

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

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 ren-

dering 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-

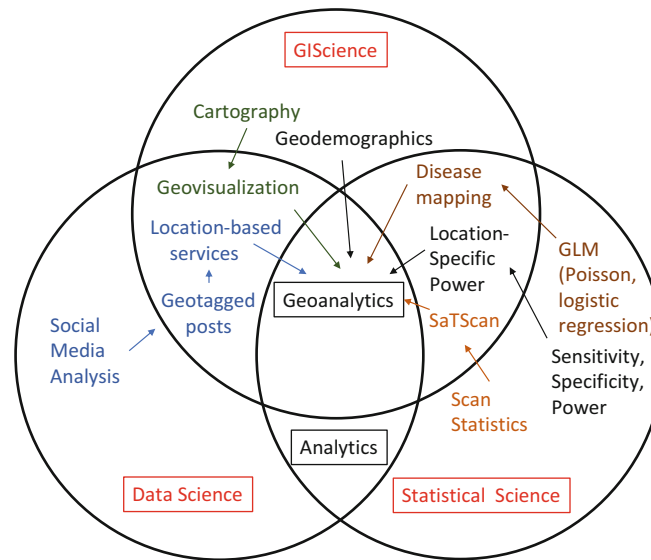


Fig. 1 Illustration of system approach of GIScience, Data Science, and Statistical Science and their components to achieve data-driven goals. Arrows indicate related areas of research moving from discipline-

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 geanalytics. 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 geanalytics. 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 geanalytic 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 geospatial 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 geospatial 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 geospatial 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 geospatial 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 geospatial analytics, incorporating time will allow the expansion of geospatial analytics 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 geospatial analytics, 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 geospatial analytic 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. *Cybergeog: European Journal of Geography* (online). <http://journals.openedition.org/cybergeog/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. Grunewald. 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.