned to shapefiles in GIS); existing descriptive statistics are coarsely aggregated at the district level
	Trend analyses are not currently possible: data from earlier censuses cannot be matched with most recent census due to changes in district names and boundaries and unavailability of some census datasets
Delivery	Presentation of the data in 216 individual PDFs makes it difficult to compare a single variable across many districts or via spatial analyses
	Standard method not included to account for population growth; population for each year is not standardized
	Current chart and table formats in PDFs not always appropriate for the data types
	Tables and graphs lacked uniformity of design, content (statistics presented), and location within reports
Comprehension and interpretability	It is not clear how "urban" or "rural" status is determined
	It is not clear whether/how data were normalized
	Codebook and description of methods of data analysis for district reports not available
	District names and their spellings were not standardized and did not consistently match across other data sources

- Ensure public access to raw or preprocessed data in an analysis-ready format (e.g., CSV files).
- Ensure transparency of the data request process and reserve "limited access" for data where it is absolutely necessary.
- Devote sufficient resources to processing and preparing datasets for public use and further research.
- Have a national-level entity collect all tables and figures from district reports to compile and standardize them, checking for errors, prior to release to the public.
- Use widely accepted standard methods and cite them.
- Promote governmental collaborations to create a standard set of terminology.
- Improve metadata quality and completeness.
- Train internal and external users in spatial data analysis techniques to improve research and professional practice capacity.
- Distribute key results to local, regional, national, and international organizations and stakeholders for feedback and further dissemination.

Conclusions

This chapter takes the first step toward systematically defining the challenges faced when using multiple types of nationally representative data in low-income countries. By illustrating the use of nationally representative data sources from Ghana, we focused on the ways in which these challenges can be resolved. The proposed potential solutions will require the collaboration of many agencies, institutions, and community organizations and offer a "leapfrogging" opportunity to countries where national data researchers could produce models that accurately describe the relationship between

Table 12 Data challenges that remain to be addressed with the DHIMS data from Ghana

Availability	Cases of a given disease missing due to errors, under-reporting, poor case finding, lack of healthcare-seeking by patients, etc.
	Data unavailable at the level of the healthcare facility
	Data regarding geographic origins of infectious disease cases (e.g., home location, contact with other cases, water contact sites) are not available
	Estimates of % cases who are likely not seeking healthcare are not available
	Estimates of cure rates and reinfection are not available
	Estimates of lags in case reporting are not available
Access	The list of diseases and health conditions reported to the DHIMS is not readily available, so requests for data are challenging to formulate
	Lack of obvious data request mechanism
Completeness	Varying levels of missing data in case counts over time (e.g., for diarrheal disease, there were 11 blank weeks and for UGS, 6599 blanks for the same 5-year period); there are no distinctions between an absence of cases needed to be reported during reporting period and no reporting occurred; any process that may have been used to validate data completeness was not described and remains unknown
Quality	District names not universally compatible with other datasets (GIS, Census)
	In some instances, case definitions are not provided
	Unclear meaning of blank cells in the DHIMS system (e.g., could indicate "missing," "0," etc.)
	Data reporting and receiving entities and reporting timeframes are unclear, with implications for understanding lulls/spikes and temporal aggregation
Complexity	Years of complete data are limited
	Temporal aggregation (not sure—annual, quarterly, monthly?) is too high for some types of analyses
	Spatial aggregation (?) is too high for many types of analyses
Delivery	Researchers working with DHIMS data lack an obvious mechanism to report findings back to GHS
	Communities from which data are drawn lack an obvious mechanism to receive reports with findings
	A link between research findings and practical applications is not obvious
Comprehension and interpretability	Maps showing temporal trends in geographic location of cases do not yet exist

environmental factors and health outcomes. The continuing widespread adoption of geospatial technologies is likely to increase national capacity to develop environmental health policies that reflect local and country-level needs and improve public health research and practice.

Table 13 Data challenges that remain to be addressed with the GIS data from the Centre for Remote Sensing and Geographic Information Services (CERSGIS)

Availability	Shapefile with all 216 districts in Ghana unavailable and was created manually based on an older version with 137 districts
	Central repository for GIS data for all of Ghana does not yet exist
	Some types of data layers of key importance to health (e.g., health centers, WASH infrastructure) are unavailable
Access	Paywalls limit access, especially for researchers in low- and middle-income settings
Completeness	Some data layers had missing/incomplete components, such as streams and roads that did not connect to branches and tributaries
Quality	Some district and region names in shapefile were spelled incorrectly and/or did not match the names of districts from DHIMS or the 2010 Census
	No mechanism readily available to share corrections and changes to data made by individual research groups and governmental agencies
	Misalignments exist among various programs (ArcGIS, Google Earth, Open Street Map, etc.) and across various countries for features such as town outlines
	Ground-truthing methods, if any, are unclear
Complexity	Temporal data are lacking for sociodemographic and WASH variables; limits more complex spatiotemporal analyses
	Infrastructure shapefiles (road networks, households, health centers, schools) are unreliable; limits analyses such as network analysis
	Data are mostly too coarsely aggregated to assess key spatiotemporal relationships (e.g., surface water access points are missing, latitude/longitude coordinates for clinics are missing, or home locations are unknown)
Delivery	Target audience is unclear
Comprehension and interpretability	Urban-rural delineation within each district is not clear

References

Adenowo, A.F., et al. 2015. Impact of human schistosomiasis in sub-Saharan Africa. *Brazilian Journal of Infectious Diseases* 19 (2): 196–205.

Agyepong, K.A. 2010. Integrated management approach to municipal solid waste treatment in peri-urban areas of Sub-Saharan Africa, Case Study Ghana. In *Engineering & physical sciences*. Guildford: University of Surrey.

Alagidede, P., W. Baah-Boateng, and E. Nketiah-Amponsah. 2013. The Ghanaian economy: An overview. *Ghanaian Journal of Economics* 1 (1): 4–34.

Al-Delaimy, A.K., et al. 2014. Epidemiology of intestinal polyparasitism among orang Asli school children in rural Malaysia. *PLoS Neglected Tropical Diseases* 8 (8): e3074.

Arku, R.E., et al. 2016. Geographical inequalities and social and Environmental risk factors for under-five mortality in Ghana in 2000 and

2010: Bayesian spatial analysis of Census data. *PLoS Medicine* 13 (6): e1002038.
Ayanian, J.Z., et al. 1993. The relation between health insurance coverage and clinical outcomes among women with breast cancer. *New England Journal of Medicine* 329 (5): 326–331.
Baker, D.W., et al. 1998. Health literacy and the risk of hospital admission. *Journal of General Internal Medicine* 13 (12): 791–798.
Barcus, M.J., et al. 2007. Demographic risk factors for severe and fatal vivax and falciparum malaria among hospital admissions in northeastern Indonesian Papua. *American Journal of Tropical Medicine and Hygiene* 77 (5): 984–991.
Bartley, M. 2001. Pros and cons of data dollection using a census: The experience of the Caribbean countries. In *International seminar on the measurement of disability*. United Nations Statistical Division.
Basu, R., and J.M. Samet. 2002. Relation between elevated ambient temperature and mortality: A review of the epidemiologic evidence. *Epidemiologic Reviews* 24 (2): 190–202.
Bhagavan, M.R. n.d. Technological leapfrogging by developing countries. In *Encyclopedia of Life Support Systems (EOLSS): Globalization of technology*.
Bhalla, K., et al. 2010. Availability and quality of cause-of-death data for estimating the global burden of injuries. *Bulletin of the World Health Organization* 88: 831–838c.
Bhaskaran, K., et al. 2010. Short term effects of temperature on risk of myocardial infarction in England and Wales: Time series regression analysis of the Myocardial Ischaemia National Audit Project (MINAP) registry. *British Medical Journal* 341: c3823.
Cairncross, S., and R. Feachem. 1993. *Environmental health engineering in the tropics*. 2nd ed, 320. New York: Wiley.
Centers for Disease Control and Prevention (CDC). *GIS and Public Health at CDC*. 2016 cited 2019. Available from: https://www.cdc.gov/gis/index.htm.
Centre for Remote Sensing and Geographic Information Services (CERSGIS). *GIS & GPS*. 2019 cited 2019. Available from: http://www.cersgis.org/#.
Chui, K.K., et al. 2009. Geographic variations and temporal trends of Salmonella-associated hospitalization in the U.S. elderly, 1991–2004: a time series analysis of the impact of HACCP regulation. *BMC Public Health* 9: 447.
Chui, K., et al. 2011a. Hospitalization of the elderly in the US for nonspecific gastrointestinal diseases: A search for etiological clues. *American Journal of Public Health* 101 (11): 2082–2086.
Chui, K.K., S.A. Cohen, and E.N. Naumova. 2011b. Snowbirds and infection–new phenomena in pneumonia and influenza hospitalizations from winter migration of older adults: A spatiotemporal analysis. *BMC Public Health* 11: 444.
Codjoe, S., et al. 2013. *2010 Population & housing census: National analytical report.* Ghana.
Cohen, S.A., et al. 2008. The SEEDs of two gastrointestinal diseases: Socioeconomic, environmental, and demographic factors related to cryptosporidiosis and giardiasis in Massachusetts. *Environmental Research* 108 (2): 185–191.
———. 2010. Trends for influenza and pneumonia hospitalization in the older population: Age, period, and cohort effects. *Epidemiology and Infection* 138 (8): 1135–1145.
Cohen, S.A., K.K. Chui, and E.N. Naumova. 2011. Influenza vaccination in young children reduces influenza-associated hospitalizations in older adults, 2002-2006. *Journal of the American Geriatric Society* 59 (2): 327–332.
Conlon, K.C., et al. 2011. Preventing cold-related morbidity and mortality in a changing climate. *Maturitas* 69 (3): 197–202.
Crocker, J., et al. 2016. Impact evaluation of training natural leaders during a community-led total sanitation intervention: A cluster-randomized field trial in Ghana. *Environmental Science and Technology* 50 (16): 8867–8875.
Cummins, S., et al. 2005. Neighbourhood environment and its association with self rated health: Evidence from Scotland and England. *Journal of Epidemiology and Community Health* 59 (3): 207–213.
Curriero, F.C., et al. 2002. Temperature and mortality in 11 cities of the eastern United States. *American Journal of Epidemiology* 155 (1): 80–87.
Dewalt, D.A., et al. 2004. Literacy and health outcomes: A systematic review of the literature. *Journal of General Internal Medicine* 19 (12): 1228–1239.
Ekpo, U.F., et al. 2008. Geographical information system and predictive risk maps of urinary schistosomiasis in Ogun state, Nigeria. *BMC Infectious Diseases* 8: 74.
ESRI. 2019. Available from: https://www.esri.com/en-us/about/about-esri/overview.
Ferraz, F., et al. 2017. Differences and inequalities in relation to access to renal replacement therapy in the BRICS countries. *Ciencia & Saude Coletiva* 22 (7): 2175–2185.
Fink, G., I. Günther, and K. Hill. 2011. The effect of water and sanitation on child health: Evidence from the demographic and health surveys 1986–2007. *International Journal of Epidemiology* 40 (5): 1196–1204.
Fradelos, E.C., et al. 2014. Health based geographic information systems (GIS) and their applications. *Acta Informatica Medica* 22 (6): 402.
Ghana Open Source Data Initiative. 2019. Available from: https://sustainabledevelopment-ghana.github.io/.
Ghana Statistical Service. 2013. In *2010 Population and Housing Census: National Analytical Report*, ed. K. Awusabo-Asare, 409. Ghana: Ghana Statistical Service.
———. 2014. *2010 Population and Housing Census.* Ghana.
———. *About Us*. 2019a cited 2019. Available from: http://www.statsghana.gov.gh/.
———. *Central Data Catalog*. 2019b. Available from: http://www2.statsghana.gov.gh/nada/index.php/catalog.
———. *Ghana data for Sustainable Development Goal indicators*. 2019c. Available from: https://sustainabledevelopment-ghana.github.io/.
Grimes, J.E., et al. 2014. The relationship between water, sanitation and schistosomiasis: A systematic review and meta-analysis. *PLoS Neglected Tropical Diseases* 8 (12): e3296.
Hajak, G. 2001. Sine Study Group, Epidemiology of severe insomnia and its consequences in Germany. *European Archives of Psychiatry and Clinical Neuroscience* 251 (2): 49–56.
Hobcraft, J.N., J.W. McDonald, and S.O. Rutstein. 1984. Socio-economic factors in infant and child mortality: A cross-national comparison. *Population Studies* 38 (2): 193–223.
Howard, G., et al. 2003. *Domestic water quantity, service level and health*. Geneva: World Health Organization.
Jagai, J., and E. Naumova. 2009. Clostridium difficile-associated disease in the elderly, United States. *Emerging Infectious Diseases* 15 (2): 343–344.
Jagai, J.S., et al. 2009. Seasonality of cryptosporidiosis: A meta-analysis approach. *Environmental Research* 109 (4): 465–478.
———. 2010. Patterns of protozoan infections: Spatiotemporal associations with cattle density. *EcoHealth* 7 (1): 33–46.
———. 2012a. Seasonality of rotavirus in South Asia: A meta-analysis approach assessing associations with temperature, precipitation, and vegetation index. *PLoS One* 7 (5): e38168.
———. 2012b. Seasonal patterns of gastrointestinal illness and streamflow along the Ohio River. *International Journal of Environmental Research and Public Health* 9 (5): 1771–1790.
Joy, R., et al. 2008. Impact of neighborhood-level socioeconomic status on HIV disease progression in a universal health care setting. *Journal of Acquired Immune Deficiency Syndromes* 47 (4): 500–505.
Keatinge, W.R., and G.C. Donaldson. 2001. Mortality related to cold and air pollution in London after allowance for effects of associated weather patterns. *Environmental Research* 86 (3): 209–216.

Kosinski, K.C., et al. 2011a. A novel community-based water recreation area for schistosomiasis control in rural Ghana. *Journal of Water, Sanitation and Hygiene for Development* 1 (4): 259–268.

———. 2011b. Diagnostic accuracy of urine filtration and dipstick tests for Schistosoma haematobium infection in a lightly-infected population of Ghanaian schoolchildren. *Acta Tropica* 118: 123–127.

———. 2012. Effective control of Schistosoma haematobium infection in a Ghanaian community following installation of a water recreation area. *PLoS Neglected Tropical Diseases* 6 (7): e1709.

———. 2016a. A mixed-methods approach to understanding water use and water infrastructure in a schistosomiasis-endemic community: Case study of Asamama, Ghana. *BMC Public Health* 16: 322.

———. 2016b. Agreement among four prevalence metrics for urogenital Schistosomiasis in the eastern region of Ghana. *BioMed Research International: Advances in Emerging and Neglected Infectious Diseases* 2016: 1–11.

Kulinkina, A., et al. 2017. Indicators of improved water access in the context of schistosomiasis transmission in rural eastern region, Ghana. *Science of the Total Environment* 579: 1745–1755.

———. 2019. Contextualizing Schistosoma haematobium transmission in Ghana: Assessment of diagnostic techniques and individual and community water-related risk factors. *Acta Tropica* 194: 195–203.

Liss, A., et al. 2017. Heat-related hospitalizations in older adults: An amplified effect of the first seasonal Heatwave. *Scientific Reports* 7: 39581.

Liu, L., et al. 2011. Associations between air temperature and cardio-respiratory mortality in the urban area of Beijing, China: a time-series analysis. *Environmental Health* 10: 51.

Lofgren, E., et al. 2007. Influenza seasonality: Underlying causes and modeling theories. *Journal of Virology* 81 (11): 5429–5436.

Lofgren, E.T., et al. 2010. Disproportional effects in populations of concern for pandemic influenza: Insights from seasonal epidemics in Wisconsin, 1967-2004. *Influenza Other Respiratory Viruses* 4 (4): 205–212.

Madon, S., et al. 2018. The role of community participation for sustainable integrated neglected tropical diseases and water, sanitation and hygiene intervention programs: A pilot project in Tanzania. *Social Science & Medicine* 202: 28–37.

Martel, R.A., et al. 2019. Assessment of urogenital schistosomiasis knowledge among primary and junior high school students in the Eastern Region of Ghana: A cross-sectional study. *PLoS One* 14 (6): e0218080.

Mata-Alvarez, J. 2002. *Biomethanization of the organic fraction of municipal solid wastes*. London: IWA Publishing.

Michelozzi, P., et al. 2007. Assessment and prevention of acute health effects of weather conditions in Europe, the PHEWE project: Background, objectives, design. *Environmental Health* 6: 12.

Mihelcic, J.R., et al. 2009. *Field guide to environmental engineering for development workers: Water, sanitation, and indoor air*. Reston: American Society of Civil Engineers.

Moorthy, M., et al. 2012. Deviations in influenza seasonality: Odd coincidence or obscure consequence? *Clinical Microbiology and Infection* 18 (10): 955–962.

Mor, S.M., et al. 2009. Cryptosporidiosis in the elderly population of the United States. *Clinical Infectious Diseases* 48 (6): 698–705.

———. 2011. Pneumonia and influenza hospitalization in HIV-positive seniors. *Epidemiology and Infection* 139 (9): 1317–1325.

Mor, S.M., A. DeMaria Jr., and E.N. Naumova. 2014. Hospitalization records as a tool for evaluating performance of food- and water-borne disease surveillance systems: A Massachusetts case study. *PLoS One* 9 (4): e93744.

Mrkić, S. n.d. *Principles and recommendations: essential features and census methodologies*. Lagos: United Nations Statistics Division.

National Research Council. 2012. In *The case for international sharing of scientific data: A focus on developing countries: Proceedings of a symposium, in Part two: Status of access to scientific data, P.F.U.*, ed. Kathie Bailey Mathae, 17–32. Washington D.C.: The National Academies Press.

Naumova, E.N., et al. 2007. Seasonality in six enterically transmitted diseases and ambient temperature. *Epidemiology and Infection* 135 (2): 281–292.

———. 2009. Pneumonia and influenza hospitalizations in elderly people with dementia. *Journal of the American Geriatrics Society* 57 (12): 2192–2199.

Nori-Sarma, A., et al. 2017. Opportunities and challenges in public Health data collection in southern Asia: Examples from Western India and Kathmandu Valley, Nepal. *Sustainability* 9 (7): 1106.

Pandey, A., et al. 2010. Health information system in India: Issues of data availability and quality. *Demography India* 39 (1): 111–128.

President's Malaria Initiative. 2014. Ghana's innovative health information management system gains African recognition.

Rehman, I.H., et al. 2005. Availability of kerosene to rural households: A case study from India. *Energy Policy* 33 (17): 2165–2174.

Rogot, E., and S.J. Padgett. 1976. Associations of coronary and stroke mortality with temperature and snowfall in selected areas of the United States, 1962-1966. *American Journal of Epidemiology* 103 (6): 565–575.

Rutstein, S.O. 2000. Factors associated with trends in infant and child mortality in developing countries during the 1990s. *Bulletin of the World Health Organization* 78: 1256–1270.

Sallis, J.F., et al. 2009. Neighborhood built environment and income: Examining multiple health outcomes. *Social Science and Medicine* 68 (7): 1285–1293.

Samwine, T. 2017. Challenges and prospects of solid waste management in Ghana. *International Journal of Environmental Monitoring and Analysis* 5 (4): 96.

Sudore, R.L., et al. 2006. Limited literacy and mortality in the elderly: the health, aging, and body composition study. *Journal of General Internal Medicine* 21 (8): 806–812.

Timæus, I.M., and M. Jasseh. 2004. Adult mortality in sub-Saharan Africa: Evidence from demographic and Health surveys. *Demography* 41 (4): 757–772.

U.S. Census Bureau, Labor Force Statistics. 2017.

UNESCO. 2008. *Using a literacy module in household surveys: a guidebook*. Bangkok: UNESCO Bangkok.

UNICEF. *Universal Access to Sanitation - UNICEF Data*. 2017 cited 2019. Available from: https://data.unicef.org/topic/water-and-sanitation/sanitation/.

United Nations General Assembly. 2015. *Transforming our world: the 2030 Agenda for Sustainable Development*, in *70/1*, United Nations General Assembly, ed. United Nations.

United States Environmental Protection Agency. 2009. *Municipal solid waste in the United States: 2009 facts and figures*. Office of Solid Waste.

USAID. *District Health Information Management System (DHIMS) Digitization*. 2016 cited 2019. Available from: https://partnerships.usaid.gov/partnership/district-health-information-management-system-dhims2-digitization.

Varenne, B., P.E. Petersen, and S. Ouattara. 2004. Oral health status of children and adults in urban and rural areas of Burkina Faso, Africa. *International Dental Journal* 54 (2): 83–89.

WHO/UNICEF. 2017. *Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines*. Geneva.

WHO/UNICEF Joint Monitoring Programme (JMP). 2015. *Progress on sanitation and drinking water – 2015 update and MDG assessment*. Geneva: World Health Organization.

Wilkinson, R.G., and M.G. Marmot. 2003. Social determinants of health: The solid facts. In *World Health Organization*. Copenhagen: Regional Office for Europe.

Wolf, M.S., J.A. Gazmararian, and D.W. Baker. 2005. Health literacy and functional health status among older adults. *Archives of Internal Medicine* 165 (17): 1946–1952.

World Bank. 2003. Household energy use in developing countries: a multicountry study (English). In *Energy Sector Management Assistance Programme (ESMAP) technical paper series*. Washington, DC.

———. 2018. *What a waste 2.0: A global snapshot of solid waste management to 2050*. Washington DC.

World Health Organization. 1993. *Guidelines for drinking-water quality: Volume 1 Recommendations*. 2nd ed. Geneva: World Health Organization.

———. *Sanitation*. 2018 cited 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/sanitation.

———. n.d. *Drinking water*.

Wrable, M., et al. 2019. The use of remotely sensed environmental parameters for spatial and temporal schistosomiasis prediction across climate zones in Ghana. *Environmental Monitoring and Assessment* 191 (Suppl 2): 301.

Modeling Distributional Potential of Infectious Diseases

Abdallah M. Samy, Carlos Yáñez-Arenas, Anja Jaeschke, Yanchao Cheng, and Stephanie Margarete Thomas

Introduction

Most epidemiological studies involve mapping the spread of diverse diseases in space and time placing a methodological suite of geospatial and ecological "toolkits" into better use to lead to greater opportunities for disease surveillance and prevention. Mapping disease risk has a crucial role in epidemiology and public health to illuminate details on spatial patterns of disease cases and their potential to spread across diverse regions of the world. Historically, the field of epidemiology has applied various geospatial methods to summarize geographic patterns for mapping disease transmission risk. These methods are based on generalizing information or averaging data to summarize geographic patterns in smoothed surface maps. Although these approaches provide basic information about spatial patterns for disease mapping, they are strongly biased by surveillance efforts (e.g., if most data were collected close to roads; Kadmon et al. 2004), and they do not usually give details of the complexity and heterogeneity of biological systems (Peterson 2006). These systems are determined by the interaction of different sets of species contributing to the complex transmission cycle of a particular disease in question (e.g., hosts, vectors, and pathogens). Important insights can be obtained from better understanding of the ecology of each species in the complex disease transmission cycle (Peterson 2007). By the means of mathematical formulas describing these processes within the transmission cycle, epidemiological models (EM) can project the spread of infectious diseases.

Ecological niche modeling (ENM) combines occurrence records and environmental information to project the habitat suitability for a certain species and to identify ecological drivers of its distribution. It has revolutionized the field of epidemiology for understanding diverse patterns of disease spread and dynamics across the world (Peterson 2014). The epidemiological applications of ENM include disease forecast during outbreak events to anticipate areas at risk of disease transmission for guiding surveillance and control program (e.g., detailed map of Zika virus spread; Samy et al. 2016a). This chapter also provides the conceptual framework of EM and offers new opportunities for integrating both modeling approaches (EM and ENM) to better assess the distributional potential of infectious diseases, particularly vector-borne diseases.

Ecological Niche Modeling

Ecological niche models (ENMs) – also termed as species distribution models, habitat suitability models, or environmental envelope models – are applied to estimate the potential distribution of a species in space and time. ENMs are used to identify and characterize statistical patterns linking

A. M. Samy (✉)
Entomology Department, Faculty of Science, Ain Shams University, Cairo, Egypt

C. Yáñez-Arenas
Laboratorio de Ecología Geográfica, Unidad de Conservación de la Biodiversidad, Parque Científico y Tecnológico de Yucatán, UMDI-Sisal, Facultad de Ciencias-Universidad Nacional Autónoma de México, Mérida, Yucatán, Mexico
e-mail: cyanez@comunidad.unam.mx; carlos_yanez@ciencias.unam.mx

A. Jaeschke · Y. Cheng
Department of Biogeography, University of Bayreuth, Bayreuth, Germany
e-mail: anja.jaeschke@uni-bayreuth.de; yanchao1.cheng@uni-bayreuth.de

S. M. Thomas
Department of Biogeography, University of Bayreuth, Bayreuth, Germany

Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Bayreuth, Germany
e-mail: stephanie.thomas@uni-bayreuth.de

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_18

occurrence data to environmental predictors and fitting values to model parameters (Elith and Leathwick 2009; Franklin and Miller 2010; Peterson et al. 2011). This approach transfers the geographical space, i.e., occurrence records, into environmental space, represented by environmental variables that are important drivers of the species' distribution. Operationally, the study area is generally represented as a grid of equally sized raster cells containing information on the occurrence of the species and the corresponding environmental predictors. Theoretically, the objective of this type of modeling is to characterize the ecological niche of a focal species. Therefore, formalizing and clarifying on this concept are essential to properly understand the principles of this approach.

Most ecologists agree that the concept of niche plays a crucial role in ecology (Real and Levin 1991). However, since it was first used in an ecological context (Johnson 1910), this term has acquired a wide variety of definitions and interpretations. The different niche definitions can be separated into two main categories: Eltonian and Grinnellian niches (Soberón 2007). The Eltonian class focuses on interspecific interactions, resource-consumer dynamics, and other aspects that can generally be measured at local scales (Elton 1927; MacArthur 1968; Vandermeer 1972; Leibold 1996). On the other hand, the Grinnellian class is defined by non-interactive variables and environmental conditions that influence species' distributions on broad scales (Grinnell 1917; Whittaker et al. 1973; Austin and Smith 1989; Peterson 2003). This Grinnellian perspective represents the theoretical basis underlying the construction of correlative species' distribution model (SDM) and ecological niche model (ENM). Recent studies introduced the concept of BAM framework (i.e., the framework that summarizes the three sets of conditions shaping the species niche; B refers to biotic conditions, A refers to abiotic conditions, and M refers to movement) to incorporate dispersal capacity as an additional parameter to estimate a species' distribution (Barve et al. 2011; Qiao et al. 2017; Gherghel et al. 2019). A summary of the key definitions of the ecological niche modeling and the BAM framework is presented in Box 1.

Box 1 A Summary of the Key Definitions of the Ecological Niche Modeling Concept

Species: Refers to any organism or component in the transmission cycle (e.g., pathogen, vector, or reservoir host).

BAM diagram: A diagram that summarizes the three sets of conditions that together shape a species' distribution: B, biotic conditions; A, abiotic conditions; and M, movement of the species.

Ecological niche modeling: Estimation of the suitable conditions from occurrence records considering the assumptions of the factors in the BAM diagram.

Eltonian niche: A niche concept concerned with community ecology questions, defined at small spatial extents at which experimental manipulations are feasible, emphasizing the functional role of species in communities and including models of resource consumption and impacts.

Grinnellian niche: Niche concepts defined based on environmental space of noninteracting scenopoetic environmental variables that the species can tolerate.

Occupied niche space: The subset of environmental space that the species inhabits; it is equivalent to the set of environments in the occupied distributional area.

Invadable niche space: The subset of environmental space corresponding to the elements of geographic space that the species could occupy if distributional constraints were to be overcome.

Model calibration (model training): The step involved in building a model when the species' niche is estimated based on primary occurrence data and values of environmental variables (i.e., calibrated area may also depend on the accessible area of the species in question).

Model projection: The steps involved in transferring the model from calibrated area to another region or another time period.

Ensemble prediction: A consensus prediction of a niche based on combining results of different algorithms, alternative parameterizations of the same method, or multiple iterations of stochastic methods to generate a composite value of suitability.

Thresholding: The process of selecting a threshold of occurrence, for converting continuous model output to a binary prediction of "present" versus "absent."

Threshold-dependent: An approach for evaluating model performance or model robustness based on a binary prediction, typically obtained by applying a threshold to a continuous prediction of suitability.

Threshold-independent: An approach for evaluating model performance or model robustness based on continuous prediction of suitability and without applying the thresholding process.

Extrapolation: Prediction into environmental values beyond the range of the values in the calibration area. The process is very common in modeling projection into different times or different regions.

Model evaluation (model testing): The process of model testing based on different approaches (e.g., AUC, partial ROC, and independent data records).

Uncertainty: Estimation of an index indicating the possible level of error regarding data (i.e., occurrence records and environmental data).

(continued)

To date, different model algorithms are available to estimate niche models. These algorithms were developed for different types of distribution data (e.g., presence-absence data or presence-only data) and different modeling purposes. Maximum entropy (MaxEnt), generalized linear models (GLM), generalized additive models (GAM), and boosted regression trees (BRT) are commonly applied in ecology and geospatial analyses owing to their superior performance compared to other algorithms (Elith and Leathwick 2009; Carvalho et al. 2017). To date, applying diverse algorithms to the same data is not a common practice in studies which model vector-borne diseases (Carvalho et al. 2017; Semenza and Suk 2018). A recent study also recognized the idea of "no silver bullets," indicating that there is no single best algorithm in ecological niche modeling under all circumstances (Qiao et al. 2015). Some studies used ensemble modeling to avoid any possible uncertainties from applying different algorithms to model the ecological niche of the species in question (Buisson et al. 2010; Zhu and Peterson 2017; Eneanya et al. 2018; Hao et al. 2019); however, there is limited unambiguous information on the performance of individual models versus ensemble models (Hao et al. 2019).

Beyond the estimation of the importance of individual drivers, ENMs calculate the environmental suitability for each raster cell and transfer it to the geographical space. These outputs are mainly visualized as maps showing areas with high to low suitability or binary information reflecting the potential presence or absence of species in question in the thresholded models (Fig. 1).

A Practical Framework of ENM

Building ecological niche models consists of executing a concrete list of steps (Box 2). This list of modeling processes was recently updated to include additional steps seeking better improvements in estimating species' ecological niches. This section of the chapter presents detailed steps of building ecological niche models.

Step 1: Data Collection and Cleaning

Niche modeling requires two types of data: (1) occurrence records documenting sampling sites where a species has been observed and (2) covariates or predictors presenting the environmental conditions that elaborate the ecological requirements of the species in question. Occurrence records are point localities defined by x and y coordinates or longitude and latitude; on the other hand, predictors are Geographic Information System (GIS) raster layers that are used to characterize variation in environmental conditions across the study area. There are several sources from which occurrence data can be obtained. Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) and iNaturalist (https://www.inaturalist.org/) provided comprehensive online repositories to search data for any species, VertNet data portal (http://vertnet.org/) for vertebrate biodiversity data, Walter Reed Biosystematics Unit VectorMap portal for disease vector data (e.g., mosquito and tick vectors; http://vectormap.si.edu/), and HealthMap (https://www.healthmap.org/) to provide information on historical and recent disease occurrences across the world. Other sources included routine active surveillance of the species or disease under the study and museum collections. Data from any of these sources should pass through a careful error-check to avoid any associated problems (e.g., taxonomic misidentification or errors in assigning the geographic location of the species). The environmental datasets may be interpolated weather stations climate layers (e.g., WorldClim; https://www.worldclim.org/) or remotely sensed data layers (e.g., Moderate Resolution Imaging Spectroradiometer (MODIS); https://modis.gsfc.nasa.gov/). The satellite data may require further geospatial processing to produce cloud-free satellite imagery or obtaining data with a specific temporal or seasonal coverage to obtain time-specific species niche model.

Occurrence data may require data thinning to reduce potential bias in the dataset. Some of the procedures used to achieve this step include removing duplicates, rarifying data based on a distance filter to omit all redundant records occurring in a single pixel of environmental raster data, and balancing the density of occurrence between countries to account for marked differences in sampling efforts. All these steps lead to a balanced thinned dataset and candidate covariates required for calibrating the ecological models.

Step 2: Model Calibration

This step describes the selection of calibration area and the modeling algorithm (see the ecological niche modeling section) to estimate the species' ecological niche by correlating occurrence records and environmental variables. Calibration area may represent the entire accessible area (M; see Box 1) or a subset of M based on occurrence availability and the question under the investigation.

Step 3: Model Evaluation

The next step to model calibration is to test the model predictions and evaluate how robust the model is to predict species' occurrences in unsampled areas. This step is essential before interpreting model results, projecting the

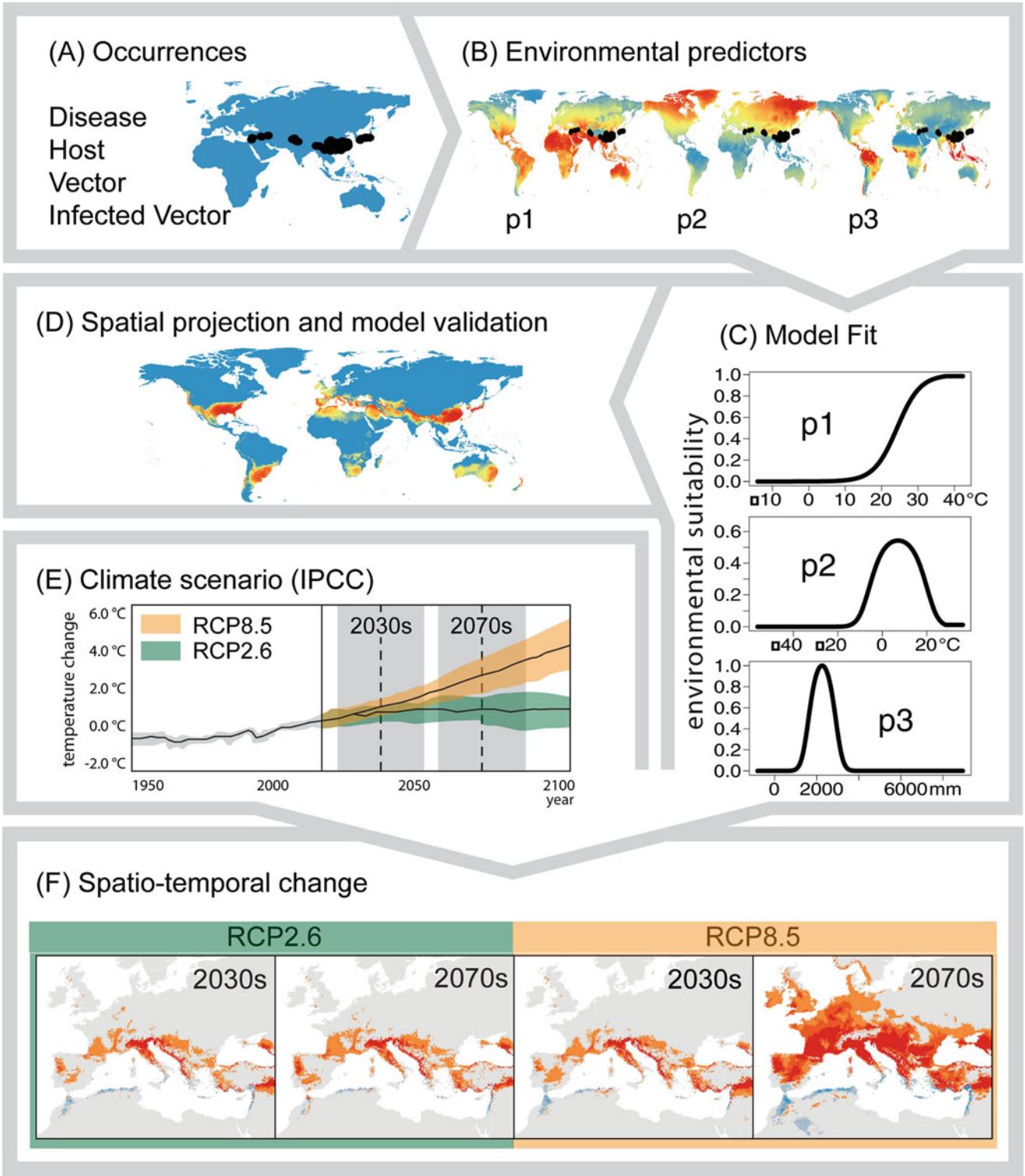
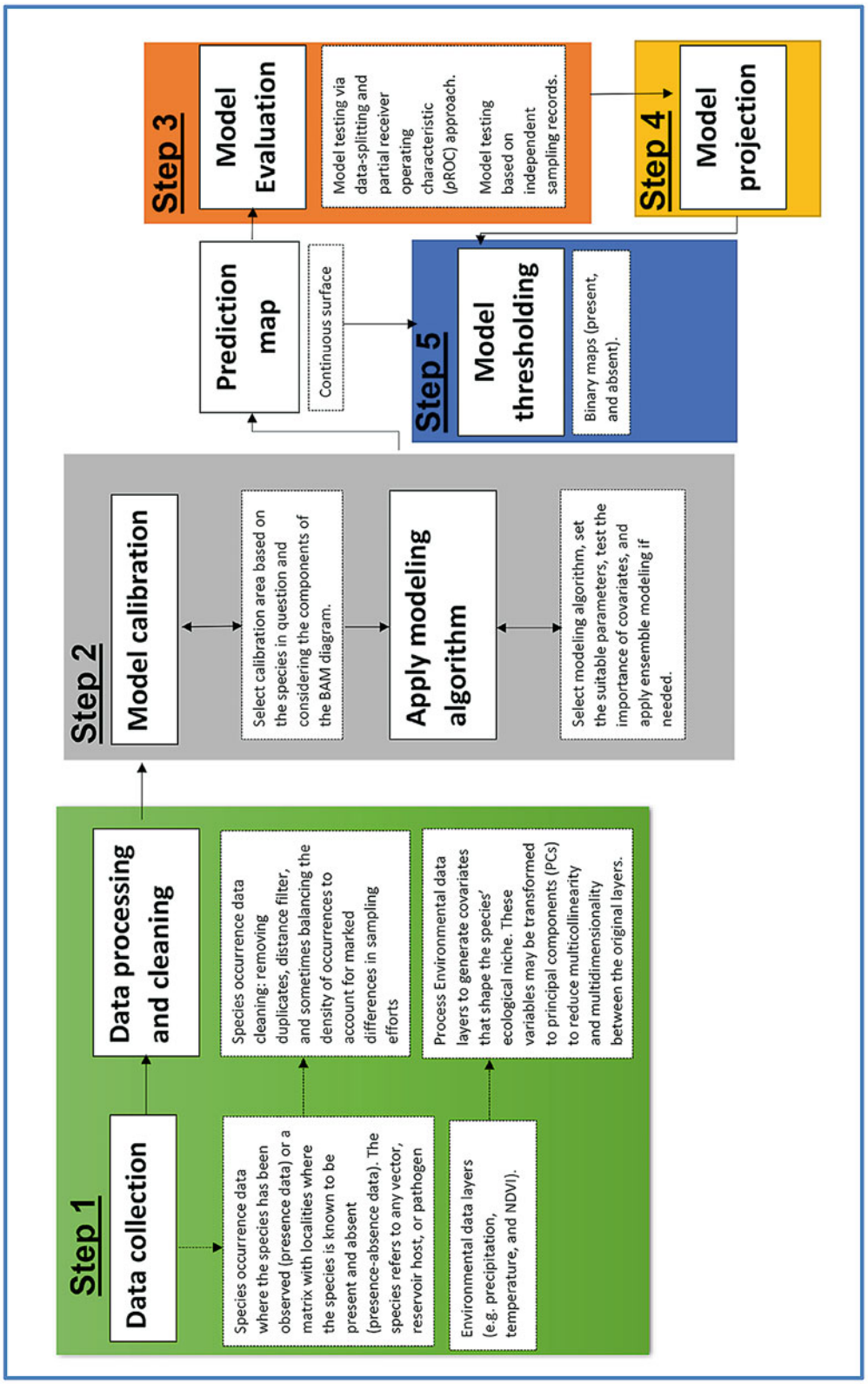


Fig. 1 Characteristic workflow of ecological niche model (ENM). (**a**) Occurrence records for the vector, infected vector, host, or disease are compiled. (**b**) Variable selection: environmental predictors – p1 = summer temperature in °C, p2 = winter temperature in °C, and p3 = annual precipitation in mm. (**c**) The best fit model that describes the probability of species occurrence in multivariate environmental space is developed (i.e., environmental suitability based on p1–3). Several different algorithms can be utilized such as maximum entropy (MaxEnt), boosted regression tree (BRT), or generalized linear model (GLM). (**d**) A spatial projection and validation of the model is made based on the environmental predictors. (**e**) Climate change scenarios mainly derived from the Intergovernmental Panel on Climate Change (IPCC) (representative concentration pathways (RCP) 2.6 and 8.5) for different time frames (mainly two or three decades) are chosen. Results shown here are the average temperature change over time for an optimistic RCP 2.6 (green) and a pessimistic RCP 8.5 (orange) climate change scenario. An estimate of uncertainty is shown as color shadings around the black lines. (**f**) Using data from global or regional climate models, further projections for the selected scenarios and time frames (gray vertical bars in panel **e**) are made. Ideally, different climate models are used to derive the ENM. (Adapted from Tjaden et al. 2018)

Box 2 A Summary of the Routine Steps of the Ecological Niche Modeling Approach



model to other regions or transferring the model through time. Model evaluation approaches are divided into two main classes: approaches designed for thresholded binary predictions (i.e., threshold dependent) and approaches used for continuous predictions (i.e., threshold-independent). Many modeling studies used the threshold-independent area under the receiver operating characteristic curve (AUC) as a standard approach to assess the robustness of model predictions (Pigott et al. 2014; Moyes et al. 2016; Messina et al. 2016; Messina et al. 2019). However, the AUC approach is known to have many problems (Lobo et al. 2008; Peterson et al. 2008). Major problems associated with AUC-based evaluation are the equal weighting of omission and commission errors and lack of information on the spatial distribution of model errors (Lobo et al. 2008). An alternative method in which some of the limitations of traditional AUC are resolved is the partial receiver operating characteristic (*p*ROC) (Peterson et al. 2008). *p*ROC is based on random data-splitting to obtain two subsets for calibrating and evaluating model predictions. The threshold-dependent approach uses records collected independently and applies the binomial test to assess if the evaluation data fall into regions of a thresholded binary prediction more often than expected by chance.

Step 4: Model Projection

Some studies used ecological niche modeling to transfer a calibrated model into a new region or into a different time period. These applications are very common in predicting the potential spread of invasive species from calibrated range area (i.e., endemic range area) to new regions (i.e., invasive range) that may be at risk of species invasion. Another application is to transfer the model to identify potential distributional shifts under alternative climate scenarios. Model projection generates additional problems related to extrapolation beyond the range of environmental conditions in the calibration area; however, areas of strict extrapolation can be defined by Mobility-Oriented Parity (MOP) analysis (Owens et al. 2013; see details in strengths and limitations of ENM section).

Step 5: Model Thresholding

Model thresholding refers to the conversion of a continuous model output to binary maps representing the species potential presence and absence. The selection of a thresholding value is based on numerous methods (Manel et al. 2001; Pearson et al. 2004; Liu et al. 2005). The most common practice for model thresholding is based on a maximum allowable omission error which is commonly set as 5% (Peterson et al. 2008); however, these values can be changed based on the accuracy of occurrence records. The defined percentage is assumed to have misrepresented environmental data.

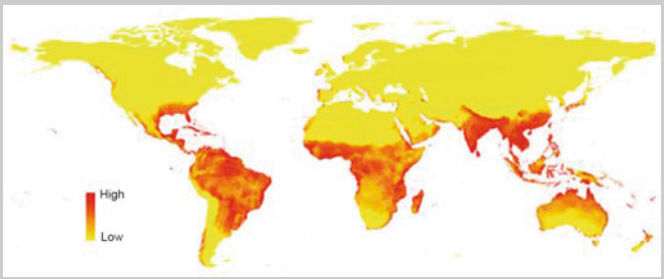
Applications of ENM in Disease Mapping

Ecological niche models have become a very popular tool in biogeography, conservation biology, ecology, paleoecology, and wildlife management over the last two decades (Araújo and Guisan 2006). The applications of ENMs to the world of diseases are somewhat more recent; however, there are currently a very common suite of toolkits in epidemiology and public health (Peterson 2006). Common implementations of ENMs include modeling actual geographic patterns in disease incidence, mapping potential distribution of candidate vectors and reservoir hosts, identifying major drivers of disease emergence and dynamics, and forecasting disease risk under climate change in future (Rogers and Randolph 2003; Eisen and Eisen 2011; Hay et al. 2013). Specifically, the ENMs have also been used to anticipate the geographic potential of species invasion in other regions, anticipate their distributional responses to environmental changes, and infer likely interactions in diseases transmission systems (Peterson 2007). Several studies mapped diverse vector species (Fischer et al. 2011; Samy et al. 2016b; Kraemer et al. 2015; Alkishe et al. 2017; Samy et al. 2018; Kamal et al. 2018; Kraemer et al. 2019), reservoir hosts (Gholamrezaei et al. 2016; Samy et al. 2016c, 2018), and etiological agents (Samy et al. 2016a, 2018; Carlson et al. 2016; Tjaden et al. 2017). Another study modeled infected vectors to project into areas at disease transmission risk and revealed the most important environmental drivers for their distributions (Mweya et al. 2016).

Calibrated models are used for spatial and temporal projections to estimate the spread of invasive species into new regions or to project future climate change impacts on species distributions. Temporal projections can also be done backward to estimate the past distributional potential of a particular species (i.e., this is known also as hindcasting to look for the species distribution in the past). Climate change projections have gained a special importance in recent years with projecting inter alia impacts of climatic changes on the distribution of disease vectors and pathogens (see section "Climate Change Impact: A Case Study of Vector-Borne Diseases"). These geospatial and ecological analyses are used to inform policymakers, veterinarians, researchers, and local human populations about areas of high disease risk. Finally, they identify the potential spread and possible shifts of vector populations to better place measures to avoid successful establishment of vector populations in new invaded areas. Box 3 summarizes different applications of ecological niche modeling.

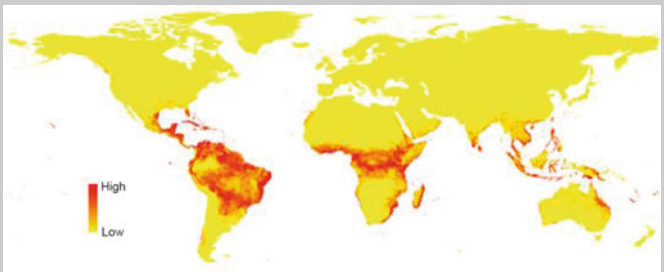
Box 3 Applications of Ecological Niche Modeling to Identify Disease Transmission Risk

1) *Estimating distributional potential of infectious diseases*
Ecological niche modeling (ENM) is widely used for estimating the distributional potential of diverse diseases including ones with complex transmission cycles (e.g., vector-borne diseases (VBDs)). ENM is used to map several VBDs including dengue, Zika, chikungunya, leishmaniasis, lymphatic filariasis (Samy et al. 2016a, c, Tjaden et al. 2017; Eneanya et al. 2018; Messina et al. 2019). These mapping efforts include modeling the distributional potential of diverse vector species too. A recent study of *Aedes aegypti* and *Ae. albopictus* mapping (Kamal et al. 2018) offer a good example for the application of ENM to estimate the ecological niche of the two arboviral species.



The map estimates the global distributional potential of arbovirus vector *Ae. aegypti*. The yellow color depicts areas with unsuitable conditions of *Ae. aegypti* occurrence. The probability of *Ae. aegypti* occurrence increases from light red to dark red.

2) *Predicting disease emergence, and the invasion of vectors and reservoir hosts to new areas*
Some vectors and reservoir hosts invade new areas based on natural drivers (Tsiamis et al. 2013; Huestis et al. 2019). Other species may invade new areas via transportation with international trades, particularly with the recent expansion of transportation networks. (Thomas et al. 2014) These activities allow the exchange of goods (e.g., tires harboring mosquito breeding sites) across country borders. ENMs provide a sensitive tool to anticipate new suitable areas. An important application for ENMs in this context is the strategy applied for mapping Zika virus during the recent Brazilian outbreak (Samy et al. 2016a). This study calibrated the model based on the data available from South and Central America and then projected the model to the entire world to characterize new areas at risk of Zika virus transmission. The projected Zika model anticipated disease risk in Africa (i.e., where disease is historically known and data is scarce), and in Asia, and continental USA where disease is recently emerged in outbreak potential.

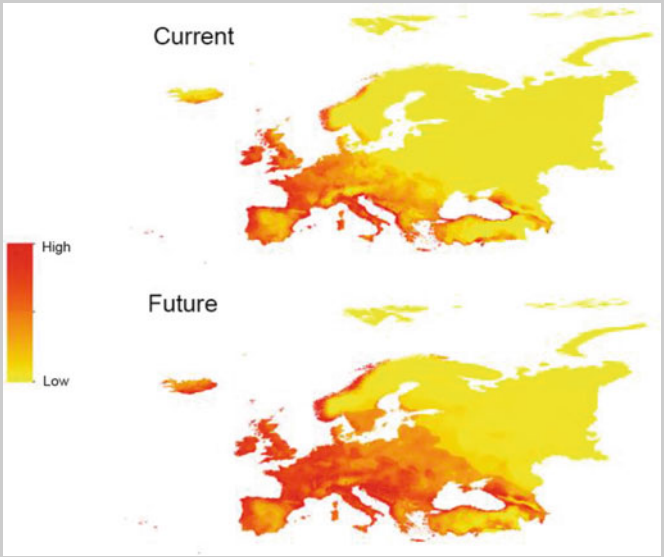


The projected global suitability of Zika virus occurrence. This model is calibrated only in South and Central America and projected to global climate. The yellow color depicts areas with unsuitable conditions of Zika occurrence. The probability of Zika occurrence increases from light red to dark red.

3) *Forecasting the potential impacts of climate change on pathogens, vectors, and reservoir hosts*
Many studies used ENMs to assess the influences of climate changes on the distributional potential of deadly diseases, their vectors, and reservoir hosts (Samy et al. 2016b, Tjaden et al. 2017; Alkishe et al. 2017; Kamal et al. 2018). Chikungunya virus is anticipated to expand under climate changes to cover broader ranges in USA, South America, Europe, Sub-Saharan Africa, and China (Tjaden et al. 2017). A recent model of arboviral vector *Aedes albopictus* suggested possible population expansion to the East of Europe to include most of Europe under the influences of climate change owing to increase in carbon dioxide (CO_2) emission (Kamal et al. 2018).

(continued)

Box 3 (continued)



The distributional potential of *Aedes albopictus* in Europe based on current and future climate conditions. The model is calibrated based on current climate and projected to the meteorological research institute (MRI-CGCM3) general circulation model and representative concentration pathway (RCP) 8.5 in 2070. The yellow color presents areas with unsuitable conditions of *Ae. albopictus* occurrence. The probability of *Ae. albopictus* occurrence increases from light red to dark red.

4) *Assessing the impacts of land use change on species distribution*
ENMs are used to assess the effect of land use changes on the distributions of etiological agent, vector, and reservoir host of an infectious disease (Wimberly et al. 2008; Santos 2017; Hess et al. 2018; Chavy et al. 2019). These studies identified significant influences of land use on the potential distributions of diverse infectious diseases. The incidence of West Nile fever elevates as rural and irrigated areas increase in the Northern Great Plains of the United States (Wimberly et al. 2008). Changes in land use and vegetation cover are major drivers to the spread of the main hantavirus reservoir *Necromys lasiurus* in Brazil (Santos 2017).

Climate Change Impact: A Case Study of Vector-Borne Diseases

Environmental conditions, including temperature and precipitation, are changing globally under the current climate change regime (Fig. 2). The maximum temperature of the warmest month is expected to increase most in northern parts of North America, Brazil, the Mediterranean region, and the polar and tundra regions of Russia by more than 5 °C by 2050 relative to the 1970s. A decline in precipitation of more than 100 mm per year is projected for the northern parts of South America, the Chilean Andean cordilleras, and the southwestern areas of the Mediterranean region.

The distribution of vectors and disease transmission is anticipated to shift in space and time under climate change conditions. In temperate zones, vector species will likely broaden their distribution and a northward shift of their range is projected (Carvalho et al. 2017; Thomas et al. 2018). In tropical climates, warming can constrain a species geographic range to areas with higher elevations or lead to local or complete extinction of vector species (Escobar and Craft 2016).

Several studies have anticipated climate change influences on the distributional potential of vector-borne diseases and their vectors (Samy et al. 2016a, b; Samy and Peterson 2016; Alkishe et al. 2017; Kamal et al. 2018). For example, the distributional potential of bluetongue virus (BTV) was anticipated to broaden in Central Africa, United States, and western Russia (Samy and Peterson 2016). The BTV was estimated based on diverse representative concentration pathways (RCPs); there is a 9% increase in the BTV range from current climate to RCP 8.5. Other studies anticipated possible shifts under climate changes in mosquito vectors (Samy et al. 2016b; Kamal et al. 2018). The potential distribution of *Culex quinquefasciatus* increased by 5% from present-day conditions to RCP 6.0; however, it decreased by 1.5% from RCP 6.0 to RCP 8.5 (Samy et al. 2016b). The arbovirus vector *Aedes albopictus* was anticipated to expand broadly in northern USA, Southern Canada, North Africa, and Europe under climate changes in 2050 and 2070 (Kamal et al. 2018). The response of these species to climate change was different; each species showed a characteristic pattern under climate changes; however, range expansion is common among these species.

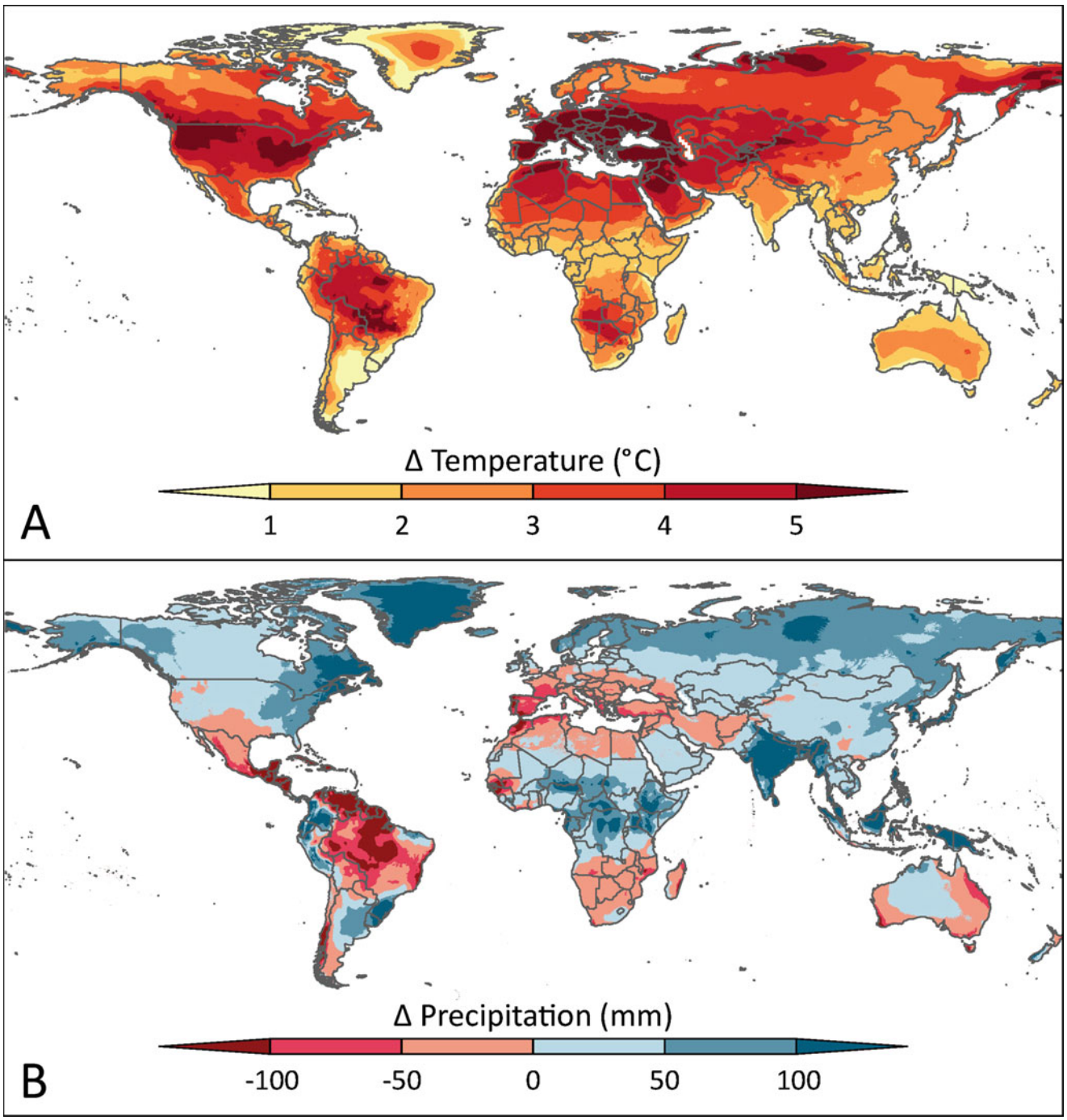


Fig. 2 Predicted future change in climatic key variables. (**a**) Maximum temperature of the warmest month. (**b**) Annual precipitation. Expected changes for 2041–2060 relative to 1960–1990 long-term averages under the RCP 8.5 climate change scenario. Data from WorldClim 1.4, based on 12 different General Circulation Models (BCC-CSM1-1, CCSM4, GFDL-CM3, GISS-E2-R, HadGEM2-ES, MPI-ESM-LR, HadGEM2-AO, HadGEM2-CC, INMCM4, IPSL-CM5A-LR, MRI-CGCM3, NorESM1-M, http://www.worldclim.org)

Strengths and Limitations of ENM

ENM represents a very powerful tool for modeling and mapping infectious disease dynamics. ENM is used to test several ecological and distributional questions related to epidemiological research. However, the use of this technique should be based on a clear understanding of ecological and biogeographical concepts, as well as mathematical aspects of methods and algorithms (Escobar and Craft 2016). A key issue for the correct use and interpretation of ecological niche model results is to fully understand their strengths and short-

comings. The main strengths of ENM are (1) the approach can implicitly include any process statistically related with the environmental predictors, (2) the approach is easier to implement than any other type of models, (3) it is more likely to identify limiting factors, and (4) availability of input data.

ENM also has several limitations, for example, it is not possible to differentiate correlation from causality, and variable selection is generally subjective (Kearney and Porter 2009). A model always represents a simplification of the reality aiming at identifying general patterns and trends. The quality of input data, i.e., the way of data collection, the reliability of the occurrence records, as well as the availability of absence information, contains the largest portion of uncertainty. Further sources of uncertainty are related to the choice of the modeling algorithm, climate models (e.g., general circulation models (GCMs)), and emission scenarios (Buisson et al. 2010). These uncertainties can be considered and reduced by using more than one algorithm (e.g., ensemble modeling); however, diverse climate models and emission scenarios provide opportunities for potential developments. The improvement of input data plays another major role in providing the required quality to project reliable potential developments.

Non-analogue climate conditions pose another uncertainty related to model projections. Non-analogue climate refers to climatic conditions that have not been experienced before in a specific study area, possibly leading to strong underestimations of the potential spread of a particular species. For example, if a species is known to occur in areas with an annual mean temperature range between 5 °C and 35 °C, climate change projections of temperature increase beyond 35 °C might lead to the exclusion of these areas as potential habitat as the model was trained up to 35 °C. Nevertheless, the species of interest might be able to tolerate temperatures above 35 °C, even though there is no empirical evidence. Hence, non-analogue climate should be considered in ENMs to identify areas of high uncertainty. A commonly applied method to address non-analogue climate is the so-called Multivariate Environmental Similarity Surface (MESS) analysis (Elith et al. 2010). MESS identifies regions of strict extrapolation and provides an index of environmental similarity between each pixel and the median of the most dissimilar variable in the calibration area. The use of the most dissimilar variable as an indicator of overall similarity in MESS analysis marks its limitations (Owens et al. 2013). A recent study introduced a Mobility-Oriented Parity (MOP; Owens et al. 2013) analysis as a modification and extension of MESS to identify regions of strict extrapolation and better characterize the degrees of novelty in projection regions.

Epidemiological Modeling

Epidemiological models (EMs) are typically used to investigate the transmission mechanics between vectors and hosts. EMs are used to describe the health state change (e.g., susceptible-exposed-infected-removed, each health state is a compartment) of both vectors and hosts with mathematical equations (Box 4). Consequently, EMs are also referred to as compartmental models, mathematical models, compartmental epidemic models, mechanic models, and mechanistic models, though these terms do not share the exact same field. Nevertheless, the output from these EMs is a threshold quantity parameter, namely the basic reproduction number (R_0; Fig. 3).

R_0 is defined as the average secondary cases caused by an infected individual during its lifetime in a completely susceptible population (Diekmann et al. 1990; Heffernan et al. 2005). When $R_0 > 1$, an outbreak of the investigated vector-borne disease can take place. The concept of the "basic reproduction" can be dated back to the "net reproduction rate" from demography (Dietz 1993), and EMs calculating R_0 can be applied to investigate infectious disease such as malaria (Dietz 1993; Heffernan et al. 2005; Delamater et al. 2019). The calculating methods and the respective interpretations of R_0 have evolved and are still evolving (Ridenhour et al. 2018; Delamater et al. 2019).

There are mainly two methods used to build an EM: the survival function and the next generation matrix (NGM) (Heffernan et al. 2005). The survival function method, as its name suggests, describes through mathematical functions whether a new infection can survive or not. It is a straightforward approach and sticks to the definition of average secondary infections resulting from a single introduction in the infected individual's lifetime. An infection can be sustained if $R_0 > 1$. A simple equation example can explain the process:

R_0 = (probability of an individual getting infected for a unit of time)
× (probability of an infected individual being infectious for a unit of time)
× (average number of secondary cases an infected individual will produce per unit of time)

Each compartment in this equation is a variable that either varies over time or remains a constant parameter. In practice, the EMs using survival functions can be more complicated (Dietz 1993; Moraga et al. 2015). However, the key concept of this method should still be clear: the R_0 calculated in this way is the average secondary infections, a straightforward biological interpretation.

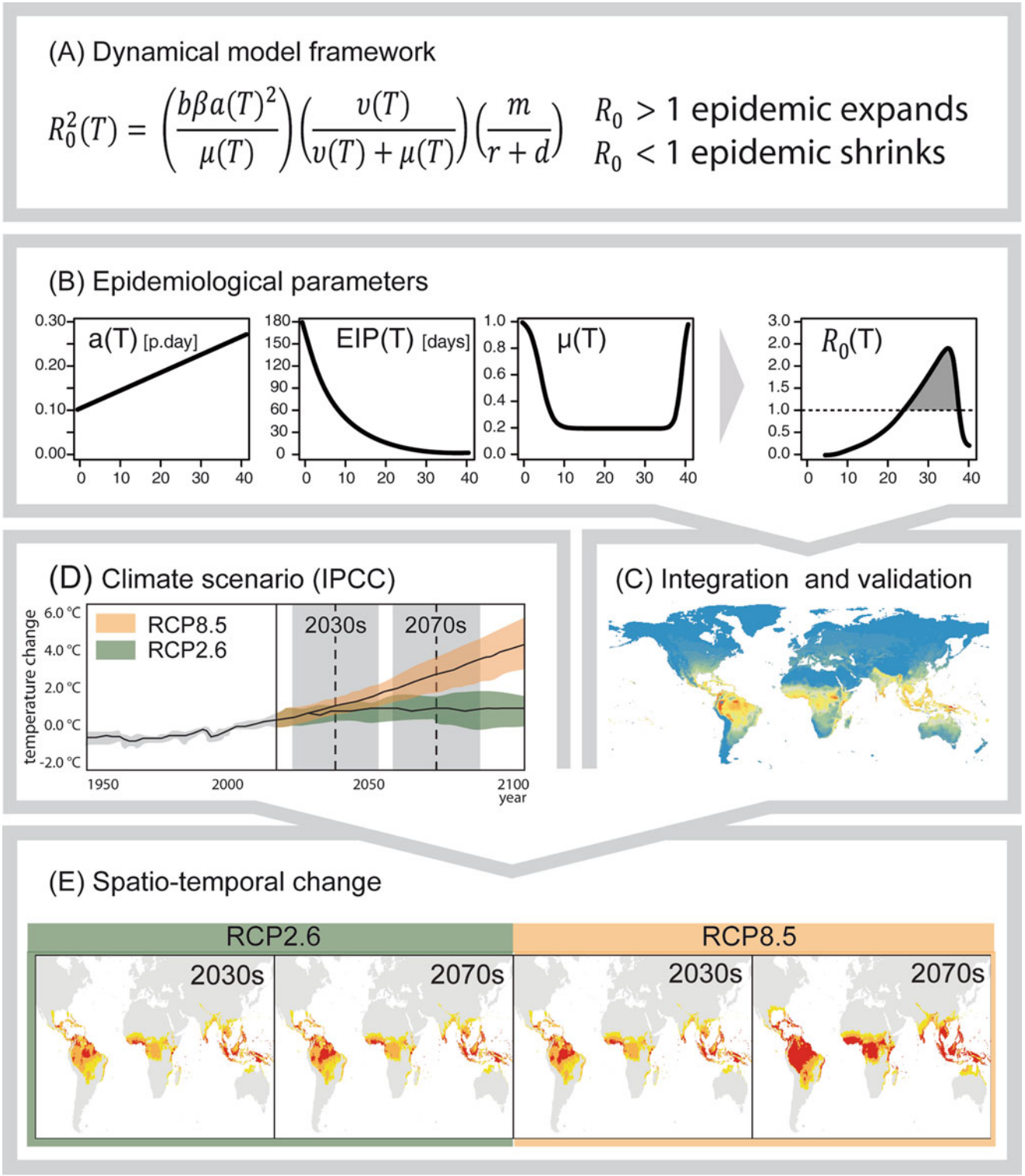


Fig. 3 Characteristic workflow of an epidemiological model derived from the Ross–MacDonald framework [$R_0(T)$ model]. (**a**) Dynamical model framework. T = temperature (°C); b = vector–host transmission probability; β = host–vector transmission probability; m = vector-to-host ratio; r = recovery rate; d = infectious recovery rate; $a(T)$ = vector biting rate per day; EIP(T) = $1/y(T)$ = extrinsic incubation period in days; $m(T)$ = vector mortality rate. (**b**) Epidemiological parameters derived from laboratory experiments or field data are fed into the model to gain an estimate of $R_0(T)$. (**c**) A risk map is derived from the model. (**d**) Climate change scenarios mainly derived from IPCC scenarios (representative concentration pathways (RCP) 2.6–8.5) for different time frames (mainly two or three decades) are chosen. Results shown here are observed and expected future average temperature increase over time for an optimistic RCP 2.6 (green) and a pessimistic RCP 8.5 (orange) climate change scenario. An estimate of uncertainty is shown as color shadings around the black lines. (**e**) Using data from global or regional climate models, further projections for the selected scenarios and time frames (gray vertical bars in panel E) are made. Ideally, different climate models are used to drive the EM. (Adapted from Tjaden et al. 2018)

On the other hand, the NGM method is not straightforward in any case. The EMs using this method typically describe the different health states of vectors and hosts or other subpopulations (e.g., sex and age group) with ordinary differential equations (ODEs). Based on these ODEs, an NGM is constructed. Each compartment in the NGM denotes health state change of vectors or hosts. The basic reproduction number $R_{0\,\mathrm{NGM}}$ is defined as the dominant eigenvalue of this NGM (Diekmann et al. 1990), where only the "states-at-infection" are contributing to the calculation.

The $R_{0\,\mathrm{NGM}}$ calculated with the NGM method has a different interpretation from the ones calculated with the survival function method. As mentioned above, the $R_{0\,\mathrm{NGM}}$ is the dominant eigenvalue of the respective NGM (Diekmann et al. 1990; Diekmann et al. 2010). It has been discussed that $R_{0\,\mathrm{NGM}}$ is indeed a "ratio." For a certain time period, each "generation" is $R_{0\,\mathrm{NGM}}$ times as big as the preceding generation. This $R_{0\,\mathrm{NGM}}$ is a geometric mean through "generations." The "generation" is not a natural generation of vectors or hosts, but for a new generation of infections (Diekmann et al. 2010).

Box 4 A Summary of the Key Definitions of the Epidemiological Model

Epidemiological model: An approach used to simplify the transmission mechanics between vectors and hosts. EM is used to describe the health state change (e.g., susceptible, exposed, infected, and removed health states) of both vectors and hosts with mathematical equations.

Compartmental models: An approach to simplify the mathematical modeling of infectious disease. In this approach, the population is divided into compartments where every individual in the same compartment has the same characteristics (see susceptible, exposed, infected, and removed).

Basic reproduction number (R_0): The expected number of secondary cases produced by a single infection in a completely susceptible population. It is used to estimate the transmission potential of a disease.

Susceptible (S compartment): Refers to the model compartment where all individuals are susceptible if they have contact with a disease.

Exposed (E compartment): Refers to the compartment where individuals are infected by the disease but do not have the visible clinical symptoms of the disease and cannot transmit the disease to susceptible S individuals (see S compartment).

Infected (I compartment): Refers to the compartment where all individuals are infected by the disease and infectious to spread it.

Removed (R compartment): All the individuals are removed from the susceptible-infective interaction by recovery via immunity, isolation, or death (i.e., all recovered, immune, or dead individuals).

Next-generation matrix (NGM): The natural basis used to define and derive the basic reproduction number for a compartmental model of infectious disease spread.

Dominant eigenvalue: The expected number of secondary cases produced by a typical infected individual during its entire period of infectiousness in a completely susceptible population.

Survival function: The function denotes the probability that a patient will survive beyond any specified time.

Pros and Cons of Epidemiological Modeling

EMs provide temporal outbreak risk information of an investigated infectious disease, particularly in fine temporal resolutions (e.g., daily observations from meteorological stations) (Rubel et al. 2008). In the EMs, temperature observations play an important role as many processes within the chain of infections are temperature dependent (e.g., mosquito survival and viral dissemination). The temperature observations are very often acquired from the meteorological stations with daily temporal resolution or satellite images of temperatures (Hartemink et al. 2011; Hartley et al. 2012; Calistri et al. 2016). Consequently, they have the advantage of capturing weather changes such as extreme weather events like drought or frost. For disease control and public health, this enables a more detailed and advanced short-term scale management.

Besides the temporal outbreak risk, spatial risk maps can also be produced by EMs, resulting in spatial-temporal risk maps. There are several methods available for this application: (1) calculating the R_0 for some scattered places (e.g., the location of weather stations in several cities) and then using this value for a certain area (Hartemink et al. 2009), (2) calculating the correlation function between R_0 and a highly correlated environmental variable, then applying this function into the gridded raster variable layer (Wu et al. 2013), (3) calculating the R_0 for each gridded cell of a raster file, which is often a temperature observation layer, resulting directly in spatial risk maps (Holy et al. 2011; Cadar et al. 2017).

EMs generally do not include non-temperature variables that may affect disease circulation compared to ENMs (e.g., land use type or climate type). The latter marks a limitation to the EMs, particularly if these additional variables play a crucial role in disease transmission. Rainfall was included in addition to temperature in a dynamic process-based model that follows a deterministic compartmental approach to the epidemiology of Rift Valley fever transmission (Leedale et al. 2016) and in the R_0 model for Zika (Caminade et al. 2017).

Integrating ENMs and EMs: A Case Study

As with the developing of both model disciplines, it is possible to use these two separate modeling disciplines simultaneously and draw a conclusion from both approaches. Here, we provided an example of interdisciplinary modeling (Cheng et al. 2018) for Usutu virus (USUV) (i.e., a mosquito-borne *Flavivirus* affecting avian host populations (Cadar et al. 2017). The EM concerning the USUV is available from Rubel and colleagues (Rubel et al. 2008) and is adapted by Cheng and colleagues (Cheng et al. 2018). This EM was applied annually for 2017 (January 1 to December 31) using the gridded temperature dataset from gridded observational dataset (E-OBS). The temporal risk duration was estimated with the EM (Fig. 4a).

A MaxEnt algorithm was applied in parallel to estimate the environmental suitability of the USUV across Europe (Fig. 4b). The occurrence records observed from 2003 to 2016 were used as input data. Five ecological bioclimatic variables from WorldClim archive (10 arc-minutes spatial resolution) were used: annual mean temperature, minimum temperature of the coldest month, mean temperature of the coldest quarter, precipitation seasonality, and precipitation of the warmest quarter.

The ENM depicts the USUV occurrences in Northern Italy, western Germany, in Benelux and at the Austrian Hungarian border in 2017 better than the EM inference. But there is a chance for underestimation of the risk because USUV is still emerging in Europe and the ecological niche is not fully occupied yet. As ENMs use occurrence records and respective explanatory variables as input data, they can capture the recorded occurrence well. However, sampling bias is inevitable in this case. Though spatial rarifying methods can be applied to cut down sampling bias or spatial auto-correlation to certain degree, overall ENMs still highly depend on the quality of occurrences data. R_0 values predicted by the EM in areas without evidence for real-life transmission such as in Spain suggest that it may tend toward over-estimation of the risk. One explanation for this might be that temperature but not precipitation, or humidity is used in the EM. EMs, on the other hand, are highly restricted by the up-to-date understanding of the investigated disease. For instance, in our USUV-EM, mosquito-relevant variables play an important role in estimating the final R_0. However, as the spatial heterogeneous mosquito birth/death rate is not available, the same mosquito-relevant variable settings (e.g., birth rate and mortality rate of mosquitoes) and a constant vector-to-host ratio were applied across the whole study area.

Interestingly, the model results from the different approaches may have areas of agreement and disagreement; it is useful to compare the results of both modeling approaches (Cheng et al. 2020). There is no one single approach to be preferred for every pathogen, area, or timespan. As the disease circulation across the study area has not been fully understood yet, it is difficult to evaluate which model performed better. Relying on a single model concerning VBDs may lead to biased conclusions.

Conclusion

Ecological niche modeling has improved disease mapping and has been widely applied to map complex diseases including vector-borne diseases. Epidemiological models have also been used to infer disease spread in diverse applications. Both models are usually simple representations of complex disease systems; they do not typically provide details on the complexity and heterogeneity of these biological systems. However, although both models contributed to better understanding of disease epidemiology, they also remain attached to some limitations. These limitations render results based on a single model for assessing vector-borne disease risk incomplete. Thus, an integrated model could benefit from the strengths of both models. In an integrated modeling approach, spatial distribution of potential risk could first be estimated by an ENM, followed by an investigation of temporal risk patterns in high-risk areas through an EM (Tjaden et al. 2021). In this case, both spatial and temporal aspects of potential risk can be included. The finer temporal scale available through EM, and use of daily weather data or weather forecast data, can work as a live, early warning forecast. In addition, the output of an ENM that estimates the potential spatial distribution of vectors and hosts could be used as input data in an EM. While having the advantage of investigating potential risk at fine temporal scale, EMs do not typically consider spatial heterogeneity. By using estimated spatial distribution of vectors and hosts, the spatial aspect of potential risk can be thus better understood. For instance, a varying vector-to-host ratio can be assigned accordingly. To benefit from the strengths of both model disciplines, early warning systems can be built by integrating them and generating risk maps with fine spatio-temporal resolution. The public health sector and policymakers will benefit from risk maps available online based on automated model runs.

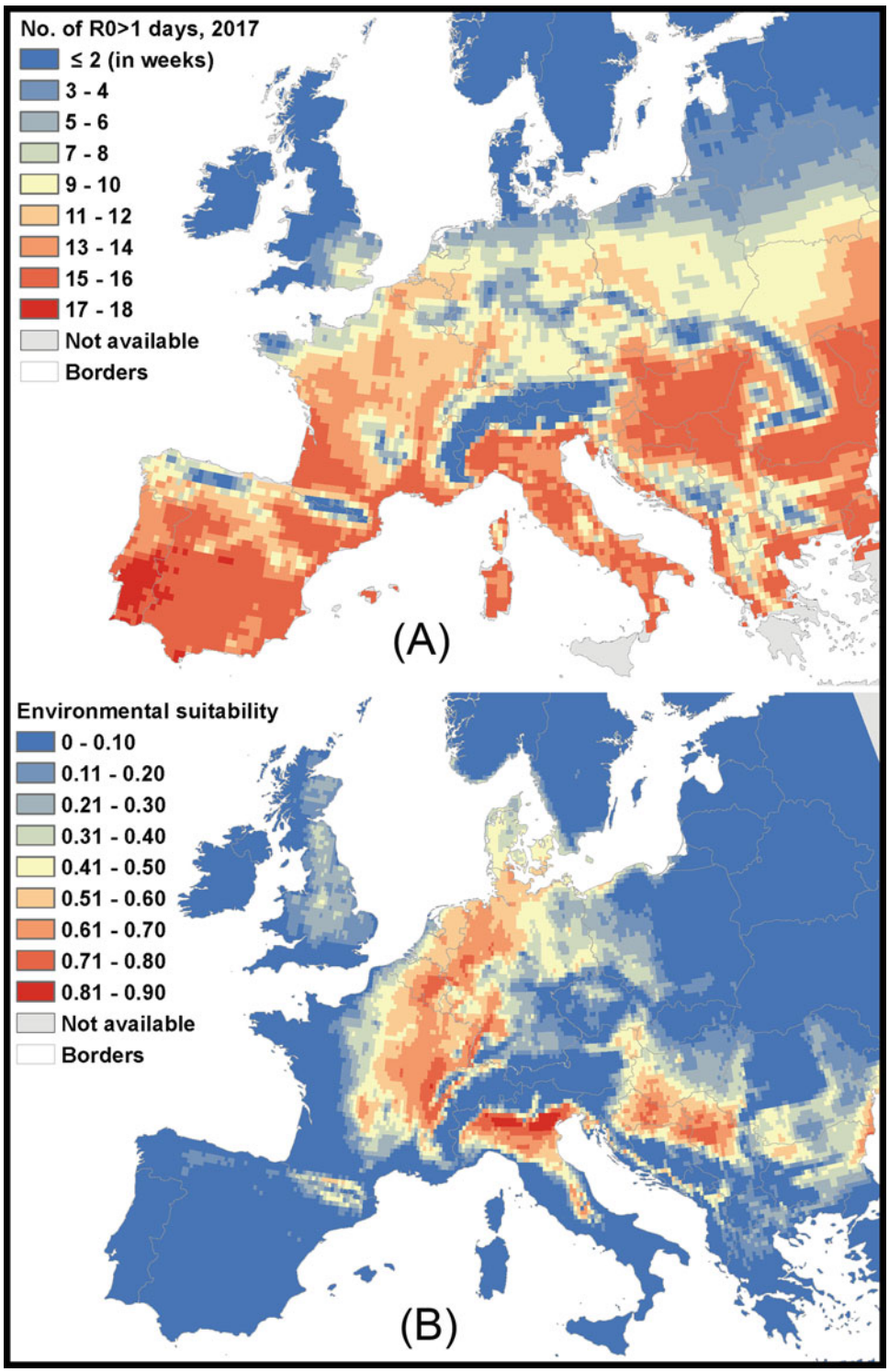


Fig. 4 Spatial risk maps of the Usutu virus from (**a**) the EM showing the total number of days of R_0 >1 in 2017 and from (**b**) the ENM showing the environmental suitability of the area for the transmission of the Usutu virus which ranges from 0 (unsuitable) to 1 (suitable). While the ENM also includes the precipitation seasonality, and the precipitation of the warmest quarter as environmental variables, the EM is solely based on temperature

References

Alkishe, A.A., A.T. Peterson, and A.M. Samy. 2017. Climate change influences on the potential geographic distribution of the disease vector tick *Ixodes ricinus*. *PLoS One* 12: e0189092.

Araújo, M.B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33: 1677–1688.

Austin, M.P., and T.M. Smith. 1989. A new model for the continuum concept. *Vegetatio* 83: 35–47.

Barve, N., V. Barve, A. Jiménez-Valverde, A. Lira-Noriega, S.P. Maher, A.T. Peterson, J. Soberón, and F. Villalobos. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling* 222: 1810–1819.

Buisson, L., W. Thuiller, N. Casajus, S. Lek, and G. Grenouillet. 2010. Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* 16: 1145–1157.

Cadar, D., R. Luhken, H. Van Der Jeugd, M. Garigliany, U. Ziegler, M. Keller, J. Lahoreau, L. Lachmann, N. Becker, M. Kik, B.B. Oude Munnink, S. Bosch, E. Tannich, A. Linden, V. Schmidt, M.P. Koopmans, J. Rijks, D. Desmecht, M.H. Groschup, C. Reusken, and J. Schmidt-Chanasit. 2017. Widespread activity of multiple lineages of Usutu virus, western Europe, 2016. *Euro Surveillance* 22: 30452.

Calistri, P., L. Savini, L. Candeloro, D. Di Sabatino, F. Cito, R. Bruno, and M.L. Danzetta. 2016. A transitional model for the evaluation of West Nile virus transmission in Italy. *Transboundary and Emerging Diseases* 63: 485–496.

Caminade, C., J. Turner, S. Metelmann, J.C. Hesson, M.S. Blagrove, T. Solomon, A.P. Morse, and M. Baylis. 2017. Global risk model for vector-borne transmission of Zika virus reveals the role of El Nino 2015. *Proceedings of the National Academy of Sciences of the United States of America* 114: 119–124.

Carlson, C.J., E.R. Dougherty, and W. Getz. 2016. An ecological assessment of the pandemic threat of Zika virus. *PLoS Neglected Tropical Diseases* 10: e0004968.

Carvalho, B.M., E.F. Rangel, and M.M. Vale. 2017. Evaluation of the impacts of climate change on disease vectors through ecological niche modelling. *Bulletin of Entomological Research* 107: 419–430.

Chavy, A., A. Ferreira Dales Nava, S.L.B. Luz, J.D. Ramírez, G. Herrera, T. Vasconcelos Dos Santos, M. Ginouves, M. Demar, G. Prévot, J.-F. Guégan, and B. De Thoisy. 2019. Ecological niche modelling for predicting the risk of cutaneous leishmaniasis in the Neotropical moist forest biome. *PLoS Neglected Tropical Diseases* 13: e0007629.

Cheng, Y., N.B. Tjaden, A. Jaeschke, R. Luhken, U. Ziegler, S.M. Thomas, and C. Beierkuhnlein. 2018. Evaluating the risk for Usutu virus circulation in Europe: Comparison of environmental niche models and epidemiological models. *International Journal of Health Geographics* 17: 35.

Cheng, Y., N. Tjaden, A. Jaeschke, S.M. Thomas, and C. Beierkuhnlein. 2020. Deriving risk maps from epidemiological models of vector borne diseases: State-of-the-art and suggestions for best practice. *Epidemics* 33: 100411.

Delamater, P.L., E.J. Street, T.F. Leslie, Y.T. Yang, and K.H. Jacobsen. 2019. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases* 25: 1–4.

Diekmann, O., J.A.P. Heesterbeek, and J.A.J. Metz. 1990. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology* 28: 365–382.

Diekmann, O., J.A.P. Heesterbeek, and M.G. Roberts. 2010. The construction of next-generation matrices for compartmental epidemic models. *The Journal of the Royal Society Interface* 7 (47): 873–885.

Dietz, K. 1993. The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research* 2: 23–41.

Eisen, L., and R.J. Eisen. 2011. Using geographic information systems and decision support systems for the prediction, prevention, and control of vector-borne diseases. *Annual Review of Entomology* 56: 41–61.

Elith, J., and J.R. Leathwick. 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697.

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1: 330–342.

Elton, C.S. 1927. The nature and origin of soil-polygons in Spitsbergen. *Quarterly Journal of the Geological Society* 83: 163-NP.

Eneanya, O.A., J. Cano, I. Dorigatti, I. Anagbogu, C. Okoronkwo, T. Garske, and C.A. Donnelly. 2018. Environmental suitability for lymphatic filariasis in Nigeria. *Parasites & Vectors* 11: 513.

Escobar, L.E., and M.E. Craft. 2016. Advances and limitations of disease biogeography using ecological niche modeling. *Frontiers in Microbiology* 7: 1174.

Fischer, D., S.M. Thomas, F. Niemitz, B. Reineking, and C. Beierkuhnlein. 2011. Projection of climatic suitability for *Aedes albopictus* Skuse (Culicidae) in Europe under climate change conditions. *Global and Planetary Change* 78: 54–64.

Franklin, J., and J.A. Miller. 2010. *Mapping species distributions: Spatial inference and prediction*.

Gherghel, I., F. Brischoux, and M. Papeş. 2019. Refining model estimates of potential species' distributions to relevant accessible areas. In *Progress in Physical Geography: Earth and Environment*, 0309133319881104.

Gholamrezaei, M., M. Mohebali, A.A. Hanafi-Bojd, M.M. Sedaghat, and M.R. Shirzadi. 2016. Ecological niche modeling of main reservoir hosts of zoonotic cutaneous leishmaniasis in Iran. *Acta Tropica* 160: 44–52.

Grinnell, J. 1917. Field tests of theories concerning distributional control. *The American Naturalist* 51: 115–128.

Hao, T., J. Elith, G. Guillera-Arroita, and J.J. Lahoz-Monfort. 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions* 25: 839–852.

Hartemink, N.A., B.V. Purse, R. Meiswinkel, H.E. Brown, A. De Koeijer, A.R. Elbers, G.J. Boender, D.J. Rogers, and J.A. Heesterbeek. 2009. Mapping the basic reproduction number (R(0)) for vector-borne diseases: A case study on bluetongue virus. *Epidemics* 1: 153–161.

Hartemink, N., S.O. Vanwambeke, H. Heesterbeek, D. Rogers, D. Morley, B. Pesson, C. Davies, S. Mahamdallie, and P. Ready. 2011. Integrated mapping of establishment risk for emerging vector-borne infections: A case study of canine leishmaniasis in Southwest France. *PLoS One* 6: e20817.

Hartley, D.M., C.M. Barker, A. Le Menach, T. Niu, H.D. Gaff, and W.K. Reisen. 2012. Effects of temperature on emergence and seasonality of West Nile virus in California. *The American Journal of Tropical Medicine and Hygiene* 86: 884–894.

Hay, S.I., K.E. Battle, D.M. Pigott, D.L. Smith, C.L. Moyes, S. Bhatt, J.S. Brownstein, N. Collier, M.F. Myers, D.B. George, and P.W. Gething. 2013. Global mapping of infectious disease. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368: 20120250.

Heffernan, J.M., R.J. Smith, and L.M. Wahl. 2005. Perspectives on the basic reproductive ratio. *Journal of the Royal Society, Interface* 2: 281–293.

Hess, A., J.K. Davis, and M.C. Wimberly. 2018. Identifying environmental risk factors and mapping the distribution of West Nile virus in an endemic region of North America. *GeoHealth* 2: 395–409.

Holy, M., G. Schmidt, and W. Schroder. 2011. Potential malaria outbreak in Germany due to climate warming: Risk modelling based on temperature measurements and regional climate models. *Environmental Science and Pollution Research International* 18: 428–435.

Huestis, D.L., A. Dao, M. Diallo, Z.L. Sanogo, D. Samake, A.S. Yaro, Y. Ousman, Y.M. Linton, A. Krishna, L. Veru, B.J. Krajacich, R.

Faiman, J. Florio, J.W. Chapman, D.R. Reynolds, D. Weetman, R. Mitchell, M.J. Donnelly, E. Talamas, L. Chamorro, E. Strobach, and T. Lehmann. 2019. Windborne long-distance migration of malaria mosquitoes in the Sahel. *Nature* 574: 404–408.

Johnson, R.H. 1910. *Determinate evolution in the color-pattern of the lady-beetles*. Washington: Carnegie Inst.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14: 401–413.

Kamal, M., M.A. Kenawy, M.H. Rady, A.S. Khaled, and A.M. Samy. 2018. Mapping the global potential distributions of two arboviral vectors *Aedes aegypti* and *Ae. albopictus* under changing climate. *PLoS One* 13: e0210122.

Kearney, M., and W. Porter. 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters* 12: 334–350.

Kraemer, M.U., M.E. Sinka, K.A. Duda, A.Q. Mylne, F.M. Shearer, C.M. Barker, C.G. Moore, R.G. Carvalho, G.E. Coelho, W. Van Bortel, G. Hendrickx, F. Schaffner, I.R. Elyazar, H.J. Teng, O.J. Brady, J.P. Messina, D.M. Pigott, T.W. Scott, D.L. Smith, G.R. Wint, N. Golding, and S.I. Hay. 2015. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *eLife* 4: e08347.

Kraemer, M.U.G., R.C. Reiner, J.R. Brady, O.J. Messina, J.P. Gilbert, M. Pigott, M. D., D. Yi, K. Johnson, L. Earl, L.B. Marczak, S. Shirude, N. Davis Weaver, D. Bisanzio, T.A. Perkins, S. Lai, X. Lu, P. Jones, G.E. Coelho, R.G. Carvalho, W. Van Bortel, C. Marsboom, G. Hendrickx, F. Schaffner, C.G. Moore, H.H. Nax, L. Bengtsson, E. Wetter, A.J. Tatem, J.S. Brownstein, D.L. Smith, L. Lambrechts, S. Cauchemez, C. Linard, N.R. Faria, O.G. Pybus, T.W. Scott, Q. Liu, H. Yu, G.R.W. Wint, S.I. Hay, and N. Golding. 2019. Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. *Nature Microbiology* 4: 854–863.

Leedale, J., A.E. Jones, C. Caminade, and A.P. Morse. 2016. A dynamic, climate-driven model of Rift Valley fever. *Geospatial Health* 11: 394.

Leibold, M.A. 1996. A graphical model of keystone predators in food webs: Trophic regulation of abundance, incidence, and diversity patterns in communities. *The American Naturalist* 147: 784–812.

Liu, C., P.M. Berry, T.P. Dawson, and R.G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385–393.

Lobo, J.M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151.

MacArthur, R.H. 1968. The theory of the niche. In *Population biology and evolution*, ed. R.C. Lewontin, 159–176. Syracuse: Syracuse University Press.

Manel, S., H.C. Williams, and S.J. Ormerod. 2001. Evaluating presence–absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology* 38: 921–931.

Messina, J.P., M.U. Kraemer, O.J. Brady, D.M. Pigott, F.M. Shearer, D.J. Weiss, N. Golding, C.W. Ruktanonchai, P.W. Gething, E. Cohn, J.S. Brownstein, K. Khan, A.J. Tatem, T. Jaenisch, C.J. Murray, F. Marinho, T.W. Scott, and S.I. Hay. 2016. Mapping global environmental suitability for Zika virus. *eLife* 5: e15272.

Messina, J.P., O.J. Brady, N. Golding, M.U.G. Kraemer, G.R.W. Wint, S.E. Ray, D.M. Pigott, F.M. Shearer, K. Johnson, L. Earl, L.B. Marczak, S. Shirude, N. Davis Weaver, M. Gilbert, R. Velayudhan, P. Jones, T. Jaenisch, T.W. Scott, R.C. Reiner, and S.I. Hay. 2019. The current and future global distribution and population at risk of dengue. *Nature Microbiology* 4: 1508–1515.

Moraga, P., J. Cano, R.F. Baggaley, J.O. Gyapong, S.M. Njenga, B. Nikolay, E. Davies, M.P. Rebollo, R.L. Pullan, M.J. Bockarie, T.D. Hollingsworth, M. Gambhir, and S.J. Brooker. 2015. Modelling the distribution and transmission intensity of lymphatic filariasis in sub-Saharan Africa prior to scaling up interventions: Integrated use of geostatistical and mathematical modelling. *Parasites & Vectors* 8: 560.

Moyes, C.L., F.M. Shearer, Z. Huang, A. Wiebe, H.S. Gibson, V. Nijman, J. Mohd-Azlan, J.F. Brodie, S. Malaivijitnond, M. Linkie, H. Samejima, T.G. O'brien, C.R. Trainor, Y. Hamada, A.J. Giordano, M.F. Kinnaird, I.R.F. Elyazar, M.E. Sinka, I. Vythilingam, M.J. Bangs, D.M. Pigott, D.J. Weiss, N. Golding, and S.I. Hay. 2016. Predicting the geographical distributions of the macaque hosts and mosquito vectors of *Plasmodium knowlesi* malaria in forested and non-forested areas. *Parasites & Vectors* 9: 242.

Mweya, C.N., S.I. Kimera, G. Stanley, G. Misinzo, and L.E. Mboera. 2016. Climate change influences potential distribution of infected *Aedes aegypti* co-occurrence with dengue epidemics risk areas in Tanzania. *PLoS One* 11: e0162649.

Owens, H.L., L.P. Campbell, L.L. Dornak, E.E. Saupe, N. Barve, J. Soberón, K. Ingenloff, A. Lira-Noriega, C.M. Hensz, C.E. Myers, and A.T. Peterson. 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling* 263: 10–18.

Pearson, R.G., T.P. Dawson, and C. Liu. 2004. Modelling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography* 27: 285–298.

Peterson, A.T. 2003. Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology* 78: 419–433.

———. 2006. Ecologic niche modeling and spatial patterns of disease transmission. *Emerging Infectious Diseases* 12: 1822–1826.

———. 2007. Ecological niche modelling and understanding the geography of disease transmission. *Veterinaria Italiana* 43: 393–400.

———. 2014. *Mapping disease transmission risk: Enriching models using biogeography and ecology*. Baltimore: Johns Hopkins University Press.

Peterson, A.T., M. Papeş, and J. Soberón. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling* 213: 63–72.

Peterson, A.T., J. Soberón, R.G. Pearson, R.P. Anderson, E. Martínez-Meyer, M. Nakamura, and M.B. Araújo. 2011. *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press.

Pigott, D.M., N. Golding, A. Mylne, Z. Huang, A.J. Henry, D.J. Weiss, O.J. Brady, M.U.G. Kraemer, D.L. Smith, C.L. Moyes, S. Bhatt, P.W. Gething, P.W. Horby, I.I. Bogoch, J.S. Brownstein, S.R. Mekaru, A.J. Tatem, K. Khan, and S.I. Hay. 2014. Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* 3: e04395–e04395.

Qiao, H., J. Soberón, and A.T. Peterson. 2015. No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution* 6: 1126–1136.

Qiao, H., L.E. Escobar, and A.T. Peterson. 2017. Accessible areas in ecological niche comparisons of invasive species: Recognized but still overlooked. *Scientific Reports* 7: 1213.

Real, L.A., and S.A. Levin. 1991. *Theoretical advances*, 177. Foundations of Ecology: Classic Papers with Commentaries.

Ridenhour, B., J.M. Kowalik, and D.K. Shay. 2018. Unraveling R(0): Considerations for public health applications. *American Journal of Public Health* 108: S445–S454.

Rogers, D.J., and S.E. Randolph. 2003. Studying the global distribution of infectious diseases using GIS and RS. *Nature Reviews. Microbiology* 1: 231–237.

Rubel, F., K. Brugger, M. Hantel, S. Chvala-Mannsberger, T. Bakonyi, H. Weissenbock, and N. Nowotny. 2008. Explaining Usutu virus dynamics in Austria: Model development and calibration. *Preventive Veterinary Medicine* 85: 166–186.

Samy, A.M., and A.T. Peterson. 2016. Climate change influences on the global potential distribution of bluetongue virus. *PLoS One* 11: e0150489.

Samy, A.M., S.M. Thomas, A.A. Wahed, K.P. Cohoon, and A.T. Peterson. 2016a. Mapping the global geographic potential of Zika virus spread. *Memórias do Instituto Oswaldo Cruz* 111: 559–560.

Samy, A.M., A.H. Elaagip, M.A. Kenawy, C.F. Ayres, A.T. Peterson, and D.E. Soliman. 2016b. Climate change influences on the global potential distribution of the mosquito *Culex quinquefasciatus*, vector of West Nile virus and lymphatic Filariasis. *PLoS One* 11: e0163863.

Samy, A.M., B.B. Annajar, M.R. Dokhan, S. Boussaa, and A.T. Peterson. 2016c. Coarse-resolution ecology of etiological agent, vector, and reservoirs of zoonotic cutaneous Leishmaniasis in Libya. *PLoS Neglected Tropical Diseases* 10: e0004381.

Samy, A.M., A.A. Alkishe, S.M. Thomas, L. Wang, and W. Zhang. 2018. Mapping the potential distributions of etiological agent, vectors, and reservoirs of Japanese Encephalitis in Asia and Australia. *Acta Tropica* 188: 108–117.

Santos, J.P.D. 2017. Does land cover influence the spatial distribution of reservoir rodent *Necromys lasiurus? SOJ Microbiology & Infectious Diseases* 5: 1–5.

Semenza, J.C., and J.E. Suk. 2018. Vector-borne diseases and climate change: A European perspective. *FEMS Microbiology Letters* 365: fnx244.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters* 10: 1115–1123.

Thomas, S.M., N. Tjaden, S. van den Bos, and C. Beierkuhnlein. 2014. Implementing cargo movement into climate based risk assessment of vector-borne diseases. *International Journal of Environmental Research and Public Health* 11 (3): 3360–3374.

Thomas, S.M., N.B. Tjaden, C. Frank, A. Jaeschke, L. Zipfel, C. Wagner-Wiening, M. Faber, C. Beierkuhnlein, and K. Stark. 2018. Areas with high Hazard potential for autochthonous transmission of *Aedes albopictus*-associated arboviruses in Germany. *International Journal of Environmental Research and Public Health* 15: 1270.

Tjaden, N.B., J.E. Suk, D. Fischer, S.M. Thomas, C. Beierkuhnlein, and J.C. Semenza. 2017. Modelling the effects of global climate change on chikungunya transmission in the 21(st) century. *Scientific Reports* 7: 3813.

Tjaden, N.B., C. Caminade, C. Beierkuhnlein, and S.M. Thomas. 2018. Mosquito-borne diseases: Advances in modelling climate-change impacts. *Trends in Parasitology* 34: 227–245.

Tjaden, N., Y. Cheng, C. Beierkuhnlein, and S.M. Thomas. 2021. Chikungunya beyond the tropics: Where and when do we expect disease transmission in Europe? *Viruses* 13: 1024. https://doi.org/10.3390/v13061024.

Tsiamis, C., E. Poulakou-Rebelakou, and S. Marketos. 2013. Earthquakes and plague during byzantine times: Can lessons from the past improve epidemic preparedness. *Acta Medico-Historica Adriatica* 11: 55–64.

Vandermeer, J.H. 1972. Niche theory. *Annual Review of Ecology and Systematics* 3: 107–132.

Whittaker, R.H., S.A. Levin, and R.B. Root. 1973. Niche, habitat, and Ecotope. *The American Naturalist* 107: 321–338.

Wimberly, M.C., M.B. Hildreth, S.P. Boyte, E. Lindquist, and L. Kightlinger. 2008. Ecological niche of the 2003 West Nile virus epidemic in the northern great plains of the United States. *PLoS One* 3: e3744.

Wu, X., V.R. Duvvuri, Y. Lou, N.H. Ogden, Y. Pelcat, and J. Wu. 2013. Developing a temperature-driven map of the basic reproductive number of the emerging tick vector of Lyme disease *Ixodes scapularis* in Canada. *Journal of Theoretical Biology* 319: 50–61.

Zhu, G.-P., and A.T. Peterson. 2017. Do consensus models outperform individual models? Transferability evaluations of diverse modeling approaches for an invasive moth. *Biological Invasions* 19: 2519–2532.

Spatially Integrating Microbiology and Geochemistry to Reveal Complex Environmental Health Issues: Anthrax in the Contiguous United States

Erin E. Silvestri, Steven H. Douglas, Vicky A. Luna, C. A. O. Jean-Babtiste, Deryn Pressman-Mashin née Harbin, Laura A. Hempel, Timothy R. Boe, Tonya L. Nichols, and Dale W. Griffin

Maxent models were run using the *B. anthracis* presence data and/or the animal outbreak presence data. Models run using the animal outbreak data alone utilized two scales: the Outbreak State scale which included only states reporting animal anthrax outbreaks from 2001 to 2013 and the National scale which included all states in the contiguous United States. Three iterations of the environmental data were used and included the Sample Location dataset which utilized the environmental variable data with assigned latitude and longitude locations from the USGS NASGLP project; the Normalized dataset which scaled the environmental variables so that the values fell between 0 and 1; and the Interpolated dataset which provided an interpolation of the environmental variables averaged for each county and assigned to a point for that county at the centroid (rather than using the NASGLP latitude and longitude location). Two metrics were used to measure model performance including the widely used area under the curve (AUC) and an alternative method, the True Skill Statistic (TSS). The AUC gives the probability that a randomly chosen presence location has been correctly ranked higher than the absence/background site. AUC values at 0.5 or lower mean the ranking is no better than random, while the AUC values nearer to 1 mean the model is a better predictor. The TSS provides a comparison of how well the background predictions made by the model match the model results at the test dataset (presence) locations. TSS values near +1 means the model approaches perfect agreement, while values near −1 indicate the model is no better than random.

Maxent models to determine the influence of environmental factors on the *B. anthracis* distribution using the PCR data yielded a low TSS, which suggested the model might

E. E. Silvestri
U.S. Environmental Protection Agency, Office of Research and Development, Center for Environmental Solutions and Emergency Response, Homeland Security and Materials Management Division, Disaster Characterization Branch, Cincinnati, OH, USA
e-mail: silvestri.erin@epa.gov

S. H. Douglas
Geospatial Analyst, Versar, Inc., Newport News, Virginia, USA
e-mail: sdouglas89.sd@gmail.com

V. A. Luna
University of South Florida, Center for Biological Defense, Tampa, FL, USA
e-mail: vluna@bt.usf.edu

C. A. O. Jean-Babtiste
Cindy (or Cynthia) Ogolla Jean-Baptiste, DrPH, MA, MPH, CPH
Former research assistant for Dale Griffin, USGS, Current Public Health Analyst, United States Air Force, Suffolk County, Massachusetts, IL, USA
e-mail: caogolla@gmail.com

D. Pressman-Mashin née Harbin
MBA Hydrological technician and assistant to Dale Griffin, Director of Community Engagement and Communications, Epstein Hillel School, Marblehead, Massachusetts, DeBary, FL, USA
e-mail: derynharbin@gmail.com

L. A. Hempel
Former research assistant for Dale Griffin, USGS. Current Hydrologist U.S. Geological Survey, Colorado Water Science Center, Pueblo, CO, USA
e-mail: lhempel@usgs.gov

T. R. Boe
Geographer, US EPA, Office of Research and Development, Center for Solutions and Emergency Response, Homeland Security and Materials Management Division, Research Triangle Park, NC, USA
e-mail: boe.timothy@epa.gov

T. L. Nichols
U.S. Environmental Protection Agency, Threat and Consequence Assessment Division, National Homeland Security Research Center, Washington, DC, USA
e-mail: nichols.tonya@epamail.epa.gov

D. W. Griffin (✉)
U. S. Geological Survey, Coastal and Marine Science Center, St. Petersburg, FL, USA
e-mail: dgriffin@usgs.gov

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_19

be underfitting the data. This was not surprising due to the difficulty in recovering *B. anthracis* in soil samples as well as the samples themselves being discrete in nature and only capturing a snapshot in time. Therefore, the distribution of *B. anthracis* and its niche in the contiguous United States could not be determined in this study. However, efforts to investigate environmental factors that would have a higher potential of supporting an anthrax outbreak in wildlife and livestock yielded better results. Results showed that most of the Maxent models in this study performed best when using the Outbreak State scale. When the models were scaled up to the National scale, model performance declined, except for the Normalized variable dataset. At the Outbreak State scale, a large proportion of the area was predicted to be of higher probability for wildlife/livestock anthrax outbreaks, and the statistical measures assumed the model was underfitting the data. The model with the highest AUC and TSS scores for this study was the Outbreak State scale using Sample Location dataset (AUC = 0.918 and TSS = 0.82). Some of the variables found to be closely related to the occurrence of *B. anthracis* in this study included pH, drainage potential, and concentration of elements including Na, Ca, Sr, and Mg, which have also been found to be related to animal outbreaks or to the occurrence of *B. anthracis* in previous studies.

The models in the current study indicated possible regions that have not had recent wildlife/livestock anthrax outbreaks but contained environmental conditions that could potentially support an outbreak if one were to occur (Michigan and Maine). This work provides an extension to the use of ecological niche modeling to outbreak potential in livestock/wildlife in the United States because it utilizes additional soil geochemistry data and has shown that further validation techniques, such as the TSS, should be considered in addition to AUC. Results from this study could be used by animal and public health officials to identify areas with a higher potential for anthrax outbreak in wildlife and livestock due to naturally occurring soil and environmental conditions.

Introduction

Bacillus anthracis is an aerobic, Gram-positive, endospore-forming, rod-shaped bacterium that is the causative agent for anthrax. *B. anthracis* occurs in two physical forms: in a vegetative or growing state and in a spore state. Spores are formed to enhance survival and are resistant to many environmental stressors (Titball et al. 1991). *B. anthracis* is a common community member of many soil environments and has been shown to germinate and survive in the rhizosphere (root zone) of grasses (Saile and Koehler 2006; Van Ness and Stein 1956). Surveillance and survival studies have shown that although only a fraction (10%) of a *Bacillus* sp. soil inoculum can survive after more than 60 days of incubation, spores can be detectable in the top few centimeters for many years (Manchee et al. 1994; West and Burges 1985). However, *B. anthracis* could be present in levels of the soil that are at undetectable levels, making detection difficult.

The route of exposure for *B. anthracis* infections have been recognized as a critically important issue in the United States in wildlife and livestock for over 200 years. The pathogen can be spread by a range of hosts, including herbivores, scavengers, carnivores, and insects, and has even been shown to replicate in a type of amoeba (*Acanthamoeba castellanii*) that is a ubiquitous member of soil biota (Breed 1932; Dey et al. 2012; Hugh-Jones and Blackburn 2009). Recurrent and persistent animal anthrax outbreaks typically consist of a cycle between exposure of a susceptible animal (host) to *B. anthracis* spores (which must contain both the pX01 and pX02 plasmids for virulence), deposition of the pathogen back into the soil or environment, and then acquisition by a new host (USEPA 2015; Van Ness 1971). It has been reported that recurrent anthrax outbreaks in wildlife or livestock might stop for long periods of time, even decades, before a new host is exposed and the cycle starts again (Hugh-Jones and Blackburn 2009). Briefly, the classic *B. anthracis* lifecycle consists of (Lindeque and Turnbull 1994; Schuch and Fischetti 2009; USEPA 2014; USEPA 2015; Van Ness 1971):

- Exposure of an animal host to spores via ingestion or inhalation during grazing, which, in turn, can pass the infection to carnivores through consumption (Breed 1932; Stein 1945; Stein 1950)
- Germination of spores in the host
- Multiplication of *B. anthracis* vegetative organisms inside the host on the order of multiple millions of organisms per milliliter of blood
- Production of toxins by vegetative organisms and death of the host
- Opening of the carcass by predation or other events which cause bodily fluids to drain from the infected carcass into the surrounding environment (air, soil, water), dispersing vegetative cells
- Rapid sporulation at the carcass site
- Spore acquisition by a new host

A brief review of the environmental and geographical factors shown to influence the persistence of *B. anthracis* or occurrence of an animal anthrax outbreak from the literature is listed in Table 1. It is interesting to note that factors influencing anthrax outbreaks in animals and occurrence (e.g., weather/climate, environmental, and geological factors) are consistent over a range of different studies and time periods. For example, anthrax outbreaks in wildlife and livestock typically occur: during warmer months in which dry periods follow moderate to heavy rain events; in areas that have short

Table 1 Historical observations of the environmental, weather/climate, and geographical factors associated with wildlife/livestock anthrax outbreaks and the distribution of *Bacillus anthracis*

Citation	Weather/climate	Environmental	Geological	Not favored
Pasteur (1880)			Calcareous clay soil	Schistose or granite soil
Koch (1882)	Optimal laboratory growth at 43.0 °C (in vitro)	Decaying vegetative matter and cadavers		
Higgins (1916)			Alkaline soils and soils from the edge of infected water bodies	
Breed (1932)	Wet and dry season (that makes foraging assessable to lowlands via drying of wet soils in the late summer or autumn)	Decomposing vegetation, short grasses that favor the uptake of soil particles during grazing. Blood-sucking flies, scavengers, and carnivores. Can be spread by the shipment of forage and grain crops	Alkaline and naturally wet soils, lowlands	Vaccinated herds using Pasteur's vaccine but some resistance in areas of N. America. Recent vaccine development is addressing this issue
Minett and Dhanda (1941)	25–35 °C sporulation (in vitro). Rain season		Neutral or alkaline soils, high nitrogen and calcium, wet and marsh soils. Soil moisture of 20% or over (in vitro)	
Stein (1945)	Heavy rains, floods, periodic inundations, droughts, extreme heat	An abundance of flies		
Minett (1950)	32.2–36.7 °C favors sporulation; cooler temps favor longer survival although at lower sporulation rates		Moist to wet soils (in vitro)	15.5–21.1 °C, bacilli disintegration via growth of other bacteria in blood (in vitro)
Van Ness and Stein (1956)			Neutral or alkaline soils containing adequate calcium	Acidic soils
Van Ness (1959)	Flooding followed by a period of drying	Water-damaged vegetation/drying grasses	Limestone or alluvial soils, limestone road chat, liming, water courses	Well-drained sandy or shale soils
Van Ness (1967)	Flooding followed by a period of drying, >15.5 °C	Water-damaged vegetation/drying grasses	Limestone, alluvial, clay soils, pH > 6.0	Sandy or shale soils
Wright et al. (1970)			Higher phosphate concentrations (0.01 M) resulted in higher protective antigen production (in vitro)	Reduction of protective antigen production at phosphate concentrations of 0.001 M or less (in vitro)
Turell and Knudson (1987)		Spread by flying insects		
Weinberg (1987)			Factor 1 of exotoxin may allow scavenging of manganese and thus explain why *B. anthracis* thrives in alkaline soil types (calcareous peats, soils with high organic content, limed and high-water table soils) containing what may be considered suboptimal concentrations for other *Bacillus* sp.	
Turnbull et al. (1989)		Feces of scavengers/carnivores following outbreaks		
Kochi et al. (1994)			Zinc requirement for lethal factor (in vitro)	
Lindeque and Turnbull (1994)	typo: Daily temperature at death of all animals at marked sites was >25.0 °C	Water in waterholes, feces of scavengers and carnivores during and following outbreaks	Karstveld soils had higher concentrations of spores near carcasses versus deep or sandy soils. Topical lows, i.e., waterholes	Carcasses rarely found on saline par or misc. rock type soils. Short-term exposure to UV appeared lethal to spores

(continued)

Table 1 (continued)

Citation	Weather/climate	Environmental	Geological	Not favored
Dragon and Rennie (1995)		Disturbance of an infected carcass by scavengers or carnivores	Elevated calcium, i.e., calcareous soils. Topical lows act as "storage areas" via precipitation events	Undisturbed carcasses do not allow aerobic exposure of the vegetative cells that triggers sporulation before they are destroyed by rapidly growing prokaryote decomposers
Turner et al. (1999a)	Prolonged hot, dry, humid weather		Poorly drained alluvial soils. Significant accumulation of water following moderate precipitation events	Well-drained lands via anthropogenic modifications. Vaccination programs
Smith et al. (2000)			Group A isolates (worldwide distribution) have a lower tolerance range of calcium (avg., 185.7 me/kg) and pH (avg., 6.7) concentrations versus that observed with Group B isolates (Ca avg., 274.1 me/kg, pH avg., 7.8 (restricted in distribution to southern Africa)	
Saile and Koehler (2006)		Spores can germinate and vegetative cells can survive in the rhizosphere of grasses		
Siamudaala et al. (2006)	Following the wet season	Flood deposits of organic detritus	Low-lying areas	
Griffin et al. (2009)			For *rpo*B *B. anthracis* PCR-positive samples. Moist to wet soils (range 10–57 weight %, avg., 25.4%), for a N-S El Paso to Manitoba, Canada, transect and elevated sodium and sulfur (avg., 1.2 and 0.5 weight %, respectively) in New Orleans soils following the Katrina flood	
Hugh-Jones and Blackburn (2009)	Hot-dry season disease influenced precipitation events	Necrophagic flies = case-multipliers, hemophagic flies = space-multipliers	Water/pot-holes can contain elevated concentrations of calcium (2–3X), phosphorus (6–10X), magnesium (>2X), and sodium (conc. not specified). High calcium level soils and pH > 6.1	Low pH soils
Dey et al. (2012)		Replication in amoeba (*A. castellanii*) and sporulation in the demised amoeba detritus (in vitro)		
Ahsan et al. (2013)	May to June with an average temperature of 32 °C		Low-lying areas, livestock pastures, near carcass sites. Favored loamy soils with an average pH of 6.38 ± 0.15 and elevated moisture content (16.69 ± 2.06%). Organic carbon content and calcium ranged from 0.15% to 2.35% and 448.35 to 1372.35 ppm, respectively	Clay soils
Griffin et al. (2014)			Soils with elevated concentrations (tentative thresholds) of calcium (0.43 wt %), manganese (142 mg/kg), phosphorus (180 mg/kg), and strontium (51 mg/kg)	
Summary – observational trends through time	*Hot (>15.5 °C) dry periods following large to moderate precipitation events*	*Post-flood organic detritus, short dry grazing material. Spread from carcasses by scavengers, carnivores, and insects*	*Topical lows used by grazers for waterholes. Calcareous and alluvial soils at pH values >6.0 and with elevated nutrient and spore component content*	*Schistose/rocky soil types of low nutrient content and acidic (pH < 6.0) in nature. Low temperature seasons*

dry grazing grasses or have detritus deposited after a flood; in areas containing alluvial and calcareous soils that have a pH greater than 6.0 and have elevated nutrient content (examples include but are not limited to phosphate and nitrogen); and/or in areas where the topology is low (waterholes or riverbanks).

Sites which contained animal carcasses from previous anthrax outbreaks have been shown to favor spore survival and repeated outbreaks in those areas, although the spore titers vary between carcass sites and among the areas directly surrounding the carcass sites (Ahsan et al. 2013; Lindeque and Turnbull 1994; Turnbull et al. 1998; USEPA 2014; USEPA 2015). For example, following a 2010 anthrax outbreak affecting native and exotic wildlife and livestock approximately 75 km North of Del Rio, Texas, maggots collected 10–20 days post-outbreak from the soil near a deer that had died from anthrax were culture positive for *B. anthracis*, and PCR-positive results were obtained for 80% of leafy vegetation collected within several meters of the carcasses which contained fly droppings (Blackburn et al. 2014). Twelve months following the outbreak, both pX01 and pX02 plasmids were recovered from soils surrounding a carcass (Blackburn et al. 2014). A study of enzootic areas in Etosha National Park in Namibia, Africa, found that 65% of samples taken near 106 carcass sites (of animals dying from anthrax) had at least 1 spore/g of soil and up to 10,000 spores/g soil and another 14% had more than 10,000 spores/g soil (Lindeque and Turnbull 1994). A second study in Etosha National Park also took soils near three sites of animals that had died of anthrax (two zebras and one springbok) and found that spore counts remained high in the soil for years (10^4–10^6 CFU/g soil) (Turnbull et al. 1998). Following repeated anthrax outbreaks in Sirajganj, Bangladesh, 14 of 48 soil samples taken in low-lying areas, pastures of livestock, and near burial sites were positive for *B. anthracis* spores (Ahsan et al. 2013). Of those 14 samples, all were from loamy soils with an average pH of 6.38 ± 0.15 and elevated moisture content (16.69 ± 2.06%) (Ahsan et al. 2013). Other studies have also found a similar positive relationship between soil moisture and presence of *Bacillus* sp. and/or *B. anthracis* (Griffin et al. 2009; Dragon and Rennie 1995). Geological elements that were noted near outbreak sites included soils with higher levels of magnesium (Mg), sodium (Na), and calcium (Ca) (Hugh-Jones and Blackburn 2009).

The geographic and ecological potential of all species involved in outbreak/disease transmission (vector, host, and pathogen) can affect emergence of a disease (Peterson 2008). The realized niche is the portion of the fundamental niche (in other words, all the variables/conditions that support long-term persistence) that the species is truly occupying (Hutchinson 1957; Phillips et al. 2006; Phillips and Dudik 2008). A species might not inhabit all areas of the fundamental niche due to competition with other species, historical factors, or lack of access to areas/geographic barriers (Anderson 2003; Anderson et al. 2002a, b; Peterson and Cohoon 1999; Peterson and Soberon 2012; Phillips et al. 2006; Phillips and Dudik 2008).

The realized and fundamental niches for *B. anthracis* or occurrence of anthrax outbreaks have not yet been fully defined. However, much of the anthrax outbreak data shows anthrax occurring in "hotspot" areas throughout the country. Anthrax outbreak data for the contiguous United States for the years 1915–1944 and 1944–1955 as reported by Van Ness and Stein found that many of the anthrax hotspots or "Anthrax Districts" were located along the Texas-Louisiana Gulf Coast, the eastern border region of Nebraska and South Dakota, and north-central California, while "sporadic or occasional outbreaks" were reported across widespread regions both east and west of the Mississippi River (Stein 1945; Stein and Van Ness 1955; Van Ness and Stein 1956). Their research illustrates the potential for sporadic outbreaks in wildlife and livestock over a wide geographic region. Anthrax outbreak range maps mirror historical bison range maps, and it was hypothesized that this herbivore must have played a significant role as a vector of this pathogen throughout North America once it was introduced (Hornaday 1889; Stein 1945).

Contrasting the 1915 to 1955 outbreak data, most anthrax outbreaks in animals have occurred west of the Mississippi River since 2000. This decrease in the widespread distribution of this disease is due to successful animal and public health efforts such as surveillance and vaccination (Grabenstein 2008; Hugh-Jones and de Vos 2002; Ndiva Mongoh et al. 2008; Zhang et al. 2013). Recent livestock disease occurrence data recorded by the National Animal Health Reporting System (NAHRS) (APHIS 2014) for the years 2005 through 2012 reports anthrax cases in nine states of the contiguous United States, California (2005, 2007, 2008, and 2011), Minnesota (2005, 2006, and 2008), Mississippi (2012), Montana (2005, 2007, 2008, and 2010), North Dakota (2005–2010 and 2012), Oregon (2012), South Dakota (2005 through 2009 and 2011), Texas (2005 and 2008–2012), and Nevada (2009). The NAHRS data illustrate persistence of disease in livestock in some geographic areas (southern Texas and a region that stretches north and east out of California and into Minnesota) despite some of the best immunization efforts to date.

Between 2004 and 2005, several researchers investigated the actual occurrence of *B. anthracis* across the United States. As part of a US Geological Survey (USGS) North American Soil Geochemical Landscape Project (NASGLP) pilot study, 220 soil samples were collected along two transects, a North-South transect extending from northern Manitoba, Canada, to the US border near El Paso, Texas, and a Gulf Coast transect along the I-10 corridor from Sulfur, Louisiana, to DeFuniak Springs, Florida (Griffin et al. 2009; Smith et al. 2009). Sites from each transect were spaced in 40 km intervals

using global position system to identify sample location sites (Griffin et al. 2009; Smith et al. 2009). In addition, samples were collected in downtown New Orleans and Chalmette post-Hurricane Katrina in 2005 and then again in 2007 (Griffin et al. 2009). Samples were analyzed for the presence of *B. anthracis* and *Bacillus* species using polymerase chain reaction (PCR), soil moisture content, and elemental concentrations (Griffin et al. 2009). The study detected *B. anthracis* and *Bacillus* sp. in 5% and 20% of 107 sites, respectively, that extended from Manitoba, Canada, to the Texas-Mexican border (Griffin et al. 2009). The study also found that 5 of 19 samples were PCR positive for *B. anthracis* in samples collected areas in downtown and surrounding New Orleans following flooding by Hurricane Katrina in 2005 (Griffin et al. 2009). Two years later, these same sites while rich in *Bacillus* species were negative for the presence of *B. anthracis*. Elevated concentrations of geochemicals were found in samples that were PCR positive for *Bacillus* sp. and included cobalt (Co), copper (Cu), lead (Pb), tin (Sn), thallium (Tl), and zinc (Zn) in the North-South transect; Ca, Mg, and phosphorus (P) in the Gulf Coast transect; and Na and sulfur (S) in the 2005 New Orleans sample subset (Griffin et al. 2009). Several geological factors have been noted to potentially influence or *B. anthracis* survival, as noted through in vivo or in vitro observations. These factors include are elevated phosphate (results in higher protective antigen production), Zn (for lethal factor production), and certain concentrations of manganese (Mn) (Kochi et al. 1994; Weinberg 1987; Wright et al. 1970).

While previous work has shown that soil characteristics, such as Ca content and pH, influence the occurrence of *B. anthracis*, large-scale geochemical and microbiological studies are still needed to determine constraints on the spatial distribution of the species (Smith et al. 2000). USGS conducted a full-scale NASGLP project across the conterminous United States from 2007 to 2011, collecting soil samples at 1 site per 1600 square kilometers using a generalized random tessellation stratified design to expand baseline geochemical and microbiology data (Smith et al. 2011; USGS 2013). A large subset (4770) of the soil samples were screened in a joint USGS-USEPA investigation for the presence of *Bacillus* sp. and *Bacillus anthracis* using a multiplex polymerase chain reaction assay (PCR). The NASGLP pilot study was able to detect *B. anthracis* and *Bacillus* sp. in soil samples collected using a random sampling design over different time intervals, in areas with differing geochemical make-up, and following different climatic conditions (post-flood and 2 years after). Therefore, the authors hypothesized that the available PCR data from the USGS-USEPA investigation could be evaluated against soil and environmental conditions (geochemical constituents, climate conditions, topology of the area, and animal density) using Maxent modeling to gain insight on the influence of these conditions on the distribution of *B. anthracis* in soils of the contiguous United States. In addition, animal anthrax outbreaks have been noted to occur in certain parts of the country and have been associated with specific climatic and environmental conditions (as noted in the previous discussion and Table 1). Therefore, authors also hypothesized that Maxent modeling could be used to evaluate reported wildlife and livestock outbreaks against soil and environmental conditions to identify areas that might have characteristics with a higher potential to support animal anthrax outbreaks.

Maxent is a type of ecological niche modeling (ENM). Broadly, ENMs are used to characterize the geographic distribution of a species by comparing known occurrences of the species/disease to environmental variables and ultimately to define which environmental variables are likely predictors of areas that meet the ecological requirements of the species (Peterson 2006; Silva et al. 2014). Maxent is a machine learning algorithm that can be used to predict the probability that a species occupies a given location using occurrence locations, conditional on the corresponding environmental covariates (Elith et al. 2011; Phillips et al. 2006; Phillips and Dudik 2008; Phillips et al. 2017). Maxent typically uses a presence-only (presence/background) approach for modeling, so it can be used in datasets that don't have complete information about absence (Elith et al. 2011; Phillips et al. 2006). With presence-only data, a presence denotes the locations where the species have been observed, and the probability of presence is estimated (Elith et al. 2011; Phillips et al. 2006; Phillips and Dudik 2008). Background samples from the dataset are used to train the model by providing a sample of the environmental characteristics of the study area (Elith et al. 2011; Guillera-Arroita et al. 2015). The output of Maxent modeling is either raw data, cumulative, or a logistic representation of the probability that the species that is present at each pixel in the mapping extent (Elith et al. 2011; Phillips et al. 2006; Phillips and Dudik 2008).

Maxent has been used extensively in species distribution studies for many different organisms (Carnaval and Moritz 2008; Cordellier and Pfenninger 2009; Elith et al. 2006; Elith et al. 2011; Graham and Hijmans 2006; Kharouba et al. 2009; Kumar and Stohlgren 2009; Lamb et al. 2008; Monterroso et al. 2009; Murray-Smith et al. 2009; Pearson et al. 2007; Phillips et al. 2006; Phillips and Dudik 2008; Silva et al. 2014; Tinoco et al. 2009; Tittensor et al. 2009; Tognelli et al. 2009; Verbruggen et al. 2009; Ward 2007; Williams et al. 2009; Wollan et al. 2008; Yates et al. 2010; Yesson and Culham 2006; Young et al. 2009). Recent studies have even used Maxent to examine the soil conditions suitable for anthrax outbreaks in affected counties and the potential distribution of *B. anthracis* in Minnesota (Nath and Dere 2016) as well as predicting an ecological niche for *B. anthracis* in Zimbabwe based on outbreak data and environmental variables (Chikerema et al. 2013).

This paper describes the methods used for the PCR analysis, environmental variables collected for analysis, and the Maxent (Maxent version 3.3.3K) modeling and recursive feature elimination (via Python scikit-learn 0.18.1 and ArcGIS) that was conducted. This study also used two metrics to measure model performance including the widely used area under the curve (AUC) and an alternative method, the True Skill Statistic (TSS). This work provides an extension to the use of ecological niche modeling for identifying areas with soil and environmental conditions supportive of potential anthrax outbreaks in wildlife/livestock in the contiguous United States as it utilizes additional soil geochemistry data not used in similar research and has shown that further validation techniques should be considered when evaluating the data.

Materials and Methods

USGS Geochemistry Data

Collection of Soil Samples

Soil samples utilized in this study were collected by the NASGLP project as described by Griffin et al. (2014), Smith et al. (2011), Smith et al. (2012), Smith et al. (2009), Stevens and Olsen (1999), Stevens and Olsen (2003), Stevens and Olsen (2004), and USGS (2013). Briefly, the NASGLP project collected 4857 samples across the contiguous United States using a generalized random tessellation stratified design for sample site selection, at a density of 1 site per 1600 km^2. Latitude and longitude for each sample site were recorded in decimal degrees (WGS84 datum). Layers of soil collected included Horizon A, which was the uppermost mineral soil (<2 mm); soil at a depth from 0 to 5 cm (Horizon P); and Horizon C which was comprised of partially weathered parent material at the deeper levels. Samples were analyzed for major and trace elements to identify the abundance and spatial distribution of elements and minerals.

For microbial analyses, sterilized 50 ml tubes and aseptic technique were used to collect P-horizon soils as previously described (Smith et al. 2009). The target collection depth interval was 0–5 cm, but ranged from 0 to 40 cm due to variations in site characteristics, such as heavy detritus cover. All samples were shipped from various field locations to the USGS microbiology laboratory in St. Petersburg, Florida, and then stored by refrigeration until analyzed. In addition to elemental data, site-specific land cover data for each site was extracted from the National Land Cover Database 1992 Classification System (Homer et al. 2004).

Geochemical Analysis

The analytical methods and quality control protocols utilized for the analyses of major and trace elements were previously described (Griffin et al. 2014; Smith et al. 2005; Smith et al. 2009). In short, a < 2-mm-depth fraction of each sample was analyzed for elemental concentrations and reported as percent by weight (wt. %) or parts per million (ppm) (Smith et al. 2012; USGS 2013). Mineral components were reported as present by weight (USGS 2013).

Mapping of Geochemical Data

The inverse distance weighting (IDW) method was used to generate maps of elemental concentrations, based on sampled point data collected by the USGS (USGS 2013). Interpolations were performed using ArcGIS™ (ESRI, Redlands, CA). A resolution of 400 m, with a fixed search radius of 75 km, was used for all elements except arsenic (As), barium (Ba), bismuth (Bi), mercury (Hg), potassium (K), lanthanum (La), Pb, rubidium (Rb), antimony (Sb), silicon (Si), Sr, thorium (Th), and uranium (U), which were made using a resolution of 90 m and a variable search radius of 12 points.

Microbial Data

DNA Extraction

A total of 4770 samples collected by USGS were screened for the presence of *Bacillus* species and *B. anthracis* (Fig. 1). Samples where geochemical data and/or location data were lacking were excluded from analysis. Approximately 0.25 g of soil was transferred from the 50 ml collection tubes and weighed with a plastic weigh-boat and a bench-top scale, using sterile technique. DNA was extracted from the sample using the PowerSoil™ DNA Isolation Kit and protocol (MO BIO Laboratories, Inc., Carlsbad, CA). Three µl of kit eluent (total eluent volume = 100 µl) was utilized as PCR template.

Bacillus sp. and Bacillus anthracis PCR

For *Bacillus* species, *Bacillus anthracis*, and pX02 screening, the multiplex PCR primers utilized were previously described (Ko et al. 2003) and included BA-RF (5'-GACGATCATYTWGGAAACCG-3′), BA-RR (5'-GGNGTYTCRATYGGACACAT-3′), and Ba-SF (5'-TTCGTCCTGTTATTGCAG-3′) and Cap-S (5'-ACGTATGGTGTTTCAAGATTCATG-3′). These primers amplify a 359-base pair region of *rpo*B gene (encodes the RNA polymerase ß-subunit) that is specific for *Bacillus* species at the genus level (BA-RF/BA-RR primer pair) and a 208-base pair region of the same gene that is specific for *B. anthracis* (inclusion of the additional forward primer Ba-SF) (Ko et al. 2003). Master-mix recipe per reaction was 10 µl of QIAGEN HotStarTaq Plus Master Mix Kit (QIAGEN, Valencia, CA), 2 µl of the CoralLoad concentrate (QIAGEN), 1 µl of each of the five primers (10 µm working stock), and 3 µl of template. The Tempcycler reaction profile was 15 min at 95 °C, 30 cycles of 95 °C for 30 s, 45 °C for 30 s, and 72 °C for 1 min and a final

extension at 72 °C for 10 min followed by hold at 4 °C. PCR amplicons were visualized using 10 µl of the reaction volume and SYBR Gold-stained gel electrophoresis. Samples that produced amplicons for both *rpo*B markers were logged as presumptively positive for *B. anthracis*. *Bacillus atrophaeus* DNA obtained from a liquid culture extract was utilized for PCR-positive control reactions. Negative control template was PCR-grade water. Control template spike volumes were 3 µl of water for the negative control and 2 µl of water and 1 µl of DNA for the positive control.

Confirmation of PCR B. anthracis rpoB-Positive Samples

Samples that were *rpo*B PCR positive for *B. anthracis* were sent to the University of South Florida Center for Biological Defense (USF-CBD; http://www.virtualbiosecuritycenter.org/organizations/university-of-south-florida) for confirmation. Approximately 5 g of each sample was transferred to pre-sterilized 50 ml tubes using sterile technique and sent to USF-CBD for analyses. USF-CBD utilizes published and in-house designed primers and probes and dot-blot assay to screen the samples for the presence of the pOX1 (*pag* and *lef* markers) and pOX2 (*cap* marker) plasmid virulence genes (Luna et al. 2006).

PCR and Plasmid Data

PCR data for *Bacillus* species and *B. anthracis rpoB* gene detection was reported as non-detect (0), low (1), medium (2), and high (3). Results obtained by the USF-CBD for both pag and lef genes of the pX01 plasmid were recorded. This data was recorded as negative or positive for each of the genes and included the following combinations: neg/neg (0), pos/neg (1), neg/pos (2), and pos/pos (3). Data for the pX02 plasmid were recorded as negative (0) or positive (1). PCR and USF-CBD blot data results are available through USGS data release at http://coastal.er.usgs.gov/data-release/doi- F7WW7FRJ/. Contact information for additional data is listed at the above data release site.

Mapping of PCR Data

A shapefile was generated in ArcGIS using the PCR data, Final Sample ID (State and ID combined), and latitude and longitude of the corresponding sample analyzed for geochemical data. The sample sites are illustrated in Fig. 1. Note that sites that reported PCR detects for *B. anthracis* also reported detects for *Bacillus* sp. data; however, this is not depicted on the map.

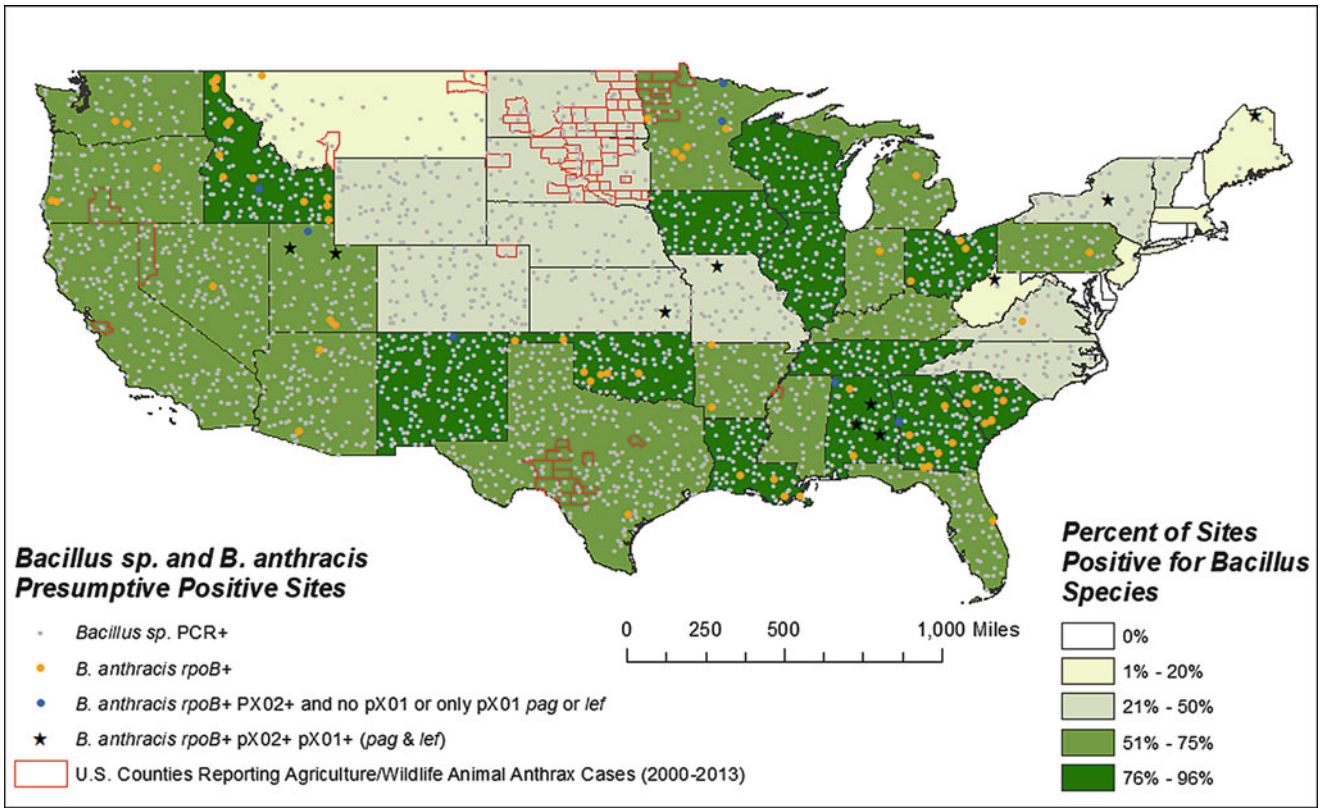


Fig. 1 Sample sites where *Bacillus* sp. and *B. anthracis* were detected

Climatic Variables

Precipitation and temperature have both been noted to influence *B. anthracis* spore survival or occurrence (see Table 1) and thus were chosen as variables for this study. Average annual precipitation data (ann-prcp-normal.txt), encompassing measurements from 1981 to 2010, was downloaded from the NOAA NCDC FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/products/precipitation/). Spatial information for all meteorological stations in the contiguous United States ($n = 8864$) is available as a separate file (prcp-inventory.txt) and was downloaded from the FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/station-inventories/). Metadata for these files and an explanation of the abbreviations and contents are available for download at ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/readme.txt.

Precipitation data and their locations were imported into ArcGIS version 10.3.1 and were used to create a point shapefile with data from each meteorological station. The point shapefile was clipped to the extent of the conterminous United States (WGS 1984 projection). Using the inverse distance weighting method, precipitation values at each point were interpolated to generate a 400 m resolution, continuous precipitation map.

Temperature records for the contiguous United States, from 1960 through 2015, were obtained from the NOAA Sate llite and Information Service site (http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp). The data were distributed by NCDC as Climate Division polygons; there are 344 climate divisions within the contiguous United States. The Climate Division Boundaries polygon layer was retrieved by downloading the CONUS_CLIMATE_DIVISIONS. shp.zip file from ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv. The final maps were manually classified into seven different choropleth classes.

Topological Characteristics

Alkaline soil, areas where topology is low, and areas where grasses or detritus has deposited after a flood are some of the factors that influence animal anthrax outbreak occurrence (see Table 1); therefore, slope, soil pH, flood frequency class, drainage class, and elevation were selected as variables for this study. Slope, soil pH, flood frequency class, and drainage class were extracted from the US Department of Agriculture's (USDA) National Resources Conservation Service (NRCS) SSURGO (Soil Survey Geographic Database) and STATSGO (State Soil Geographic Database) soil datasets. The SSURGO and STATSGO databases can be downloaded from the Web Soil Survey (http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm). SSURGO data are collected at a scale ranging from 1:12,000 to 1:31,680, while STATSGO data is collected at a scale of 1:250,000. While the finer-resolution data of SSURGO was preferred, the SSURGO data has holes in the datasets in certain areas of the country; therefore, STATSGO data was needed for those data gaps. SSURGO and STATSGO drainage class and flood frequency class are supplied in text classes (e.g., well drained, very well drained, etc.). These classes were reclassified into numerical scales from 1 to 7 for drainage class and 1 to 5 for flood frequency. Elevation data was extracted using a 100 m resolution elevation map of the conterminous United States was downloaded from the USGS 2012 100 m National Elevation Dataset (NED) (https://catalog.data.gov/dataset/100-meter-resolution-elevation-of-the-conterminous-united-states-direct-download).

Extracting Animal Outbreak Data by County

States (Texas, Minnesota, Oregon, Montana, California, North Dakota, and South Dakota) and counties with anthrax outbreaks in animals reported between 2001 and 2013 were extracted from the national dataset (APHIS 2014). In a new text field named "outbreak," counties with recorded anthrax outbreaks in wildlife and livestock between 2001 and 2013 were assigned a value of 1, and a 0 was assigned in non-outbreak counties. The ArcGIS zonal statistics tool was used to calculate the mean values for each environmental variable in each county in the states that reported animal anthrax outbreaks.

Agricultural Mammal Density by County

Because the anthrax outbreaks being used for this study are those that have been reported in livestock and wildlife, this study also wanted to capture locations and density of animal populations as a possible variable to potential anthrax outbreaks. The USDA National Agricultural Statistics Service (NASS) conducted an Agricultural Census in 2012 (USDA 2014). The density values are calculated as population/km^2 and displayed in percentile classes of 10%. Bison data (2012), all cattle (2012 – including beef, milk, and calves), equine (2012 – including donkeys, mules, horses, and ponies), farm-raised deer, and farm-raised elk were downloaded at the county level for the contiguous United States. Deer and elk refer to only farm-raised, not wild, deer and elk and were combined into one dataset. Beef cattle, milk cows, calves, elk, deer, equine, and bison population data for each county were added together. The total population for each county was divided by the county area (km^2) to get the total density.

Extracting Data Prior to Maxent Model Run

Continuous maps for each parameter (geochemical properties, climatic properties, topological properties, and animal density data extracted for each location where soil was sampled) were generated. For consistency with other datasets, all input data was converted or resampled to produce 400 m resolution raster and with a geographic coordinate system of WGS84.

Maxent Modeling

Presence Data, Model Scales, and Training Data

Presence data that was available for this evaluation included locations where *B. anthracis rpoB*-positive samples were detected ($n = 83$) and US counties reporting historic recorded anthrax outbreaks in wildlife/livestock (from 2001 to 2013). Because there were only 83 samples that were *B. anthracis rpoB* positive, initial model runs to investigate the influence of environmental factors on the distribution of *B. anthracis* used the latitude and longitude of the presumptive positives for *B. anthracis rpoB* PCR positives (yellow dots, blue dots, and black stars in Fig. 1) as well as the central latitude and longitude for any county reporting wildlife/livestock anthrax outbreaks (counties reporting outbreaks are shown in red outlines in Fig. 1) as presence data. The blue triangles in Fig. 2 show this combined data. The model utilized the same variables determined to be most important by Nath and Dere (Nath and Dere 2016) including pH, Ca, Mg, Na, total carbon (C), clay content, Sr, and Mn to estimate the likelihood of presumptive positives across the United States to see if similar results were obtained.

For the remainder of the study, to investigate the influence of environmental factors on areas with conditions which could support a higher potential for anthrax outbreaks in wildlife/livestock, only the reported animal anthrax outbreak data was used as presence data. This data was evaluated on two scales: the Outbreak State scale and the National Scale. The Outbreak State scale consisted of only the states that reported animal outbreaks from 2001 to 2013 and was used as a proof of concept. The National scale covered the full extent of the lower 48 states. Twenty percent of presence data were randomly removed for cross-validation, leaving 80% to train the model (data not described). The "write background predictions" feature was turned on in the advanced settings menu. The number of iterations (500) and output format (logistic) were left to the default settings.

Three different iterations of the available environmental variables were used: 1) The Sample Location data subset included the environmental variable data using the assigned latitude and longitude locations from the USGS NASGLP project that fell within either the Outbreak State or National scales. The Interpolated dataset included an interpolation of the environmental variables averaged for each county and assigned to a point for that county at the centroid (rather than using the NASGLP latitude and longitude) for both the Outbreak State and National scales. This was done to try to account for any possible outliers. Finally, the environmental variables were normalized (completing the Normalized dataset) by scaling the environmental variables so that the values fell between 0 and 1 using Python prior to running Maxent. This was done so the environmental variables would have the same units and order of magnitude in order to determine how different scales might change the model. Maxent version 3.3.3k was used for all modeling.

Recursive Feature Elimination

Recursive feature elimination (RFE) is a method to select a set of features by training an estimator on a set of weighted features and recursively eliminating ones with the smallest absolute weights (Pedregosa et al. 2011). For this study, multiple RFE were run used to narrow down the list of environmental variables that would be utilized during Maxent modeling, to prevent overfitting of the model. Before using RFE, minor elements and unstable elements were removed from the dataset. It was determined that several elements could be excluded because (1) they do not exist naturally in the environment, (2) they are unstable on their own in the environment, or (3) they are too minor (<500 mg/kg) and can throw off statistical models if included. RFE was run with the Python scikit-learn 0.18.1 package. The RFE was set up using the logistic regression method for all variables collected with an end goal of five features. The final list of environmental variables used for Maxent modeling is listed in Table 2.

Evaluation of Performance

This study utilized two measures of model performance, the area under the curve (AUC) and the True Skill Statistic (TSS). The AUC and TSS methods are briefly described below.

Area Under the Curve

Sensitivity is the probability that the model classifies a presence correctly, while specificity is the probability that the model classifies an absence correctly (Allouche et al. 2006). The receiver operating characteristic (ROC) curve provides a measure of model performance. The ROC plots the model's sensitivity against the proportion of false positives (commission error), which is calculated as 1- specificity and summarizes that measurement in a single number, the area under the curve (AUC) (Allouche et al. 2006; Chikerema et al. 2013; Lobo et al. 2008; Phillips et al. 2006; Phillips and Dudik 2008). Specifically, the AUC gives the probability that a randomly chosen presence location has been correctly ranked higher than the absence/background site and can be compared between model algorithms (Chikerema et al. 2013;

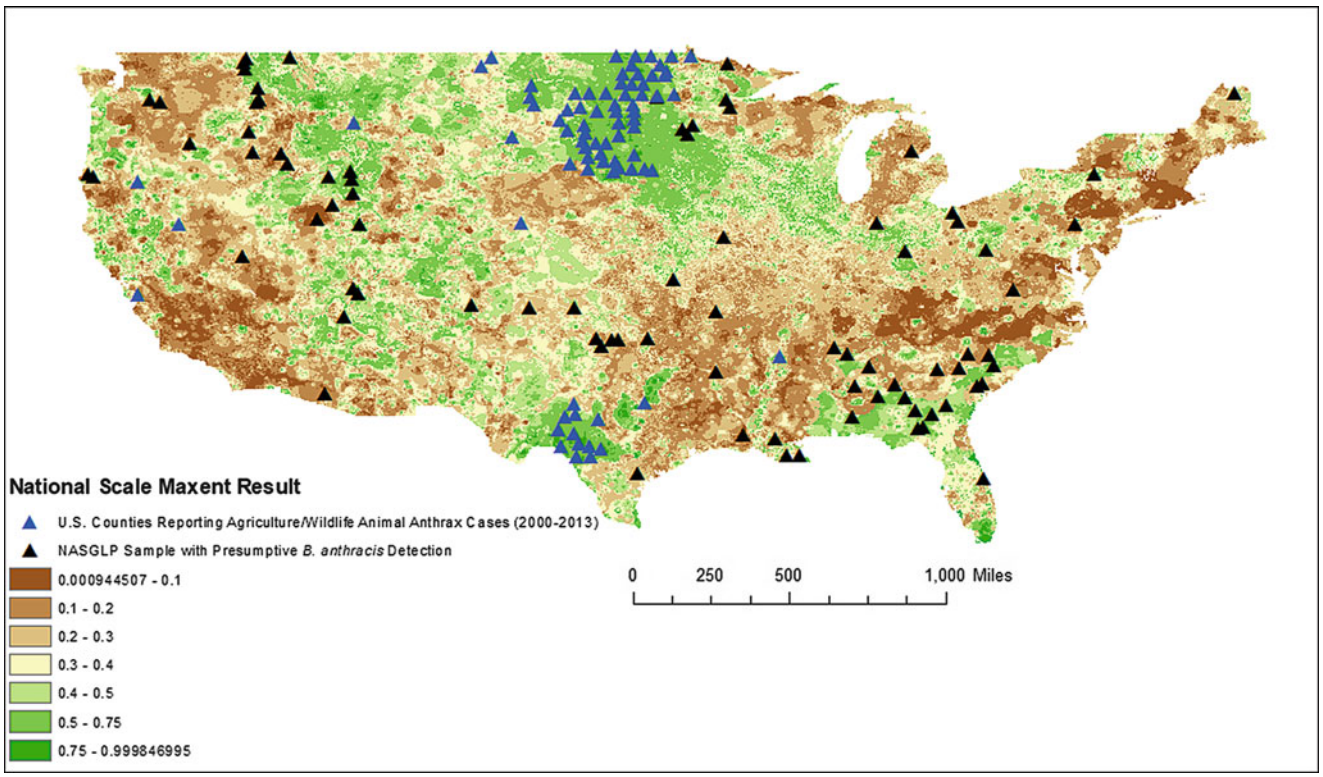


Fig. 2 Maxent result using *rpoB* presumptive positives for *Bacillus,* counties reporting wildlife/livestock outbreaks from 2000 to 2013 (center of each county) data and the variables pH, Ca, Mg, Na, Sr, Mn, clay content, and carbon content

Phillips et al. 2006; Phillips and Dudik 2008). The AUC weights the commission and omission error (false negative) the same (Lobo et al. 2008). AUC values at 0.5 or lower mean the ranking is no better than random, while those nearing 1 mean the ranking is a better predictor and nearing perfect (Chikerema et al. 2013; Phillips and Dudik 2008).

True Skill Statistic (TSS)

The TSS has been used as an alternative measure of performance for species distribution models (Allouche et al. 2006) and has recently been used as a measure of model performance for determining the realized niche of Orchid bees using Maxent modeling (Silva et al. 2014). The TSS provides a comparison of how well the background predictions made by the model match the model results at the test dataset (presence) locations. Values near +1 are perfect agreement of the model distributions and the observations, while values near −1 indicate the model is no better than a random model (Silva et al. 2014). The TSS is not affected by the size of the background dataset or prevalence and combines specificity and sensitivity in order to account for omission and commission errors and random guessing (Allouche et al. 2006).

Briefly,

1. The model saves 10,000 random background probability predictions (predictions for this study were saved in the csv file, "species_backgroundPredicitons.csv").
2. Using the 10% threshold value, the number of background predictions above and below that threshold is counted.
3. The model also creates a csv file that saves the probability predictions at each of the sample sites (saved in the csv file, "species_samplePredictions.csv").
4. The number of test location predictions above and below the 10% threshold is counted.
5. Sensitivity is the

   (Number of test cells < threshold) / (Total number of test cells)

6. Specificity is the

   (Number of background cells < threshold) / (Total number of background cells)

7. TSS is Sensitivity + Specificity – 1.

TSS score interpretations are listed in Table 3. For this study, omission and commission rates were calculated only for the National level scale because the Outbreak State scale was used as a proof of concept and did not contain enough data points to calculate an accurate omission or commission rate.

Table 2 Final environmental variables used for Maxent modeling

Variable	Abbreviation	Source
Aluminum	Al	USGS (2013)
Arsenic	As	USGS (2013)
Barium	Ba	USGS (2013)
Calcium	Ca	USGS (2013)
Cesium	Ce	USGS (2013)
Cobalt	Co	USGS (2013)
Chromium	Cr	USGS (2013)
Copper	Cu	USGS (2013)
Iron	Fe	USGS (2013)
Mercury	Hg	USGS (2013)
Potassium	K	USGS (2013)
Magnesium	Mg	USGS (2013)
Manganese	Mn	USGS (2013)
Sodium	Na	USGS (2013)
Nickel	Ni	USGS (2013)
Phosphorus	P	USGS (2013)
Lead	Pb	USGS (2013)
Sulfur	S	USGS (2013)
Strontium	Sr	USGS (2013)
Titanium	Ti	USGS (2013)
Zinc	Zn	USGS (2013)
Amorphous soil content	Amorph	USGS (2013)
Carbonate soil content	Carb	USGS (2013)
Clay content	Clay	USGS (2013)
Elevation	DEM	USGS (2012)
Drainage class	Drain	Soil Survey Staff (2017)
Total feldspar soil content	Flds	USGS (2013)
Cattle, elk, deer, equine, bison density	Mammal	USDA (2014)
Average annual precipitation	Prcp	NOAA (2010)
Slope	Slope	Soil Survey Staff (2017)
Average annual temperature	Temp	NOAA (2015)
Total carbon content	Tot_c	USGS (2013)

Table 3 TSS score interpretation

TSS score	Interpretation
0–0.4	Poor
0.4–0.5	Fair
0.5–0.7	Good
0.7–0.85	Very good
0.85–0.9	Excellent
>0.9	Perfect

Results

Bacillus sp. and *Bacillus anthracis* PCR Results

Bacillus species were detected in 2876 (60.3%) of the samples in 43 of the 48 states (% positive range of 7.2 to 95.7, Fig. 1). States where *Bacillus* sp. was detected in more than 75% of soil samples included Alabama, Georgia, Iowa, Idaho, Illinois, Louisiana, New Mexico, Ohio, Oklahoma, South Carolina, Tennessee, and Wisconsin (Fig. 1). States where *Bacillus* sp. was not detected (i.e., Connecticut, Delaware, Maryland, New Hampshire, and Rhode Island) or was detected infrequently (i.e., Massachusetts and New Jersey, where <10% of samples were positive) were all located in the northeast. These states were relatively small in size, and consequently the number of samples collected in those states ranged from only 2 to 18 (data not shown). Several states had a low incidence of *Bacillus* sp.-positive samples, despite medium- or large-sized sample sets. Those states included Maine at 13.7% positive out of 51 samples, West Virginia at 15.4% positive out of 39 samples, and Montana at 20.5% positive out of 234 samples.

Table 4 *Bacillus anthracis* presumptive PCR positives (83 total) and pX01/PX02 blot results

# PCR positive	By state (# PCR positive)	pX01 *pag*	pX01 *lef*	pXO2
10	AL (3), KS (1), ME (1), MO (1), NY (1), UT (2), WV (1)	+	+	+
66	AL (2), AR (2), AZ (2), FL (1), GA (7), ID (13), IN (1), LA (4), MI (1), MN (5), MT (1), NV (1), OH (3), OK (7), OR (3), PA (1), SC (6), TX (1), UT (2), VA (1), WA (2)	−	−	−
1	AL (1)	−	+	+
3	GA (1), MN (2)	−	−	+
3	ID (1), NM (1), UT (1)	+	−	+

+ positive, − negative

The *rpo*B gene for *Bacillus anthracis* was detected across the *United States* in 83 of the samples that ranged in origin from northern Maine to southwestern Oregon (illustrated in Fig. 1 and Table 4). The highest rates of occurrence were found in a cluster in the south (a total of 20 sites in Alabama, Georgia, and South Carolina), in 7 sample site groups in Oklahoma and Minnesota, and in a cluster of 26 sites in the northwest (Idaho, Montana, Nevada, Oregon, Utah, and Washington).

Of the 83 *rpo*B-positive sites, 10 (one in Maine, New York, West Virginia, Missouri, and Kansas, two in Utah, and three in Alabama) were positive for both the PX01 *pag* and *lef* and pX02 *cap* virulence markers (Table 4). These ten samples were confirmed to contain all three markers by USF-CBD. The pX02 *cap* virulence marker was detected in another 7 of the 83 presumptive positive samples. These samples were collected in Alabama (1), Georgia (1), Idaho (1), Minnesota (2), New Mexico (1), and Utah (1). Four of these seven samples contained one of the other two virulence markers (the Alabama sample contained the PX01 *lef* marker and the Idaho, New Mexico, and Utah samples contained the PX01

Table 5 RFE results for national scale models

Sample location dataset	Interpolated dataset	Normalized dataset
Al	Fe	Al
Na	Na	Ca
K	pH	Sr
Mg	Feldspar content	Feldspar content
Drainage class	Drainage class	Slope

pag marker). None of the virulence markers were detected in 66 of the presumptive PCR-positive samples.

All positive control PCR reactions produced the appropriate-sized amplicon, and no amplicon signal was noted in any of the negative control reactions. To address reproducibility of the PCR assay, a total of 276 samples were run in duplicate for the detection of *Bacillus* sp. and/or *B. anthracis*. Of these, 63 of the samples were negative for both reactions, 197 were positive for both reactions (primarily *Bacillus* sp. positive), and 16 were reactions where one was positive and one was negative (6.0%). Given the *Bacillus* species PCR reaction sensitivity as previously published (4 CFU) (Griffin et al. 2009), and the above PCR agreement rate of 94.0%, these data accurately reflect occurrence (at or above the limit of detection) at the time these samples were collected.

The final variables used for Maxent are included in Table 5. The RFE was able to determine which five variables best explained the differences between counties where anthrax outbreaks had occurred in wildlife/livestock and counties with no recorded outbreaks. RFE tests produced models with some similar variables, though no variable was consistently observed in all three models (Sample Location dataset vs. Interpolated dataset vs. Normalized dataset) (Table 5).

Maxent Results

Initial Maxent modeling which used both the locations where *B. anthracis rpoB* PCR was detected and the counties that reported anthrax outbreaks in wildlife/livestock as presence data did not perform well. The model (Fig. 2) utilized the variables determined to be most important by Nath and Dere (Nath and Dere 2016) including pH, Ca, Mg, Na, total carbon (C), clay content, Sr, and Mn to estimate the likelihood of presumptive positives across the United States underfitting the data. Although the AUC score was 0.83 (AUC test was 0.692), and the map was aesthetically pleasing, the TSS was only 0.484 (fair) suggesting the model appeared to underfit the data.

When animal anthrax outbreak data alone were used as the presence data, the models for the three iterations of environmental variables (Sample Location, Interpolated, and Normalized datasets) performed well in AUC and TSS tests (Table 6). Although the Sample Location dataset performed very well at the State scale, when scaled up to the National scale, performance decreased (TSS 0.82, very good, to 0.683, good). The Interpolated dataset produced models that performed about the same for both the State and National scales (TSS 0.558, good, and TSS 0.532, good, respectively). The Normalized dataset performed better at the National scale (TSS 0.538, good) than at the State scale (TSS 0.359, poor). There were larger differences in the rate of commission (true negative rate) in the Sample Location dataset (79.2%) compared to the Interpolated (63.5%) and Normalized (65.5%) datasets.

Probability maps for all three models are presented in Figs. 3, 4, and 5. For reference, brown areas on the maps represent locations with the lowest probability for occurrence, and the cream areas represent locations with a moderate probability for occurrence (0.3–0.4). The green areas represent locations with the highest probability for occurrence. In general, the highest probability of occurrence was found in North Dakota, South Dakota, Minnesota, Oregon, Montana, Texas, Michigan, Maine, and states along the Mississippi River for all three models. The largest differences in probability among the models were found in the Northwest and Great Plains regions.

Discussion

This study attempted to use Maxent modeling to examine the geochemical soil constituents and environmental conditions that could potentially influence the distributions of *B. anthracis* in soils of the contiguous *United States* and to identify the locations with environmental conditions which are potentially more supportive of anthrax outbreaks in wildlife and livestock. A discussion of the results from this study is presented below.

PCR Data Discussion

Figure 1 provides a "snapshot in time" of PCR results and illustrates the widespread occurrence of *Bacillus* species PCR-positive samples across the 48 contiguous *United States*. In northeastern and north central/western states, *Bacillus* species were not detected as frequently as they were in other areas, such as the south and southwest. Many of the *rpo*B *B. anthracis* PCR detects from this study occurred in regions that have reported cases or outbreaks of anthrax in wildlife and livestock in historical records (Stein 1945; Stein and Van Ness 1955). However, the cluster in Idaho is not consistent with historical (Stein 1945; Stein and Van Ness 1955) or recent observations (Fig. 5). Outside of the seven *rpo*B *B. anthracis*-positive sites in Minnesota (none

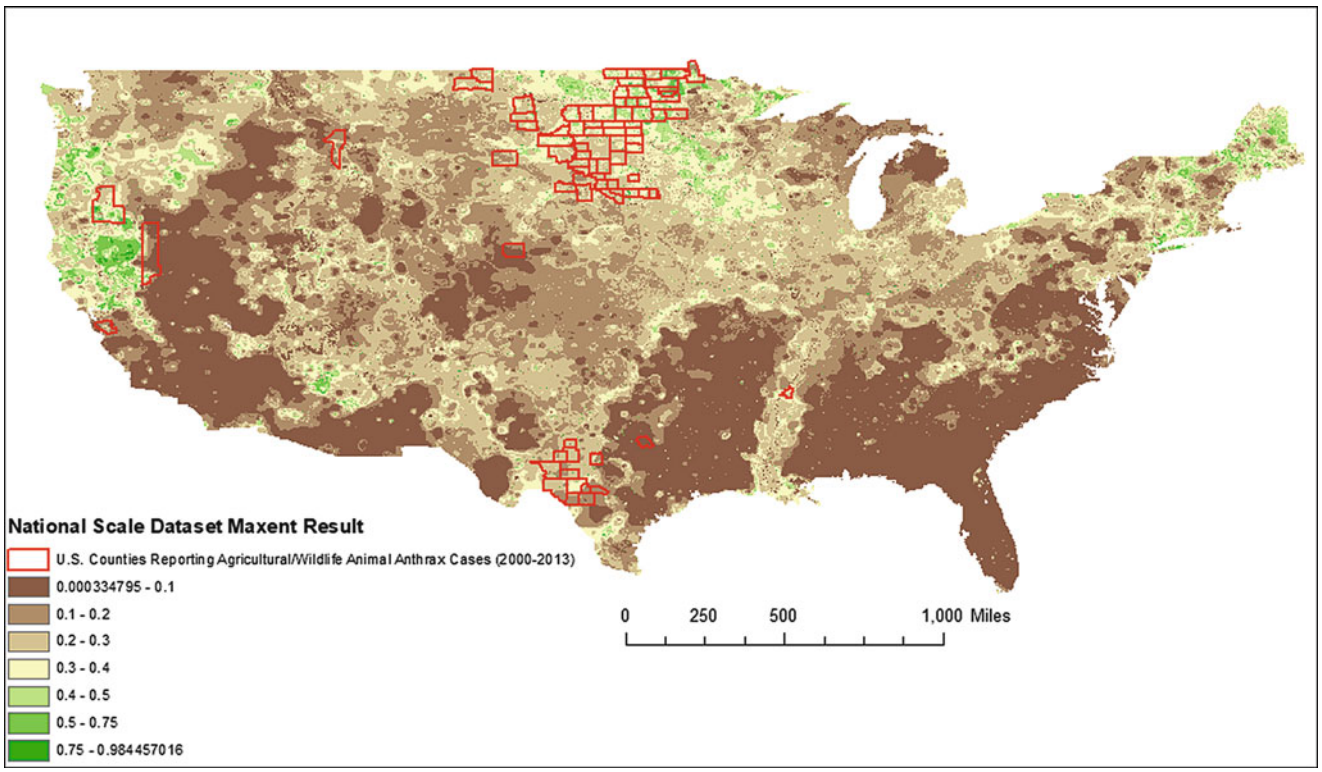


Fig. 3 Sample Location dataset Maxent result (Al, Na, K, Mg, drainage class)

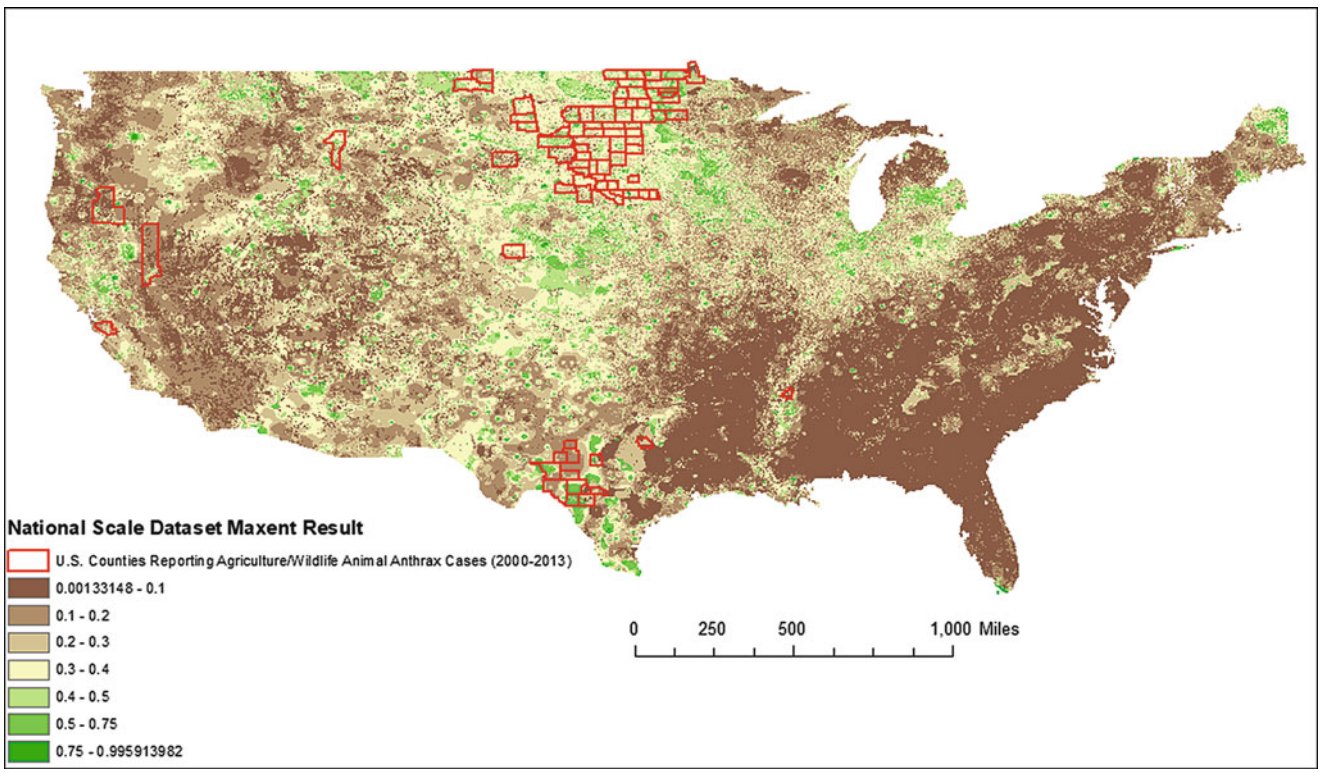


Fig. 4 Interpolated dataset Maxent results (Fe, Na, pH, feldspar content, drainage class)

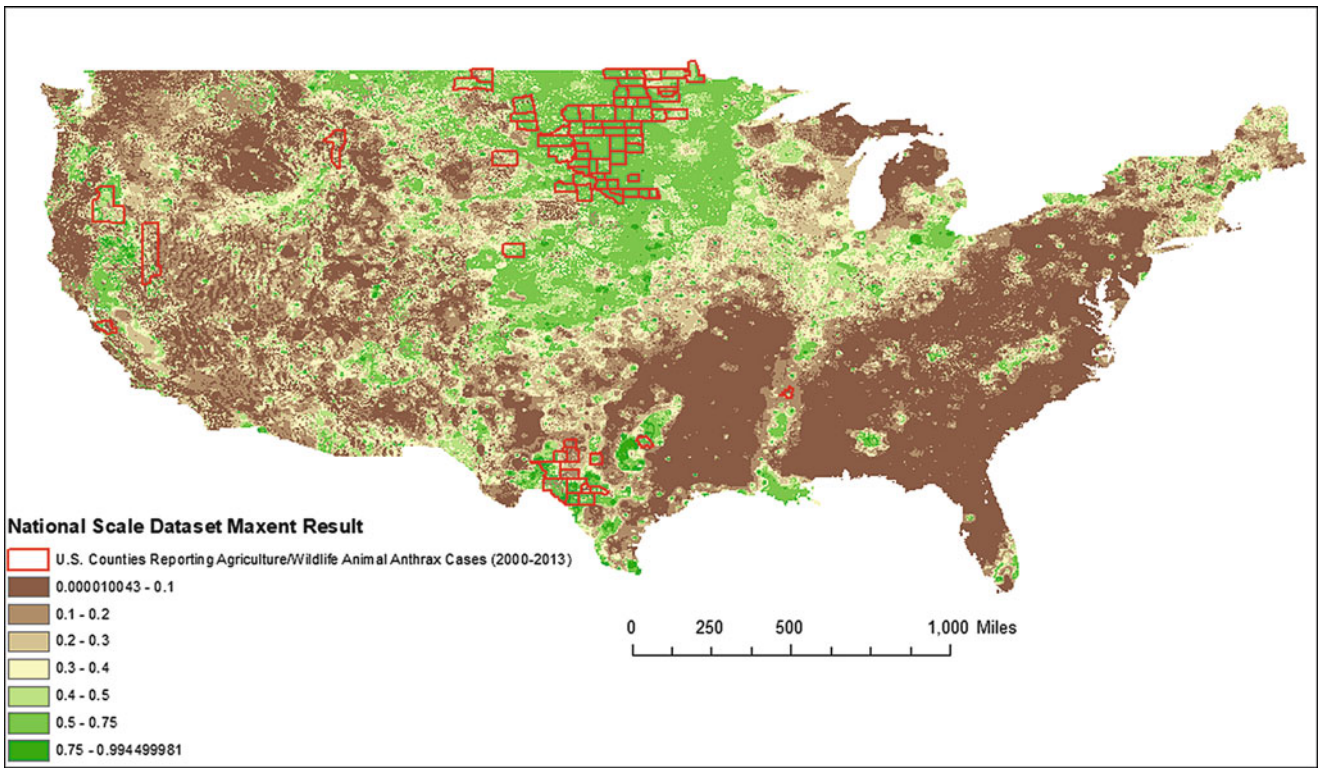


Fig. 5 Normalized dataset Maxent results (Al, Ca, Sr, slope, feldspar content)

of these were positive for all plasmid markers), the *rpo*B gene for this pathogen was not detected in what may be considered the modern "anthrax hotspots" of southern Texas, the Dakotas, or northeast Montana. The inability to detect this pathogen in animal anthrax outbreak areas, including in soil samples collected near anthrax-infected carcasses, has been previously observed and attributed to factors such as low sporulation rates, spore degradation, survival below surface soils, and assay detection limits (Beyer et al. 1999; Coker 2002; Dragon and Rennie 1995; Teshale et al. 2002). An effort was conducted to increase detection of *B. anthracis* in soil samples by optimizing processing of the soil samples (Silvestri et al. 2016). In addition, an enrichment-based PCR assay has been suggested for enhancing the detection of this pathogen in soil samples given the limitations observed in previous studies (Coker 2002; Letant et al. 2011; USEPA 2012). The presence or absence of the PCR *B. anthracis rpo*B marker in the samples only represents a snapshot in time and could reflect temporal variations in environmental conditions, variances in vaccination efforts, or the lack of optimal elemental concentrations needed to facilitate an infectious state. In addition, using a random stratified sampling strategy for this type of study might not be ideal for locating *B. anthracis* spores in soil. Given the spatial, temporal, climatic, and topological variability of the samples collected during the NASGLP study, locating this pathogen at the "right time and the right place" could be extremely difficult. A more targeted approach for sampling might need to be considered which takes into account the factors mentioned above.

Discussion of Maxent Modeling

It is not surprising the model, which utilized the *B. anthracis rpoB* in addition to the animal anthrax outbreak county point data (Fig. 2) to look at the potential distribution of *B. anthracis*, did not yield a high TSS score. The PCR data represented the presence of *B. anthracis* collected from a discrete sample during a single snapshot in time, which wasn't concurrent with the animal anthrax outbreaks themselves. Temporal variations within a sample site are difficult to capture with presence-only data (Elith et al. 2011). Although spores have been shown to survive for many years in soil, and much has been documented on the environmental, weather/climate, and geographic factors associated with animal anthrax outbreaks (see Table 1), information is scarce on spore distribution and transport in soil following an outbreak, factors affecting sporulation, natural attenuation of *B. anthracis*, and persistence in non-host microenvironments (see USEPA (2014) and USEPA (2015) for a summary). In addition, detection of spores in soil is difficult due to low concentrations, low processing efficiencies used with past sampling methods, and inhibiting compounds in soil matrices (Silvestri et al. 2015; USEPA 2014). The method utilized to

process the samples for the study had a limit of detection of 10^4 colony-forming units (CFU) per g soil (Silvestri et al. 2016); therefore, recovery of *B. anthracis* in the discrete samples that were collected was difficult. It is possible that samples taken near locations of an animal anthrax outbreak contained spores that were not detected, thus limiting the amount of presence data that was available for this study. In addition, using PCR data of this nature still has its limitations as conclusions will be limited to identifying areas that could potentially promote pathogen survival, rather than being able to concretely identify where the pathogen is located. Re-analysis of samples in potential anthrax "hotspots" has been considered utilizing an improved processing protocol (Silvestri et al. 2016; USEPA and USGS 2017) if funding allows.

Focusing on soil and environmental conditions over larger areas, such as counties that have experienced anthrax outbreaks in wildlife and livestock, provided a more realistic characterization of areas that could potentially support anthrax outbreaks. Utilizing animal anthrax outbreak data only, the model with the highest AUC and TSS scores for this study was the State model using the Sample Location dataset (AUC = 0.918 and TSS = 0.82). This was the model that included all environmental variables for each sample location within only the states that had reported anthrax outbreaks in livestock and wildlife. The Sample Location and Interpolated datasets in this study performed best when using state-level data from the animal anthrax outbreak vs. non-outbreak counties; however, when the models were scaled up to the national level, model performance declined. When using the Normalized dataset, the model performed better at the national scale; at the State scale, the statistical measures assumed the model was underfitting the data as it had high sensitivity (true positive rate) and low specificity (true negative rate). A similar scale-dependent effect was also observed by Nath and Dere (Nath and Dere 2016), who used Maxent to examine the spatial distribution of soil conditions suitable for anthrax outbreaks in affected Minnesota counties. Their statewide model had a better AUC (0.978) than the local model (0.698), and the predictors for anthrax outbreaks varied by scale (with the exception of sand content, which was noted for both). The authors suggested that the statewide model performed better due to heterogeneity of the soils across Minnesota.

The RFE evaluation for the current study looked at which five variables best explained the differences between counties where animal anthrax outbreaks occurred and counties where there were no recorded outbreaks (Table 5). Interestingly, RFE tests (Sample Location vs. Interpolated vs. Normalized datasets) produced models with some common variables (Table 5). However, it is not possible to identify a set of variables which fits all model scales and situations, because the variables varied by scale and how the data was treated. The Normalized dataset noted Ca and Sr to be important predictors, similar to previous research which looked at geochemicals present in counties reporting anthrax outbreaks in livestock and wildlife (Griffin et al. 2014). A study of anthrax outbreaks in Minnesota (Nath and Dere 2016) also found Ca (Minnesota statewide model) and Sr (Minnesota local model) to be important predictors of anthrax outbreaks. However, Ca and Sr were not in the top five predictors for the Sample Location and Interpolated datasets in this study.

Previous studies have found that soil type (Chikerema et al. 2013), sand content (Nath and Dere 2016), and clay content (Nath and Dere 2016) are potentially important predictors for anthrax outbreaks. While this study did not specifically find soil type such as sand or clay to be in the top 5 predictors, the total feldspar content was a key variable for both the Interpolated and Normalized datasets, suggesting that ENM models should expand the soil type variable to also include an analysis of the mineral content in niched model evaluations.

RFE tests determined that slope was a key indicator in the Normalized point dataset, while a related variable, drainage class, was a key indicator for both the Sample Location and Interpolated datasets. These correspondences are consistent with the assumption that natural drainage and flooding can disperse spores over a large area (Epp et al. 2010; Turner et al. 1999a, b) and the "concentrator theory" (Dragon and Rennie 1995) that posits floods and runoff are capable of washing soils into spore-bearing depressions, burying spores until subsequent surface runoff uncovers them.

These models in this study are characterized by a region of high probability stretching from Texas to Minnesota, which was also observed in several studies using Genetic Algorithm for Rule-Set Prediction (GARP) for predicting *B. anthracis* distribution in the United States and other countries (Blackburn 2010; Blackburn et al. 2007; Mullins et al. 2013). GARP utilizes a genetic algorithm to give a binary prediction using positive (suitable environmental conditions) and negative (unsuitable environmental conditions) rules (Phillips et al. 2006). However, the Maxent models from this study predicted higher-probability areas in East Texas than the GARP models while also under-predicting probability in areas such as South West Texas, where historical animal anthrax outbreaks are known to have occurred. These areas included Falls, Kinney, Val Verde, Uvalde, Edwards, and Real counties (Kenefic et al. 2008; USDA 2006) that were not modeled (Sample Location dataset) to have a high probability for potential anthrax outbreaks. However, it is worth noting that there have not been animal anthrax outbreaks in these areas since 2000, which was several years earlier than the time period encompassed by the datasets used in this study. Surprisingly, two states, Michigan and Maine, were predicted to be areas where environmental conditions could potentially support anthrax outbreaks in animals (in the National Scale Interpolated and Normalized datasets), despite not having

recent outbreaks of anthrax in wildlife or livestock. There were some similarities between the current study to the models predicted by GARP in the Midwest and some areas of the western United States (Blackburn 2010; Blackburn et al. 2007; Mullins et al. 2013). However, with the addition of new nationwide soil geochemistry data, the current models were able to map probability at a higher resolution compared to studies using GARP.

The Maxent models in this study predicted an elevated probability for anthrax outbreaks in wildlife and livestock along the Mississippi river. Previous studies that looked at *B. anthracis* distribution ENM models indicated that Mississippi and Louisiana are not areas predicted by ENM where *B. anthracis* is likely to persist (Blackburn 2010; Blackburn et al. 2007; Mullins et al. 2013). However, the predictions presented in this paper are in line with several other studies and cases, one of which found virulence markers for *B. anthracis* in 26% of the soil samples collected and analysis from New Orleans following Hurricane Katrina (Griffin et al. 2009). In addition, there were several reports of historical anthrax outbreaks in livestock (with multiple reports of cattle losses) and wildlife in the Mississippi River Delta between 1954 and 1971 in which calcareous soils in the area were hypothesized to have contributed to the outbreaks (Kellogg et al. 1970; Van Ness 1971; Van Ness and Stein 1956). However, more recent studies did not find the same link with higher concentrations of calcium in soil along Mississippi River Delta (Griffin et al. 2014; USGS 2013) as was seen in the past reported outbreaks. This suggests that Ca content alone cannot explain the occurrence.

Statistical Considerations

This study utilized Maxent modeling, a presence-only approach, to predicting potential locations of potential anthrax outbreaks in wildlife and livestock in the contiguous United States. Maxent can address some of the limitations with presence-only data (Elith et al. 2011). Several advantages to using Maxent compared to GARP have been noted and include the following: Maxent typically has lower omission rates, greater AUC scores, and is better able to discriminate between suitable and unsuitable areas compared to GARP; however, it was not possible to directly compare the performance of the current model to other ENM models because Maxent produces continuous data, while GARP values are discrete (Phillips et al. 2006). Based on the AUC and rates of omission, models from the current study performed fairly well. When using the 10% binary threshold provided by Maxent, an omission rate of 10% would be expected. Omission rates in this study were 11.3%, 11.6%, and 11.69% for the Sample Location, Interpolate, and Normalized datasets, respectively. Furthermore, model performance in this study was evaluated with an additional metric, the TSS rating, which ranged from "very good" to "good" for many of the models (Table 6).

AUC should not be the only measure used to evaluate performance and can be affected by the extent of the geographical area used for the model (Allouche et al. 2006; Lobo et al. 2008). The use of TSS in addition to the AUC, as this study used, could help provide a more accurate description of the model performance. The TSS is not affected by the size of the background dataset or prevalence and combines specificity and sensitivity in order to account for omission and commission errors and random guessing (Allouche et al. 2006), unlike the AUC, which weights the commission and omission error the same (Lobo et al. 2008). The suggestion to include TSS in addition to AUC is supported by data from the current study in which the AUC score for the State Scale Normalized dataset was 0.757, but the TSS score was only 0.359 (poor) (Table 6). Without use of the TSS, this model might have been reported as a very good fit of the data using AUC alone. One explanation for the different in AUC and TSS scores is that the AUC does not account for any transformations of probability predictions used and tends to include performance for all possible areas of the model, even those that may not be of interest (Lobo et al. 2008).

Summary of Strengths and Limitations

Strengths

This study utilized the Maxent model, and background environmental conditions, to identify areas which could potentially support anthrax outbreaks in wildlife and livestock. Like other presence-only models, Maxent only requires the locations a species has been observed. This is preferable to presence-absence data due to the difficulty of verifying absence.

Unlike other commonly used ENM models such as GARP, Maxent includes features that can address some of the limitations with presence-only data. Maxent is preferable to other models due to its logistic output format, which makes model interpretation simple, and its ability to address sample selection bias, use continuous or categorical data, and include a regularizer to prevent overfitting (Hastie et al. 2001; Phillips and Dudik 2008). Compared to GARP, Maxent typically has lower omission rates, has higher AUC scores, is better able to discriminate between suitable and unsuitable areas, and has been shown to successfully model species distributions on a relatively small number of observations (Pearson et al. 2007).

This study identified that the soil variables, used in other anthrax distribution model publications, should be expanded to include analysis of mineral content. The Maxent models in this study were able to map probability at a higher resolution

Table 6 Current study results and results for similar SDM models

Source	Model type	Presence data type	Scale	Area under the curve (AUC)[1]	Rate of omission[2]	Rate of commission[3]	True Skill Statistic (TSS)[4]	TSS interpretation
Current study	Maxent[5]	*B. anthracis rpoB* PCR data with wildlife/livestock outbreak point data	National (United States)	0.692 (0.83)	Not calculated	Not calculated	0.484	Fair
		Sample locations: Wildlife/livestock anthrax outbreaks 2001–2013	State (United States)	0.918 (0.968)	Not calculated	Not calculated	0.82	Very good
			National (United States)	0.926 (0.937)	11.30%	79.62%	0.683	Good
		Interpolated: Wildlife/livestock anthrax outbreaks 2001–2013	State (United States)	0.882 (0.912)	Not calculated	Not calculated	0.558	Good
			National (United States)	0.881 (0.905)	11.60%	63.50%	0.532	Good
		Normalized: Wildlife/livestock anthrax outbreaks 2001–2013	State (United States)	0.757 (0.789)	Not calculated	Not calculated	0.359	Poor
			National (United States)	0.825 (0.867)	11.69%	65.50%	0.538	Good
(Nath and Dere 2016)	Maxent	Livestock anthrax outbreaks 2001–2013	Statewide (Minnesota)	0.978 (0.969)	Not given	Not given	Not calculated	Not applicable
			Local (Minnesota)	0.698 (0.920)	Not given	Not given	Not calculated	Not applicable
(Chikerema et al. 2013)	Maxent	Reported animal and human anthrax outbreaks 1995–2010	National (Zimbabwe)	0.717 (0.774)	Not given	Not given	Not calculated	Not applicable
(Blackburn et al. 2007)	GARP[6]	Wildlife/livestock anthrax outbreaks 1957, 1969, 2000–2005	National (United States)	0.7916	6.8% (23.2%)[7]	41.6% (66.5%)[8]	Not calculated	Not applicable
(Blackburn 2010)	GARP	Wildlife/livestock anthrax outbreaks 1957, 1969, 2000–2005	National (United States)	0.832	4.4%[9]	Calculated but not reported	Not calculated	Not applicable
			National (Mexico)	0.846	0%[10]	Calculated but not reported	Not calculated	Not applicable
(Mullins et al. 2011)	GARP	Livestock anthrax outbreaks 1960–2000	National (Kazakhstan)	0.7188	2.6% (9.9%)[7]	32.7% (58.4%)[8]	Not calculated	Not applicable
		Outbreak isolates for *B. anthracis* A1.a lineage from livestock, humans, and soil	Regional (Southeast Kazakhstan)	0.6964	0.0% (13.1%)[7]	19.18% (66.11%)[8]	Not calculated	Not applicable
(Mullins et al. 2013)[9]	GARP	*B. anthracis* isolates for A1.a and TEA/WNA lineages	National (United States)	0.93	0.0% (6.7%)[7]	8.38% (22.88%)[8]	Not calculated	Not applicable
			National (Italy)	0.84	0.0% (1.4%)[7]	27.86% (50.21%)[8]	Not calculated	Not applicable
			National (Kazakhstan)	0.90	0.0% (0.0%)[7]	19.77% (46.3%)[8]	Not calculated	Not applicable

[1]Area under the curve test– used for validation (training – used for building); [2]sensitivity; [3]specificity; [4]sensitivity + specificity − 1; [5]maxium entropy; [6]Genetic Algorithm for Rule-Set Prediction; [7]total omission (average omission across the ten best models); [8]total commission (average commission across the ten best models); [9]only the native models run by Mullins et al. (2013); [10]total omission are reported in this table. Transferred model evaluation results are not included

overall compared to studies using GARP when the nationwide soil geochemistry data was included.

In addition, the Maxent models used in this study predicted an elevated probability for anthrax outbreaks in wildlife and livestock in regions not previously predicted by others, even though several other studies and cases have found virulence markers for *B. anthracis* and/or historical reports of anthrax outbreaks in livestock have been noted in these areas. The results from this study also indicated regions, which had not reported recent anthrax outbreaks in wildlife or livestock but contained environmental and soil conditions, that could potentially support an animal anthrax outbreak.

Limitations

One of the most significant limitations of species distribution models is the inherent imperfection of observations. When presence data is imperfect, model results may estimate where a species is more likely to be observed, rather than where it occurs (Guillera-Arroita 2017; Guillera-Arroita et al. 2015). Although the sampling effort this study relied on was robust, it was not able to predict true persistence and viability of the bacteria in the contiguous United States, which would require more complete presence data and true absence data.

The PCR results from the NASGLP could not be utilized to determine a distribution for *B. anthracis* across the *United States* due to the low incidence that *B. anthracis rpoB* was identified in the collected samples. Interpretation of these types of data can be confounded by climate, temporal, and spatial variability as well as anthropogenic factors. These same confounding variables could affect the ability to locate *B. anthracis* in environmental samples which have been collected using a random stratified sampling strategy. A more targeted sampling approach which considers the confounding variables might be needed, which could be difficult at a national scale.

The lack of detection of the pathogen in areas formerly known to have frequent animal anthrax outbreaks highlights the need for enhanced sensitivity (e.g., enrichment-PCR) to measure reproducible detection in soil surveillance studies and/or the need for use of a more defined presence variable. The model does not account for non-detects which are true non-detects versus those that were non-detect due to the bacteria being present in concentrations below the detection limit of the analytical methods used to analyze the samples. Conclusions of this study were therefore limited to identifying areas that could potentially promote pathogen survival, rather than being able to concretely identify where the pathogen was located.

Overall, the models tended to be scale-dependent and tended to produce better model performance at a finer scale of the data. Using variables which varied by scale and how the data were treated make it difficult to identify a common set of variables using RFE which fits all model scales and situations. The model might need to focus on soil and environmental conditions over more defined areas such as counties that have experienced anthrax outbreaks in wildlife and livestock to provide a more realistic characterization of areas that could potentially support anthrax outbreaks. It was not possible to directly compare the performance of the current model to other ENM models because of the differences in the type of data used for the models (Maxent produces continuous data, while GARP values are discrete).

A major challenge with using these types of models is trying to balance overfitting and underfitting of the data (Radosavljevic and Anderson 2014). Creating a model to predict events that have not occurred (and may not occur) is complicated and compounded by many factors, leading to the need to use expert knowledge and common sense to further judge the results, a method which is not always reproducible. Due to these compounding factors, these models tend to be misunderstood, misapplied, and misinterpreted. Many past research efforts fell short of fully determining the success of their models. This study confirmed that TSS should be used in addition to the AUC to help provide a more accurate description of the model performance (Allouche et al. 2006; Lobo et al. 2008).

Conclusion

This is one of the first studies to present PCR results from the NASGLP, which found *Bacillus* sp. present in 83% of samples, demonstrating the widespread prevalence of *Bacillus* sp. in soils of the contiguous United States. The PCR results could not be utilized to determine a distribution for *B. anthracis* across the United States due to the low incidence that *B. anthracis rpoB* was identified in the collected samples. In addition, the pathogen wasn't detected in areas formerly known to have frequent animal anthrax outbreaks, including areas of southern Texas or eastern North and South Dakota. This lack of detection highlights the need for enhanced sensitivity (e.g., enrichment PCR) to measure reproducible detection in soil surveillance studies. Other factors that could confound interpretation of these types of data are climate variables, temporal variables, and anthropogenic factors (vaccination programs and the geographic distribution of livestock). These same confounding variables, along with spatial variability, could affect the ability to locate *B. anthracis* in environmental samples which have been collected using a random stratified sampling strategy; a more targeted sampling approach which takes into account the confounding variables might be needed.

This study was able to use Maxent to identify areas which could potentially support anthrax outbreaks in wildlife and livestock based off the geochemical soil constituents and environmental conditions in those locations. Maxent model results from the current study also indicated regions which had not reported recent anthrax outbreaks in wildlife or livestock but contained environmental and soil conditions that could potentially support an animal anthrax outbreak (Michigan and Maine). This work provides an extension to use of econiche modeling to investigate animal anthrax outbreaks in the United States as it utilizes additional soil geochemistry data and have shown that further validation techniques, such as the TSS should be considered. Unfortunately, direct comparison with other models is not possible. With anthrax occurrence data, the environmental and soil conditions in a given location could be used to predict areas that could potentially support an anthrax outbreak in wildlife and livestock, should one occur. However, without more complete presence data, and true absence data, it is not possible to predict true persistence and viability of the bacteria in these regions.

A challenge with these models is trying to balance overfitting, underfitting, and reality. A model that perfectly fits the data can be created, but that model might overfit the fundamental and even realized niche. On the other hand, a model that underfits the data might not predict all areas where the species can survive. Thus, when creating a model to predict events that have not occurred (and may not occur), relying on the available data and statistical inferences from them is complicated and may be compounded by many factors. In this case, expert knowledge and common sense could be applied to further judge the results.

Acknowledgements and Disclaimer USGS collaborated on the analysis of samples for the presence of *Bacillus* and *B. anthracis* with the US Environmental Protection Agency (USEPA) through USEPA's Office of Research and Development under EPA IA# DW 1495774801. Maxent and statistical analysis was completed by completed under the USEPA and USGS IA # DW 1492401101. This joint agency project was supported by the USGS Geochemical Landscape Project and the USGS Environmental Health Mission Area's Contaminate Biology Program. We would like to thank John Lisle (USGS), Sarah Perkins (formerly from USEPA), Charlena Bowling (USEPA), M. Worth Calfee (USEPA), and Paul Lemieux (USEPA) for their help and assistance on this project. This content has been peer and administratively reviewed and has been approved for publication as a joint USGS and USEPA publication. Note that approval does not signify that the contents necessarily reflect the views of the USEPA or the USGS. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not constitute or imply its endorsement, recommendation, or favoring by the US government. The views and opinions expressed herein do not state or reflect those of the US government and shall not be used for advertising or product endorsement purposes.

References

Ahsan, M.M., et al. 2013. Investigation into *Bacillus anthracis* spore in soil and analysis of environmental parameters related to repeated anthrax outbreak in Sirajganj. *Bangladesh Thai Journal of Veterinary Medicine* 43: 449–454.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Anderson, R.P. 2003. Real vs. artefactual absences in species distributions: Tests for *Oryzomys albigularis (Rodentia: Muridae)* in Venezuela. *Journal of Biogeography* 30: 591–605.

Anderson, R.P., M. Gomez-Laverde, and A.T. Peterson. 2002a. Geographical distributions of spiny pocket mice in South America: Insights from predictive models. *Global Ecology and Biogeography* 11: 131–141. https://doi.org/10.1046/j.1466-822X.2002.00275.x.

Anderson, R.P., A.T. Peterson, and M. Gomez-Laverde. 2002b. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* 98: 3–16. https://doi.org/10.1034/j.1600-0706.2002.t01-1-980116.x.

APHIS. 2014. Animal and plant inspection service. U.S. Department of Agriculture. http://www.aphis.usda.gov/wps/portal/aphis/home/. Accessed 8 Aug 2014.

Beyer, W., S. Pocivalsek, and R. Bohm. 1999. Polymerase chain reaction-ELISA to detect *Bacillus anthracis* from soil samples-limitations of present published primers. *Journal of Applied Microbiology* 87: 229–236.

Blackburn, J.K. 2010. Integrating geographic information systems and ecological niche modeling into disease ecology: A case study of *Bacillus anthracis* in the United States and Mexico. In *Emerging and endemic pathogens: Advances in surveillance, detection, and identification*, ed. K.P. O'Connell, E.W. Skowronski, A. Sulakvelidze, and L. Bakanidze, 59–88. Dordrecht: Springer Science.

Blackburn, J.K., K.M. McNyset, A. Curtis, and M.E. Hugh-Jones. 2007. Modeling the geographic distribution of *Bacillus anthracis*, the causative agent of anthrax disease, for the contiguous United States using predictive ecological [corrected] niche modeling. *The American Journal of Tropical Medicine and Hygiene* 77: 1103–1110.

Blackburn, J.K., M. Van Ert, J.C. Mullins, T.L. Hadfield, and M.E. Hugh-Jones. 2014. The necrophagous fly anthrax transmission pathway: empirical and genetic evidence from wildlife epizootics. *Vector Borne and Zoonotic Diseases* 14: 576–583. https://doi.org/10.1089/vbz.2013.1538.

Breed, F. 1932. Anthrax history, diagnosis and control measures. *The Iowa Veterinary* 3: 34–38.

Carnaval, A.C., and C. Moritz. 2008. Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. *Journal of Biogeography* 35: 1187–1201. https://doi.org/10.1111/j.1365-2699.2007.01870.x.

Chikerema, S.M., A. Murwira, G. Matope, and D.M. Pfukenyi. 2013. Spatial modelling of *Bacillus anthracis* ecological niche in Zimbabwe. *Preventive Veterinary Medicine* 111: 25–30.

Coker P.R. 2002. *Bacillus anthracis* spore concentrations at various carcass sites. LSU Doctoral Dissertation, Louisiana State University and Agricultural and Mechanical College. Department of Pathobiological Sciences.

Cordellier, M., and M. Pfenninger. 2009. Inferring the past to predict the future: Climate modelling predictions and phylogeography for the freshwater gastropod *Radix balthica* (Pulmonata, Basommatophora). *Molecular Ecology* 18: 534–544. https://doi.org/10.1111/j.1365-294X.2008.04042.x.

Dey, R., P.S. Hoffman, and I.J. Glomski. 2012. Germination and amplification of anthrax spores by soil-dwelling amoebas. *Applied and En-*

vironmental Microbiology 78: 8075–8081. https://doi.org/10.1128/AEM.02034-12.

Dragon, D.C., and R.P. Rennie. 1995. The ecology of anthrax spores: Tough but not invincible. *The Canadian Veterinary Journal* 36: 295–301.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Elith, J., S.J. Phillips, T. Hastie, M. Dudik, Y.E. Chee, and C.J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17: 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x.

Epp, T., C. Argue, C. Waldner, and O. Berke. 2010. Spatial analysis of an anthrax outbreak in Saskatchewan, 2006. *The Canadian Veterinary Journal* 51: 743–748.

Grabenstein, J.D. 2008. Countering anthrax: Vaccines and immunoglobulins. *Clinical Infectious Diseases* 46: 129–136. https://doi.org/10.1086/523578.

Graham, C.H., and R.J. Hijmans. 2006. A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography* 15: 578–587. https://doi.org/10.1111/j.1466-822x.2006.00257.x.

Griffin, D.W., T. Petrosky, S.A. Morman, and V.A. Luna. 2009. A survey of the occurrence of *Bacillus anthracis* in North American soils over two long-range transects and within post-Katrina New Orleans. *Applied Geochemistry* 24: 1464–1471.

Griffin, D., E.E. Silvestri, C.Y. Bowling, T. Boe, D.B. Smith, and T.L. Nichols. 2014. Anthrax and the geochemistry of soils in the contiguous United States. *Geosciences* 4: 114–127.

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* 40. https://doi.org/10.1111/ecog.02445.

Guillera-Arroita, G., et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* 24: 276–292. https://doi.org/10.1111/geb.12268.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Higgins, C.H. 1916. *Anthrax. Bulletin no. 23*. Ottawa: Health of Animals Branch, Department of Agriculture.

Homer, C., C.Q. Huang, L.M. Yang, B. Wylie, and M. Coan. 2004. Development of a 2001 National Land-Cover Database for the United States. *Photogrammetric Engineering and Remote Sensing* 70: 829–840.

Hornaday, W.T. 1889. *Map illustrating the extermination of the American bison*. Washington: Government Printing Office.

Hugh-Jones, M., and J. Blackburn. 2009. The ecology of *Bacillus anthracis*. *Molecular Aspects of Medicine* 30: 356–367. https://doi.org/10.1016/j.mam.2009.08.003.

Hugh-Jones, M.E., and V. de Vos. 2002. Anthrax and wildlife. *Revue Scientifique et Technique* 21: 359–383.

Hutchinson, G.E. 1957. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415–427.

Kellogg, F.E., A.K. Prestwood, and R.E. Noble. 1970. Anthrax epizootic in white-tailed deer. *Journal of Wildlife Diseases* 6: 226–228.

Kenefic, L.J., et al. 2008. Texas isolates closely related to *Bacillus anthracis* Ames. *Emerging Infectious Diseases* 14: 1494–1496. https://doi.org/10.3201/eid1409.080076.

Kharouba, H.M., A.C. Algar, and J.T. Kerr. 2009. Historically calibrated predictions of butterfly species' range shift using global change as a pseudo-experiment. *Ecology* 90: 2213–2222. https://doi.org/10.1890/08-1304.1.

Ko, K.S., J.M. Kim, J.W. Kim, B.Y. Jung, W. Kim, I.J. Kim, and Y.H. Kook. 2003. Identification of *Bacillus anthracis* by *rpoB* sequence analysis and multiplex PCR. *Journal of Clinical Microbiology* 41: 2908–2914.

Koch, R. 1882. On the anthrax inoculation. In *Essays of Robert Koch*, ed. K. Carters, 97–116. Westport: Greenwood Press, Inc.

Kochi, S.K., G. Schiavo, M. Mock, and C. Montecucco. 1994. Zinc content of the *Bacillus anthracis* lethal factor. *FEMS Microbiology Letters* 124: 343–348.

Kumar, S., and T.J. Stohlgren. 2009. Maxent modeling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola* in New Caledonia. *Journal of Ecology and The Natural Environment* 1: 94–98.

Lamb, J.M., et al. 2008. Phylogeography and predicted distribution of African-Arabian and Malagasy populations of giant mastiff bats, *Otomops* spp. (Chiroptera: Molossidae). *Acta Chiropterologica* 10: 21–40. https://doi.org/10.3161/150811008x331063.

Letant, S.E., et al. 2011. Rapid-viability PCR method for detection of live, virulent *Bacillus anthracis* in environmental samples. *Applied and Environmental Microbiology* 77: 6570–6578. https://doi.org/10.1128/Aem.00623-11.

Lindeque, P.M., and P.C. Turnbull. 1994. Ecology and epidemiology of anthrax in the Etosha National Park, Namibia Onderstepoort. *Journal of Veterinary Research* 61: 71–83.

Lobo, J.M., A. Jimenez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

Luna, V.A., et al. 2006. *Bacillus anthracis* virulent plasmid pX02 genes found in large plasmids of two other *Bacillus* species. *Journal of Clinical Microbiology* 44: 2367–2377. https://doi.org/10.1128/JCM.00154-06.

Manchee, R.J., M.G. Broster, A.J. Stagg, and S.E. Hibbs. 1994. Formaldehyde solution effectively inactivates spores of *Bacillus anthracis* on the Scottish Island of Gruinard. *Applied and Environmental Microbiology* 60: 4167–4171.

Minett, F.C. 1950. Sporulation and viability of *B. anthracis* in relation to environmental temperature and humidity. *Journal of Comparative Pathology* 60: 161–176.

Minett, F.C., and M.R. Dhanda. 1941. Multiplication of *B. anthracis* and *Cl. chauvoei* in soil. *The Indian Journal of Veterinary Science and Animal Husbandry* 11: 308–328.

Monterroso, P., J.C. Brito, P. Ferreras, and P.C. Alves. 2009. Spatial ecology of the European wildcat in a Mediterranean ecosystem: Dealing with small radio-tracking datasets in species conservation. *Journal of Zoology* 279: 27–35. https://doi.org/10.1111/j.1469-7998.2009.00585.x.

Mullins, J., L. Lukhnova, A. Aikimbayev, Y. Pazilov, M. Van Ert, and J.K. Blackburn. 2011. Ecological niche modelling of the *Bacillus anthracis* A1.a sub-lineage in Kazakhstan. *BMC Ecology* 11: 32. https://doi.org/10.1186/1472-6785-11-32.

Mullins, J.C., G. Garofolo, M. Van Ert, A. Fasanella, L. Lukhnova, M.E. Hugh-Jones, and J.K. Blackburn. 2013. Ecological niche modeling of *Bacillus anthracis* on three continents: Evidence for genetic-ecological divergence? *PLoS One* 8: e72451. https://doi.org/10.1371/journal.pone.0072451.

Murray-Smith, C., N.A. Brummitt, A.T. Oliveira-Filho, S. Bachman, J. Moat, E.M.N. Lughadha, and E.J. Lucas. 2009. Plant diversity hotspots in the Atlantic Coastal Forests of Brazil. *Conservation Biology* 23: 151–163. https://doi.org/10.1111/j.1523-1739.2008.01075.x.

Nath, S., and A. Dere. 2016. Soil geochemical parameters influencing the spatial distribution of anthrax in Northwest Minnesota. *USA Applied Geochemistry* 74: 144–156. https://doi.org/10.1016/j.apgeochem.2016.09.004.

Ndiva Mongoh, M., N. Dyer, C. Stoltenow, and M.L. Khaitsa. 2008. A review of management practices for the control of anthrax in animals: The 2005 anthrax epizootic in North Dakota–case study. *Zoonoses and Public Health* 55: 279–290.

NOAA. 2010. Precipitation Data. National Oceanic and Atmospheric Administration, National Centers for Environmental Information.

Available at: ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/products/precipitation/. Last accessed 01/12/18.

———. 2015. National Centers for Environmental Information Temperature Data Search. National Oceanic and Atmospheric Administration, National Climatic Data Center. Available at: https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp#. Last accessed 01/12/18.

Pasteur, L. 1880. Sur l'étiologie du charbon [On the etiology of anthrax]. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences* 91: 86–94.

Pearson, R.G., C.J. Raxworthy, M. Nakamura, and A.T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34: 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x.

Pedregosa, F., et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Peterson, A.T. 2006. Ecologic niche modeling and spatial patterns of disease transmission. *Emerging Infectious Diseases* 12: 1822–1826.

———. 2008. Biogeography of diseases: A framework for analysis. *Naturwissenschaften* 95: 483–491. https://doi.org/10.1007/s00114-008-0352-5.

Peterson, A.T., and K.P. Cohoon. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling* 117: 159–164. https://doi.org/10.1016/S0304-3800(99)00023-X.

Peterson, A.T., and J. Soberon. 2012. Species distribution modeling and ecological niche modelling: Getting the concepts right Natureza & Conservação. *Brazilian Journal of Nature Conservation* 10: 102–107.

Phillips, S.J., and M. Dudik. 2008. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography* 31: 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x.

Phillips, S.J., R.P. Anderson, and R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, S.J., M. Dudík, and R.E. Schapire. 2017. Maxent software for modeling species niches and distributions (Version 3.4.1). https://www.gbif.org/tool/81279/maxent.

Radosavljevic, A., and R.P. Anderson. 2014. Making better Maxent models of species distributions: Complexity, overfitting, and evaluation. *Journal of Biogeography* 41: 629–643. https://doi.org/10.1111/jbi.12227.

Saile, E., and T.M. Koehler. 2006. *Bacillus anthracis* multiplication, persistence, and genetic exchange in the rhizosphere of grass plants. *Applied and Environmental Microbiology* 72: 3168–3174. https://doi.org/10.1128/AEM.72.5.3168-3174.2006.

Schuch, R., and V.A. Fischetti. 2009. The secret life of the Anthrax agent Bacillus anthracis: Bacteriophage-mediated ecological adaptations. *PLoS One* 4. https://doi.org/10.1371/journal.pone.0006532.

Siamudaala, V.M., et al. 2006. Ecology and epidemiology of anthrax in cattle and humans in Zambia. *The Japanese Journal of Veterinary Research* 54: 15–23.

Silva, D.P., B. Vilela, P.J. De Marco, and A. Nemesio. 2014. Using ecological niche models and niche analyses to understand speciation patterns: The case of Sister Neotropical Orchid Bees. *PLoS One* 9: e113246. https://doi.org/10.1371/journal.pone.0113246.

Silvestri, E.E., S.D. Perkins, D. Feldhake, T.L. Nichols, and F.W.I. Schaefer. 2015. Recent literature review of soil processing methods for recovery of *Bacillus anthracis* spores. *Annales de Microbiologie* 65: 1215–1226.

Silvestri, E.E., et al. 2016. Optimization of a sample processing protocol for recovery of *Bacillus anthracis* spores from soil. *Journal of Microbiological Methods* 130: 6–13. https://doi.org/10.1016/j.mimet.2016.08.013.

Smith, K.L., V. DeVos, H. Bryden, L.B. Price, M.E. Hugh-Jones, and P. Keim. 2000. *Bacillus anthracis* diversity in Kruger National Park. *Journal of Clinical Microbiology* 38: 3780–3784.

Smith, D.B. et al. 2005. Major- and trace-element concentrations in soils from two continental-scale transects of the United States and Canada. Department of the Interior. U.S. Geological Survey open file report, 2005-1253. Reston, VA: Department of the Interior. U.S. Geological Survey.

Smith, D.B., L.G. Woodruff, R.M. O'Leary, W.F. Cannon, R.G. Garrett, J.E. Kilburn, and M.B. Goldhaber. 2009. Pilot studies for the North American Soil Geochemical Landscapes Project - Site selection, sampling protocols, analytical methods, and quality control protocols. *Applied Geochemistry* 24: 1357–1368. https://doi.org/10.1016/j.apgeochem.2009.04.008.

Smith, D.B., W.F. Cannon, and L.G. Woodruff. 2011. A national-scale geochemical and mineralogical survey of soils of the conterminous United States. *Applied Geochemistry* 26: S250–S255.

Smith, D.B., W.F. Cannon, L.G. Woodruff, F.M. Rivera, A.N. Rencz, and R.G. Garrett. 2012. History and progress of the North American Soil Geochemical Landscapes Project, 2001-2010. *Earth Science Frontiers* 19: 19–32.

Soil Survey Staff. 2017. Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at the following link: https://websoilsurvey.sc.egov.usda.gov/. Last accessed 12 Jan 2018.

Stein, C.D. 1945. The history and distribution of anthrax in livestock in the United States. *Veterinary Medicine* 40: 340–349.

———. 1950. Anthrax in livestock during 1949 and incidence of the disease from 1945 to 1949. *Veterinary Medicine* 45: 205–208.

Stein, C.D., and B.G. Van Ness. 1955. A ten year survey of anthrax in livestock with special reference to outbreaks in 1954. *Veterinary Medicine* 50: 579–588.

Stevens, D.L.J., and A.R. Olsen. 1999. Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological and Environmental Statistics* 4: 415–428.

———. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14: 593–610. https://doi.org/10.1002/env.606.

———. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99: 262–278. https://doi.org/10.1198/016214504000000250.

Teshale, E.H., et al. 2002. Environmental sampling for spores of *Bacillus anthracis*. *Emerging Infectious Diseases* 8: 1083–1087. https://doi.org/10.3201/eid0810.020398.

Tinoco, B.A., P.X. Astudillo, S.C. Latta, and C.H. Graham. 2009. Distribution, ecology and conservation of an endangered Andean hummingbird: The Violet-throated Metaltail (*Metallura baroni*). *Bird Conservation International* 19: 63–76. https://doi.org/10.1017/S0959270908007703.

Titball, R.W., P.C. Turnbull, and R.A. Hutson. 1991. The monitoring and detection of *Bacillus anthracis* in the environment. *Society for Applied Bacteriology Symposium Series* 20: 9S–18S.

Tittensor, D.P., et al. 2009. Predicting global habitat suitability for stony corals on seamounts. *Journal of Biogeography* 36: 1111–1128. https://doi.org/10.1111/j.1365-2699.2008.02062.x.

Tognelli, M.F., S.A. Roig-Junent, A.E. Marvaldi, G.E. Flores, and J.M. Lobo. 2009. An evaluation of methods for modelling distribution of Patagonian insects. *Revista Chilena de Historia Natural* 82: 347–360.

Turell, M.J., and G.B. Knudson. 1987. Mechanical transmission of *Bacillus anthracis* by stable flies (*Stomoxys calcitrans*) and mosquitoes (*Aedes aegypti* and *Aedes taeniorhynchus*). *Infection and Immunity* 55: 1859–1861.

Turnbull, P.C.B., J.A. Carman, P.M. Lindeque, F. Joubert, O.J.B. Hübschle, and G.H. Snoeyenbos. 1989. Further progress in understanding anthrax in the Etosha National Park. *Madoqua* 16: 93–104.

Turnbull, P.C., P.M. Lindeque, J. Le Roux, A.M. Bennett, and S.R. Parks. 1998. Airborne movement of anthrax spores from carcass sites in the Etosha National Park, Namibia. *Journal of Applied Microbiology* 84: 667–676.

Turner, A.J., J.W. Galvin, R.J. Rubira, R.J. Condron, and T. Bradley. 1999a. Experiences with vaccination and epidemiological investigations on an anthrax outbreak in Australia in 1997. *Journal of Applied Microbiology* 87: 294–297.

Turner, A.J., J.W. Galvin, R.J. Rubira, and G.T. Miller. 1999b. Anthrax explodes in an Australian summer. *Journal of Applied Microbiology* 87: 196–199.

USDA. 2006. Epizootiology and Ecology of Anthrax.

———. 2014. *2012 census of agriculture*. U.S. Department of Agriculture, National Agricultural Statistics Service. Available at: https://www.nass.usda.gov/. Last accessed 12 Jan 2018.

USEPA. 2012. *Protocol for detection of Bacillus anthracis in environmental samples*. Cincinnati: U.S. Environmental Protection Agency. EPA/600/R12/577.

———. 2014. *Literature review on mechanisms that affect persistence on Bacillus anthracis in soils*. Cincinnati: U.S. Environmental Protection Agency, EPA 600/R-14/216.

———. 2015. *Distinguishing intentional releases from natural occurrences and unintentional releases of Bacillus anthracis: Literature search and analysis*. Cincinnati: U.S. Environmental Protection Agency. EPA/600/R-15/066.

USEPA and USGS. 2017. *Processing protocol for soil samples potentially contaminated with Bacillus anthracis spores*. Cincinnati: U.S. Environmental Protection Agency. EPA/600/R17/028.

USGS. 2012. *National elevation dataset*. U.S. Geological Survey. Available at: https://catalog.data.gov/dataset/100-meter-resolution-elevation-of-the-conterminous-united-states-direct-download. Last accessed 12 Jan 2018.

———. 2013. *Geochemical and mineralogical data for soils of the conterminous United States*, Data series 801. Reston: U.S. Geological Survey. Available at: http:pubs.usgs.gov/ds/801/pdf/ds801.pdf Last accessed 6 Mar 2018.

Van Ness, G.B. 1959. Soil relationship in Oklahoma-Kansas anthrax outbreak of 1957. *Journal of Soil and Water Conservation* 14: 70–71.

———. 1967. Geologic implications of anthrax. *The Geological Society of America Special Paper* 90: 61–64.

———. 1971. Ecology of anthrax. *Science* 172: 1303–1307.

Van Ness, G.B., and C.D. Stein. 1956. Soils of the United States favorable for anthrax. *Journal of the American Veterinary Medical Association* 128: 7–12.

Verbruggen, H., et al. 2009. Macroecology meets macroevolution: Evolutionary niche dynamics in the seaweed Halimeda. *Global Ecology and Biogeography* 18: 393–405. https://doi.org/10.1111/j.1466-8238.2009.00463.x.

Ward, D.F. 2007. Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions* 9: 723–735. https://doi.org/10.1007/s10530-006-9072-y.

Weinberg, E.D. 1987. The influence of soil on infectious disease. *Experientia* 43: 81–87. https://doi.org/10.1007/Bf01940358.

West, A.W., and H.D. Burges. 1985. Persistence of *Bacillus thuringiensis* and *Bacillus cereus* in soil supplemented with grass or manure. *Plant and Soil* 83: 389–398.

Williams, J.N., C.W. Seo, J. Thorne, J.K. Nelson, S. Erwin, J.M. O'Brien, and M.W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. *Diversity and Distributions* 15: 565–576. https://doi.org/10.1111/j.1472-4642.2009.00567.x.

Wollan, A.K., V. Bakkestuen, H. Kauserud, G. Gulden, and R. Halvorsen. 2008. Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography* 35: 2298–2310. https://doi.org/10.1111/j.1365-2699.2008.01965.x.

Wright, G.G., L.H. Angelety, and B. Swanson. 1970. Studies on immunity in anthrax XII. Requirement for phosphate for elaboration of protective antigen and it's partial replacement by charcoal. *Infection and Immunity* 2: 772–777.

Yates, C.J., A. McNeill, J. Elith, and G.F. Midgley. 2010. Assessing the impacts of climate change and land transformation on Banksia in the South West Australian Floristic Region. *Diversity and Distributions* 16: 187–201. https://doi.org/10.1111/j.1472-4642.2009.00623.x.

Yesson, C., and A. Culham. 2006. A phyloclimatic study of Cyclamen. *BMC Evolutionary Biology* 6: 72. https://doi.org/10.1186/1471-2148-6-72.

Young, B.E., I. Franke, P.A. Hernandez, S.K. Herzog, L. Paniagua, C. Tovar, and T. Valqui. 2009. Using spatial models to predict areas of endemism and gaps in the protection of Andean Slope Birds. *Auk* 126: 554–565. https://doi.org/10.1525/auk.2009.08155.

Zhang, J., et al. 2013. An adenovirus-vectored nasal vaccine confers rapid and sustained protection against anthrax in a single-dose regimen. *Clinical and Vaccine Immunology* 20: 1–8. https://doi.org/10.1128/CVI.00280-12.

A Probabilistic Approach to Assess the Risk of Groundwater Quality Degradation

Giuseppe Passarella, Rita Masciale, Sabino Maggi, Michele Vurro, and Annamaria Castrignanò

Introduction

The progressive reduction of natural water resources availability, due to the increasing demand for water in all sectors and the ongoing climate changes, has the effect of the qualitative and quantitative deterioration of this resource and launches a severe alarm aimed to protect and preserve such natural resources.

When considering Mediterranean areas, where predicted scenarios of climate change indicate trends to a warmer and arid climate, problems of scarcity and degradation of the quality of natural water resources become more severe. In such areas, the reduction in the availability of water resources significantly influences sustainable growth. Human activities, in these areas, often produce threats for groundwater safety due to direct or incidental water overexploitation and pollution; the former is particularly critical in nearby coastal areas, where it causes the groundwater hydraulic head lowering and seawater intrusion (Zaccaria et al. 2016). Furthermore, the almost total absence of perpetual surface watercourses, which characterizes the Mediterranean environments, forces water resource managers to draw to the more precious reservoir of the groundwater to satisfy water needs. Groundwater is the major source of worldwide freshwater supply, which is currently used to meet nearly half of the drinking water needs (Machiwal et al. 2018) and around 43% of the irrigation demand (Siebert et al. 2010). Considering what said, quantitative saving and qualitative safeguarding of water resources must be the main objective of any effective environmental policy that, in turn, often implies the assessment of the risk of qualitative-quantitative groundwater degradation.

The word "risk" has two distinct meanings (Burton and Whyte 1980; Rue et al. 1999). In some contexts, it is considered as a threat that is exposure to mischance or peril. In other contexts, the risk is interpreted, more narrowly, as the possibility or chance of suffering an adverse consequence or of encountering some loss (Duckett 1983; Mishra and Sarkar 2017).

Generally, according to Varnes (1984), *the risk* is defined as functionally related to two independent variables: (i) the vulnerability V of the natural system, that is, the degree of the intrinsic weakness of the considered system, and (ii) the hazard H associated with a specific (natural or human) event (Barca and Passarella 2008), that is, the possibility that a potentially detrimental event of given characteristics occurs in a given area, for a given time period (IPCC 2012). When assessing the hazard associated with an event, it is necessary to evaluate the temporal trend, frequency, and spatial extent of the past events and determine the severity of the effects produced by them (deterministic approach). On the other hand, the elements contributing to the vulnerability of a natural system are sensitivity, adaptive capacity (resilience and renewability), and weaknesses. A system is more or less sensitive to external events, depending on how it changes in response to it. The adaptive capacity measures the degree to which a natural system can adjust in response to changing external conditions. Depending on whether the adaptation occurs autonomously or it requires some sort of external intervention, it is usually named resilience or renewability, respectively. Finally, the presence of weak points in a natural system can produce uncontrolled harmful effects in even very large areas.

G. Passarella (✉) · R. Masciale · M. Vurro · A. Castrignanò
CNR-IRSA Water Research Institute, Bari, Italy
e-mail: giuseppe.passarella@ba.irsa.cnr.it; rita.masciale@ba.irsa.cnr.it; michele.vurro@ba.irsa.cnr.it; annamaria.castrignano@ba.irsa.cnr.it

S. Maggi
CNR-IIA Institute of Atmospheric Pollution, c/o Interateneo Physics Department, Bari, Italy
e-mail: sabino.maggi@iia.cnr.it

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_20

Particularly, environmental risk is defined as the "actual or potential threat of adverse effects on living organisms and the environment by effluents, emissions, wastes, resource depletion, etc., arising out of an organization's activities" (Power and McCarty 1998). Environmental exposures, whether physical, chemical, or biological, can induce a harmful response and may affect soil, water, air, natural resources or entire ecosystems, as well as the plants and animals – including humans – and the surroundings where they live (Mishra and Sarkar 2017). The assessment of the environmental risk due to human activities is very important in order to plan and start actions aimed to reduce impacts, recreating the co-evolutionary process between human and natural components of the environment.

Concerning groundwater, this reservoir, even though somewhat protected by soil layers and vadose zone, is strongly subjected to various quantitative and qualitative risks due to human activities. The overexploitation of such a resource often heavily reduces freshwater availability and worsens its qualitative status. In fact, it increases the risk of aquifer's salinization, particularly in coastal areas, where freshwater withdrawal from wells causes an increase of natural seawater intrusion. Furthermore, groundwater is often subjected to point-type pollution, usually due to localized human activities, such as industrial sites, landfills, and wastewater treatment plants, to which chemical- and microbiological-type environmental risks are usually related. Finally, diffuse pollution of groundwater, such as the one due to nutrients spreading over the ground surface in agriculture, is one of the main groundwater degradation risks because of its intrinsic characteristic of affecting wide land areas.

In this chapter, a typical environmental risk is addressed, since it concerns the expected qualitative degradation of groundwater due to the possible increase of dissolved nitrates' concentration above critical thresholds. In fact, one of the most worrying causes of diffuse pollution of groundwater is the infiltration, through the unsaturated layers of the subsoil, of nutrients and fertilizers widely used in intensive agriculture (Goodchild 1998; Liu et al. 2005; Almastri 2007; Menció et al. 2016). Generally, nitrate is a low-toxic compound, but it becomes dangerous to human health when it reduces to nitrite. In fact, ingested nitrites from polluted drinking waters can induce *blue baby syndrome*, by blocking the oxygen-carrying capacity of haemoglobin, or also have a potential role in developing cancers of the digestive tract through their contribution to the formation of nitrosamines (Camargo and Alonso 2006). Although nitrate in groundwater can derive from many sources, such as industrial, municipal, residential, and agricultural sources, the largest cause of this kind of groundwater pollution, on a global scale, is the use of chemical fertilizers in agriculture (Haller et al. 2013). The contamination occurs by nitrogen leaching (i.e., the downward transport of nitrate by water percolating) from the soil to aquifers, when the amount of this nutrient contained in fertilizers greatly exceeds the nitrogen crop requirements (Libutti and Monteleone 2017).

The need to control and mitigate the negative impact of nitrogen leaching from agricultural activities on water quality has prompted, in 1991, the European Union to adopt the Nitrates Directive 91/676/EEC (EU 1991). The Directive sets the acceptable threshold of nitrate concentration in groundwater at 50 mg/l and involves the definition of "nitrate-vulnerable zones" (NVZs) as large areas of land draining into waters, which exceed or are at risk of exceeding the threshold. The Directive establishes mandatory action programs for these areas by means of the adoption of a code for "good agricultural practices" and providing training and information for farmers. Although the Nitrates Directive was issued and implemented long ago, the problem of nitrate pollution in the aquifers are still persistent throughout Europe (EEA 2012) with many EU Member States whose groundwater quality monitoring stations exhibit NO3 concentration over 50 mg/l (EU 2000, 2006). Therefore, different methodologies have been implemented to support groundwater protection by nitrate pollution and its management (Kronvang et al. 2009; Bouraoui et al. 2009; Barca et al. 2015).

The scope of this work has then been to investigate ways of fusing information of different types and to implement a conditional stochastic simulation algorithm for risk assessment of groundwater quality degradation. The proposed method has been applied to the aquifer of the "Tavoliere di Puglia" located below the homonym valley in the northern part of the Apulia Region (south Italy).

Materials and Methods

Methodological Framework

Considering groundwater systems pollution, vulnerability is rather easy to estimate because their transport characteristics do not change appreciably with time. However, the hazard is difficult to quantify since a number of time-dependent processes, involving several non-homogeneous variables, affect it. By its nature, this issue requires a probabilistic approach.

In fact, it allows the risk of qualitative groundwater degradation to be assessed directly, overcoming the need of evaluating its components, separately and represents a cost-effective and fast alternative to the deterministic methodologies (Burton and Whyte 1980; Duckett 1983). This kind of approach to risk analysis allows using computational tools, based on setting critical thresholds of pollutant concentration according to given standards of groundwater quality (Passarella et al. 2002). In this contest, risk assessment is based on the estimation of variables, which are subject to extreme uncertainty. This uncertainty depends on both the intrinsic

nature of the variables and the cost of obtaining information about them. Many chemical and geotechnical variables (major constituents, porosity, permeability, transmissivity, etc.), which may affect groundwater quality, can be assessed and quantified only based on sparse sampling and punctual field tests. This in turn requires spatial modelling of those variables that are deemed affecting groundwater quality.

Combinations of all variables at hand interpreted as spatial random variables, describe the possible "states of nature." Providing multiple stochastic realizations of spatial variables can form the basis of quantitative risk assessment. Treating these realizations as possible realities, risk assessment consists essentially in observing the frequency (probability) with which specified criteria are exceeded or fail to be met (Dowd and Pardo-Igúzquiza 2002).

Geostatistics is a set of probabilistic and statistical tools suitable to characterize and estimate attributes distributed in space (Castrignanò et al. 2000a). The complex spatial and temporal variability observed in the groundwater attributes, the multiplicity of the factors involved in the natural phenomena, and the limited understanding of their complex interactions are the main reasons for the shift from strictly deterministic modelling to a more statistical and probabilistic approach which assesses the uncertainty of prediction (Castrignanò et al. 2000b).

In the Bayesian formalism, any early existing information about the considered variable is called the *a priori* distribution of the variable. Any additional information, which comes from taking into account the proximity of the observations through geostatistics, produces a change in the data distribution, called a posteriori distribution, and affects the determination of the variable's uncertainty.

In the proposed case, average values of nitrate concentration measured in samples collected during eight monitoring campaigns carried out from 2007 to 2011, within the regional groundwater monitoring network, have been used as quantitative (measurements) a priori information, while hydrogeological and land use of the considered area have been used as qualitative (map) a priori information.

The proposed approach of risk assessment is essentially based on this updating process of an a priori distribution into an a posteriori distribution and will be discussed in detail hereinafter.

The main goal of a probabilistic approach consists in assessing uncertainty related to available information, which requires the estimation of conditional distributions over the domain of interest (Castrignanò et al. 2007). Stochastic conditional simulation is an effective technique in this respect since it generates multiple realizations of the random variable over the domain of interest, which allows both a direct vision of spatial and temporal uncertainty and an assessment of joint conditional distributions (Castrignanò and Buttafuoco 2004). Such simulations are conditional, because each realization honors the available data, and stochastic because the estimated spatial statistics are reproduced. The critical step of this approach is the estimation of conditional distributions, which can be realized in various ways since different stochastic simulation algorithms exist: Gaussian-related algorithms, which are more suited to continuous variables, and non-parametric indicator algorithms that are better suited to categorical variables. Given the objectives of this study, the indicator approach is more appropriate, even because it accounts for additional qualitative (soft) information. The conditional distributions at each node of a regular grid covering the domain can be estimated sequentially by indicator kriging, which is a non-parametric technique, not requiring any assumption about the type of data distribution function, and is quite flexible to any experimental condition.

Differently, estimation procedures, such as kriging or splines, provide a unique image, which is optimal for some a priori optimization criterion (minimum error variance or minimum curvature) but does not reproduce spatial statistics (histogram and variogram). Compared to stochastic conditional simulation, such estimated maps look more smoothed with the elimination of the minima and maxima observed. They are therefore unsuitable for risk analysis where the extreme classes are generally of greatest interest.

Since geostatistical techniques are data-driven, i.e., they rely on actual observations, their applications are severely constrained by the lack of data. In hydrogeological studies, direct measurements of groundwater quality attributes are usually quite expensive and then rare. However, an appreciable amount of more qualitative though imprecise indirect information related to groundwater quality, as geological, pedological, or geophysical maps, exists which could be effectively correlated with the actual attributes of interest. Therefore, when direct information is sparse, poorer but more extensively accessible information should be used, in order to improve the prediction accuracy of the spatial attributes deemed influencing groundwater quality.

Probabilistic Approach

To implement the probabilistic approach, the following steps are needed:

1. Record and analyze the existing sets of both quantitative (measurements) and qualitative (map units) information.
2. Code such information for quantitative use. The coding has to be flexible enough so that the various (both direct and indirect) types of information can be jointly processed. Moreover, it must be consistent with the probabilistic basic principles and techniques of geostatistics.
3. Define an algorithm for stochastic non-parametric simulation which provides multiple realizations of risk categories conditioned to the various types of information.

4. Post-process the simulations so as to calculate synthetic indicators of the local risk of groundwater quality degradation and assess the level of uncertainty using the concept of entropy.

Each step will be individually described in detail.

Step 1

The data will be distinguished in quantitative and qualitative data. A quantitative datum corresponds to a precise measurement of the attribute(s) (nitrate) of groundwater quality at some well of known location. The uncertainty attached to this value is assumed negligible and will not be considered in the following analyses. Using the terminology of Journel (1986), such data will be referred to as *hard data*. All the other types of data, represented by maps (raster images), will be considered qualitative and called *soft data* by opposition to hard data (Journel 1986).

Step 2

The variable to be simulated is categorical and represents the different risk levels (categories) of groundwater quality degradation. Each class corresponds to a particular interval of the attribute (nitrate) selected for characterizing groundwater quality. In particular, the thresholds of 10 mg/L and 50 mg/L have been used to define the three risk level categories. The former value refers to a guide value for good groundwater quality, while the latter indicates the maximum allowable concentration of nitrates in groundwater for civil uses.

For both hard and soft data, the indicator formalism is used that was introduced in the field of spatial statistics by Switzer (1977) and later extended and developed by Journel (1983, 1986).

Consider the K mutually exclusive categories s_k $k = 1,\ldots,K$ of the categorical variable, risk. This list of categories is intended to be exhaustive, which means that any location u belongs to one and only one of the K categories. Let $i_k(u)$ be the indicator associated with the class s_k, which takes the value 1 if $u \in s_k$ and 0 if otherwise. The following relations express mutual exclusion and exhaustivity:

$$i_k(u)i'_k(u) = 0, \quad \forall k \neq k' \tag{1}$$

$$\sum_{k=1}^{K} i_k(u) = 1 \tag{2}$$

The auxiliary information (soft data) is coded as a set of prior distributions (relative frequencies) of each category of risk within each map unit. In order to synthesize the multiple auxiliary information, one combined map was obtained by overlapping the maps of each auxiliary variable.

Step 3

The stochastic simulation approach is based on a conditional sequential indicator simulation procedure that enables to perform multiple simulations of a categorical variable. The simulations are constrained to adhere to all the hard information (conditional simulation), and the method does not correspond to a particular distribution model (model-free method). The procedure consists in evaluating the conditional probability of each category s_k at each location u as the conditional expectation of the corresponding indicator $i_k(u)$, according to the following relation (Journel 1983):

$$\text{Prob}\{I(u) = 1|(n)\} = E\{I(u)|(n)\} \tag{3}$$

where (n) stands for conditional information.

Direct kriging of the indicator variable $i_k(u)$ provides an estimate for the probability that s_k prevails at location u. Using simple indicator kriging (SIK) (Goovaerts 1997) under the assumption of second-order stationarity, the probability of the category s_k is estimated by the following relation:

$$\text{Prob}^*\{I_k(u) = 1|(n)\} = p_k + \sum_{\alpha=1}^{n} \lambda_\alpha [I_k(u_\alpha) - p_k] \tag{4}$$

where $p_k = E\{I_k(u)\} \in [0, 1]$ is the prior marginal frequency of category s_k, calculated as a proportion of data of type s_k from both hard and soft data, as described in step 2. The weights λ_α are calculated by the SIK system for each category or better by SI cokriging system taking into account the spatial correlations between the various categories (Wackernagel 1996). However, the latter requires a multivariate consistent model of both direct and cross-indicator variograms of all categories s_k. Therefore, for the sake of simplicity, the model is built starting from the univariate model fitted to the experimental variogram of the most frequent category s_k and then tuning the sills of the other variograms according to the proportion of each category. Since it is assumed that the average proportions p_k of the K categories vary locally on the basis of the local auxiliary information, simple indicator kriging with varying local means (local proportions p_k's) (Goovaerts 1997) is used. The approach consists in re-estimating these proportions from the indicator data available in the neighborhood of location u. As regards the calculation of these average proportions, taking into account all prior information from both hard and soft data, see step 2. The estimates may not necessarily lie between 0 and 1; therefore, they are truncated

to positive values and normalized a posteriori. The next step requires defining any ordering of the K categories so as to build a cumulative density function (cdf) of the probability interval [0, 1] with the following K intervals (Deutsch and Journel 1998):

$$\left[0,\, p_1^*(\cdot)\right],\left[p_1^*(\cdot)\,,\, p_2^*(\cdot)+p_1^*(\cdot)\,\right],\ldots,\left[\left(1-\sum_{k=1}^{K-1}p_k^*(\cdot)\right),1\right] \tag{5}$$

A random number p uniformly distributed in the interval [0, 1] is then drawn, and the interval in which p falls determines the simulated category at location u. After updating all K indicator data with this new simulated value, the procedure proceeds to the next location u' along the random path until the output grid is completely filled with the simulated category. The arbitrary ordering of the K probabilities $p_k^*(\cdot)$ does not affect which category is drawn nor the spatial distribution of categories (Deutsch and Journel 1998; Alabert and Massonnat 1990). When the neighborhood search around a target node contains no data (neither hard data nor already simulated nodes), the simulation becomes non-conditional, and the simulated category is drawn from the theoretical proportions previously entered as prior information.

Completed the path, the next simulation will be generated repeating all the procedures described in step 3 but with its own random path so as to avoid systematic errors in providing the different realizations (Alabert 1987).

Step 4

This step is designed to perform statistical calculations on the results of stochastic indicator simulations corresponding to the risk category simulations. Running multiple simulations, at each node of the output grid superimposed on the study area, several risk categories are generally generated, one for each realization (simulation). Using these multiple realizations, at each node, the following numerical quantities can be calculated:

1. The probability of each category, simply by counting how many times each category occurs over the total number of realizations.
2. The most probable category, i.e., the one associated with the greatest probability.
3. The least probable category, i.e., the one associated with the least probability.
4. The corrected most probable category after Soares correction (Soares 1992, 1998), which tries to compensate for the failures of sequential indicator simulation to reproduce the global proportion of each category. The problem occurs when areas, where the categories of low proportion prevail, are under- or over-sampled by the random path at the beginning of the simulation procedure. In this case, the simulated data tend to have biased proportions in all categories especially the ones with low proportions, from the beginning of the simulation. The idea of the algorithm is to correct the local probabilities by adding for each category the deviation between the global prior probability and the corresponding global proportion obtained at each realization.
5. Local standardized entropy (Journel and Deutsch 1993) at each location u, to assess the uncertainty attached to the categorical risk variable discretized in K categories with probabilities p_k, $k = 1, \ldots, K$, according to the following formula:

$$H(u) = -\frac{\sum_{k=1}^{K}\left[\ln p_k(u)\right]p_k(u)}{\ln K} \tag{6}$$

where $p_k(u)$ is the probability of each category k of the location u previously determined at point 1. The standardization of local entropy to the interval [0, 1] is obtained by dividing the common definition of entropy by its upper bound, $\ln K$, corresponding to the uniform distribution with each category k characterized by the same probability $p_k = K^{-1}$. The above approach was implemented for the case study using 3 categories for the risk variable, defined at step 2, and 1000 realizations deemed sufficient to stabilize the local probability of occurrence for each category.

Case Study

Study Area

The study area belongs to the "Tavoliere di Puglia" (hereinafter simply referred to as Tavoliere), the largest alluvial plain of Southern Italy which extends for about 2830 km^2 in the Northern Apulia Region. A wide and shallow aquifer is hosted in the sediment of the alluvial plain (shallow porous aquifer of Tavoliere) (Fig. 1). Tavoliere is characterized by an intensive agricultural activity, mostly consisting of durum wheat production, which requires large amounts of nitrate fertilizer, and vineyard, olive, and fruit trees, which are also irrigated. Not by chance, most of the nitrate-vulnerable zones of agricultural origin in the Apulia fall in this area. In the last decades, the collected data, coming from the groundwater monitoring networks of the Apulia Region, shows the high level of pollution from nitrates often exceeding the standard values in most of the area, particularly following the periods of mineral fertilizer application. Then, in this area, intensive agricultural practices, based on the non-rational use of chemicals, represent the critical event affecting the water quality standard. The regional groundwater monitoring network of

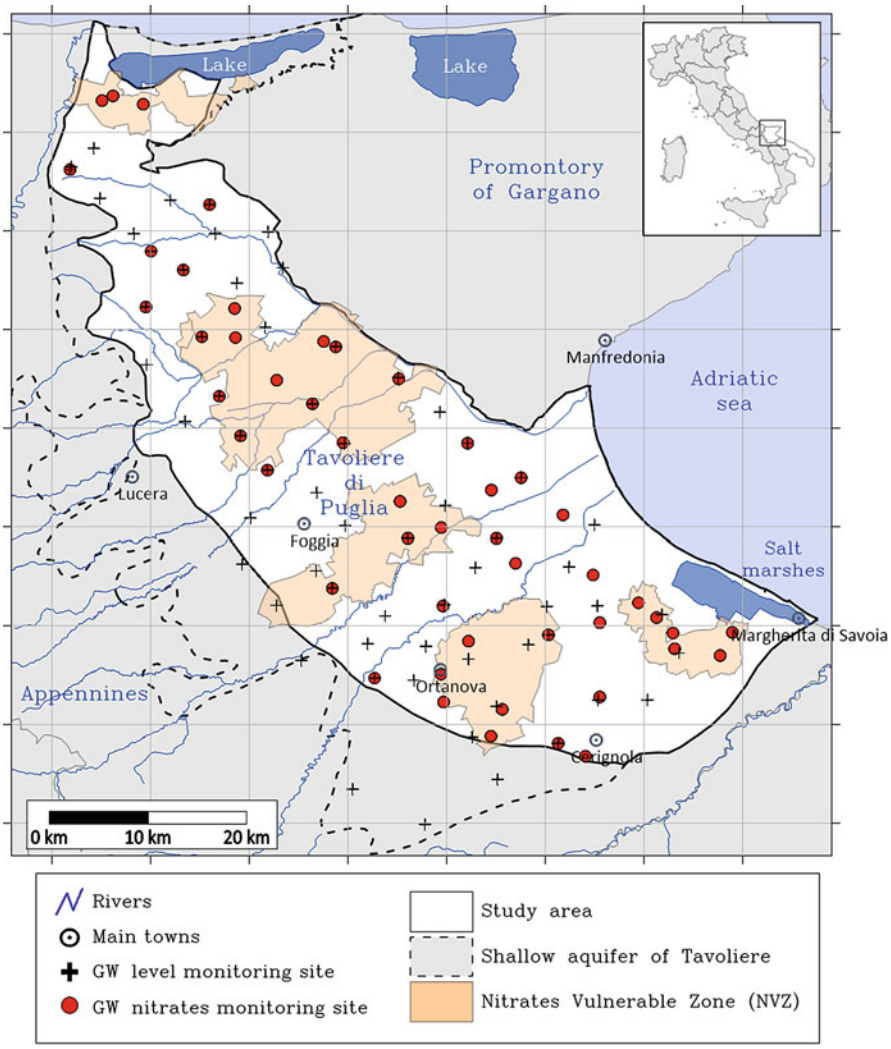


Fig. 1 Study area, groundwater (GW) monitoring networks, and nitrate-vulnerable zones (NVZ)

the shallow aquifer of Tavoliere consists of 49 quality monitoring wells and 63 quantity monitoring wells. The quality monitoring wells are all located in the downstream part of the aquifer on its eastern side where agriculture and anthropic activities are more widespread. No wells for water quality monitoring exist in the western part of the plain where the piezometric level increases, while aquifer thickness and man activities decrease. In this view, the western boundary of the study area has been set at the average piezometric level of 100 m above sea level, as reported by the regional Water Protection Plan (Apulia Region 2009).

Geology

From the geological point of view, Tavoliere represents the Northern sector of the Bradanic Foredeep, bounded to the West by the Apennine Chain and to the East by the Apulian Foreland, locally represented by Gargano Promontory. Hundreds of meters thick, younger deposits cover the Cretaceous calcareous substratum belonging to Apulian Foreland, dislocated by graben-type structures toward the Apennines.

The older units (Pliocene–early Pleistocene) consist of shallow-marine carbonate deposits (Calcarenite di Gravina Fm) passing upward to thick silty-clayey layer deposits of the argille subappennine Fm. A regional uplift phase of Bradanic Foredeep, combined with the glacioeustatic sea level changes which took place during the Quaternary period, determined the sedimentation of terraced deposits, consisting of both continental and marine synthems grouped in Tavoliere supersynthem (Gallicchio et al. 2014) (Fig. 2).

Quaternary deposits, which widely outcrop all over the plain, are characterized by a variable thickness, generally increasing from the Apennines toward the eastern edge of the plain. The grain size and texture of these deposits also vary all over the area. In fact, the western sector of the plain is characterized by debris flow and coarse-grained sediments deposited in alluvial fun settings, whereas the eastern one by gravel and sand-gravel deposited in a braided alluvial plain setting. Changes in grain size and texture are also observed

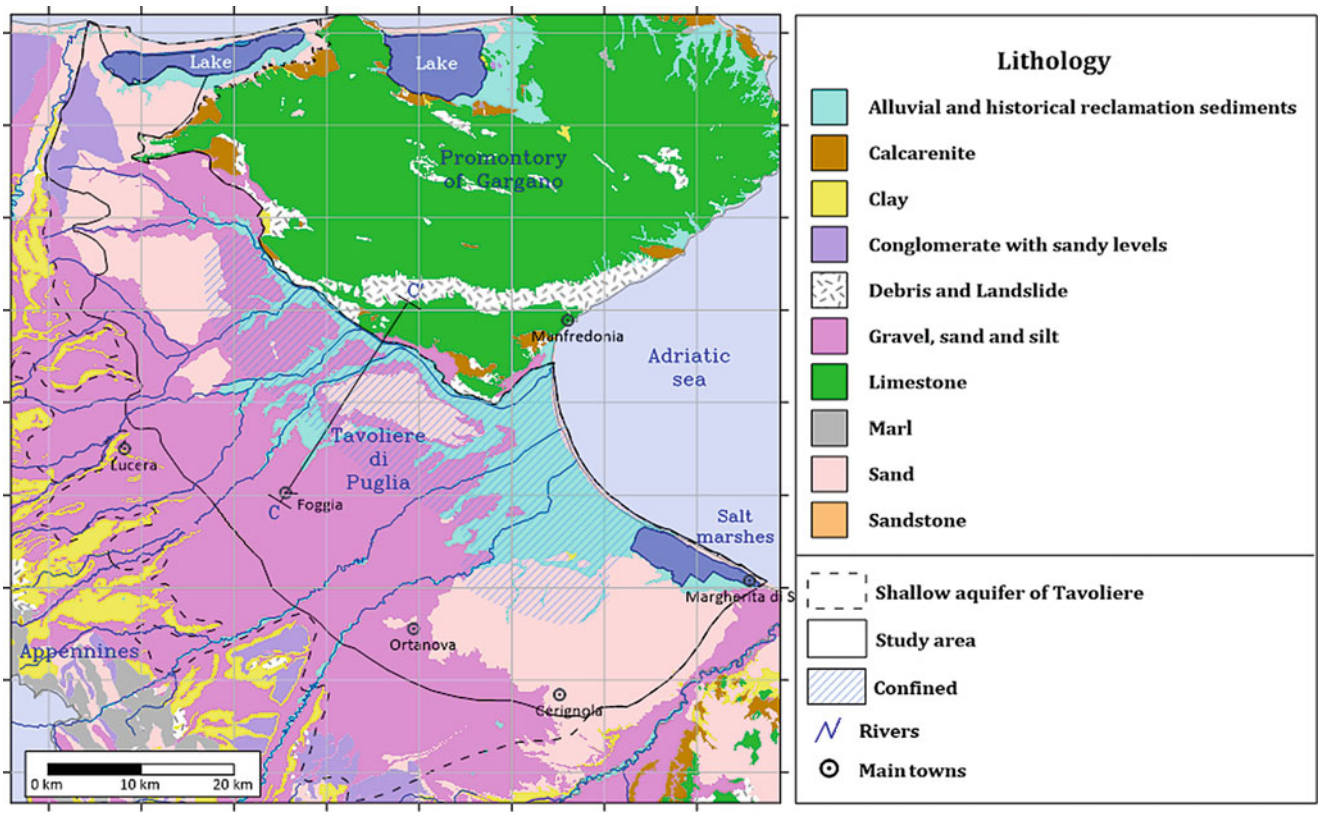


Fig. 2 Lithological characteristics of the study area (Apulia Region 2009)

among deposits of different ages, being the older alluvial deposits coarser and poorly sorted compared to younger ones.

Hydrogeology

According to the above outlined geological settings, three distinct aquifers can be detected from the bottom upward (Fig. 3): (a) the deep karst aquifer located in the cretaceous calcareous substratum, which can be found down to 300–600 m deep; (b) the deep porous aquifer located in the sandy portions within the argille subappennine Fm; and (c) the shallow porous aquifer in the quaternary deposits.

No vertical hydraulic connection exists among the three overlapping aquifers so that important differences are shown both in the flow patterns and in the geochemical features of groundwater (Maggiore et al. 1996). The shallow aquifer, defined as a significant groundwater body by the Apulian Water Protection Plan, is the most exploited for agricultural uses, representing the main source of water supply in the area. Figure 2 shows the hydrogeological boundaries of Tavoliere shallow aquifer and the portion considered in this study. Due to its geological settings, it is a multi-layered aquifer, consisting of a complex alternation of alluvial gravel and sand lenses interbedded to sandy loam and silty clay loam sediments. Nevertheless, due to limited lateral continuity of the confining beds, the different water-bearing layers are hydraulically interconnected to each other so that a single complex groundwater flow system results.

Moving to eastern sectors of the aquifer, clay and silt layers tend to prevail in the upper part of the sequence, confining the lower water-bearing layers and reducing the groundwater recharge. On the contrary, the coarse-grained sediments prevail in the upstream sector of the aquifer, where the high permeability of these outcropping sediments allows direct groundwater recharge by infiltration of rain and surface water during the wet periods (Tadolini et al. 1989; Cotecchia 1956; Maggiore et al. 1996). Consequently, groundwater flows in unconfined conditions in the upper part of the plain and confined conditions in the middle-low part (Fig. 2).

The groundwater flows mainly in the SW-NE direction under an average hydraulic gradient of about 0.5%. Proceeding toward the coast, due to the gradually deepening of the top of the clay formation, some aquifer layers are found below the sea level, and they are affected by seawater intrusion.

Because of the intense exploitation of this aquifer, since the 1990s, regional water authorities have implemented different groundwater monitoring plans. The monitoring network was changed in the course of time regarding both the number of monitoring points and their locations. A large

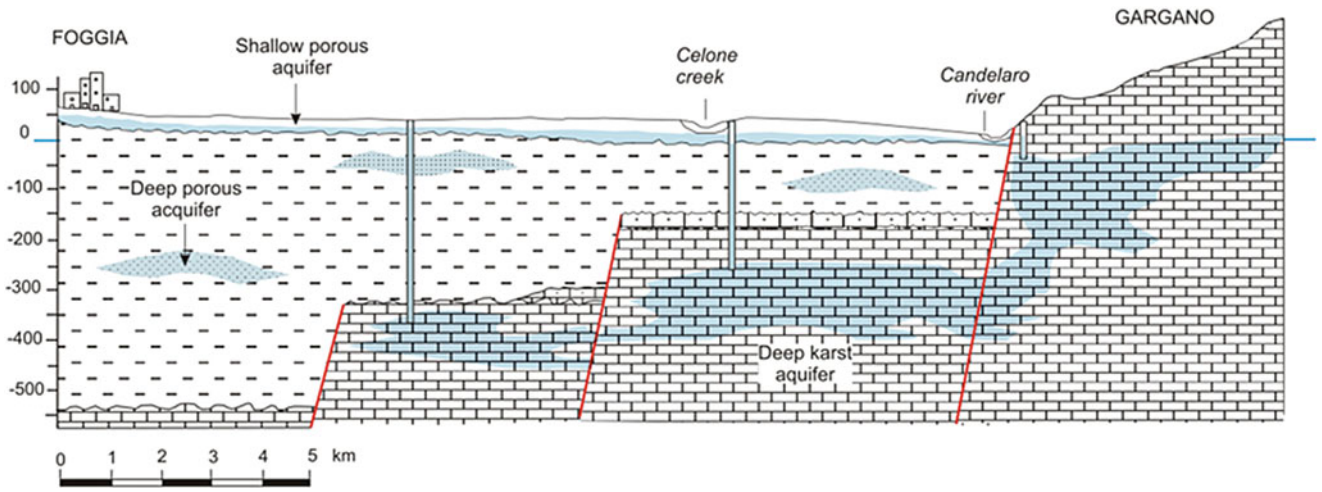


Fig. 3 Cross section C-C′: Hydrogeological complexes in Tavoliere area (Masciale et al. 2011, modified)

amount of data available for this aquifer made it an ideal experimental field in the framework of several national and international research projects (Barca et al. 2006a, 2006b; Passarella et al. 2006; Lo Presti et al. 2010; Maggiore et al. 2005; Masciale et al. 2011) which have allowed to precisely assess the aquifer size and its quantitative status and hydrogeochemical characteristics. Moreover, observed seasonal fluctuations the water table recently led to reliable studies of the groundwater system balance and an improved assessment of the freshwater availability (Passarella et al. 2017).

In this work, hydrogeology has been taken into account as auxiliary soft variables. In particular, hydrogeological information has been simplified by grouping the outcropping lithologies according to their hydraulic conductivity class, which affects the leaching of the contaminant in the subsoil and the time required to reach the water table. According to the Water Protection Plan of Apulia Region (2009), two main classes of hydraulic conductivity have been distinguished over the study area, which are low-medium and very low. The low-medium class corresponds to coarser-grained sediments (well-graded gravel with silt and silty sandy) prevailing in the upstream sector of the aquifer, while the low class corresponds to finer-grained layers (clay, silt, and silty clay) tending to prevail in the downstream sector of the aquifer (Fig. 4).

Land Use

Land use of the study area was derived from the map, updated to 2011, available at the Territorial Information System (SIT) of Apulia Region, conforming to the specifications of the Corine Land Cover project (EEA 2007, 2017). Most of Tavoliere's land covering (about 87%) is classified as agricultural area (Fig. 5).

At a more detailed level, the two main types of land cover in agricultural areas are arable land (67.4%) and permanent arboreal crops (20.7%), the latter made of vineyards, olive, and fruit trees, which cover about 88% of the total area. The remaining part of the area is represented by heterogeneous agricultural land (about 0.2%); artificial surfaces (about 5.8%), comprising infrastructures, human settlements, and urban green areas; and natural and semi-natural environments (about 5.9%), comprising woods, wetlands, water bodies, natural pasture, natural vegetation, beaches, and dunes (Fig. 6).

The paper deals with nitrate pollution of groundwater resources, which mostly comes from agricultural practices, where it is used as a crop nutrient. Land use information then represents a piece of precious information in terms of agricultural exploitation of the considered area. Considering the two prevailing agricultural categories in the study area, which is arable land and permanent crops, the land coverage was simplified in these two main classes as resulting from Fig. 7.

Nitrate Data

Nitrate concentration data, used in this paper, have been provided by Apulia Region. 310 water samples were taken from 49 wells of the regional groundwater monitoring network, established in 2007 within the Project Tiziano.

The samples were collected from autumn 2007 to spring 2011, for a total of eight seasonal monitoring campaigns named from $t = 6$ months (6M-first campaign) to $t = 48$ months (48M-last campaign). The spring campaigns correspond to the maximum water level of the aquifer, while the autumn ones correspond to the minimum. Each water sample has been analyzed, and about 100 chemical and microbiolog-

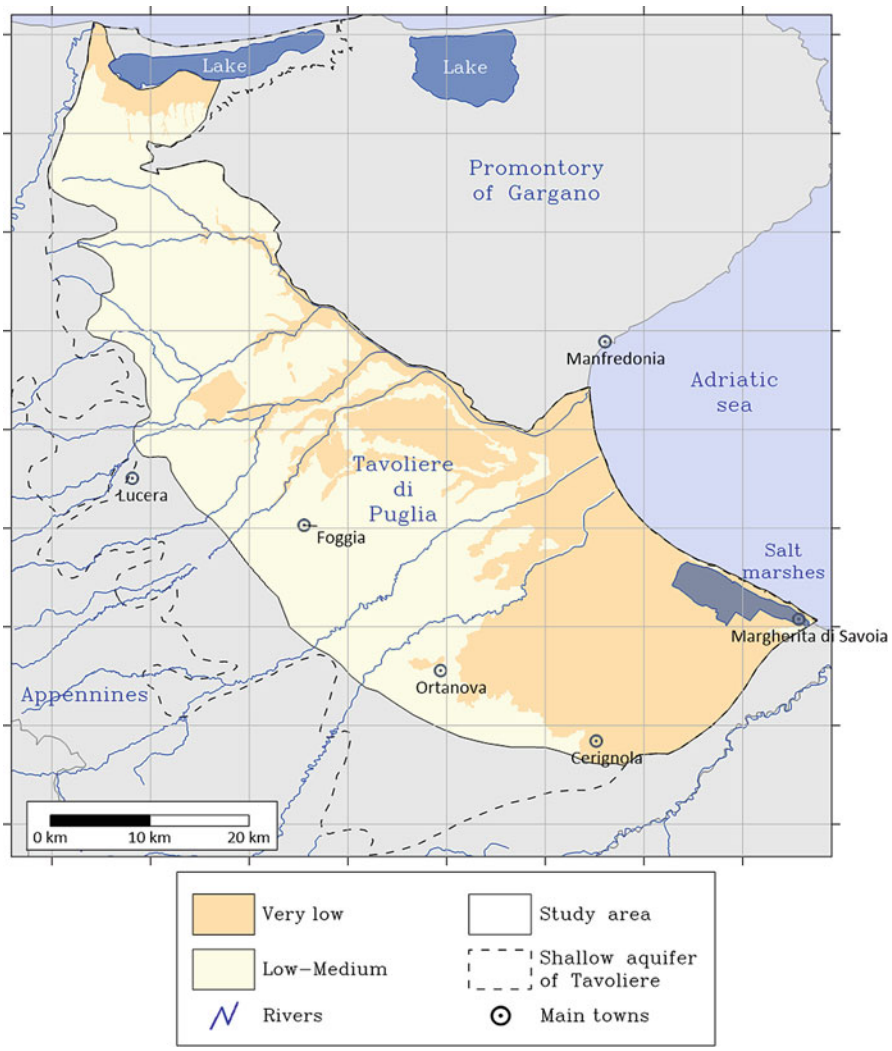


Fig. 4 Simplified map of hydraulic conductivity classes over the study area (Apulia Region 2009)

ical parameters were assessed per campaign, with different analytical methods. In particular, nitrate concentrations were measured by ion chromatography (UNI 9813:1991). Three risk classes were set for increasing nitrate values: class 1 corresponding to $NO_3 < 10$ mg/L; class 2 corresponding to $10 \geq NO_3 < 50$ mg/L; and class 3 corresponding to $NO_3 \geq 50$ that is the acceptable threshold of NO_3 concentration in groundwater set by the Nitrates Directive 91/676/EEC.

Results and Discussion

The technique described in section "Probabilistic Approach" has been applied to the real data set with the goal to show the potential and the weakness of the approach but also to stress the importance of using qualitative/auxiliary information in addition to quantitative information.

The distribution of the three risk classes in the sample data set shows a clear prevalence of the highest risk class (class 3) over the other two (classes 1 and 2), with frequencies of 57%, 21%, and 22%, respectively. These results, while highlighting the danger of qualitative degradation of the groundwater, do not give any spatial information on the location of the areas at greatest risk, which is crucial for the implementation of any recovery strategy. Due to this limitation of the "classical statistical analysis," we turned to the geostatistical techniques of simulation.

A three-dimensional spatiotemporal model was adapted to the experimental variogram of the indicator of the most frequent class (3). The model has been assumed isotropic in space, not having a sufficiently high number of locations to allow the calculation of directional variograms for the identification of any anisotropies in space. The resulting model included three structures: a nugget, a spherical spatial model with a range of 5000 m, and a spherical temporal model with a range of 24 months (Table 1). Moreover, the resulting model is isotropic in space (total sill = 0.25) and time (sill = 0.095).

The results related to the ranges suggest that groundwater quality data can be considered associated (auto-correlated)

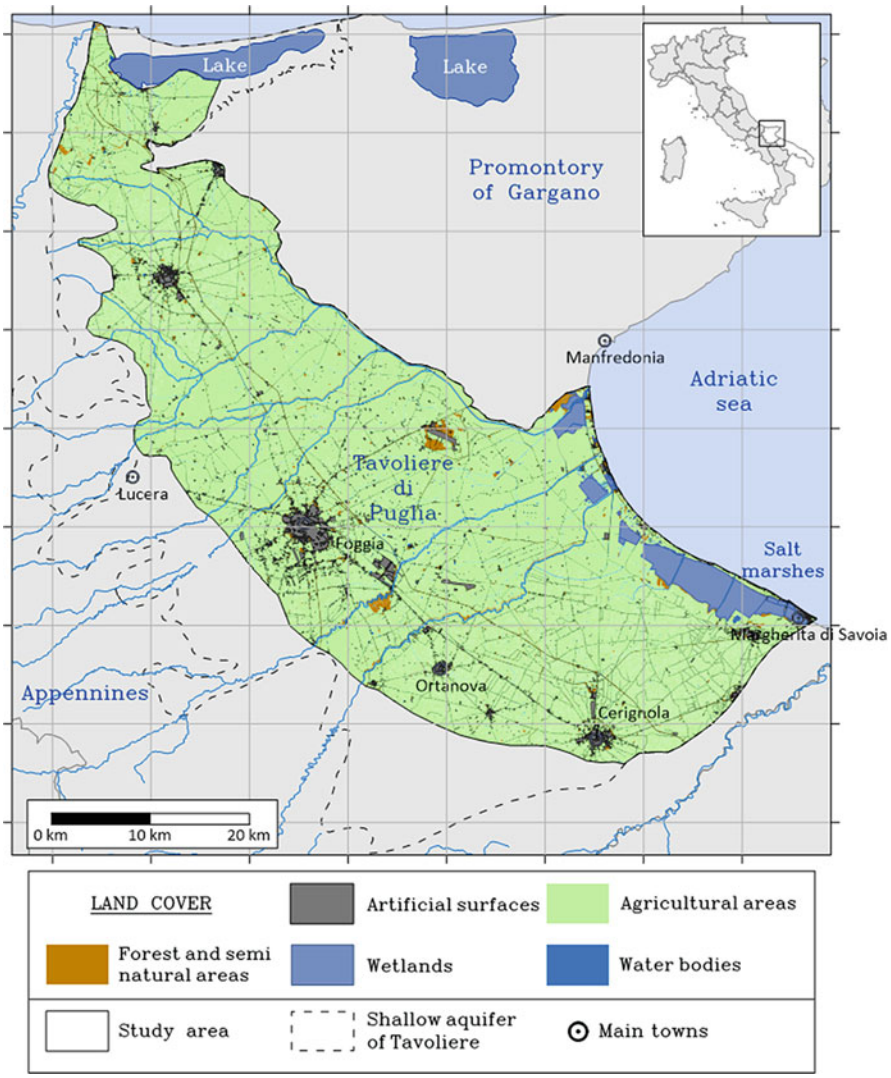


Fig. 5 Land cover of the study area (EEA 2007, 2017)

within a spatial distance of 5000 m and a period of 2 years. Obviously, these results strongly depend on the number and spatiotemporal displacement of the sample measurements (Barca et al. 2017). Given the small number of measurement sites, it is to be expected that the structural characteristics and prediction models will change with the set of sample data. This influence would obviously be less appreciable if a sufficiently exhaustive sample data set were available (approximately 100 samples within a range of 5000 m). For the reasons explained above, we have tried to compensate the missing primary information (nitrate concentration in groundwater) with the auxiliary (hydrogeological and land cover) information but assumed strictly influential the first. The auxiliary information has been combined overlapping the two simplified maps of hydraulic conductivity and land cover shown in Figs. 4 and 7. Figure 8 shows the resulting map.

Four different types of sub-areas can be identified: (1) areas characterized by the presence of arable land with a medium-low hydraulic conductivity class; (2) areas characterized by the presence of arable land with a very low hydraulic conductivity class; (3) areas characterized by the presence of permanent crops with a medium-low hydraulic conductivity class; and (4) areas characterized by the presence of permanent crops with a very low hydraulic conductivity class. Furthermore, even quantitative a priori information has been considered in this study. As described in section "Methodological Framework", such information consists in the per cent number of observed nitrate values belonging to each class over each of the areas of Fig. 8, per monitoring campaign. As an example, Fig. 9 shows the a priori probability maps related to each nitrate threshold, obtained by crossing together quantitative and qualitative information, at the sixth month monitoring campaign. The choice of the sixth month, which is the first of the considered monitoring series, is due to the practical equivalence of all the probability maps during the entire observed period. This equivalence in space and time of all the a priori probability maps confirms the

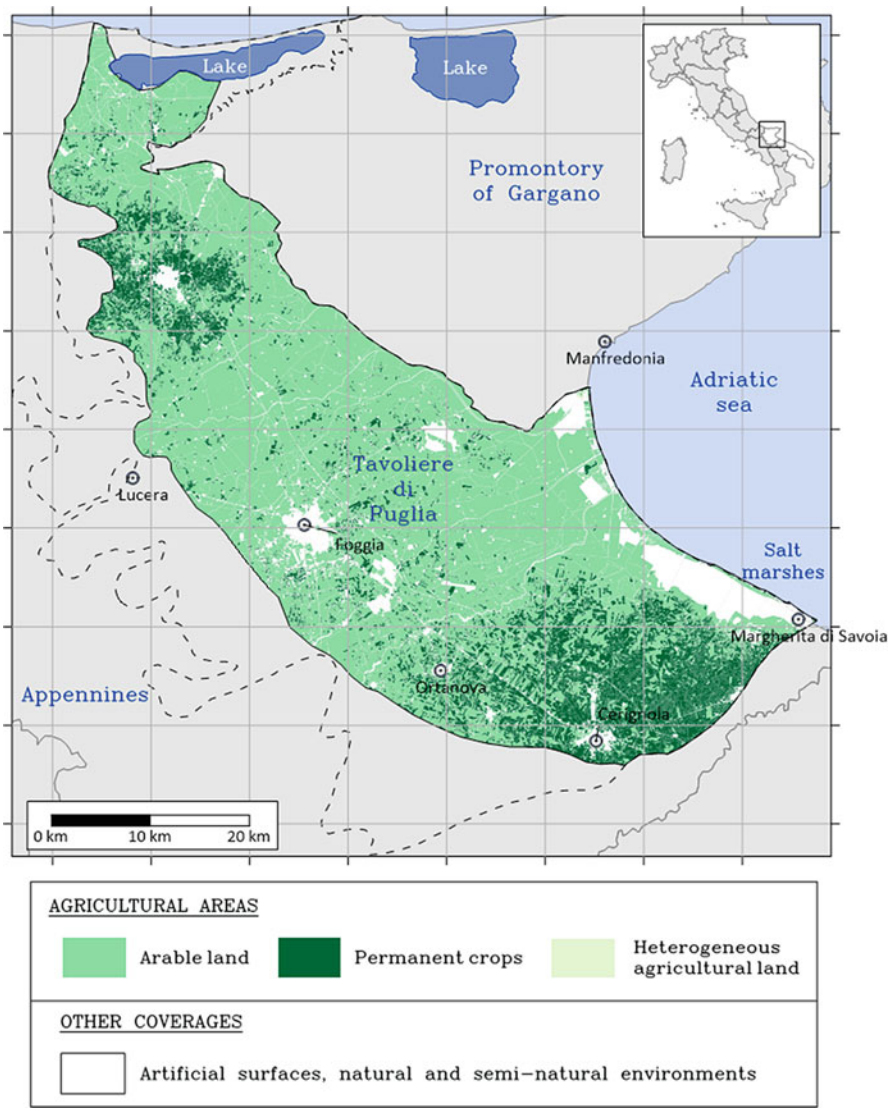


Fig. 6 Agricultural sub-zones of the study area (EEA 2007, 2017)

observed constant in space and persistent in time presence of nitrates within the considered aquifer.

More in detail, Fig. 9 shows an almost equal probability of getting nitrate values in each of the three classes in the south-eastern part of the study area, characterized by the presence of permanent crops with a very low hydraulic conductivity class.

On the other side, the remaining parts of the study area exhibit a higher probability of exceeding the threshold of 50 mg/L, being this probability around 0.6 in the areas characterized by arable land and permanent crops with a medium-low hydraulic conductivity.

The probability increases to 0.8 in the easternmost part of the study area, which is characterized by arable land and very low hydraulic conductivity.

This last zone is the most downstream part of the Tavoliere plain. It covers a wide portion of the coastal area and most of the riverbed proximity bands. The very low hydraulic conductivity of outcropping lithology does not allow nitrate concentration dilution by freshwater recharge. Furthermore, the riverbeds cut the outcropping sediments facilitating the infiltration of surface water, often rich in nitrate content due to soil leaching, toward the groundwater system.

By processing the proximal (spatial) information by the stochastic simulation techniques, the a priori probability maps have been transformed into a posteriori probability maps.

Figure 10 shows the temporal evolution of the a posteriori probability maps of occurrence of the three classes, provided by the proposed method.

From a visual inspection, it is clear that the areas, characterized by a certain level of risk, remain stable over time. This would indicate that the main cause of groundwater quality degradation, even being surely anthropogenic, is deep-seated over the area and rather permanent in time, at least during the observed period.

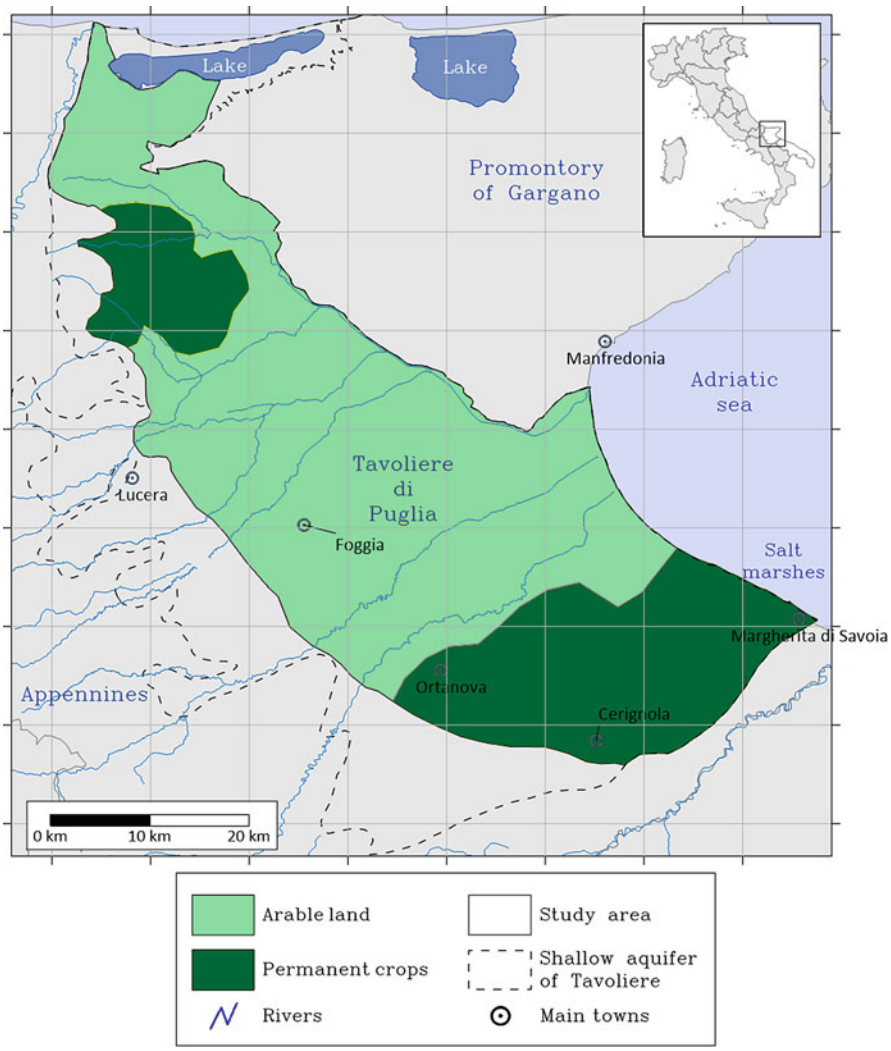


Fig. 7 Simplified map of land cover of the study area

Table 1 Basic structure of the resulting spatiotemporal variogram model

S1	Nugget effect
	Sill = 0.099
S2	Spherical spatial
	Range = 5000.00 m
	Sill = 0.150
S3	Spherical temporal
	Range = 24.00 m
	Sill = 0.095

Sparse hot-spots which appear in the a posteriori maps may have been caused by a low sampling density and a relatively short range with respect to the minimum sampling distance.

The comparison of Figs. 9 and 10 evidences that the a priori and a posteriori maps are very similar apart from a greater variability of the latter obviously due to the additional proximal information.

As already described above with regard to the a priori maps, the probability of occurrence of classes 1 and 2 ($NO_3 \leq 50$ mg/L) in the a posteriori maps is confirmed, generally, very low and reaches its maximum, about 0.4, in the southeastern part of the study area, characterized by permanent crops and very low hydraulic conductivity. At the same time, class 3 ($NO_3 \geq 50$ mg/L) reaches high probability values almost everywhere in the northern part, becoming particularly worrying, with values around 0.8, all along the eastern boundary.

The triple representation of the a posteriori maps, per monitoring campaign, has been summarized into a single map of the most probable class, after the Soares correction (Fig. 11). Even in this case no clear seasonal dynamics is evidenced and the spatial location of the probabilities is the same as described above.

The scene proposed by Fig. 11 is somewhat worrying, as most of the considered area is at risk of exceeding the critical nitrate concentration value of 50 mg/L.

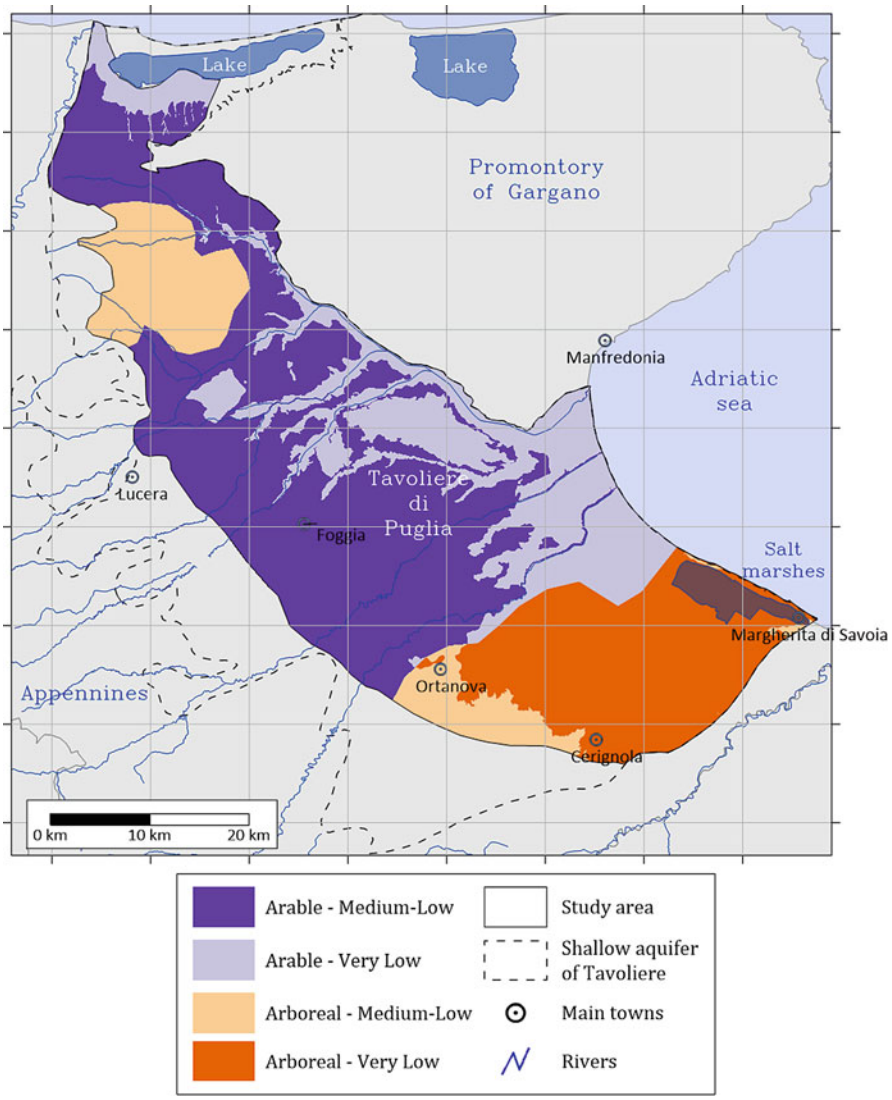


Fig. 8 Final map obtained by overlapping the two auxiliary information (hydraulic conductivity classes and land cover)

Actually, the portions of the study area with permanent crops are mostly characterized by a high probability of being below the critical threshold, and at the southern boundary, where the hydraulic conductivity is very low, the probability is high of finding values below the first threshold of 10 mg/L.

Any risk stochastic modelling is affected by errors. Quantifying it is an important task in order to define and develop reliable plans and actions to restore the environmental status of part or the entire groundwater body. The proposed method provides a suitable tool for measuring the uncertainty associated with the probability estimation. In practice, computing the local standardized entropy (H) at each estimation cell, the uncertainty associated with the risk variable discretized in three categories with probabilities p_k ($k = 1,2,3$) can be univocally assessed in that cell. As described in section "Step 4", the local standardized entropy varies from 0 to 1, which indicates the minimum and maximum uncertainty of the estimated values, respectively.

Figure 12 shows the maps of the computed entropies at each simulated time. All the maps evidence an entropy value near to 1 in the southeastern part of the study area, characterized by arable land and medium-low hydraulic conductivity. This indicates the three classes have about the same probability of occurring. The minima values of $H \cong 0.65$, which means lower uncertainty of the estimated probabilities, always appear at the eastern edge of the study area, which is characterized by arable land and very low hydraulic conductivity. Finally, Fig. 12 shows persistent uncertainty values, $H \cong 0.8$, all over the remaining part of the study area entirely characterized by medium-low hydraulic conductivity, even though by different land uses.

The causes of these high uncertainty values can be attributed to low sampling density and uneven/clustered distribution of the wells under investigation, as well as to intrinsic variability, due to natural (hydrogeological properties) and anthropogenic (land use) factors. The results of these analyses seem to suggest activating a plan of optimization of the monitoring network by intensifying the monitoring locations in areas of higher uncertainty (e.g., entropy >0.8) and more densely populated.

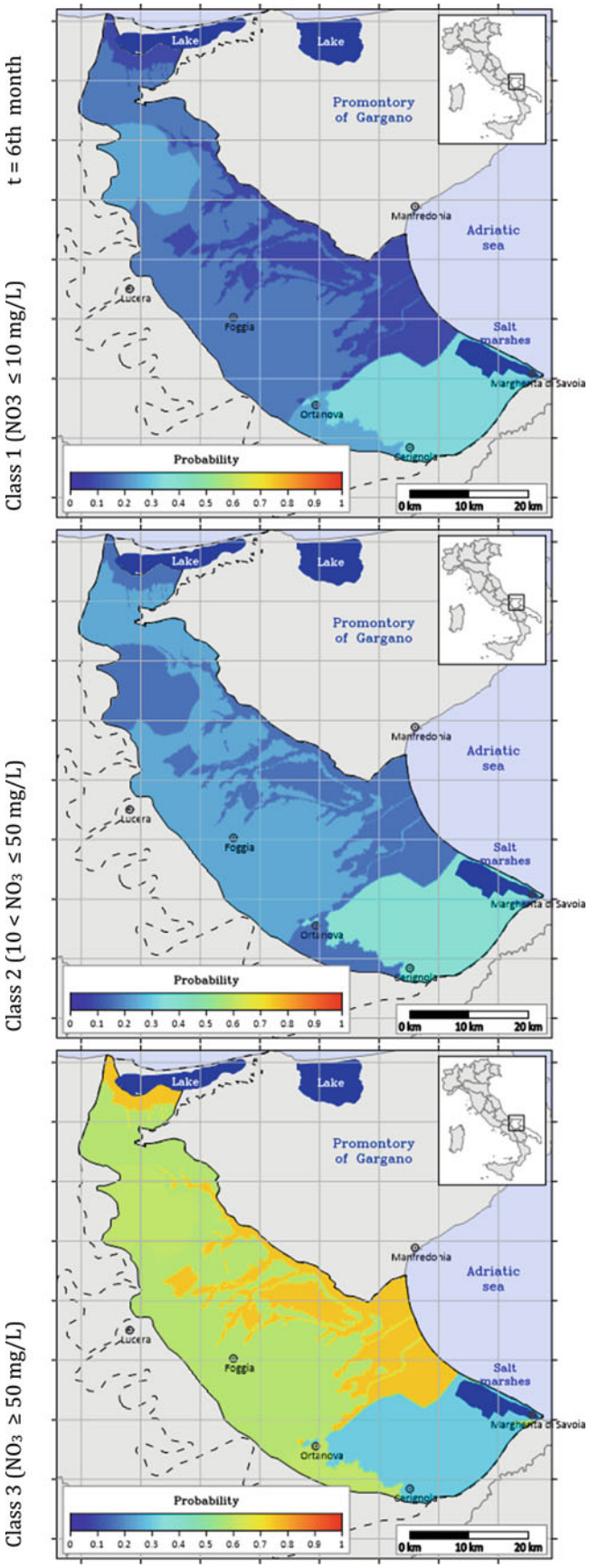


Fig. 9 A priori probability maps of the three classes of risk at the first monitoring campaign

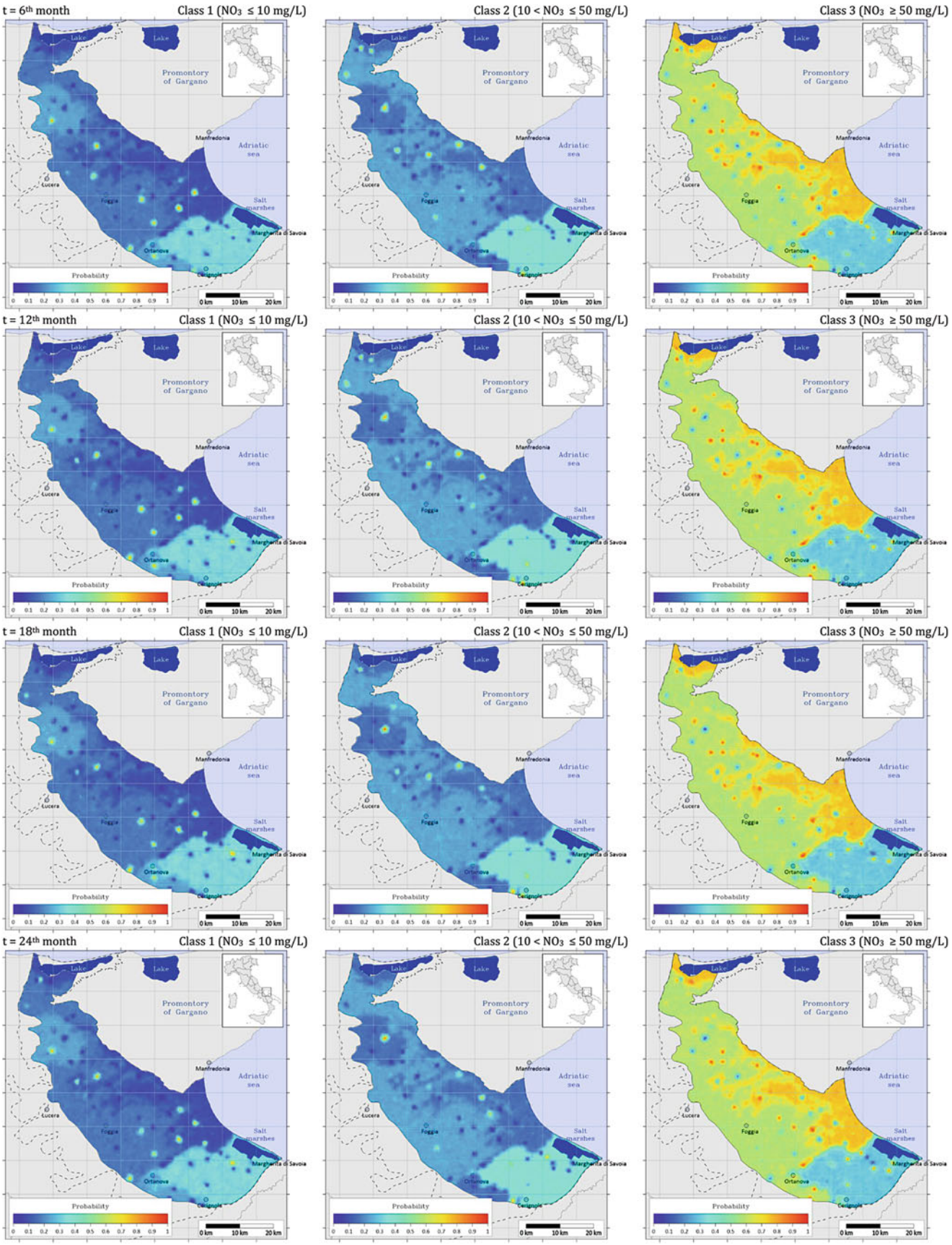


Fig. 10 A posteriori probability maps of the three classes for each monitoring campaign

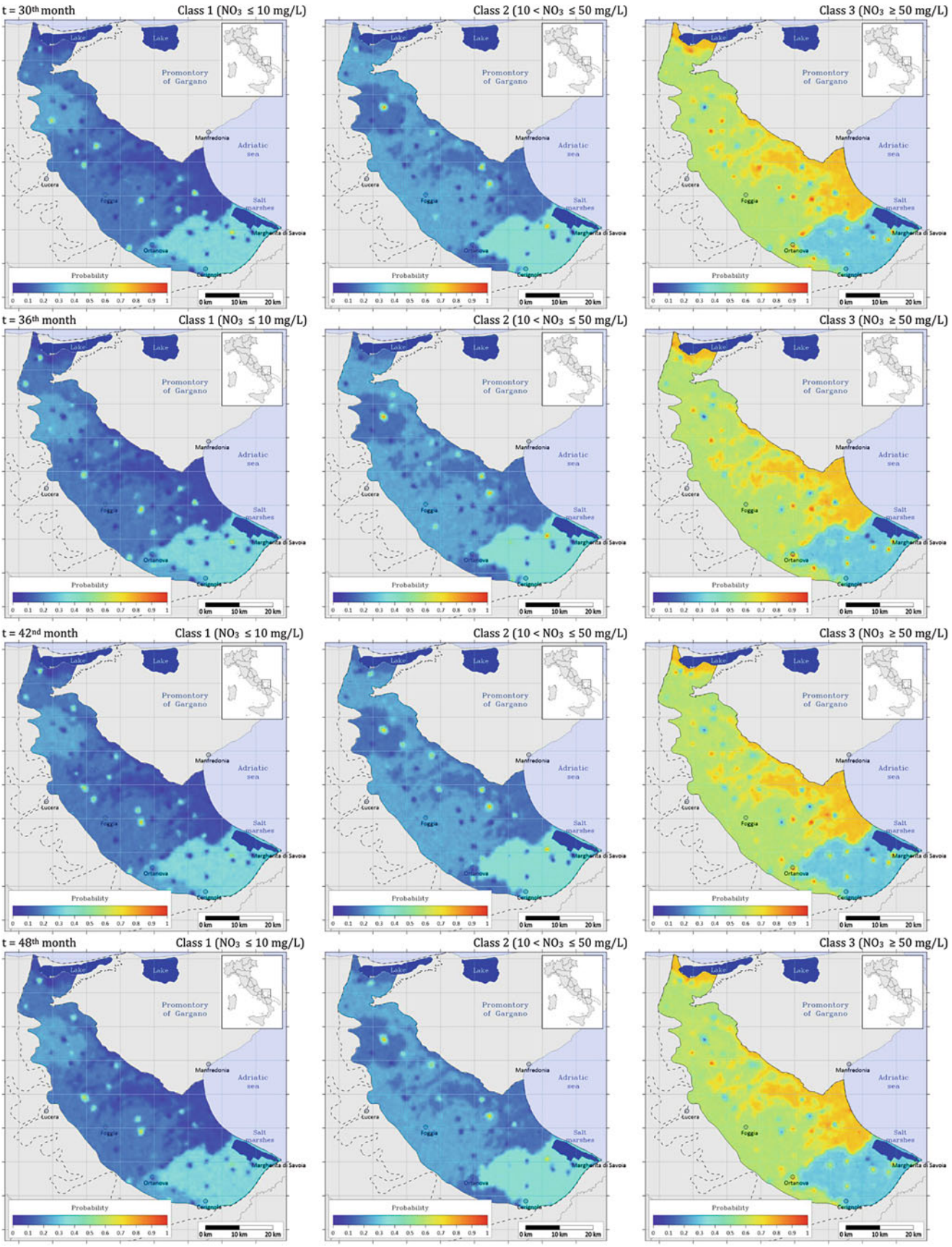


Fig. 10 (continued)

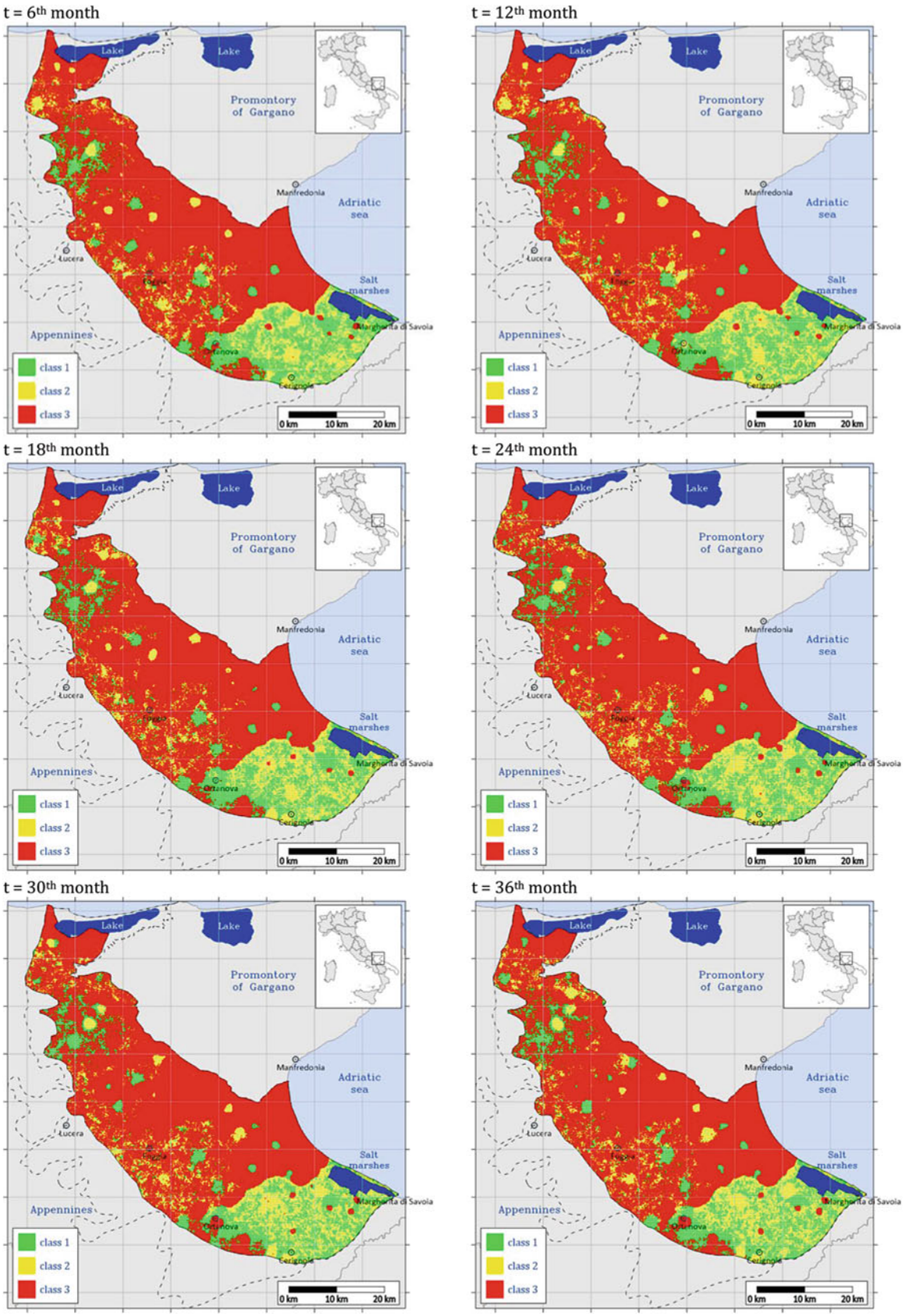


Fig. 11 Maps of the most probable class over time, after the Soares correction

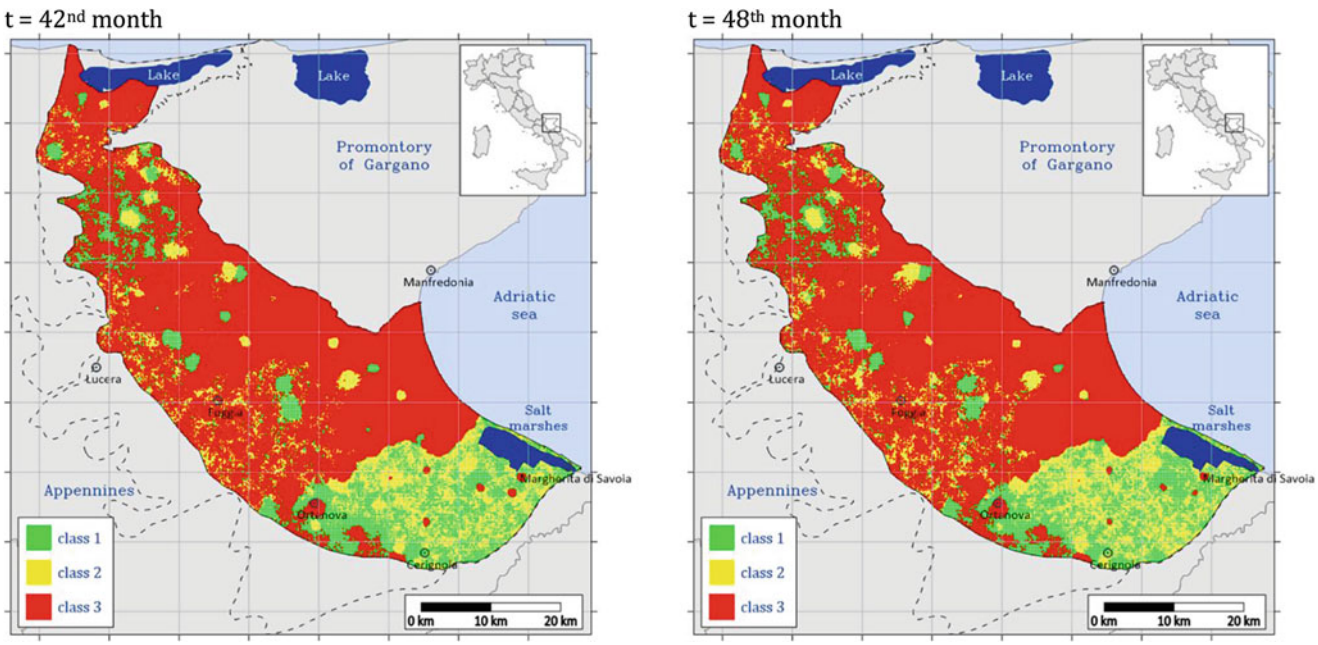


Fig. 11 (continued)

The time series of the eight risk maps of Fig. 11 has been summarized in time by evaluating the most frequent class in each estimation cell. Figure 13 shows the resulting map of the most frequent classes of risk over the entire simulated period.

In Fig. 13, also the four different areas characterized by different land uses and hydraulic conductivities, considered as auxiliary information, have been displayed by different hatchings. It is evident that the areas with the lowest risk, where classes 1 and 2 prevail, are those cultivated with permanent crops (mostly olive groves and vineyards), while the areas with the highest risk (class 3 prevalence) are those cultivated with arable crops (mostly cereals and durum wheat), owing to the more intense fertilization.

Fertilizer's use in agriculture is reported to be the highest anthropogenic source of nitrate contamination in groundwater (Shukla and Saxena 2018). In fact, it is known that nitrogen is one of the major components of fertilizers, whose usage in agriculture has increased in time to escalate the crop yield. This often causes an over-application of fertilizers by farmers, other than an improper timing of applying, in the belief believing that more fertilizer is equivalent to higher crop yield. Fertilizer application and subsequent leaching cause the nitrates to reach the groundwater.

Still concerning the results in Fig. 13, it is also possible to establish a relationship between classes of risk and subsoil hydrogeological characteristics. In fact, within the areas with permanent crops, class 3 appears only in the portions characterized by medium-low hydraulic conductivity. This is because, in these portions, the greater hydraulic conductivity of the subsoil facilitates the percolation of nitrates to groundwater.

On the contrary, within the areas characterized by more intensely fertilized arable crops, a higher frequency of class 2 is evident along the whole western strip characterized by medium-low hydraulic conductivity. Otherwise, on the eastern side, the most worrisome class 3 prevails. As detailed in section "Hydrogeology", groundwater mainly flows from the southwest to the northeast, and then the main natural recharge area is located just along the western strip.

In this context, even though the western area of the aquifer is the main infiltrating way of nitrates from soils to groundwater, here, the transfer potential, which is the groundwater capability of allowing migration of nitrogen, is higher. This would justify the larger frequency of class 2 in this area.

Instead, the eastern part of the aquifer is characterized by a very low hydraulic conductivity and results to be moderately confined. This, somehow, prevents the direct infiltration of nitrates from soils to groundwater. Nevertheless, here, riverbeds cut the outcropping less permeable sediments and allow surface water to recharge groundwater facilitating the infiltration of nitrates.

Moreover, in this part of the aquifer, which is the most downstream, the groundwater flow velocity decreases, and, consequently, the accumulation potential, which is its capability of retaining pollutants, increases allowing nitrates, flowing from the recharge area, to accumulate and increase in concentration.

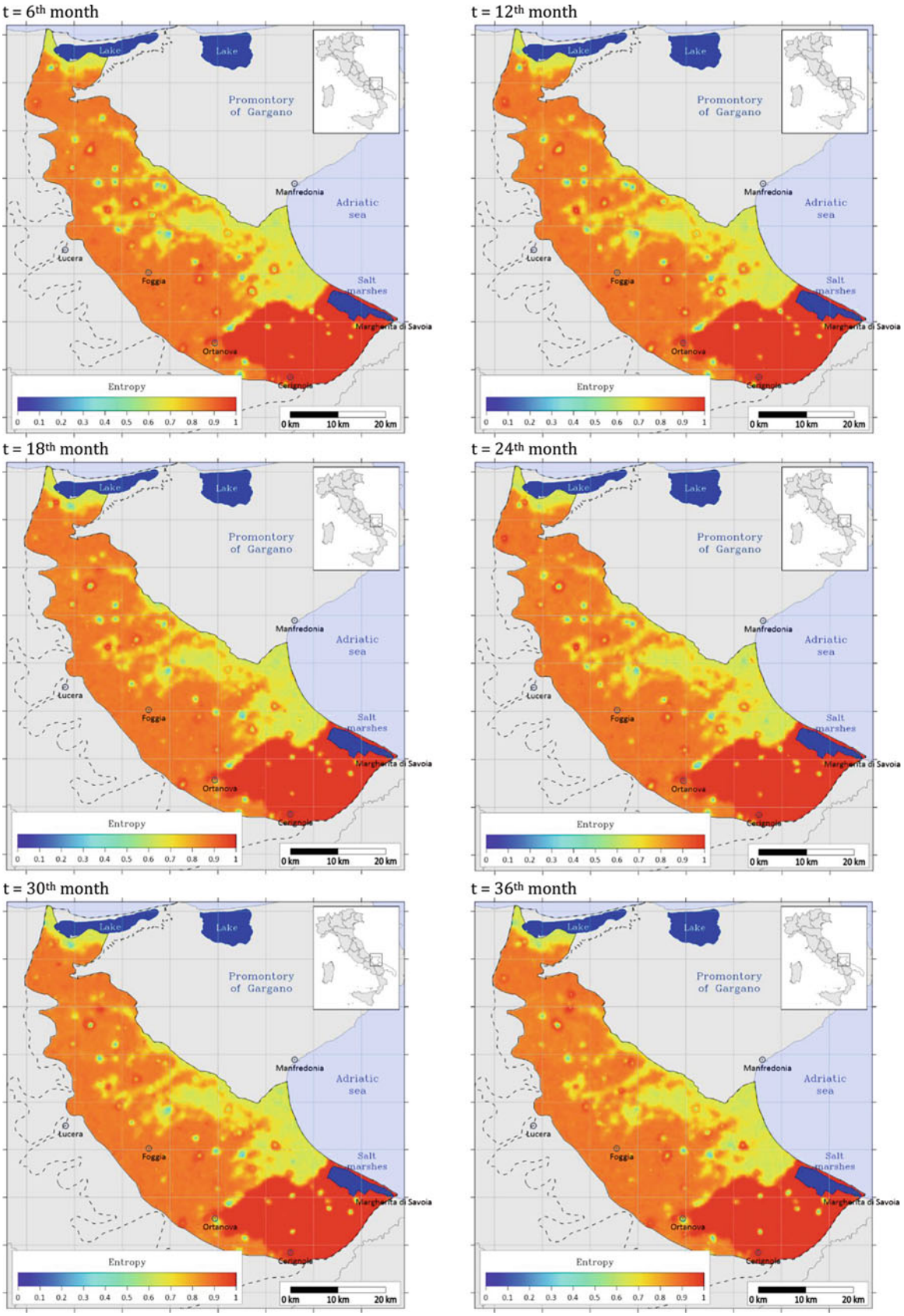


Fig. 12 Map of entropy and its evolution over time

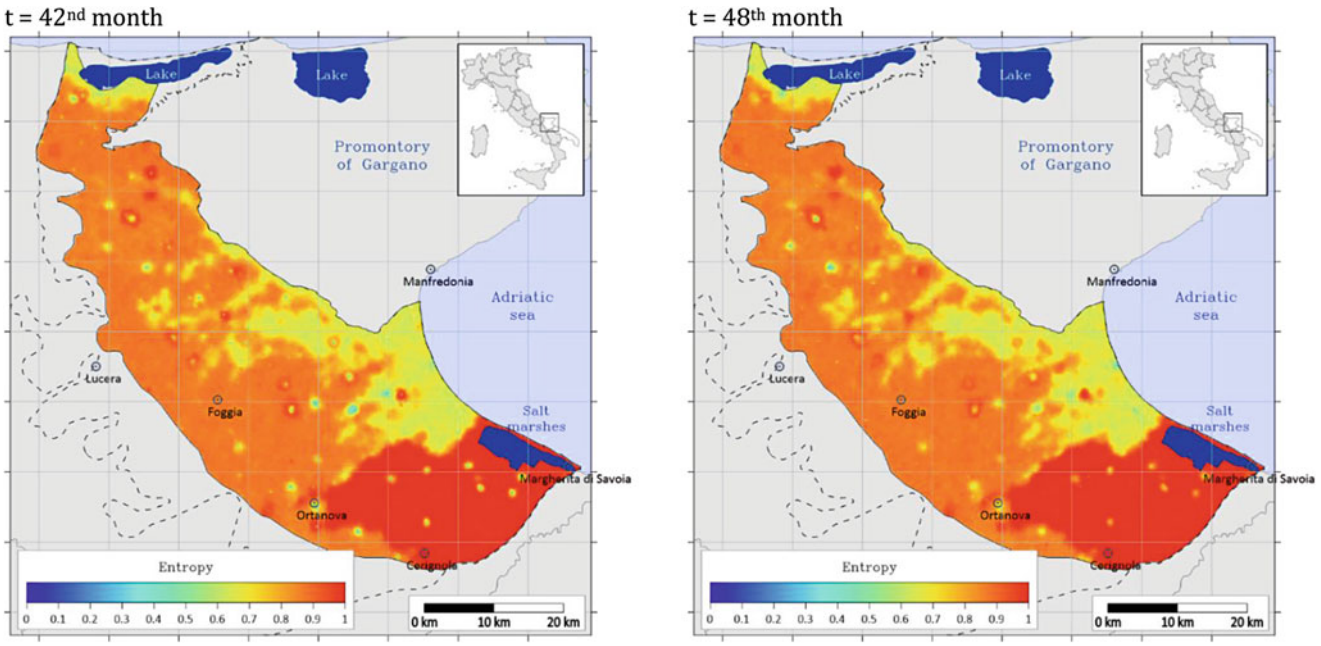


Fig. 12 (continued)

Conclusions

In arid and semiarid areas, groundwater is often the primary source of freshwater supply. Usually, focus on groundwater quality concerns its use as drinking water. Nevertheless, groundwater should be protected for its environmental value because of its important role within the hydrological cycle through the maintenance of wetlands and river flows, acting as a buffer through dry periods. From a managerial standpoint, actions are required in order to reduce the risk of groundwater quality degradation in regions such as the study area where the climate is mostly semiarid (Maggi et al. 2018; Passarella et al. 2020) and land is intensively cultivated. The European legislation (EU 1991, 2000, 2006) recommends groundwater protection, improvement, and restoration, even by promoting good agricultural practices, such as precision farming and irrigation, and rational use of fertilizers.

The goal of this chapter has been to define an approach for performing the risk analysis of groundwater quality degradation due to nitrates leaching from soils, in probabilistic terms.

The Indicator Simulation algorithm, a stochastic simulation technique provided by geostatistics, has been applied, which allows considering together both quantitative (hard) and qualitative (soft) data. The results of the proposed method have been maps showing the probability of exceeding increasing assigned thresholds of nitrate concentration combined to different patterns of land use and hydrogeological features. The more the threshold, the more the risk of groundwater quality degradation increases. The proposed approach allowed also quantifying the uncertainty associated with any estimated risk class, in terms of entropy.

GIS processing of the probability maps allowed producing summary maps of the most probable risk class per estimation cell, over the whole considered area, at each observed time. Further elaboration produced a final map of the most frequent over time, among the most probable risk classes. All three levels of risk representation should provide land and water resource managers with useful and reliable tools, capable of supporting decisions, and define priorities in launching actions. This method could also be decisive to more reliably support the delimitation of nitrate-vulnerable zones.

The proposed computational tool proves to be quite flexible and applicable in different experimental scenarios, due to its capability to treat simultaneously data of the different type and quality related to various sources of information. Further, the approach is non-parametric; therefore, it can be particularly suitably applied even in the case of highly skewed, multimodal distributions, as often found in nitrate concentrations in groundwater.

However, the proposed technique shows also some limitations: first, it is computationally challenging, and the reliability of the results increases with conditioning (sampling data) and absence of spatial clustering in the data. In general, the proposed methodology provides good and reliable results even compared to qualitative, knowledge-based evaluations of the risk of groundwater quality degradation due to nitrate. Improvements could be achieved with spatial and temporal optimization of monitoring networks and sampling procedures that can provide more detailed information for conditioning the development of the whole proposed methodology.

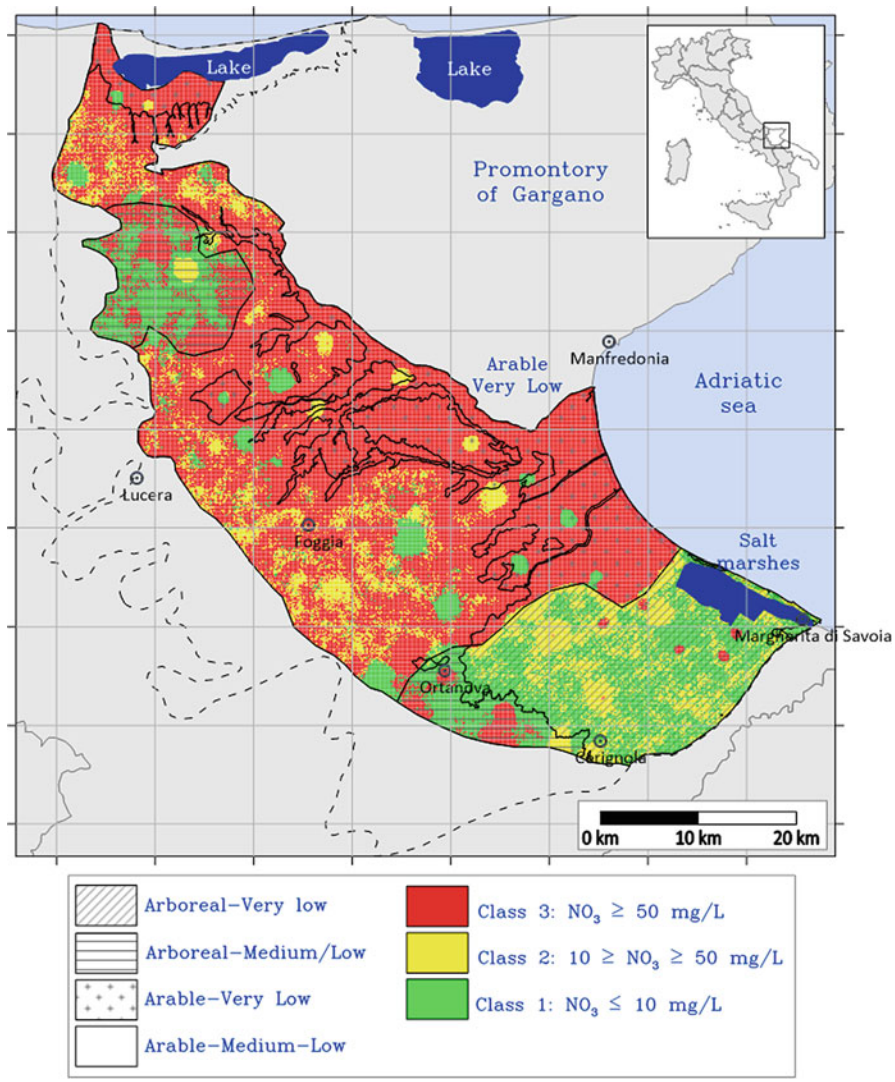


Fig. 13 Map of the most frequent classes of risk over the entire simulated period

Acknowledgments The data used in the case study were collected within the Project Tiziano for qualitative and quantitative monitoring of the Apulia groundwater bodies funded by the Regional Operational Programme (POR) 2000–2006, "Improvement of basic knowledge, adaptation and extension of the soil, surface water, groundwater and coastal waters monitoring system."

The authors' gratitude goes to the Water Resources Section of the Department of Agriculture and Environmental Development of the Apulia Region who provided the data and a valuable technical support.

References

Alabert, F.G. 1987. Stochastic imaging of spatial distributions using hard and soft information (Doctoral dissertation, Stanford University Press).

Alabert, F.G., and G.J. Massonnat. 1990, January. Heterogeneity in a complex turbiditic reservoir: Stochastic modelling of facies and petrophysical variability. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Almastri, M.N. 2007. Nitrate contamination of groundwater: A conceptual management framework. *Environmental Impact Assessment Review* 27 (3): 220–242.

Apulia Region. 2009. Water protection plan. Regional Council resolution N. 230 of October 20 2009.

Barca, E., and G. Passarella. 2008. Spatial evaluation of the risk of groundwater quality degradation. A comparison between disjunctive kriging and geostatistical simulation. *Environmental Monitoring and Assessment* 137 (1–3): 261–273.

Barca, E., G. Passarella, R. Lo Presti, R. Masciale, and M. Vurro. 2006a. HarmoniRiB River Basin data documentation. Chapter 7 – Candelaro River Basin. *HarmoniRiB Project Deliverables* D6 (3): 1–47.

———. 2006b. Candelaro River Basin, Italy – A HarmoniRiB case study. *HarmoniRiB Project Deliverables* D7 (6): 1–54.

Barca, E., A. Castrignanò, G. Buttafuoco, D. De Benedetto, and G. Passarella. 2015. Integration of electromagnetic induction sensor data in soil sampling scheme optimization using simulated annealing. *Environmental Monitoring and Assessment* 187 (7): 422.

Barca, E., E. Porcu, D. Bruno, and G. Passarella. 2017. An automated decision support system for aided assessment of variogram models. *Environmental Modelling & Software* 87: 72–83.

Bouraoui, F., B. Grizzetti, G. Adelsköld, H. Behrendt, I. De Miguel, M. Silgram, et al. 2009. Basin characteristics and nutrient losses: The EUROHARP catchment network perspective. *Journal of Environmental Monitoring* 11 (3): 515–525.

Burton, I., and A. Whyte. 1980. Environmental risk assessment. Ed. Wiley, Scope 15, Toronto, 157 p.

Camargo, J.A., and Á. Alonso. 2006. Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: A global assessment. *Environment International* 32 (6): 831–849.

Castrignanò, A., and G. Buttafuoco. 2004. Geostatistical stochastic simulation of soil water content in a forested area of South Italy. *Biosystems Engineering* 87 (2): 257–266.

Castrignanò, A., L. Giugliarini, R. Risaliti, and N. Martinelli. 2000a. Study of spatial relationships among soil physical-chemical properties using Multivariate Geostatistics. *Geoderma* 97: 39–60.

Castrignanò, A., P. Goovaerts, L. Lulli, and G. Bragato. 2000b. L. A geostatistical approach to estimate probability of occurrence of Tuber melanosporum in relation to some soil properties. *Geoderma* 98: 95–113.

Castrignanò, A., G. Buttafuoco, A. Canu, C. Zucca, and S. Madrau. 2007. Modelling spatial uncertainty of soil erodibility factor using joint stochastic simulation. *Land Degradation & Development* 19 (2): 198–213.

Cotecchia, V. 1956. Gli aspetti idrogeologici del Tavoliere delle Puglie. *L'Acqua* 11-12: 168–180.

Deutsch, C.V., and A.G. Journel. 1998. *Geostatistical software library and user's guide*. New York: Oxford University Press.

Dowd, P.A., and E. Pardo-Igúzquiza. 2002. The incorporation of model uncertainty in geostatistical simulation. *Geographical and Environmental Modelling* 6 (2): 147–169.

Duckett, E.J. 1983. Environmental risk assessment. Edited by Anne V. Whyte and Ian Burton. Published by John Wiley & Sons (Chichester, UK) on behalf of The Scientific Committee on Problems of the Environment of the International Council of Scientific Unions, 1980, xxiii+ 157 pp., $12.

EEA (European Environment Agency). 2007. CLC2006 technical guidelines. Technical report No. 17/2007. EEA (p. 66). ISBN: 978-92-9167-968-3. https://www.eea.europa.eu/publications/technical_report_2007_17

———. 2012. European Waters – Assessment of Status and Pressure. Report No. 8/2012. EEA, Copenhagen, Denmark (pp. 97).

———. 2017. Landscapes in transition — An account of 25 years of land cover change in Europe. EEA Report No. 10/2017. EEA (pp. 85). ISBN: 978-92-9213-882-0. https://www.eea.europa.eu/publications/landscapes-in-transition

EU (European Union). 1991. Council Directive 91/676/EEC of 12 December 1991 concerning the protection of water against pollution caused by nitrates from agricultural sources. *Official Journal of the European Communities* L375: 1–8.

———. 2000. Council directive 2000/60/EC of 23 October 2000 establishing a framework for community action in the field of water policy. *Official Journal of the European Communities* L327: 1–71.

———. 2006. Council directive 2006/118/EC of 12 December 2006 on the protection of groundwater against pollution and deterioration. *Official Journal of the European Communities* L372: 19–31.

Gallicchio, S., M. Moretti, L. Spalluto, and S. Angelini. 2014. Geology of the middle and upper Pleistocene marine and continental terraces of the northern Tavoliere di Puglia plain (Apulia, southern Italy). *Journal of Maps* 10 (4): 569–575.

Goodchild, R.G. 1998. EU policies for the reduction of nitrogen in water: The example of the Nitrates Directive. In *Nitrogen, the Confer-Ns* (pp. 737–740).

Goovaerts, P. 1997. *Geostatistics for natural resources evaluation*. New York: Oxford University Press. Geostatistics for natural resources evaluation. Oxford Univ. Press, New York.

Haller, L., P. McCarthy, T. O'Brien, J. Riehle, and T. Stuhldreher. 2013. Nitrate pollution of groundwater. 2014: Alpha water systems INC. *Google Scholar*.

IPCC. 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.

Journel, A.G. 1983. Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology 15* (3): 445–468.

———. 1986. Constrained interpolation and qualitative information—The soft kriging approach. *Mathematical Geology* 18 (3): 269–286.

Journel, A.G., and C.V. Deutsch. 1993. Entropy and spatial disorder. *Mathematical Geology* 25 (3): 329–355.

Kronvang, B., S.A. Borgvang, and L.J. Barkved. 2009. Towards European harmonised procedures for quantification of nutrient losses from diffuse sources—The EUROHARP project. *Journal of Environmental Monitoring* 11 (3): 503–505.

Libutti, A., and M. Monteleone. 2017. Soil vs. groundwater: The quality dilemma. Managing nitrogen leaching and salinity control under irrigated agriculture in Mediterranean conditions. *Agricultural Water Management* 186: 40–50.

Liu, A., J. Ming, and R.O. Ankumah. 2005. Nitrate contamination in private wells in rural Alabama, United States. *Science of the Total Environment* 346 (1–3): 112–120.

Lo Presti, R., E. Barca, and G. Passarella. 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment* 160 (1–4): 1–22.

Machiwal, D., V. Cloutier, C. Güler, and N. Kazakis. 2018. A review of GIS-integrated statistical techniques for groundwater quality evaluation and protection. *Environmental Earth Sciences* 77 (19): 681.

Maggi, S., D. Bruno, A. Lay-Ekuakille, R. Masciale, and G. Passarella. 2018. Automatic processing of bioclimatic data in the space and time domains. *Journal of Physics: Conference Series* 1065 (19): 9.

Maggiore, M., G. Nuovo, and P. Pagiarulo. 1996. Caratteristiche idrogeologiche e principali differenze idrochimiche delle falde sotterranee del Tavoliere di Puglia. *Memorie Società Geologica Italiana* 51: 669–684, Roma.

Maggiore M., R. Masciale, G. Passarella, and M. Vurro. 2005. A preliminary assessment of the groundwater environmental state in the shallow aquifer of the Tavoliere di Puglia (Southern Italy). In Proceedings 3rd Symposium "Quality and Management of Water Resources". June 16–18, 2005, St. Petersburg, Russia, pp: 146–156 -ISBN 5-88749-002-0.

Masciale, R., E. Barca, and G. Passarella. 2011. A methodology for rapid assessment of the environmental status of the shallow aquifer of "Tavoliere di Puglia" (Southern Italy). *Environ Monit Assess* 177 (1–4): 245–261.

Menció, A., J. Mas-Pla, N. Otero, O. Regàs, M. Boy-Roura, R. Puig, et al. 2016. Nitrate pollution of groundwater; all right . . . , but nothing else? *Science of the Total Environment* 539: 241–251.

Mishra, R.K., and S. Sarkar. 2017. Addressing social and environmental risks through cSr–an Indian perspective. *IPE Journal of Management* 7 (1): 149–158.

Passarella, G., M. Vurro, V. D'agostino, G. Giuliano, and M.J. Barcelona. 2002. A probabilistic methodology to assess the risk of groundwater quality degradation. *Environmental Monitoring and Assessment* 79 (1): 57–74.

Passarella, G., E. Barca, and A. Lo Porto. 2006. Collection and elaboration of data for the pilot area of the Candelaro River (Italy). Development of completed database accompanied by appropriate software package for the retrieval, analysis and processing of flood related data. Description of the Candelaro catchment. *FloodMed Project Deliverables* 2 (2): 1–57.

Passarella, G., E. Barca, D. Sollitto, R. Masciale, and D.E. Bruno. 2017. Cross-calibration of two independent groundwater balance models and evaluation of unknown terms: The case of the shallow aquifer of "Tavoliere di Puglia" (South Italy). *Water Resources Management* 31 (1): 327–340.

Passarella, G., D.E. Bruno, A. Lay-Ekuakille, S. Maggi, R. Masciale, and D. Zaccaria. 2020. Spatial and temporal classification of coastal regions using bioclimatic indices in a Mediterranean environment. *Science of the Total Environment* 700: 134415.

Power, M., and L.S. McCarty. 1998. Peer reviewed: A comparative analysis of environmental risk assessment/risk management frameworks. *Environmental Science & Technology* 32 (9): 224A–231A.

Rue, M., A. Poulin, E. Tremblay, and L. Provencher. 1999. *An environmental risk management approach: Zone of influence and cooperation of Kouchibouguac National Park*. New Brunswick.

Shukla, S., and A. Saxena. 2018. Global status of nitrate contamination in groundwater: Its occurrence, health impacts, and mitigation measures. In *Handbook of environmental materials management*, ed. C. Hussain. Cham: Springer.

Siebert, S., J. Burke, J.M. Faures, K. Frenken, J. Hoogeveen, P. Döll, and F.T. Portmann. 2010. Groundwater use for irrigation–a global inventory. *Hydrology and Earth System Sciences* 14 (10): 1863–1880.

Soares, A. 1992. Geostatistical estimation of multi-phase structures. *Mathematical Geology* 24 (2): 149–160.

———. 1998. Sequential indicator simulation with correction for local probabilities. *Mathematical Geology* 30 (6): 761–765.

Switzer, P. 1977. Estimation of spatial distributions from point sources with applications to air pollution measurement. Proceeding of the 41st ISI session, New Delhi. *Bulletin of the International Statistical Institute* 47 (2): 123–137.

Tadolini T., F. Sdao, and G. Ferrari. 1989. "Valutazioni sul grado di protezione della falda superficiale del Tavoliere di Foggia nei confronti dei rilasci in superficie di corpi inquinanti e sulle modalità di propagazione degli stessi in seno all'acquifero. Atti delle giornate di studio su Analisi Statistica di Dati Territoriali. 461–472, 1989, Bari.

Varnes, D.J. 1984. *Commission on landslides and other mass-movements - IAEG. Landslide hazard zonation: A review of principles and practices*. Paris: The UNESCO Press.

Wackernagel, H. 1996. Multivariate geostatistics: an introduction with applications. *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts* 33 (8): 363A.

Zaccaria, D., G. Passarella, D. D'Agostino, R. Giordano, and S.S. Solis. 2016. Risk assessment of aquifer salinization in a large-scale coastal irrigation scheme, Italy. *CLEAN–Soil, Air, Water* 44 (4): 371–382.

Correction to: Geospatial Technology for Human Well-Being and Health

Fazlay S. Faruque

Correction to:F. S. Faruque (ed.),*Geospatial Technology for Human Well-Being and Health*,https://doi.org/10.1007/978-3-030-71377-5

In chapters 12 & 13, the names of one of the authors have been inadvertently published as '*Lakitha Omal Harindha Wijerante*' which has been updated as '*Lakitha Omal Harindha Wijeratne*'.

The updated version of these chapters can be found athttps://doi.org/10.1007/978-3-030-71377-5_12https://doi.org/10.1007/978-3-030-71377-5_13

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_21

Index

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5

T