Tourism on the Verge

Zheng Xiang
Daniel R. Fesenmaier *Editors*

Analytics in Smart Tourism Design

Concepts and Methods

Springer

Tourism on the Verge

Series editors
Pauline J. Sheldon
University of Hawaii, Honolulu, Hawaii, USA

Daniel R. Fesenmaier
University of Florida, Gainesville, Florida, USA

More information about this series at http://www.springer.com/series/13605

Zheng Xiang • Daniel R. Fesenmaier
Editors

Analytics in Smart Tourism Design

Concepts and Methods

Springer

Editors

Zheng Xiang
Department of Hospitality and Tourism Management
Virginia Polytechnic Institute and State University
Blacksburg, Virginia
USA

Daniel R. Fesenmaier
National Laboratory for Tourism & eCommerce
Department of Tourism, Recreation and Sport Management
University of Florida
Gainesville, Florida
USA

ISSN 2366-2611 ISSN 2366-262X (electronic)
Tourism on the Verge
ISBN 978-3-319-44262-4 ISBN 978-3-319-44263-1 (eBook)
DOI 10.1007/978-3-319-44263-1

Library of Congress Control Number: 2016955413

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Acknowledgments

I wrote my doctoral thesis nine years ago under the supervision of Dan Fesenmaier at Temple University. In it I used search results from Google and user queries from several search engines to examine the structure and characteristics of the so-called online tourism domain. Looking back, my thesis was purely "descriptive" using "secondary" data, which would most likely be viewed as "unorthodox" back then. Today, many of the analytical approaches to understanding the new reality, which is constantly being shaped by information technology, have grown to dominate our everyday conversations about the meaning of knowledge creation. Since my graduation, I have been working with a number of colleagues worldwide on different types of research problems related to IT in travel and tourism, many of which can now be characterized as "data analytics." While I have benefited a lot from my collaborators in the works we published together, Dan's influence and support has been tremendous throughout my intellectual development. Notwithstanding his relentless pursuit of rigor and excellence, Dan has huge impact on my way of looking at the world, particularly with his open-mindedness to research and willingness to learn new things no matter how outlandish they appear at the beginning. This book embodies, primarily, Dan's idea of "moving forward" within the realms of technology, data, design of tourism experience, and the emerging topic of smart tourism.

Besides, I would also like to thank the contributors of this book. While some of them are well-established scholars around the world, several authors are actually quite young, who represent the future of research. I am grateful for the privilege of working with them on this project.

Zheng Xiang
Virginia Tech, USA

The origins of this book lie with my early years at Texas A&M University where in 1985 we designed something called the Texas Travel Research Information System (TTRIP), over twenty years of the research conducted by students and staff of the

National Laboratory for Tourism & eCommerce (NLTeC) and with the many researchers associated with the International Federation of Information Technology and Tourism (IFITT) and its annual ENTER conference. Indeed, the foundations of big data, smart systems, and tourism design were imagined by Clare Gunn and others long ago but now have been actualized by many scholars including Hannes Werthner, Arno Scharl, Matthias Fuchs, Wolfram Höpken, Zheng (Phil) Xiang some years ago, and others included in this book, wherein this work has coalesced into a defined field. In this acknowledgment, I would like to thank all the Ph.D. students associated with Texas A&M University and NLTeC during this time including Seong Il Kim, Wes Roehl, James Jeng, Christine Vogt, Kelly MacKay, Yeong-Hyeon Hwang, Ulrike Gretzel, Raymond Wang, Bing Pan, Dan Wang, Florian Zach, Sangwon Park, Jamie Kim, Jason Stienmetz, and Yeongbae Choe for all their hard work, creativity, and support and for their dedication to helping shape the future of tourism research. And, I would like to thank all my colleagues at IFITT and ENTER who I have had the privilege to meet and to learn from during this time. Last, I thank Phil for coordinating this particular volume and all the excellent scholars giving voice to the visions set forth so long ago.

Daniel R. Fesenmaier
The University of Florida, USA

Contents

Analytics in Tourism Design 1
Zheng Xiang and Daniel R. Fesenmaier

Part I Travel Demand Analytics

Predicting Tourist Demand Using Big Data 13
Haiyan Song and Han Liu

Travel Demand Modeling with Behavioral Data 31
Juan L. Nicolau

Part II Analytics in Everyday Life and Travel

Measuring Human Senses and the Touristic Experience: Methods and Applications 47
Jeongmi (Jamie) Kim and Daniel R. Fesenmaier

The Quantified Traveler: Implications for Smart Tourism Development 65
Yeongbae Choe and Daniel R. Fesenmaier

Part III Tourism Geoanalytics

Geospatial Analytics for Park & Protected Land Visitor Reservation Data 81
Stacy Supak, Gene Brothers, Ladan Ghahramani, and Derek Van Berkel

GIS Monitoring of Traveler Flows Based on Big Data 111
Dong Li and Yang Yang

Part IV Web and Social Media Analytics: Concepts and Methods

Sensing the Online Social Sphere Using a Sentiment Analytical Approach 129
Wolfram Höpken, Matthias Fuchs, Th. Menner, and Maria Lexhagen

Estimating the Effect of Online Consumer Reviews: An Application of Count Data Models 147
Sangwon Park

Tourism Intelligence and Visual Media Analytics for Destination Management Organizations 165
Arno Scharl, Lidjia Lalicic, and Irem Önder

Online Travel Reviews: A Massive Paratextual Analysis 179
Estela Marine-Roig

Conceptualizing and Measuring Online Behavior Through Social Media Metrics 203
Bing Pan and Ya You

Part V Case Studies in Web and Social Media Analytics

Sochi Olympics on Twitter: Topics, Geographical Landscape, and Temporal Dynamics 215
Andrei P. Kirilenko and Svetlana O. Stepchenkova

Leveraging Online Reviews in the Hotel Industry 235
Selina Wan and Rob Law

Evaluating Destination Communications on the Internet 253
Elena Marchiori and Lorenzo Cantoni

Market Intelligence: Social Media Analytics and Hotel Online Reviews 281
Zheng Xiang, Zvi Schwartz, and Muzaffer Uysal

Part VI Closing Remarks

Big Data Analytics, Tourism Design and Smart Tourism 299
Zheng Xiang and Daniel R. Fesenmaier

List of Contributors

Zheng Xiang is Associate Professor in the Department of Hospitality and Tourism Management at Virginia Polytechnic Institute and State University. His research interests include travel information search, social media marketing, and business analytics for the tourism and hospitality industries. He is a recipient of Emerging Scholar of Distinction award by the International Academy for the Study of Tourism and board member of International Federation for IT and Travel & Tourism (IFITT). He is currently Director of Research and Awards for the International Federation for IT and Travel & Tourism (IFITT).

Daniel R. Fesenmaier is Professor and Director of the National Laboratory for Tourism & eCommerce, Eric Friedheim Tourism Institute, Department of Tourism, Recreation and Sport Management, University of Florida. He is author, coauthor, and coeditor of several books focusing on information technology and tourism marketing including *Tourism Information Technology*. He teaches and conducts research focusing on the role of information technology in travel decisions, advertising evaluation, and the design of tourism places.

Gene Brothers, Ph.D. is Associate Professor in the Equitable and Sustainable Tourism Management Program at North Carolina State University in the USA. His career has been focused on university teaching, natural resource management, and destination planning. Over the years, his focus has evolved into a study of tourism resource management of both the natural and human dimensions of resource assessment, planning, and monitoring. A research thread which ties together his 37-year career is the evaluation of changes in destinations and the critical tourism metrics for assessment of these changes: tourism and destination analytics.

Lorenzo Cantoni graduated in Philosophy and holds a Ph.D. in Education and Linguistics. He is full professor at USI—Università della Svizzera italiana (Lugano, Switzerland), Faculty of Communication Sciences, where he served as Dean of the Faculty in the academic years 2010–2014. He is currently director of the Institute for Communication Technologies and scientific director of the laboratories webatelier.net, NewMinE Lab: New Media in Education Lab, and eLab: eLearning Lab. L. Cantoni is chairholder of the UNESCO chair in ICT to develop and promote sustainable tourism in World Heritage Sites, established at USI, and president of IFITT—International Federation for Information Technologies in Travel and Tourism. His research interests are where communication, education, and new media overlap, ranging from computer-mediated communication to usability, from eTourism to eLearning, and from ICT4D to eGovernment.

Yeongbae Choe is a Ph.D. Candidate in the Department of Tourism, Recreation & Sport Management at the University of Florida and works as a research assistant at the National Laboratory for Tourism & eCommerce and Eric Friedheim Tourism Institute, University of Florida. His research interest includes the role of ICT in travel decisions, tourist's decision-making process, smart tourism, and advertising evaluation. He has authored several research manuscripts published in internationally renowned peer-reviewed journals such as *Journal of Travel Research*, *Journal of Travel & Tourism Marketing*, *Tourism Economics*, *Asia Pacific Journal of Tourism Research*, and *Tourism Analysis*. He also received the Best Ph.D. Proposal Award from the International Federation for IT and Travel & Tourism (IFITT) and the Best Research Paper from the Academy of Global Hospitality and Tourism Conference (AGHTC).

Matthias Fuchs, Ph.D. is Professor of Tourism Management and Economics at the European Tourism Research Institute (ETOUR), Mid-Sweden University, Östersund, Sweden. Prior to this, he was the director of the *e-Tourism Competence Centre Austria* (ECCA). His main research areas include Electronic Tourism (i.e., mobile services, e-business readiness studies, online auctions, business intelligence, and data mining in tourism and destinations), destination management, destination branding, and tourism impact analysis. Matthias serves on the Editorial Board of the *Journal of Travel Research, Annals of Tourism Research*, and *Tourism Analysis*. He is also Associate Editor of the *Journal of Information Technology & Tourism.* Matthias is Education Director of IFITT (*International Federation for Information Technology and Travel & Tourism*) and has been Research Track Chair of the *ENTER Conference* 2012.

Ladan Ghahramani is a first-year Ph.D. student in Department of Parks, Recreation, and Tourism Management and Center for Geospatial Analytics at North Carolina State University, USA. Her doctoral research explores applying the emerging methodology to better understand when, where, how, and why visitors move throughout the cultural and natural heritage sites. Her work also includes understanding why and how site managers are integrating technology into their sites. She is eager to improve the experience of visitors and local communities to cultural and natural heritage sites through decreasing the negative sociocultural and environmental impacts of tourism applying education technology.

Wolfram Höpken is professor for Business Informatics and eBusiness at the University of Applied Sciences Ravensburg-Weingarten and director of the eBusiness Competence Centre eBLSIG. His main fields of interest are business intelligence and data mining, semantic web and interoperability, and mobile services. He has been involved in several research projects in the area of semantic web and seamless data interchange in tourism (EU-funded projects Harmonise, Harmo-TEN, Euromuse, and HarmoSearch) as well as in the area of knowledge discovery and management within tourism destinations. Wolfram Höpken has been vice president and commercial director of IFITT for 10 years. He has been research track chair of the ENTER conference 2009 and overall chair of ENTER 2014. He has chaired the CEN/ISSS workshop eTOUR dealing with harmonization in the field of tourism.

Jeongmi (Jamie) Kim is a Ph.D. candidate in Fox School of Business, Temple University, and a Visiting Scholar at NLTeC, the University of Florida. She is an active researcher with research interests in tourism experience, experience (service and place) design, information communication technology, and in situ measurements (e.g., mobile eye-tracker and EDA-based emotion recognition) and application. She worked for Korea National Tourism Organization for 9 years, managing international exhibitions, special events, online marketing, and contents development.

Andrei Kirilenko, Ph.D. is an associate professor at the Department of Tourism, Recreation, and Sport Management at the University of Florida. The area of Dr. Kirilenko's research is broadly described as interaction between humans and environment with concentration on the impacts of climate change and sustainability issues. He is especially interested in the research of social and mass media and big data analysis. His current research projects include (1) communication on mega-sports events in social networks; (2) public discourse on climate change in social media and newspapers; (3) people as sensors: flood monitoring through Twitter communication data mining; and (4) climate change, land-use change, and agriculture on the Northern Great Plains.

Lidija Lalicic is a Researcher and Lecturer at the Department of Tourism and Service Management at MODUL University Vienna. Her research interests are related to technology-enhanced tourist experiences, innovative marketing, and entrepreneurial practices in the field of tourism. For her Ph.D. dissertation (a three-paper design), she looked into various innovation opportunities for the tourism industry enhanced by social media. The dissertation sheds light on how tourism marketers can benefit from social media spaces in order to innovate and improve existing products or services. In particular, the dissertation provides an understanding of the usability of social media spaces for tourism marketers to engage their customers for innovation purposes.

Rob Law, Ph.D. is a Professor at the School of Hotel and Tourism Management, the Hong Kong Polytechnic University. He is also an Honorary Professor of several other universities. Dr. Law's research interests are information management and technology applications.

Maria Lexhagen is an Associate professor and the Head of the discipline of Tourism Studies at Mid Sweden University where she is also part of the Business Intelligence in Tourism group. She has a Ph.D. in business administration and tourism with a special interest in marketing and new technology. Her research covers business practice, destination management, and consumer behavior, and she has published internationally in both tourism journals and technology-focused journals. Her current research interests include the use, impact, potentials, and challenges with information technology in the tourism industry; destination management; branding and social media; as well as pop culture tourism induced by film, music, and literature.

Dong Li received his Ph.D. in Urban Ecology from the Chinese Academy of Science in 2008. Before he joined Beijing Tsinghua Tongheng Urban Planning & Design Institute (THUPDI) as deputy director of the newly established Technology Innovation Center in 2015, Dr. Li spent several years as a senior engineer in China Academy of Urban Planning & Design (CAUPD). His major research areas include environmental planning, infrastructure, and disaster prevention on urban and regional level. In recent years, he adopted the trend of data-driven planning, testing new data sources, algorithms, and tools for various issues in urban and regional planning. He strongly advocates a progressive transition from the traditional static paradigm in planning toward a more dynamic and integrated one in the new era of big data.

Han Liu is an Associate Professor in Quantitative Economic at Jilin University in China. He obtained his Ph.D. from the same University and is currently working as a Postdoctoral Fellow in the School of Hotel and Tourism Management at the Hong Kong Polytechnic University. His research interests are focused on tourism demand forecasting and has presented research papers at such international conferences as the 5th Conference of the International Association for Tourism Economics (IATE 2015) and the 2nd Global Tourism and Hospitality Conference Hong Kong 2016.

Elena Marchiori, Ph.D. is Postdoctoral Researcher and Lecturer at USI—Università della Svizzera italiana (Lugano, Switzerland), Faculty of Communication Sciences. She holds an M.Sc. in Media Management and a Ph.D. in Communication Sciences. She is the executive director of webatelier.net, the eTourism Lab at USI, and works for the Institute of Communication Technologies at USI. She is member of IFITT (International Federation for Information Technologies in Travel and Tourism) and general secretary of the IFITT Swiss Chapter. Her research interests are online tourism communication, reputation in online media, maturity of destinations and web adoption, and media effects.

Estela Marine-Roig is a Serra Húnter Fellow at the Faculty of Law, Economics and Tourism, University of Lleida, Catalonia, Spain, an Assistant Professor of Social Media and Smart Tourism at the Open University of Catalonia and a post-doctoral researcher in the GRATET research group of the Rovira i Virgili University, Catalonia. She holds a European PhD in Tourism and Leisure, an MSc in Tourism Management and Planning, a BA in Humanities, and a BA in Tourism. In 2015, the International Federation for Information Technologies and Travel & Tourism (IFITT) awarded her the Thesis Excellence Award for a Doctoral Thesis, and the Spanish Agency for Quality Assessment and Accreditation (ANECA) accredited her as Associate Professor in Social and Juridical Sciences in recognition of her academic career. Her research interests include the analysis of the image and identity of tourist destinations through tourism online sources, especially user-generated contents.

Thomas Menner successfully completed his master's degree in business informatics at the University of Applied Science Ravensburg-Weingarten, Germany. Within the fields of Business Intelligence and Data Mining, his main researches are Text Mining and more specific the field of Sentiment Analysis. At the moment, T. Menner participates in a touristic research project of Mid-Sweden University.

Juan L. Nicolau is a Full Professor of Marketing and Ph.D. in Economics and Business Administration. He is currently Dean of the Faculty of Economics and Business Administration at the University of Alicante. He has been visiting scholar at the National Laboratory for Tourism and eCommerce and at the Coggin College of Business (University of North Florida). He has won the Prize for Teaching Excellence as the best professor of the year awarded by the Valencian Regional Government and the University of Alicante and has received more than ten research prizes. He has published in *Strategic Management Journal*, *Omega*, *European Journal of Operational Research*, *Journal of Business Research*, *European Journal of Marketing, Economics Letters*, *Marketing Letters*, *Annals of Tourism Research*, *Tourism Management*, *Journal of Travel Research*, *Tourism Economics*, *International Journal of Hospitality Management*, *Journal of Hospitality & Tourism Research*, *Tourism Geographies*, *International Marketing Review*, Journal of Services Marketing, *Technology Analysis & Strategic Management*, and *Journal of Cultural Economics*.

Irem Önder is Assistant Professor at the Department of Tourism and Hospitality Management. She obtained her Ph.D. from Clemson University, South Carolina, where she worked as a research and teaching assistant from 2004 until 2008. She obtained her master's degree in Information Systems Management from Ferris State University, Michigan. She has two main research interests, which are information technology and tourism economics. Her specific information technology-related interests include social media, user-generated content, big data analysis, decision support systems, and online travel information search. Her tourism economics interests are about tourism forecasting, comparison of accuracy of various forecasting models, and city tourism.

Bing Pan, Ph.D. is Associate Professor in the Department of Hospitality and Tourism Management and Head of Research in the Office of Tourism Analysis within the School of Business at the College of Charleston, USA. He has published in the area of information technologies and their adoption in the hospitality and tourism industries. His research publications include using online data to understand, predict, monitor, and forecast tourism economic activities, tourist online behavior, social media, search engine marketing, and research methodologies. Dr. Pan has consulted with the Charleston Area Convention and Visitors Bureau for ten years.

Sangwon Park is a Senior Lecturer at the School of Hospitality and Tourism Management in the University of Surrey, UK. His research includes information search behaviors, travel decision-making process, hospitality and tourism marketing, and influence of information technology on travel behaviors.

Arno Scharl heads the Department of New Media Technology at MODUL University Vienna and is the Managing Director of webLyzard technology. Previously, he held professorships at the University of Western Australia and Graz University of Technology and was a Visiting Fellow at Curtin University of Technology and the University of California at Berkeley. Arno Scharl completed his doctoral research and habilitation at the Vienna University of Economics and Business. Additionally, he holds a Ph.D. from the University of Vienna, Department of Sports Physiology. He has authored more than 170 refereed publications and edited two books in Springer's Advanced Information and Knowledge Processing Series. His research interests focus on Web intelligence and big data analytics, human–computer interaction, and the integration of semantic and geospatial Web technology.

Zvi Schwartz, Ph.D. is a Professor of hotel management at Lerner's College of Business and Economics, University of Delaware. Prior positions include a Marriott Senior Faculty Fellow for Hospitality Finance and Revenue Management at Virginia Tech, associate professor at the University of Illinois, and over a decade of lodging industry experience as a manager and an entrepreneur. His scholarly research and industry consulting focus on the core technical and strategic elements of hospitality revenue management. He is a recipient of numerous research awards, including three times ICHRIE's best published paper of the year, and over $600,000 in research grants.

Haiyan Song is Chair Professor of Tourism in the School of Hotel and Tourism Management at the Hong Kong Polytechnic University. His research interests include tourism demand modeling and forecasting, impact assessment, and tourism supply chain management. He has published in such journals as *Annals of Tourism*

Research, *Tourism Management*, *Journal of Travel Research*, *International Journal of Forecasting*, *Journal of Applied Econometrics*, and *Tourism Economics*.

Svetlana Stepchenkova, Ph.D. is an assistant professor at the Department of Tourism, Recreation, and Sport Management at the University of Florida. The area of her research interests is destination management, marketing, and branding, with the focus on quantitative assessment of destination image and brand communications using unstructured and qualitative data. She is especially interested in influence of user-generated content and media messages on image formation and destination image as a factor in explaining destination choice. Svetlana also studies applications of information technologies in travel and tourism, particularly virtual travel communities, destination websites, and user-generated content as a means of obtaining a competitive advantage.

Stacy Supak is a Teaching Assistant Professor at North Carolina State University. Starting in 2014, she has held an appointment within the Center for Geospatial Analytics, where she teaches both graduate and undergraduate courses on Geospatial Information Science. She holds a Bachelor of Science in Environmental Civil Engineering from Columbia University, a Master of Science in Geophysics from the University of California at Santa Barbara, and a Ph.D. in the Department of Parks, Recreation and Tourism Management from North Carolina State University. This diverse background guides her current teaching and research interests including geographic information systems (GIS), spatial analysis, and geocomputational techniques applied for park and protected land management decision support as well as prospective visitor planning. She has previously published work on topics including geology, geology tourism, open-source web-based mapping, geospatial analytics, and marketing for tourism destinations.

Muzaffer Uysal is a professor in the Department of Hospitality and Tourism Management at Virginia Tech. He is a member of International Academy for the Study of Tourism, the Academy of Leisure Sciences. He is cofounder of *Tourism Analysis: An Interdisciplinary Journal* and sits on the editorial boards of more than ten journals, including *Journal of Travel Research* and *Annals of Tourism Research*. He has authored and coauthored numerous articles, monographs, and several books related to tourism research methods, tourist service satisfaction, tourism and quality of life, experience value in tourism, tourism-related scales, and management science applications in tourism and hospitality. Dr. Uysal has received a number of awards for research, excellence in international education, teaching excellence, and best paper awards. His current research interests focus on tourism demand/supply interaction, tourism development, and quality-of-life research in tourism.

Derek Van Berkel is a researcher investigating geospatial solutions to sustainability challenges, with a focus on modeling the social components of feedbacks between land-use change and the provision of ecosystem services across various landscapes. Derek received his Ph.D. in Environmental Spatial Analysis from the VU University Amsterdam where he developed methodologies for mapping and quantifying ecosystem services. There he also made use of geospatial visualization techniques for collaborative resource management and land-use model simulations of agricultural landscape change. He most recently held a postdoctoral position at the Department of Geography, the Ohio State University with the Appalachian Ohio Research group. This research employed socio-spatial techniques to examine forest return uncovering a complex socio-ecological system where diverse land management, environmental suitability, and regional political dynamics drive forest dynamics. Derek has published on rural development, ecosystem services, forest dynamics, agriculture, and tourism.

Selina Wan, D.HTM is an instructor in the Department of Marketing at the City University of Hong Kong. She is currently teaching a variety of marketing-related courses in the university, including Marketing, Consumer Behavior, Marketing Research, Public Relations, and Sustainable Business. Her extensive working experience in the hospitality and tourism industry ranges from a market analyst at New World Renaissance Hotel to senior researcher at Hong Kong Tourism Board. Her research interests are information and communication technology, sustainable tourism, destination marketing, and hospitality management. Dr. Wan received her doctorate degree from the School of Hotel and Tourism Management, the Hong Kong Polytechnic University.

Yang Yang, Ph.D. joined Temple University in 2013 from the University of Florida, where he earned his Ph.D. in Geography, with a minor in Econometrics, as well as two master's degrees in Statistics and Economics. A winner of four Best Paper Awards, Dr. Yang has published more than 20 English academic articles in top-tier peer-reviewed journals such as the *Annals of Tourism Research*, *Journal of Travel Research*, *Tourism Management*, and *International Journal of Hospitality Management*. He has also delivered 25 conference presentations globally and authored three book chapters. Dr. Yang has served as a reviewer for fourteen journals, including *Annals of Tourism Research* and *Journal of Travel Research*. His research interests include big data analytics in tourism and hospitality as well as location and financial analysis in the hospitality industry.

Ya You, Ph.D. is Assistant Professor of Marketing at the School of Business, College of Charleston. Her research interests focus on online word-of-mouth effectiveness and social media strategies. Her research has been published in *Journal of Marketing* and featured prominently in the book *Empirical Generalizations about Marketing Impact*, published by the Marketing Science Institute. In addition, her work has received extensive publicity in the business press, in outlets such as Science Daily and Phys.org.

Analytics in Tourism Design

Zheng Xiang and Daniel R. Fesenmaier

1 Introduction

In 2008 Chris Anderson, the American author and entrepreneur, made a bold claim in an article published in WIRED magazine that we are seeing the "end of theory" due to the deluge of data which will make conventional scientific methods obsolete. While his claim is extremely provocative and obviously debatable, Anderson challenged our understanding of the construction of knowledge, the processes of research, as well as how we should engage with the real world in the so-called era of Big Data. Big data is being generated at tremendous speed through numerous sources including Internet traffic, mobile transactions, online user-generated content, business transactions, various sensor systems embedded in the environment, as well as many operational domains such as finance and bioinformatics. Big data analytics, therefore, aims to discover novel patterns and business insights that can meaningfully and, oftentimes in real time, complement traditional approaches of research such as experiments, focus group studies and consumer surveys.

There is huge potential in developing big data analytics in travel and tourism. Particularly, as an experience-based product the design and development of tourism requires a profound understanding of what today's travelers need and want, how they move through and interact with physical and social spaces, and what leads to their enjoyment, happiness, and the realization of personal values. Increasingly, the focus on creating this knowledge is shifting toward the capabilities of capturing,

Z. Xiang (✉)
Department of Hospitality and Tourism Management, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
e-mail: philxz@vt.edu

D.R. Fesenmaier
National Laboratory for Tourism & eCommerce, Department of Tourism, Recreation and Sport Management, University of Florida, Gainesville, Florida, USA

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_1

storing, measuring, and interpreting data generated through different stages of the travel process in a timely fashion. As discussed in the first book of this series, today's tourism marketers and managers have increasingly realized the needs to make sense of the world and to design the tourism experience based upon scientific, data-driven approaches. In recent years we have seen advancements in several important areas of analytics, ranging from mapping the digital footprint of travelers to understanding their sentiments and preferences using online user-generated content, which can be best characterized as Analytics in Tourism Design.

Analytics in travel and tourism is its infancy and existing publications are scattered around fairly limited topics. In order to advance this line of research, this book brings together some of the leading authors with a variety of backgrounds, interests and expertise in data analytics to shed light on the nature, scope and characteristics of Analytics in Tourism Design. With this in mind, this very first chapter opens the discussion by introducing our readers to the foundations, needs, and research directions in the development of analytics in travel and tourism.

2 Foundations of Big Data Analytics

While there is a lack of formal definition, analytics is generally understood as the discovery and communication of meaningful patterns in data. Although conventional statistical tools are widely utilized, analytics often takes the form as a simultaneous combination of statistics, computer programming and data visualization to quantify findings to generate and communicate useful insights, predictions, and decisions for business problems. In many cases, analytics is connected with large quantities of data. The classic example is the pioneering study using Google search queries to identify pandemic diseases in the society (Ginsberg et al., 2009). As demonstrated by Ginsberg et al. (2009), analytics using large datasets can lead to an epistemological change which enables us to reframe key questions about the constitution of knowledge, the processes of research and how we should engage with reality (Boyd & Crawford, 2012). One of the application areas of growing importance is the so-called business intelligence in that big data analytics can be used to understand customers, competitors, market characteristics, products, business environment, impact of technologies, and strategic stakeholders such as alliance and suppliers. Many examples and cases illustrate the applications of big data analytics to discover and solve business problems (Mayer-Schönberger & Cukier, 2013). Importantly, although big data analytics does not preclude hypothesis testing, it is often applied to explore novel patterns or predict future trends (Aiden & Michel, 2014).

While it is to a great degree intended to address business needs (Chen, Chiang, & Storey, 2012), big data analytics has been propelled by the recent developments in computer engineering especially in areas such as data storage and access, machine learning, data mining, and data visualization. In particular, machine learning has progressed dramatically over the past two decades, from laboratory exploration to a practical analytical tool with widespread applications in both commercial and non-commercial domains (Jordan & Mitchell, 2015). For example, online interactions, mobile devices and embedded computing generate large amounts of data for us to understand human behavior, and machine-learning algorithms can be developed to learn from these data to customize products and services to the needs and circumstances of each individual. Any businesses or organizations with data-intensive issues such as customer relationship management and the diagnosis of problems in complex systems can benefit from the implementation of analytics with the aid of machine learning.

Another important driver of big data analytics is the development in computational linguistics, also known as natural language processing, which uses computational techniques to learn, understand, and produce human language content (Hirschberg & Manning, 2015). It is an increasingly critical component in big data analytics because it enables us to gain rich understanding of human experience within social contexts by applying text analysis to the rapidly growing social media sphere. Linguistic data available from social media sites such as Facebook, Twitter, blogs, and online review sites allow us to examine various aspects of human communication and behavior (Ruths & Pfeffer, 2014). And by combining Web crawling and natural language process with statistical and machine learning techniques, we are now able to track trending topics and popular sentiments, identify opinions and beliefs about products, predict disease or food-related illnesses spreading from symptoms mentioned in tweets, and identify social networks of people who interact together online. Social media analytics, therefore, aims to develop informatics tools to collect, monitor, summarize, and visualize social media data to extract useful patterns and business intelligence (Fan & Gordon, 2014). Due to its unique nature and characteristics of data, social media analytics can be applied throughout the product life cycle from need recognition, to design, to implementation, to its evaluation and redesign.

3 Analytics in Tourism Design: Needs and Opportunities

Tourism is an important component of many national and local economies. While the success of tourism management hinges on many policy and managerial areas, it is increasingly reliant upon a deep understanding of the ever-changing consumer behavior in order to mobilize necessary resources to satisfy their needs and wants. As discussed in the first book in this series, design science in tourism supports a framework for designing systems and artefacts to improve people's daily lives as well as their travel experiences. Different from conventional perspectives on product development, tourism design has the emphasis on a scientific, data-driven approach to supporting and enhancing the tourism experience. It has been widely documented that today's information technology, on the one hand, has fundamentally changed the way travelers access and consume tourism products; on the other hand, it has also generated new needs and opportunities for us to gain access to data and a better understanding of travel behavior (Gretzel, Sigala, Xiang, & Koo, 2015; Xiang, Schwartz, Gerdes, & Uysal, 2015). From this perspective, travel and tourism is a rich and also ideal domain for applications of big data analytics because the capabilities of any business or destination to capture, monitor, analyze, and interpret travelers' behaviors are critical.

Technology has transformed the tourism experience (Gretzel, Fesenmaier, & O'Leary, 2006). For example, MacKay and Vogt (2012) and Wang, Xiang, and Fesenmaier (2016) argue that our use of technology links our daily lives with the way we experience travel. Technology restructures the experience, as it is manifest in many ways, none more so than "travel in the network" as a metaphor to describe the ways today's travelers interact with various systems and environments (Gretzel, 2010). Importantly, technology-supported networks are social and community-based. Facebook, Twitter, Youtube, and Pinterest are quintessential Web 2.0 applications in that they are novel ways to facilitate exchange of information and social networking (Xiang & Gretzel, 2010). Technology-supported networks are also mobile with smartphones (and tablet computers) to facilitate transactions and strengthen traveler's social ties on the go. For many people, a mobile phone is far beyond a communication tool or an accessory to everyday living; in fact, the smartphone has become an inseparable part of one's life or even body (Turkle, 2011; Tussyadiah & Wang, 2014). Mobile technology arguably leads toward more hedonic and creative use; indeed, it has been argued that the development in location-based services (LBS) are making places more immersive and captivating for travelers (Hannam, Butler, & Paris, 2014). Geo-based technologies have been suggested to help tourists have more meaningful and even more playful experiences (e.g., in the form of location-based social gaming) (Tussyadiah & Zach, 2012). This suggests that change in the tourism experience as result of its interaction with information technology is multi-faceted and takes place within several technological and social domains. And, the change in the technological environment generates new needs and opportunities to understand and describe the conditions of travel and the tourism experience.

Information technology also directly leads to new patterns in travelers' decision making behavior (Wang and Xiang 2012, 2014). It is generally understood that the travel process involves three stages, i.e., pre-trip, en route and on-site, and post-trip, wherein the traveler engages with different activities in terms of information use and interaction with the environment. Recent studies find that use of the smartphone in travel is likely to change the travel experience by "unlocking" the three-stage process (Wang et al., 2016); that is, the tasks which tourists fulfill in the pre-trip and post-trip stages are now increasingly fulfilled in the en route and on-site stage due to the pervasive connection to the Internet using the smartphone, leading to several important behavioral changes. The level of decision making flexibility during the actual travel experience is likely to become higher in that the traveler can easily change the original plan due to the availability of new information. Eventually, travel activities may become more spontaneous, resulting in more unplanned trips or activities. Decision-making in the *en route* phase is dynamic in that it involves a number of interdependent decisions within which the contexts of later decisions are contingent upon earlier ones. Thus, the use of mobile devices such as smartphones changes the decision environment for *en route* and on-site decisions, especially when we consider the availability of search engines and social media (almost) anytime anywhere (Lamsfus, Wang, Alzua-Sorzabal, & Xiang, 2015). This challenges our conventional wisdom on travel decision making and requires a new set of analytical tools that can truthfully capture and measure the process and structure of travel behavior.

Further, technological innovations continue to emerge and requires new visions for tourism development (Gretzel et al., 2015). For example, the Internet of Things (IoT) signifies the pervasive presence around us of a variety of objects such as radio-frequency-identification (RFID) tags, sensors, actuators, mobile devices, etc., which are able to interact with each other and cooperate with their neighboring objects to achieve common goals (Atzori, Iera, & Morabito, 2010). These objects are connected to the Internet which consequently bridges the gap between the real world and the digital realm. Further, the development of mobile computing supports a plethora of applications by combing visual tagging of physical objects and near field communication (NFC) devices that contribute to the development of the IoT (Borrego-Jaraba, Ruiz, & Gómez-Nieto, 2011). Importantly, the emergence of the IoT provide a shift in service provision, moving from the current vision of always-on services, typical in the Web era, to always-responsive situated services, built and composed at run-time to respond to a specific need and able to account for the user's context. Thus, it is predicted that within the next decade the Internet will realize the vision long dreamed—a seamless fabric of classic networks and networked objects which can be identified, located, monitored, and managed anytime and anyplace. Content and services will all be around us, always available, paving the way to new applications and enabling new ways of working, interacting, entertainment, and living (Miorandi, Sicari, De Pellegrini, & Chlamtac, 2012).

This new technological infrastructure creates new connectivity and modalities of interaction within and outside travel and thus, is likely to impact on the way we understand the travel process (Xiang et al. 2015). As such, it is clear that advances in mobile, social, communication, and location-based technologies have augmented and mediated tourists' senses and experiences of place through emotional, aesthetical, informational, playful and social engagement, enabling tourists to be more creative and spontaneous (Richards, 2011; Wang et al., 2012). These recent developments require the formation of new models of travel behavior, new models for product design, and new models for research and evaluation which, in turn, establishes a new paradigm of tourism management.

4 Directions for Research

Analytics in tourism design supports a new type of inquiry into the nature and process of the tourism experience. There are many applications of analytics which give new meanings to travel and tourism. For example, "smart" systems will grow to be aware of, and be able to address, the contextual needs of the traveler in a pervasive yet non-intrusive way. Computer chips embedded in tourist attractions will enable tourism service providers to track tourists' locations and their behavior so that location-based services could be offered. Tourists can use their smartphones to explore the destination and events of interest using in-situ data collection and reporting. Online activities leave digital 'traces' resulting in rich multidimensional data which enable tourism organizations to develop new business models supporting traveler experiences. Within a social setting we will be able to collect and monitor information about 'events' of people and places which is gathered and uploaded to provide information about traveler. This implies that travel will no longer be an individual experience, but rather a shared experience wherein time, space, as well as interaction with one's physical environment is seamlessly (and instantly) distributed (i.e., shared with friends and colleagues) among many digitally connected social networks. More opportunities will unfold as we further engage in these inquiries to understand the market conditions as well as the true connections between the supply and demand of tourism.

The collection of chapters in this book reflects the cutting-edge research on the development of analytics in travel and tourism including new conceptual frameworks, new measurement tools, as well as applications and case studies for destination marketing and management. The chapters can be roughly grouped in to five parts. Part I, which can be called Travel Demand Analytics, focuses the attention on conceptualizing and implantation of travel demand modeling using big data. There are two chapters in this part with the first titled "predicting tourism demand using big data" by Haiyan Song and Han Liu, fills the void that there is very limited academic research has been conducted into tourism forecasting using big data due to the difficulties in capturing, collecting, handling, and modeling this type of data. To address these issues, a framework of tourism forecasting with big data is proposed. The second chapter, entitled "travel demand modeling with behavioral data" contributed by Juan Nicolau, discusses new developments and analytical approaches to travel demand modeling with behavioral big data, with the ultimate goal of generating customer-based knowledge through tourists' feedback and information traces. These two chapters illustrate new ways to identify, generate, and utilize large quantities of data in tourism demand forecasting and modeling. This part reflects the emerging tools which can be used to establish the link between demand and supply in tourism using large data (e.g., Yang, Pan, & Song, 2014).

Part II, also consisting of two chapters, can be characterized as Analytics in Travel and Everyday Life. This part focuses on the recent developments in wearable computers and physiological measurement devices and the implications for our understanding of on-the-go travelers and tourism design. The first chapter, entitled "the quantified traveler: implications for smart tourism development" by Yeongbae Cho and Daniel R. Fesenmaier, posits that technologies related to the quantified self, particularly wearable devices connected to the Internet, perfectly matches the needs of context-relevant information and therefore offer opportunities to create and shape tourism experiences. The second chapter, entitled "Measuring human senses and the touristic experience: methods and applications" by Jeongmi Kim and Daniel R. Fesenmaier, identifies emerging measurement techniques which enable researchers to examine the role of human senses in touristic experiences in natural environments. A human-centered approach for extracting contextual sense information using various wearable human-traits sensors is proposed to gain a better understanding of how a traveler creates touristic experiences. In this chapter, it is argued that capturing 'human sensing' data offers the potential to transform the way tourism researchers measure traveler's experiences and therefore design touristic environments.

Part III can be characterized as Tourism Geoanalytics consisting of two chapters. The first chapter, entitled "geospatical analytics using travel reservation data" by Supak, Brothers, Ghahramani and van Berkel, examines approximately 12.5 million reservation records from the US Parks and Protected Lands (PPL) database with 3272 distinct destinations between January 1, 2007 and December 30, 2015 to understand longitudinal destination usage attributes including total reservation count, median distance travelled by park users, media lead-time between order and start date, and cumulative nights of human occupancy, etc. This chapter summarizes literature related to geospatial analytics of PPLs, highlights ways to enrich PPL reservation data for enhanced analysis, and outlines how spatiotemporal databases could be used by Federal, State and County agencies tasked with tourism and resource management. The second chapter, entitled "GIS monitoring of traveler flows based on big data" contributed by Dong Li and Yang Yang, investigates the spatial patterns of Chinese domestic tourist flows during a major national holiday season. Geo-coded origin-destination information from a Chinese social media site similar to Twitter was collected and analyzed to create a dyadic matrix of inter-province tourist flows. The results show that social media data were highly correlated with tourism statistics published by official tourism administrations, and they highlight several factors that contribute to tourist flows as reflected in classic tourism geography literature.

Part IV, with five chapters on Web-Based and Social Media Analytics—Concepts and Methods, represents the recent developments in utilizing user-generated content on the Internet to understand a number of managerial problems. The chapter entitled "sensing the online social sphere—the sentiment analytical approach" by W. Höpken, M. Fuchs, Th. Menner, and M. Lexhagen, provides an overview of different approaches to tackle the problem of sentiment analysis and discusses current applications in the field of tourism. Each of the techniques are demonstrated and validated based on a prototypical implementation as part of a destination management information system for a leading mountain destination in Sweden. The second paper, entitled "estimating the effect of online consumer reviews: an application of a count data model" contributed by Sangwon Park, uses sample data from Yelp to examine the utilities of count models such as Negative Binomial regression in analyzing reviewer data. The third chapter, entitled "Tourism Intelligence and Visual Media Analytics for Destination Management Organizations" by A. Scharl, I. Önder and Lalicic, presents the structure and analytical framework of a tourism web intelligence platform that acquires, analyzes and visualizes Web-scale information flows in real time. The fourth chapter in Part IV, entitled "Online Travel Reviews: A Massive Paratextual Analysis" by Estela Marine Roig, presents an analytical framework for understanding the effects of paratextual features, i.e., an author's name, a title, a preface, illustrations alongside with online reviews. The fifth chapter, entitled "Conceptualizing and Measuring Online Behavior through Social Media Metrics", reviews and discusses measurement frameworks that connect online behavior to business performances in travel and tourism.

The next part (Part V) is a collection of case studies using Web-Based and Social Media Analytics. The first chapter, entitled "Sochi Olympics on Twitter: geographical landscape and temporal dynamics" by A. Kirilenko and S. Stepchenkova, focuses on Twitter as a new medium and investigates how mega events such as the Sochi Olympics were portrayed on Twitter by hosts and guests in terms of geographical representation and salient topics before, during, and after the event. The second chapter, entitled "leveraging online reviews in the hotel industry" written by S. Wan and R. Law, reviews literature on issues related to the use of online reviews as well as their impact on hotel performance. The successful and poor responses of hotel management to online reviews are presented to highlight the best practices in enhancing hotel guest experiences and reputation management. The next chapter, entitled "Evaluating destination communications on the Internet" by E. Marchiori and L. Cantoni, provides an overview of different approaches for the evaluation of destination communications on the Internet. In particular, it proposes two analytical frameworks, namely UsERA—User Experience Risk Assessment Model, and DORM—Destination Online Reputation Model. The last chapter, entitled "market intelligence with online hotel reviews" contributed by Z. Xiang, Z. Schwartz and M. Uysal, applies several dimensions related to hotel guests' experiences in relation to satisfaction ratings developed based upon a large quantity of online reviews to the hotel market in the United States. The results clearly show that the market can be segmented into distinct value categories based upon these factors. These chapters, collectively, describe a range of different approaches to understanding market dynamics in the tourism and hospitality industries.

With this introduction we hope our readers will have a better understanding of the foundations, needs, as well as possible research directions in analytics in tourism design. As can be seen from this collection of research ideas and case studies, analytics in tourism design does not always engage with the so-called big data. However, these chapters clearly demonstrate the growing importance of new data sources, new measurement tools, and emerging frameworks that enable us to discover meaningful patterns in travel behavior. This does not necessarily suggest that theory is dead as proclaimed by Anderson (2008); rather, it signifies new ways to engage with travel behavior and tourism experiences, particularly in its interface with today's information technology and new media. As such, it is hoped that the following chapters help to inspire you to appreciate the growing opportunities to engage with analytics in tourism design.

References

Aiden, E., & Michel, J.-B. (2014). The predictive power of big data. *Newsweek.* Retrieved April 22, 2014, from http://www.newsweek.com/predictive-power-big-data-225125

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired.* http://www.wired.com/2008/06/pb-theory/

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks, 54*(15), 2787–2805.

Borrego-Jaraba, F., Ruiz, I. L., & Gómez-Nieto, M. Á. (2011). A NFC-based pervasive solution for city touristic surfing. *Personal and Ubiquitous Computing, 15*(7), 731–742.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74–81.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Gretzel, U. (2010). Travel in the network: Redirected gazes, ubiquitous connections and new frontiers. In M. Levina & G. Kien (Eds.), *Post-global network and everyday life* (pp. 41–58). New York: Peter Lang.

Gretzel, U., Fesenmaier, D. R., & O'Leary, J. T. (2006). The transformation of consumer behaviour. In D. Buhalis & C. Costa (Eds.), *Tourism business frontiers* (pp. 9–18). Oxford: Butterwork–Heinemann.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets, 25*(3), 179–188.

Hannam, K., Butler, G., & Paris, C. M. (2014). Developments and key issues in tourism mobilities. *Annals of Tourism Research, 44*, 171–185.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Lamsfus, C., Wang, D., Alzua-Sorzabal, A., & Xiang, Z. (2015). Going mobile defining context for on-the-go travelers. *Journal of Travel Research, 54*(6), 691–701.

MacKay, K., & Vogt, C. (2012). Information technology in everyday and vacation contexts. *Annals of Tourism Research, 39*(3), 1380–1401.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* New York: Houghton Mifflin Harcourt.

Miorandi, D., Sicari, S., De Pellegrini, F., & Chlamtac, I. (2012). Internet of things: Vision, applications and research challenges. *Ad Hoc Networks, 10*(7), 1497–1516.

Richards, G. (2011). Creativity and tourism: The state of the art. *Annals of Tourism Research, 38*(4), 1225–1253.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064.

Turkle, S. (2011). *Life on the screen*. New York: Simon and Schuster.

Tussyadiah, I. P., & Wang, D. (2014). Tourists' attitudes toward proactive smartphone systems. *Journal of Travel Research, 0047287514563168.*

Tussyadiah, I. P., & Zach, F. J. (2012). The role of geo-based technology in place experiences. *Annals of Tourism Research, 39*(2), 780–800.

Wang, D., & Xiang, Z. (2012). The new landscape of travel: A comprehensive analysis of smartphone apps. *Information and Communication Technologies in Tourism, 2012*, 308–319.

Wang, D., Park, S., & Fesenmaier, D. R. (2012). The role of smartphones in mediating the touristic experience. *Journal of Travel Research, 51*(4), 371–387.

Wang, D., Xiang, Z., & Fesenmaier, D. R. (2014). Adapting to the mobile world: A model of smartphone use. *Annals of Tourism Research, 48*, 11–26.

Wang, D., Xiang, Z., & Fesenmaier, D. R. (2016). Smartphone use in everyday life and travel. *Journal of Travel Research, 55*(1), 52–63.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*(2), 179–188.

Xiang, Z., Schwartz, Z., Gerdes, J., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*(1), 120–130.

Xiang, Z., Wang, D., O'Leary, J. T., & Fesenmaier, D. R. (2015). Adapting to the Internet: Trends in travelers' use of the web for trip planning. *Journal of Travel Research, 54*(4), 511–527.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447.

Part I
Travel Demand Analytics

Predicting Tourist Demand Using Big Data

Haiyan Song and Han Liu

1 Introduction

Big data is one of the most popular and most frequently used terms to describe the exponential growth and availability of data in the modern age, which is likely to be maintained or even accelerate in the foreseeable future (Hassani & Silva, 2015). It is a broad term for datasets that are so large in size or complex that traditional data processing applications and software tools are inadequate to capture, curate, manage, and process the data within a reasonable period of time (Snijders, Matzat, & Reips, 2012). There are challenges regarding the analysis, capture, search, sharing, storage, transfer, visualization, and information privacy of big data, and these challenges require new technologies to uncover hidden values from large datasets that are diverse, complex, and massive in scale (Hashem et al., 2015). Big data brings new opportunities to modern society (Fan, Han, & Liu, 2014) since these vast new repositories of information can provide researchers, managers, and policymakers with the data-driven evidence needed to make decisions on the basis of numbers and analysis rather than anecdotes, guesswork, intuition, or past experience (Frederiksen, 2012), and it may lead to more accurate analysis, more confident decision-making, and greater operational efficiencies, cost reductions, and risk reductions (De Mauro, Greco, & Grimaldi, 2015).

Nowadays, people try to use the insights gained from big data to uncover new opportunities for their businesses (Mayer-Schönberger & Cukier, 2013). The process of discovering and determining insights from large, complex, and unstructured datasets attracted our attention. So, what is big data? There is no unified definition

H. Song (✉)
The Hong Kong Polytechnic University, Hung Hom, Hong Kong
e-mail: haiyan.song@polyu.edu.hk

H. Liu
Jilin University, Jilin, People's Republic of China

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_2

of big data. The basic definition is “datasets which could not be captured, managed, and processed by general computers within an acceptable scope” (Chen, Mao, & Liu, 2014). More and more researchers and institutes are exploring the characteristics of big data in order to define it. These definitions always include the characteristics of volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Laney (2001), for example, used the above “3V’s” model to define big data. In this model, volume means that with the generation and collection of masses of data, the scale of the data becomes increasingly big; velocity means that the collection and analysis of big data must be rapidly and timely conducted so as to maximally utilize its commercial value; and variety indicates the various types of data, which include semi-structured and unstructured data, such as audio, video, web page, and text data, as well as traditional structured data. Beyer and Laney (2012) updated the definition of big data by adding another “V”: veracity. Chen, Mao, Zhang, and Leung (2014) added “value” (huge value but very low density) to make the definition perfect. Recently, a consensual definition was produced: “Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value” (De Mauro et al., 2015).

Big data is not simply defined by the 4V’s: it is about complexity. Beyond the definition of big data, we should be concerned about the details of it. Hashem et al. (2015) classified big data into five categories: data sources, content format, data stores, data staging, and data processing. In each category, there are numerous subcategories, as shown in Fig. 1. In this chapter, we focus on tourism forecasting using big data, and we will therefore pay special attention to the data sources, data staging, and data processing categories.

2 What Is Tourism Big Data?

The tourism industry thrives on information (Benckendorff, Sheldon, & Fesenmaier, 2014; Poon, 1988). The vast new big data repositories of information—far greater than what is captured in standard databases—can provide researchers, managers, and policymakers with the data-driven evidence needed to make decisions on the basis of numbers and analysis rather than anecdotes, guesswork, intuition, or past experience (Frederiksen, 2012). The bounty of tourism big data has the potential to deliver new and more highly informed inferences about human activity and behavior that will give the tourism industry a big boost and benefit not only customers but also those who participate in the tourism industry (Fuchs, Höpken, & Lexhagen, 2014).

Travelers leave different digital traces behind on the Web when using mobile technologies. Through every traveler, large amounts of data are available about anything that is relevant to any travel stage: prior to, during, and after travel (Hendrik & Perdana, 2014). Most of this data is of an external nature: for example, in the form of Twitter or other social networking feeds. Due to the large amounts of

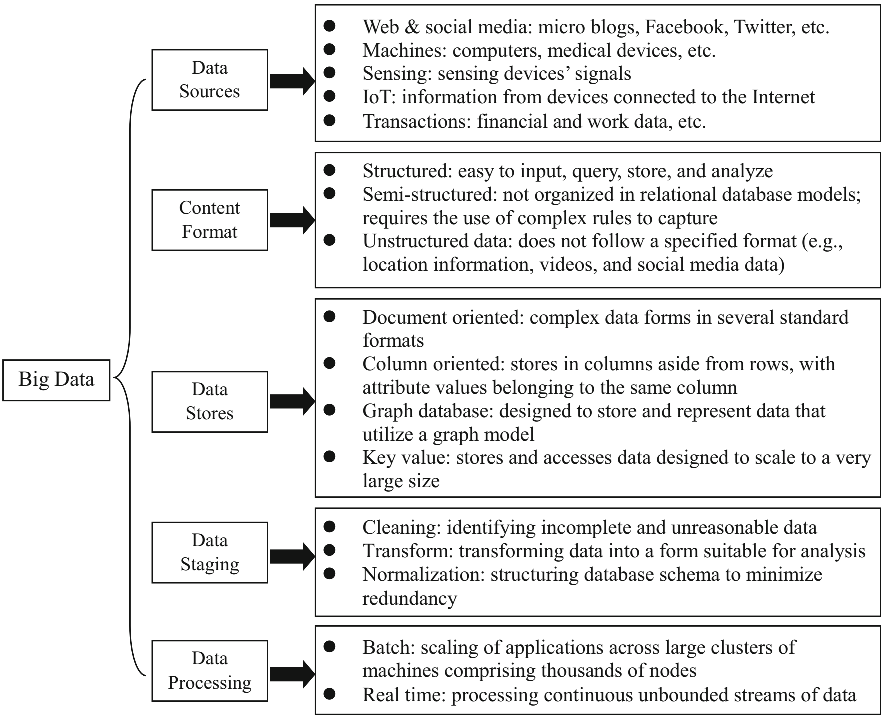

Big Data
Data Sources
● Web & social media: micro blogs, Facebook, Twitter, etc.
● Machines: computers, medical devices, etc.
● Sensing: sensing devices' signals
● IoT: information from devices connected to the Internet
● Transactions: financial and work data, etc.
Content Format
● Structured: easy to input, query, store, and analyze
● Semi-structured: not organized in relational database models; requires the use of complex rules to capture
● Unstructured data: does not follow a specified format (e.g., location information, videos, and social media data)
Data Stores
● Document oriented: complex data forms in several standard formats
● Column oriented: stores in columns aside from rows, with attribute values belonging to the same column
● Graph database: designed to store and represent data that utilize a graph model
● Key value: stores and accesses data designed to scale to a very large size
Data Staging
● Cleaning: identifying incomplete and unreasonable data
● Transform: transforming data into a form suitable for analysis
● Normalization: structuring database schema to minimize redundancy
Data Processing
● Batch: scaling of applications across large clusters of machines comprising thousands of nodes
● Real time: processing continuous unbounded streams of data

Fig. 1 Big data classification (Hashem et al., 2015)

available data stored in the cloud, analytics are needed in order to make sense of the information within the data. If you are a potential customer planning a trip, you probably get more than a little help from the Internet when you are searching for inspiration, buying tickets, reserving accommodation, or researching attractions. Participants in the tourism industry are increasingly turning to big data to discover new ways to improve decision-making, opportunities, and overall performance (Irudeen & Samaraweera, 2013): for example, big data can be used to interconnect the dispersed information from different systems and then improve decision-making capability.

Big data provides unprecedented insights into customers' decision-making processes by allowing companies to track and analyze shopping patterns, recommendations, purchasing behavior, and other drivers that are known to influence sales. Agencies and merchants involved in tourism can find innovative ways to use a variety of data resources to connect with potential visitors at every stage of a trip and use these big data sources to better and timely understand the fastest growing visitor demographics. They can also remarket to target shoppers who have looked at a specific destination on an online travel agency website (Sust et al., 2014). Through the use of big data, industries become more efficient. More and more companies have started specializing in the storage and evaluation of the large amounts of data on travelers' hotel stays, purchase transactions, and customer information in order to provide more efficient and high quality services.

3 Advantages of Using Big Data in Tourism

We are confident that consumers and tourism product providers will see the benefits of using big data. Personalized marketing and targeted product designs are extremely powerful opportunities for both groups. It is crystal clear that big data can provide better, targeted, and profitable services and products to consumers (Pries & Dunnigan, 2015). For instance, big data analysts can capture information of consumer interests from photos posted on Facebook or other social networks (e.g., a tourism provider could push information about local biking destinations or biking clubs when they obtain a picture of a mountain bike).

Previous studies on tourism have mostly been based on surveys or experts' views, which mean that they have taken samples from the population as a whole and do not have real data about all tourists. In contrast, one study on tourism big data tried to introduce data based on real actions by all users instead of drawing information from survey samples (Irudeen & Samaraweera, 2013). In this chapter, we introduce a framework that incorporates big data produced by tourists themselves (e.g., through mobile phones connecting to the telecom network or bank cards connecting to POS terminals) that increases knowledge of the industry's target market into tourism demand forecasting.

Tourism big data using innovative methods has advantages over traditional methodologies, as discussed below.

(1) Reliability
Big data are based on users' real actions, not on surveys. In other words, real actions have been analyzed rather than stated intentions or answers to questions. Taking all information sources together, it can be stated that big data increases the sample base on which conventional research tends to be based by several orders of magnitude (Meeker & Hong, 2014). The reliability of big data analysis allows us to consider all aspects of the information in order to provide comprehensive results instead of biased conclusions due to information loss in the sample data.

(2) New information flows
Tourism big data is a type of information produced by tourists themselves; it enriches the knowledge of tourism businesses' target market and is very useful for analyzing the consumers' demand for different tourism products and services (Hendrik & Perdana, 2014). Since tourism big data are structured and repositioned data, it is possible to cross-reference them with other sources such as social media and open public data, whether these are sources currently in production or potential information sources that may be created or released in the future. The analysis of tourism big data can be contrasted with internal data

from each tourism business with a view to determining whether the supply of tourism products/ services in each area of a city is in tune with the tourists who demand for these products and services.

(3) Real-time data and nowcasting
One of the innovative uses of big data is nowcasting, that is, the use of real-time data to describe contemporaneous activities before official data sources are made available (Bollier & Firestone, 2010): for example, Varian (2014) argued that real-time Google search queries are a good way to nowcast consumer activities, as the contemporaneous correlation analysis obtained from the Google Correlate data is still a 6-week lead on reported values. A notable example of using Google search queries for nowcasting is Google Flu Trends, which identifies possible flu outbreaks 1–2 weeks earlier than the official health reports by tracking the incidence of flu-related search terms in the Google search engine.

There are many studies that have used structured search-engine data for tourism nowcasting and forecasting (Artola, Pinto, & Pedraza, 2015; Bangwayo-Skeete & Skeete, 2015; Yang, Pan, Evans, & Lv, 2015). Besides search engine queries, there are other types of real-time data streams that can be assembled and analyzed: for example, data on credit card purchases, the trucking and shipping of packages, and mobile phone usage are all useful bodies of information. Much of these data is becoming available on a near real-time basis, which can be used to predict the macro data that will be compiled at some point in the future (Jeng & Fesenmaier, 2002; Yang, Pan, & Song, 2014).

The ultimate objective of using real-time big data is to develop applications that are able to respond as soon as the economic pulse has been taken and provide suggestions; of course, this should be done under controlled conditions and be capable of being switched on and off at any time.

4 Characteristics of Tourism Big Data

Having scoured the literature and found the 4V's characteristics of big data, we used these and added another V (value) to ascertain the unique characteristics of tourism big data.

(1) Volume
Volume always seems to top the list of big data characteristics, and is a key contributor to the problem of why traditional relational database management systems fail to handle big data (Prajapati, 2013). The volume of tourism big data always comes from points of sales or other traditional channels of distribution (i.e., call centers, websites, premises, newsletters, customer relations, etc.). The content of tourism big data is created on a daily, or even hourly, basis, and we are interested in making sense of the information, transforming big data into smart data and then using it for tourism planning.

(2) Variety

Another key characteristic of big data, both in terms of cost and ease of use, is the variety of data that stems from all accessible technologies. Variety describes the different formats of data that do not lend themselves to storage in structured relational database systems. The formats of big data include a long list of data such as documents, e-mails, text messages, images, graphs, videos, and the output from all types of machine-generated data from cell phones, GPS signals, sensors, machine logs, and DNA analysis devices (Li, Jiang, Yang, & Cuzzocrea, 2015). This type of data is characterized as unstructured or semi-structured and has always existed. 80 % of tourism-relevant information originates in unstructured form, and organizations can only count on the 20 % of structured data: for example, property management systems (PMS), Web or blog content management systems (CMS), or customer relationship management (CRM) systems can only deal with structured data, while the data on customer preferences at various points of contact are in the form of unstructured or semi-structured data, which require novel technologies to analyze them in order to develop new or improved products and services.

(3) Velocity

The third key characteristic of big data is velocity, which is referred to as the speed of responsiveness. There are three important aspects of the velocity of tourism big data (Chen, Mao, Zhang, et al., 2014). The first aspect is the consistent and complete capture, storage, and analysis of the fast moving streams of big data: for example, the stream of readings taken from a sensor or the weblog history of page visits and the clicks by each visitor to a website. The second aspect is the characteristics of timeliness or latency. We should capture, store, and use big data within a certain lag time depending on the type of the information since some of the data are permanently valuable while some would no longer be meaningful after a very short period of time. The third aspect is the speed with which big data must be stored and retrieved; the architecture of capture, analysis, and deployment must support real-time turnaround (in this case, fractions of a second); and must do this consistently over thousands of new customers. In tourism, for instance, we are concerned about how to send the right offer to the right person at the right moment when he or she arrives at a destination and what you should do if someone checks in to your hotel and is disappointed with the room and decides to tweet about it rather than call the front desk. Take the airlines in the travel business as an example, the dynamic revenue management could make a timely price change according to complex algorithms based on real-time or near-real-time customer online behaviors.

(4) Veracity

Veracity means the truthfulness and accuracy of data given the context, the variety of communication “touch points”, and the speed at which things happen. Big data veracity refers to the biases, noise, and abnormality in data: Is the data being stored and mined meaningful to the problem being analyzed? Compared with volume and velocity, veracity in data analysis is the biggest challenge. In developing a big data strategy, you need your team and partners to help you keep your data clean and to have processes to keep "dirty data" from accumulating in your systems.

(5) Value

Value is frequently seen as another important characteristic of big data. The value of tourism big data can be described by its novel application in the tourism industry. First, there is the personalized application of tourism big data. Personalized marketing and targeted product design are extremely powerful opportunities that can be obtained from big data (Jani, Jang, & Hwang, 2014). Using a series of interviews conducted within the travel industry, Radovich (2015) showed how big data can be used to increase impact and reduce friction across disciplines, both within a company and within the industry. Personalization is a key tenet of big data. In order to most effectively win at true personalization, large travel companies must work across information databases to gather the myriad data points created by a consumer at different points. The second valuable application of tourism big data is the customer-centric experience. The customer should be at the center of all big data efforts. If big data gathering is seen as creepy or invasive, the consumer will not be pleased and loyalty will be lost. However, all signs point to consumers being willing to accept big intrusions into their behaviors if the resulting product is more targeted and able to anticipate their needs throughout.

5 Benefits of Big Data to Tourism Businesses

Big data analysis is changing all sorts of industries, not just the usual retail, logistics, and high-tech industries. It is also transforming the worlds of hospitality and travel since hospitality and tourism companies deal with a slew of user data covering all sorts of different information (e.g., flight confirmations or a customer's room preferences), and it creates all sorts of opportunities for correlating data to find otherwise unknown insights (Turner, 2014). In addition, there are some significant changes for big data because the cost of analytics platforms keeps dropping and employees are becoming more familiar with what big data can do. Essentially, big data can be used to tailor marketing campaigns and find business model inefficiencies. Big data analysis can deliver much needed business insights and can be the source of innovation for tourism organizations and the industry in general. The potential for big data in tourism is huge, and tourism organizations should not underestimate its importance (Pries & Dunnigan, 2015).

With the right approach, the tourism industry can learn a lot about consumer preferences and use this information and insight to build connections with individual travelers. Being able to offer travelers the right service or product at the right time is crucial. Without the right information and a very good targeting strategy, advertising will not result in any conversions and there will be no value. Travel is a fast-paced industry, and this drives the need for speedy data analytics and quick decisions. In tourism, any demand needs to be addressed instantly in order to remain relevant to travelers, and this is what makes big data so important. With the vigorous growth of the amount and applications of big data, traditional tourism data and methods are going to be interfacing with the novel data and methodologies: for example, call centers are going to be interfacing with online consumer reviews; loyalty programs are going to be linking with booking histories; and "property preferences" are going to be combined with social media chatter.

5.1 Consumer Behavior

We are in a time of unprecedented flux in consumer behavior, customer expectations, and company business models created by technologies that is simultaneously disrupting established businesses and spawning new ones (Marko, 2015; McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). However, tourism big data show significant changes in the relationship between businesses and their customers. So, we can use big data to provide superior buying and support experiences with a view to enhancing customer choice and expectations. The catalyst of using big data to re-recognize consumer behavior is the pervasive use of mobile devices, apps, and other social media, which play an ever-increasing role in the collection of raw information and easy access to the relevant tasks at hand.

Big data holds many insights into customers' behavior, some of which is already delivering while others yet to be realized. The potentials created by big data is particularly acute in retail since industries and business processes can successfully exploit new communication channels, service delivery options, and unprecedented sources (Marko, 2015). Collecting, correlating, and analyzing tourism big data from customer interactions across channels is the key to transforming the customer experience from a nightmare to nirvana (Chase, 2013). The nexus between big data and machine learning in all its forms, including predictive analytics and even neural network deep learning, is the foundation of well informed, highly efficient, and deeply satisfying interactions that benefit both customers and businesses.

The aim of using tourism big data is to create an authentic emotional connection between customers and partners of the tourism industry in order to make a significant improvement in customer service and support. The exploration of tourism big data has huge implications and provides opportunities for the seamless meshing of consumer experiences across mobile devices, websites, and personal interactions using multiple communication channels (e.g., phone, instant messaging, e-mail, web chat, and social networks). The key to the goal of using tourism big data is to be proactive, not just provide an integrated service. We need to anticipate customer needs and prevent problems. In other words, we can anticipate problems and queries by using statistical modeling and forecasting before we are asked the same question or asked to explain the same situation again and again.

5.2 Feedback Mechanisms

Feedback in the tourism industry is important in the quest to identify customer preferences and deliver positive experiences. Soliciting customer feedback is one of the most important elements in achieving high company growth and building a strategy around better meeting customer needs. Feedback based on tourism big data from customers, employees, partners, suppliers, and communities has also improved the capabilities of big data analytics. Data-driven business and consumer apps are the most common ways to collect feedback anytime and anywhere. A growing set of cloud services gives us the immediate and ubiquitous ability to interact using smart phones, tablets, or even watches (Chen, Mao, Zhang, et al., 2014).

The increase in gathering feedback using modern techniques has led to traditional feedback marketing being progressively replaced by commercial messages which are quick, unique, focused, and personal. One of the applications of the feedback mechanisms applied by the providers of tourism-related goods and services is price adjustment, in which a change in travel demand obtained from big data analysis and forecasting can provide useful information for quick and effective price adjustment.

Machine learning is one of the main technical methods used in the tourism industry to construct the feedback mechanism between customers and tour operators (Bajari, Nekipelov, Ryan, & Yang, 2015): for example, through cooperation between tour services providers, financial institutions, and telecom operators, machine learning can identify whether a person has just changed his/her residential address or travel internationally through checking for unusual charges. Machine learning with big data on customer experience can enable travel businesses and tour operators to proactively send text messages or calls to customers with new offers after they purchased their services. Specifically, machine learning could modify the feedback system by identifying the attempted user tasks and measuring their rates of success. Using this information, tourism businesses could then provide solutions to process inefficiency, customer frustration, and cross-channel breakdowns.

Predictive analytics are often presented as a cure-all for companies and can be incredibly useful. The predictive analytics with tourism big data used in modern feedback mechanisms represent a major improvement over old-fashioned human feedback. Predictive analytics can give marketing professionals more insight into customer preferences, which can be used to understand customers better and improve sales (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). However, the success of predictive analytics depends on both the quality of the big data and the customer feedback mechanisms.

Customer feedback mechanisms must be well designed and comprehensive to deliver actionable data in a timely fashion and acted upon immediately. Timely and reliable tourism big data can provide a rich portrait of customers and potential customers and subsequently lead to marketing efforts with advertising dollars being more precisely targeted toward the most fruitful channels. The framework shown above converts feedback into data that can be incorporated into broader analytics.

6 How to Use Big Data in Tourism Forecasting

We now turn to the key step of using big data in tourism forecasting, since we know that big data could bring many benefits to the tourism industry.

6.1 Capturing Big Data for Tourism Forecasting

Companies that effectively capture and implement big data strategies gain a competitive advantage since the technology required to process big data is a hindrance for many business users because of its complexity and cost. There are several steps in the process of capturing big data before we use it.

(1) Objective
The first step is the objective of using big data, which is to make sure that business benefits are derived from it (Pellegrini, 2013). When we capture big data, we should be able to access it and know what is available and determine where the business value lies. In other words, we should know the capability of big data and exactly what we are looking for and look to see what its values are. It is important to set specific business goals rather than just dealing with the big data itself.

(2) Visualizing big data
The second step is to make the big data visible to users within a company/organization. This will enable tourism forecasters to determine the optimal quantities of a product and to adjust logistical processes to maximize efficiency (Weiler & Black, 2014). The purpose of data visualization is to find the ways in which data could be effectively collected from different sources (visual and non-visual) and presented so that users could easily understand them. This will also help forecasters to better utilize big data in fulfilling their forecasting tasks.

(3) Structuring big data
The third step is to structure the unstructured data. This means to arrange big data according to traditional data length and format so that they can be fitted neatly into rows and columns in the spreadsheet. Structured data generally resides in a relational database and, as a result, is sometimes called relational data (Akerkar, 2013). The unstructured data can be easily mapped into predesigned fields: for example, a call center's structured data include numbers, dates, and groups of words and numbers called strings. It is commonly agreed that this kind of data accounts for about 20 % of the total amount of big data. Unstructured data are very difficult to analyze, since most of the big data is

unstructured or semi-structured data that contains a wealth of valuable information and does not fit into predefined data models. Thus, a number of different software solutions have been designed to search unstructured data and extract important information. In this chapter, we use pre-cleaned structured big data for tourism forecasting.

7 Selecting and Shrinking Big Data

Big data contains lots of information, which creates not only a storage issue but also a massive analysis problem. How to use these large datasets is the biggest problem in tourism forecasting using structured big data. The two most popular methods used in selecting and shrinking large amounts of structured data are the factor and LASSO (least absolute shrinkage and selection operator) modeling approaches.

(1) The factor model
The factor model is the most commonly used method in selecting and shrinking structured big data. A number of recent studies in the economics literature have focused on the usefulness of factor models in the context of forecasting related to the use of large datasets (Bai & Ng, 2006; Bańbura & Rünstler, 2011; Forni, Giannone, Lippi, & Reichlin, 2009; Hallin & Liška, 2011; Schumacher & Breitung, 2008; Stock & Watson, 2002; Stock & Watson, 2006; Teixeira, Klotzle, & Ness, 2008). We particularly analyze the predictive benefits associated with the use of dimension reducing independent component analysis (ICA) and sparse principal component analysis (SPCA), coupled with a variety of other factor estimation and data shrinkage methods, including, amongst others, bagging, boosting, and the elastic net. To assess the success of using big data, we could carry out a forecasting "competition" involving the estimation of different baseline model types, each constructed using a variety of specification approaches, estimation approaches, and benchmark econometric models (Stock & Watson, 2012).

(2) The LASSO method
The LASSO method is a covariates selection method in a linear regression framework (Tibshirani, 1996). It operates by penalizing the optimization problem associated with the regression with a term that involves the L1-norm of coefficients. It belongs to the family of penalized regression models that involve performing least squares with some additional constraints on the coefficients, the L1-norm in the case of LASSO. The literature has shown that LASSO tends to have a lower misspecification risk in forecasting models when compared with the usual information criteria (Ng, 2012). The LARS method (Efron, Hastie, Johnstone, & Tibshirani, 2004) can be combined with the factor model to shrink large datasets and used for forecasting economic series (Bai & Ng, 2008; Bessec, 2013; Schumacher, 2010).

8 A Framework for Predicting Tourism Demand Using Big Data

There is a widespread belief that big data can aid the improvement of forecasts provided we can analyze and discover hidden patterns and that predictions can be improved through data-driven decision-making (Shi, 2014). Some researchers believe that data mining techniques can be exploited to help forecasting with big data (Rey & Wells, 2013; Varian, 2014). However, data mining techniques always use static data as opposed to time series and are seldom used in tourism demand forecasting. When we turn to traditional forecasting methods for tourism demand forecasting with big data, the biggest problem is that the traditional forecasting tools cannot handle the size, speed, and complexity inherent in big data (Madden, 2012) even when it has been structured.

In order to apply a traditional forecasting method to big data, we have to simplify the structured big data (Hassani & Silva, 2015). One of the solutions is to shrink the big data and get the most important information in a suitable format that can be easily applied to the traditional forecasting model. Factor models are the most common and popular statistical and data mining technique used for big data forecasting; neural networks and Bayesian models are two other popular choices. In this chapter, we focus on the factor models.

(1) Mixed frequency model with big data
There has been some research success using big data for tourism forecasting. Choi and Varian (2012) aggregated Google data for Hong Kong's tourism demand forecasting and suggested that Google Trends' data about a destination may be useful in predicting visits to that destination. Yang et al. (2015) used web search query volume to predict visitor numbers for a popular tourist destination in China, and their results showed a significant decrease in forecasting errors when search engine data were used. However, these studies always aggregated or ignored weekly observations in order to make the datasets suitable for the traditional forecasting methods. Choi and Varian (2012), for example, only used the first two weekly observations of the month, discarding information for the latter 2 weeks, to predict total monthly visitors. Yang et al. (2015) aggregated weekly search engine data for forecasting. As a matter of fact, these researches could be improved by using a novel forecasting method, the mixed-data sampling (MIDAS) approach (Ghysels, Santa-Clara, & Valkanov, 2005), to fully utilize the high frequency search engine data (Bangwayo-Skeete & Skeete, 2015). Another mixed frequency model that fulfills the mixed frequency data job is the mixed frequency VAR model (Kuzin, Marcellino, & Schumacher, 2011; Qian, 2010), which treats low frequency data as high frequency data with missing data and then uses the state space model to deal with it.

(2) Factor model and forecasting combination

As a matter of fact, the best way to forecast the low frequency series (such as tourism demand) using high frequency data is to combine the shrinkage method with the mixed frequency models. Some studies used the mixed frequency model with factor high frequency data to forecast the macroeconomic indicators and obtained improved forecasting performance (Frale & Monteforte, 2011; Kuzin et al., 2011; Marcellino & Schumacher, 2010). The existing literature shows that compared with single model forecasts, forecast combination can improve forecasting accuracy in many practical situations (Bates & Granger, 1969; Chu, 1998; Coshall & Charlesworth, 2011; Deutsch, Granger, & Teräsvirta, 1994; Stock & Watson, 2004). In order to reduce the risk of forecasting failure (Wong, Song, Witt, & Wu, 2007), we suggest using forecast combination after obtaining different forecasting results from different methods and data.

Figure 2 displays the framework of tourism forecasting with big data. There are three important steps: (1) data exploration, which is the data processing that prepares the proper data for the model; (2) use modeling techniques to predict user behavior on the basis of their previous business transactions and

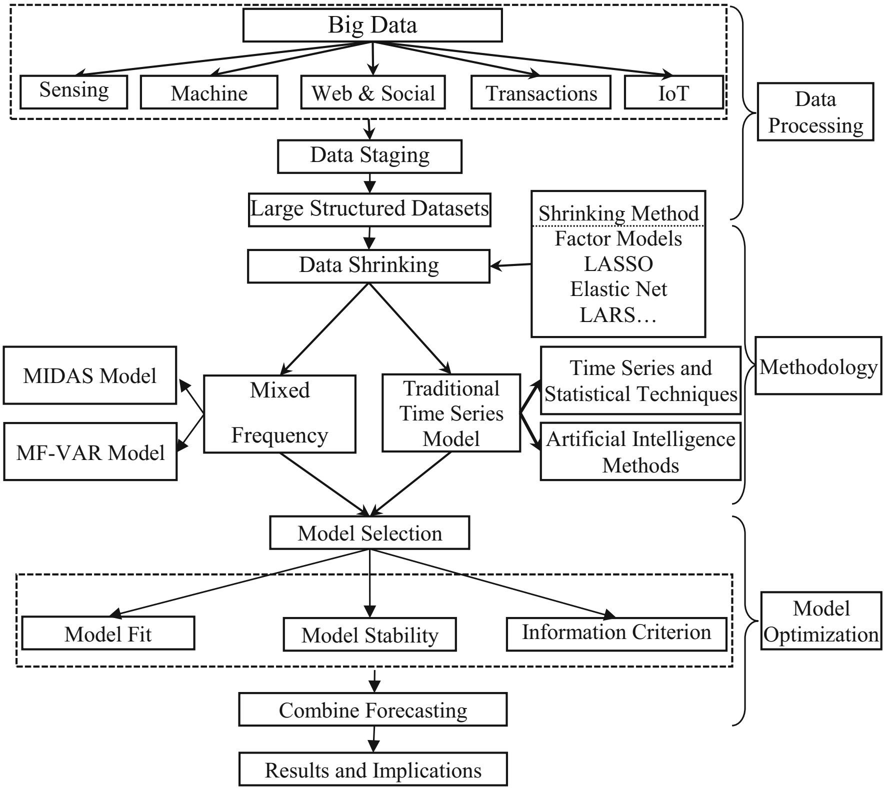


Fig. 2 The framework of tourism forecasting with big data

preferences; (3) optimize the forecast results and decrease the forecast failure risk by model selection and combination forecasting.

9 Conclusions

Big data is a social, cultural, technological, and ethical phenomenon that is not all good, all bad, or consistently neutral. With the proliferation and explosive increase in the application of big data, it has become a common tool in corporate decisions and a number of new social perils have arisen. At the same time, as data technologies become more pervasive, there are also privacy concerns and the potential for the abuse and misuse of big data (Bollier & Firestone, 2010). The use of tourism big data for forecasting has some visible and hidden pitfalls (Chareyron, Da-Rugna, & Raimbault, 2015). There are questions about the stability of the analysis and interpretation when the tools and techniques that we used in analyzing the big data have changed: Can the patterns that emerged from big data analysis or forecasting be generalized? How can information and privacy be controlled when anything and everything is systematically counted and recorded? In other words, there are challenges when tourism big data is used for forecasting. The first challenge is the difficulty of identifying the right data and determining how to best use it. The second challenge is to find the right talent capable of both working with the new technologies and interpreting the data to find meaningful business insights, and the third is to overcome the obstacle of data access and connectivity, which requires the right platforms to aggregate and manage big data. The fourth problem is how to find new ways of leveraging big data. The final concern is the security of big data and how to keep the advantage of using such data.

There are many potential solutions to overcome these challenges. First of all, the results of big data forecasting must promptly meet the need of business decisions. The purpose of tourism forecasting is to find and analyze the relevant data quickly and accurately. Visualization is a good way to present results and help those involved in tourism to make rapid decisions. We can also explore huge data volumes and gain business insights in near real time by improving the hardware and forecasting models. The second solution is to gain an overall understanding of the big data, which is crucial for visualizing and interpreting the data. To be specific, we need to have a deep understanding of where the data come from, what audience will be consuming the data, and how that audience will interpret the information. It is worth noting that outliers are important for tourism; therefore, we should pay more attention to the distribution and pattern of outliers and identify their influence. A third solution is to proactively take advantages of big data, as most of the information contained in big data is real time and huge in volume. Hence, the timely use of big data for forecasting and decision-making using proper approaches and methods is the best way to capitalize the benefits of big data.

All in all, the use of big data in the tourism and hospitality industry is still in its infancy, but the potential growth in application is huge. There is a lot of behind-the-scenes work to be done, including sequencing for synchronous and asynchronous events and computing elapsed times of clusters of events, latency, and time between events, before big data results are presented to users. Fortunately, solutions for big data are emerging and the costs are much lower than before. In our opinion, the use of big data by airlines, restaurants, hotels, and other tourism and hospitality related industries enables them to learn a great deal about customers' preferences on the macro level and to benefit a lot with relatively small investment in the near future.

References

Akerkar, R. (2013). *Big data computing*. Boca Raton, FL: CRC Press.

Artola, C., Pinto, F., & Pedraza, P. D. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*(1), 103–116.

Bai, J., & Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-Augmented regressions. *Econometrica, 74*(4), 1133–1150.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics, 146*(2), 304–317.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review, 105*(5), 481–485.

Bańbura, M., & Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting, 27*(2), 333–346.

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454–464.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Or, 20*(4), 451–468.

Benckendorff, P. J., Sheldon, P. J., & Fesenmaier, D. R. (2014). *Tourism information technology*. Wallingford: Cab International.

Bessec, M. (2013). Short-term forecasts of French GDP: A dynamic factor model with targeted predictors. *Journal of Forecasting, 32*(6), 500–511.

Beyer, M. A., & Laney, D. (2012). *The importance of 'big data': A definition*. Stamford, CT: Gartner.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data*. Washington, DC: Aspen Institute.

Chareyron, G., Da-Rugna, J., & Raimbault, T. (2015). *Big data: A new challenge for tourism*. Paper presented at the 2nd IEEE International conference on big data.

Chase, C. W. (2013). Using big data to enhance demand-driven forecasting and planning. *Journal of Business Forecasting, 32*(2), 27–32.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171–209.

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big data: Related technologies, challenges and future prospects*. Heidelberg: Springer.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88* (s1), 2–9.

Chu, F.-L. (1998). Forecasting tourism: A combined approach. *Tourism Management, 19*(6), 515–520.

Coshall, J. T., & Charlesworth, R. (2011). A management orientated approach to combination forecasting of tourism demand. *Tourism Management, 32*(4), 759–769.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings, 1644*(1), 97–104.

Deutsch, M., Granger, C. W., & Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting, 10*(1), 47–57.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32*(2), 407–499.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*(2), 293–314.

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the black box: Structural factor models with large cross-sections. *Econometric Theory, 25*(5), 1319–1347.

Frale, C., & Monteforte, L. (2011). *FaMIDAS: A mixed frequency factor model with MIDAS structure*. Bank of Italy Temi di Discussione (Working Paper) No.788.

Frederiksen, L. (2012). Big data. *Public Services Quarterly, 8*(4), 345–349.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations—A case from Sweden. *Journal of Destination Marketing and Management, 3*(4), 198–209.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics, 76*(3), 509–548.

Hallin, M., & Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics, 163*(1), 29–41.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.

Hassani, H., & Silva, E. (2015). Forecasting with big data: A review. *Annals of Data Science, 2*(1), 5–19.

Hendrik, H., & Perdana, D. H. F. (2014). Trip guidance: A linked data based mobile tourists guide. *Advanced Science Letters, 20*(1), 75–79.

Irudeen, R., & Samaraweera, S. (2013). *Big data solution for Sri Lankan development: A case study from travel and tourism.* Paper presented at the 2013 International Conference on Advances in ICT for Emerging Regions, ICTer 2, Colombo.

Jani, D., Jang, J. H., & Hwang, Y. H. (2014). Big five factors of personality and tourists' internet search behavior. *Asia Pacific Journal of Tourism Research, 19*(5), 600–615.

Jeng, J., & Fesenmaier, D. R. (2002). Conceptualizing the travel decision-making hierarchy: A review of recent developments. *Tourism Analysis, 7*(1), 15–32.

Kuzin, V., Marcellino, M., & Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the Euro area. *International Journal of Forecasting, 27*(2), 529–542.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety*. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–31.

Li, K.-C., Jiang, H., Yang, L. T., & Cuzzocrea, A. (2015). *Big data: Algorithms, analytics, and applications*. Boca Raton, FL: CRC Press.

Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 3*, 4–6.

Marcellino, M., & Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics, 72*(4), 518–550.

Marko, K. (2015). Using big data and machine learning to enrich customer experiences. http://www.forbes.com/sites/kurtmarko/2015/04/08/big-data-machine-learning_customer-experience/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 61–67.

Meeker, W. Q., & Hong, Y. (2014). Reliability meets big data: Opportunities and challenges. *Quality Engineering, 26*(1), 102–116.

Ng, E. C. (2012). Forecasting US recessions with various risk factors and dynamic probit models. *Journal of Macroeconomics, 34*(1), 112–125.

Pellegrini, T. (2013). The economics of big data: A value perspective on state of the art and future trends. In R. Akerkar (Ed.), *Big data computing* (pp. 343–371). Boca Raton, FL: CRC Press.

Poon, A. (1988). Tourism and information technologies. *Annals of Tourism Research, 15*(4), 531–549.

Prajapati, V. (2013). *Big data analytics with R and Hadoop*. Birmingham: Packt Publishing.

Pries, K. H., & Dunnigan, R. (2015). *Big data analytics: A practical guide for managers*. Boca Raton, FL: CRC Press.

Qian, H. (2010). *Vector autoregression with varied frequency data*. https://mpra.ub.uni-muenchen.de/34682/

Radovich, A. (2015). *Big data is fundamental in the hospitality and travel industry*. http://cliintel.com/big-data-is-fundamental-in-the-hospitality-and-travel-industry/

Rey, T., & Wells, C. (2013). *Integrating data mining and forecasting*. New York: OR/MS Today.

Schumacher, C. (2010). Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters, 107*(2), 95–98.

Schumacher, C., & Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting, 24*(3), 386–398.

Shi, Y. (2014). Big data: History, current status, and challenges going forward. *Bridge, 44*(4), 6–11.

Snijders, C., Matzat, U., & Reips, U.-D. (2012). Big data: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science, 7*(1), 1–5.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association, 97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*(6), 405–430.

Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. In C. W. J. G. G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 515–554). North Holland: Elsevier.

Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics, 30*(4), 481–493.

Sust, V. O., Illera, E. G., Berengué, A. S., García, R. G., Alonso, M. V. P., Torres, M. J. T., et al. (2014). *Big data and tourism: New indicators for tourism management*. http://telefonicacatalunya.com/wp-content/uploads/2014/05/BIG-DATA-Y-TURISMO-eng-interactivo.pdf

Teixeira, M. F., Klotzle, M. C., & Ness, W. L. (2008). Determinant factors of Brazilian country risk: An empirical analysis of specific country risk. *Brazilian Review of Finance, 6*(1), 49–67.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological), 58*(1), 267–288.

Turner, E. (2014). *How big data is changing the travel industry*. http://blog.sumall.com/journal/big-data-changing-travel-industry.html

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–27.

Weiler, B., & Black, R. (2014). *Tour guiding research: Insights, issues and implications*. Bristol: Channel View Publications.

Wong, K. K. F., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tourism Management, 28*(4), 1068–1078.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447.

Travel Demand Modeling with Behavioral Data

Juan L. Nicolau

1 Introduction

Today's travelers demand personalized and comprehensive experiences, and guided by their personal motivations, they try to back their decisions on recommendations expressed on the Internet. Besides, they write on official and unofficial websites their personal preferences, and tell other travelers about their intentions on their next destinations, plan the itinerary of the visit, compare, make reservations and pay with a few clicks from home just seating at their computer. Also, with their cell phones they build a unique story with pictures and comments on what they see and feel while at the destination. Fuchs, Abadzhiev, Svensson, Höpken, and Lexhagen (2013) indicate that this customer-generated data can be divided into explicitly-provided information through the use of surveys and e-reviews or implicitly-given information via information traces such as Internet-navigation data, online requests, booking, payment data, or tourists' spatial movements through sat navs; distinguishing between structured data (e.g. surveys) and unstructured data (e.g. e-reviews with free text) (Höpken, Fuchs, Keil, & Lexhagen, 2011). Only if firms and analysts were able to manage this amount of information—structured and unstructured—they could identify consumers' preferences and, more importantly, anticipate their decisions to adapt companies' services in real time and in a personalized way (Invat·tur Report, 2015). Certainly, in tourism more than in any other industry, the 3Vs of big data reflect purely the essence of the intricacies that entail managing such plethora of information: in line with Dolnicar and Ring (2014), "Big Data implies the availability of significantly larger, often gigantic, amounts of data (volume) on a continuous basis and often in real time (velocity) from a range of diverse data sources (variety)".

J.L. Nicolau (✉)
University of Alicante, Alicante, Spain
e-mail: jl.nicolau@ua.es

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_3

Observe that one of the great opportunities that Big Data offers the tourism industry resides in the Smart Cities and more specifically in smart destinations, which facilitate the experience and interaction between the destination and the tourist by ensuring sustainable development. As the production and consumption take place simultaneously in tourism, it means that information is being massively produced through all the stages the tourist is going through. Also, one of the great appeals of massive data is its potential to predict phenomena, anticipate behavior, expectations and future needs of tourists, so that smarter and safer business decisions are made (Invat·tur Report, 2015). For example, adjusting prices quickly and competitively in response to an analytically predictable change in travel demand represent an edge over rivals.

It is evident that these advantages come with challenges. While most companies have myriads of data (surveys, internal customer transaction data, quality data (complaints), secondary research reports about trends and markets, and online data), there is a strong need to coordinate all the information sources and put together the data in a manageable way. When it comes to a destination level, this challenge is even more acute as huge volumes of data on customer transactions, needs and behaviour are stored by different stakeholders of the destination (Vriens & Kidd, 2014). In their knowledge destination framework architecture, Fuchs et al. (2013) emphasize the fact that different data sources require different techniques for the data extraction, so that heterogeneous data from distinct data sources should be mapped into a homogeneous data format.

Another challenge is analysis. Be it through methods of data mining (e.g. techniques of machine learning and artificial intelligence) or traditional methods (e.g. regression-like procedures), detection of patterns and relationships in the data is not fully guaranteed unless some empirical-related issues are considered when the modeling of the travel demand is carried out. In the increasing use of Big Data in tourism research (Mellinas, Martínez María-Dolores, & Bernal García, 2015), the review of the literature identifies three relevant aspects of demand analysis (Radojevic, Stanisic, & Stanic, 2015): (1) tourist heterogeneity; (2) the ability to identify all the alternatives available to the tourists when they make their choices; and (3) the inherently hierarchical character of the data at the destination level (e.g. hotels are nested within destinations, destinations within countries).

2 Empirical Results

Vriens and Kidd (2014) outline the key areas where advanced analytics derived from Big Data can provide solutions with special added value. These are market forecasting (especially if a firm operates in multiple markets), quantifying customer needs and motivations (with an emphasis on quantitatively determined emotional states which leads to an improved ability to understand customer needs), analyzing drivers of brand share (e.g. the predictive power of brand perceptions), product and pricing optimization (to find the best mix of attributes to optimize volume, share or profitability), marketing efficiency modelling (to detect how well marketing efforts are working), and customer dynamics (e.g. which customers are most likely to defect and when, or how to determine lifetime value of customers). However, in all cases, when modeling individual behavior with big data three issues are to be considered: heterogeneity, choice set and information hierarchy.

2.1 Heterogeneity in Tourists

The existence of strong heterogeneous demand looking for product and service provision adapted to its specific needs, along with the intensification of competition in the market, has led to heterogeneity identification becoming fundamental to the marketing strategies of organizations and tourism destinations. As the heterogeneity of the market reflects the existence of a diversity of needs and desires and, therefore, of differentiated consumer behaviour among individuals, understanding heterogeneity in tourist preferences is of paramount importance in many tourism marketing actions. Strategically, knowing the distribution of people’s responses to destination attributes would guide the design decisions of the tourism products (this insight would not be detected if the preference is observed only at the mean). Operationally, modeling individual-level responses to marketing actions allows tourist firms to adjust allocation of resources across regions, establishments, and tourists.

Despite the fact that segmentation allows the definition of different market segments that group consumers with shared behaviour and needs, nowadays there is more and more importance attached to personalised service for each client. More pro-active consumers and an intense competition increase the demand for better service, better adapted to their individual needs and, therefore, personalized. Tourists expect to be treated as individual clients. This situation leads to the appearance of one-by-one marketing, which entails individual consideration of consumers and a one-by-one service. This approach is the basic pillar of relationship marketing -and, therefore, the application of CRM (*Customer Relationship Management*)-, which is designed to create, strengthen and maintain relationships between companies. Mass marketing has been transformed into fragmented or micro-segmented marketing to satisfy the demands of smaller and smaller segments, even down to the level of the individual customer. So, the key question in the context of Big Data is how to analyze and detect individual preferences of tourists by introducing heterogeneity.

Tourists process and integrate information to choose an alternative (e.g. destination, type of accommodation or method of transport) that maximises their utility. The objective or subjective character with which the researcher examines the result of this choice process determines the different approximations of choice analysis. The study of tourist behaviour and, therefore, of the way in which they process, evaluate and integrate the information used to make a decision, is traditionally made in two ways. The first approximation is centred on the analysis of the *real choices* made by individuals. This approach is based on the Neoclassical Economic Theory and the Theory of Discrete Choice, and assumes the existence of *preferences* that are unobservable to the analyst but that tourists implicitly consider when ranking alternatives, and which are only *revealed* through the real purchase choice. Therefore, this approximation is known as the *Revealed Preferences* approach.

The second approach examines the ranking or scoring according to *preferences*, given by individuals to hypothetical choice alternatives. This approximation is based on the Information Integration Theory and the Social Judgement Theory, and assumes that the decision maker is capable of ranking alternatives according to his/her preferences. In contrast to the previous case, the analyst does not observe the real purchase choice, given that the individual only makes a *declaration of intent* based on their preferences (i.e. which alternative would be chosen if they had to choose from the given possibilities). This approximation, therefore, is known as the *Stated Preferences* approach.

To give an example, an individual declares that Hawaii is the destination he/she would like to go to on his/her next holiday. In other words, the individual selects Hawaii from a series of destinations and, through this *declaration*, preferences are analysed. However, this aspect has been widely criticized, due the fact that this approach does not reflect reality in the sense that the declaration of the preferred alternative of an individual does not necessarily coincide with his/her real behaviour, i.e. with the alternative that is really chosen. The fact that an individual *declares* that he/she would like to go to Hawaii on his/her next summer holiday does not necessarily mean that he/she will go there in the end.

Conversely, the *Revealed Preferences Approach* analyses the real choices made by tourists in order to obtain their preferences. In the example above, the individual *reveals* his/her preferences when, from a group of destination choices, he/she chooses and goes to Hawai. However, one of the weak points of the *Revealed Preferences Approach* derives from the fact that the estimation of preferences is made at a global sample level, which does not allow representation of individual level preferences. If U_{in} is the utility of alternative i for tourist n, explained through the personal characteristic x_n of individual n and through attribute z_i of the same alternative i, then the utility function is expressed as

$$U_{in} = \alpha_i + x_n\beta_i + z_i\gamma_i + \varepsilon_{in}$$

where α_i is the utility constant, β_i and γ_i are the parameters that measure (respectively) the effects of characteristic x_n of the individual and attribute z_i on the utility of alternative i and ε_{in} is the error term.

Specifically, β_i and γ_i represent the marginal utilities of individuals of alternative i; and these parameters allow us to answer questions such as "If a destination improves one of its attributes (for example, the quality and cleanliness of its water), to what extent would preferences for this destination increase?" The value of this tool for the decision making of tourism organisations is unquestionable, as it allows them to know the responses of a series of people to this improvement.

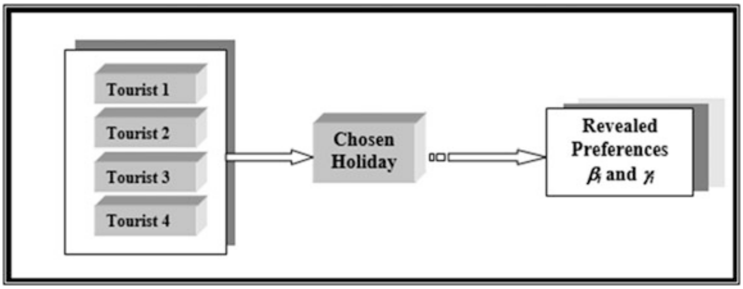


Fig. 1 Linking *sample revealed preferences* through choices made

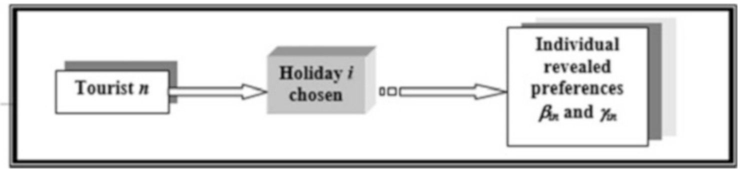

Tourist *n*
Holiday *i* chosen
Individual revealed preferences β_{in} and γ_{in}

Fig. 2 Individual revealed preferences through choices made

However, note that the estimations of parameters β_i and γ_i are made at the global sample level (see Fig. 1).

What if the estimation of these parameters could be made tourist by tourist? This way, the resulting equation would be

$$U_{in} = \alpha_i + x_n\beta_{in} + z_i\gamma_{in} + \varepsilon_{in}$$

where, in this case, β_{in} and γ_{in} represent the preferences of tourist n around alternative i. Note that now we obtain a parameter for each tourist (and not for the whole sample) (see Fig. 2).

The main implication of knowing the tourist by tourist preference structure is that it allows the adaptation of each product to each individual, as well as the formation of groups of individuals with similar preferences.

In the context of Big Data, most user-generated data is observed data, so revealed preferences can be obtained through this modeling; thus, the introduction of the heterogeneity of tourist preferences into the analysis of the choice process is a major issue.

One of the procedures proposed in the literature to incorporate heterogeneity of preferences assumes the existence of differentiated response parameters for each individual. The most used models in this approach are the random effects models, which model heterogeneity with the assumption that the coefficients of the utility functions of each individual vary according to the probability distribution, either continuous -which gives rise to the Random Coefficients Logit Model- or discrete -which leads to the Latent Class Logit Model-. Initially, the Latent Class Logit Model has been widely accepted in the literature due to the fact that the estimation of the *mass probabilities* -or points where the distribution reaches the greatest *probability masses* allows identification of *latent segments* in the market, which are represented by groups of individuals with similar response profiles. Moreover, in order to segment the market, discrete distribution has an advantage over continuous distribution in that there is no need to assume a concrete probability distribution, as the segments are obtained through empirical data. However, the discrete approach has two important limitations (Allenby & Rossi, 1999): (1) the estimation becomes complex with six or more *mass probabilities*, which hinders the capture of the complete sample heterogeneity; and (2) the impossibility of identifying the preferences of individuals situated beyond a certain threshold of the distribution function (e.g. in the distribution tails).

Because of this, some authors consider that the optimum method of capturing market heterogeneity is to estimate the parameters of each individual, as this allows the capture of any individual preference structure (Allenby & Rossi, 1999). In fact, this model has enough flexibility to provide a tremendous range within which to specify individual unobserved heterogeneity. This flexibility can even offset the specificity of the distributional assumptions.

2.2 Choice Set

One major issue in demand analysis is the definition of the choice set, that is, the alternatives from which the tourist selects the preferred option. The analyst is always uncertain about the set of alternative that the individual considered when making the decision. Obviously, the more data exists, the more alternatives, and the more the potential error of omitting alternatives considered by the tourist but not regarded by the analyst will increase. Therefore, when using Big Data, not only is important to collect information on the selected alternative, but also on the whole choice set. So, in the analysis of Big Data of hotels and airlines, all alternatives on the "screen" presented to the customer should be stored in a database. The first challenge here is to store the data. The size of the data increases 50-fold if the choice set has 50 alternatives. After storing the data in json files, the next challenge is how to analyze it. Bookings with choice sets can be used in discrete choice models. The fact that there are individuals who have been presented with different sets of alternatives can be easily managed with these models (Train, 2009).

2.3 Information Hierarchy

In the context of Big Data, researchers examine data from multiple entities (e.g. hotels and destinations). Certainly, this type of data is inherently hierarchical, as hotels are nested within destinations, and destinations within countries, and ignoring this effect might reduce the validity of results and conclusions (Radojevic et al., 2015). In an attempt to mimic and reflect the way people process information, hierarchical decision processes should be considered when analyzing travel demand. This statement is based on the idea that, when confronted with many alternatives, people tend to follow strategies of the "satisficing" type (satisfice = *satis*fy + suf*fice*), as defended by Simon (1955), where alternatives are considered sequentially. This proposal is further backed by: (1) The Associative Network Theory (Collins & Loftus, 1975) which, through "cognitive networks", explains the way the information on alternatives is represented, processed and activated in consumers' memory through nested links. Specifically, this theory proposes that information is held in the memory through an interrelated structure of "cognitive networks", in which each cognitive network has various "nodes" and "links" between different nodes. (2) The Cybernetic model of decision making (Steinbruner, 2002), which explains how the consumer can follow a hierarchical choice process to reduce uncertainty and complexity in the decision task. Destination choice has numerous factors for consideration and problems related with available information, so they are inclined to use a hierarchical strategy for their choice to reduce uncertainty to a certain manageable level.

Radojevic et al. (2015) use a four-level mixed linear model with random intercepts for country of origin, destination, and hotel, so that the implicit hierarchy is considered in the analysis; and Park, Nicolau, and Fesenmaier (2013) propose the Destination Advertising Response (DAR) model in which they examine the advertising information effects on a sequential travel decision process, including different travel products advertised. Specifically, the choice in the first stage is between visiting and not visiting a destination. Once individuals decide to visit a tourism destination in the first stage, those travelers go on to a second stage where they make a decision whether or not to purchase advertised items. People who select advertised items in the second stage go on to a third stage choice among six different advertised items.

Let us imagine a group of people has to decide the hotel where they are staying. Accordingly, the previous sections show the issues that must be considered when modelling this decision. First, as not all people behave the same way, it means that their preferences are dissimilar or, more formally, there is heterogeneity. Second, if each of them has searched the hotel availability in different periods of time (say, different days), the set of alternatives (e.g. types of hotels) that each individual has been confronted with might be different. Third, before selecting the hotel, they have had to choose the destination; and even before, they have had to decide on whether they take a vacation or not; thus, a hierarchical structure is implied.

With regard to heterogeneity of preferences, a choice model that allows the coefficients of the preferences to vary over tourists is required. Therefore, the utility of an alternative i for tourist t is defined as $U_{it} = X_{it}\beta_t + \varepsilon_{it}$ where X_{it} is a vector that represents the attributes of the alternative and the characteristics of tourists; β_t is the vector of coefficients of these attributes and characteristics for each individual t which represent personal tastes; and ε_{it} is a random term that is iid extreme value. This utility specification leads to a Random Coefficient Logit Model (RCL) in which its coefficients β_t vary over tourists with density *g(β)*. Thus, the non-conditional probability is the integral of $P_t(i/\beta_t)$ over all the possible values of β_t:

$$P_i = \int_{\beta_t} \frac{\exp\left\{\sum_{h=1}^{H} x_{ih}\beta_{th}\right\}}{\sum_{j=1}^{J_t}\exp\left\{\sum_{h=1}^{H} x_{jh}\beta_{th}\right\}} g\left(\beta_t \middle| \theta\right) d\beta_t \tag{1}$$

where J_t is the number of alternatives the tourist t has been presented to, g is the density function of β_t, and θ are the parameters of this distribution (mean and variance). So far, this model considers both heterogeneity and the existence of different choice sets for each individual. As for the hierarchical structure, the RCL model is flexible enough to represent different correlation patterns among non-independent alternatives; in fact, it does not have the restrictive substitution patterns of traditional Logit models, allowing representation of any random utility model (McFadden & Train, 2000). In particular, an RCL model can approximate a Nested Logit (NL), which is appropriate for non-independent and nested choice alternatives. Following Browstone and Train (1999), the RCL model is analogous to an NL model in that it groups the alternatives into nests by including a dummy variable in the utility function which indicates which nest an alternative belongs to. Technically, the presence of a common random parameter for alternatives in the same nest allows us to obtain a co-variance matrix with elements distinct from zero outside the diagonal, obtaining a similar correlation pattern to that of an LN model. Regarding the previous example, the analyst should consider that all the hotels in destination A belong to the same nest. So, this fact has to be included in the model. Let us assume that the utility function of alternative i is $U_{it} = \beta x_t + \mu_t z_i + \varepsilon_{it}$, where μ is a vector of random terms with zero mean and variance $\sigma^2{}_{\mu}$, and ε_{it} is independently and identically distributed extreme value with variance $\sigma^2{}_{\varepsilon}$. The non-observed random part of the utility is $\eta_i = \mu_t z_i + \varepsilon_{it}$, which can be correlated with other alternatives depending on the specification of z_i. For example, assume that four alternatives "Hotel 1 in Destination A" (H1A), "Hotel 2 in Destination A" (H2A), "Hotel 1 in Destination B" (H1B) and "Hotel 2 in Destination B" (H2B) have the following utility functions:

$$\begin{aligned} U_{H1A,t} &= \beta x_t + \mu_t + \varepsilon_{H1A,t} \\ U_{H2A,t} &= \beta x_t + \mu_t + \varepsilon_{H2A,t} \\ U_{H1B,t} &= \beta x_t + \varepsilon_{H1B,t} \\ U_{H2B,t} &= \beta x_t + \varepsilon_{H2B,t} \end{aligned}$$

If two alternatives H1A and H2A are truly correlated, their covariance is Cov $(\eta_A,\eta_B) = E(\mu_t + \varepsilon_{At})(\mu_t + \varepsilon_{Bt}) = \sigma^2{}_{\mu}$, which permits identification of correlated non-independent alternatives. Therefore, if the parameter of the variance $\sigma^2{}_{\mu}$, is significantly different from zero, it implies that the alternatives are correlated and must be "closer to each other" and even at the same level of decision. In the context of this example, it means that the two hotels belong to the same "nest", i.e. the same destination (Fig. 3) The advantage of this procedure is that you can test as many nest combinations as "paths to the final decision" the tourist might have in mind. If one were to hypothesize that a tourist, for some reason, chooses the "type of hotel" first (say, the number of stars a hotel has: for example, Hotel 1 means five stars, Hotel 2 means four stars, and so on) and then selects the destination (Fig. 4), the model can accommodate this situation just by defining the non-observed random part of the utility function. Accordingly, assuming that H1A and H1B are hotels with the same number of stars (and the same happens with H2A and H2B), the specification of the utility function would be like this:

$$
\begin{aligned}
U_{H1A,t} &= \beta x_t + \mu_t + \varepsilon_{H1A,t} \\
U_{H2A,t} &= \beta x_t + \varepsilon_{H2A,t} \\
U_{H1B,t} &= \beta x_t + \mu_t + \varepsilon_{H1B,t} \\
U_{H2B,t} &= \beta x_t + \varepsilon_{H2B,t}
\end{aligned}
$$

This way, the model tests whether the tourists follow the hierarchical decision "first the hotel type and second the destination", rather than "first the destination and then the hotel type". An illustration of testing different hierarchical structures in tourist decisions can be found in Nicolau and Mas (2008).

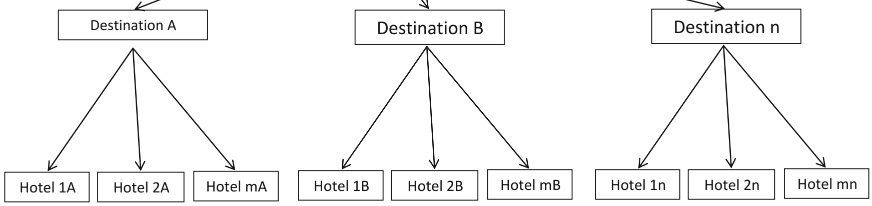


Fig. 3 Hierarchical hotel decision with n destinations and m hotels

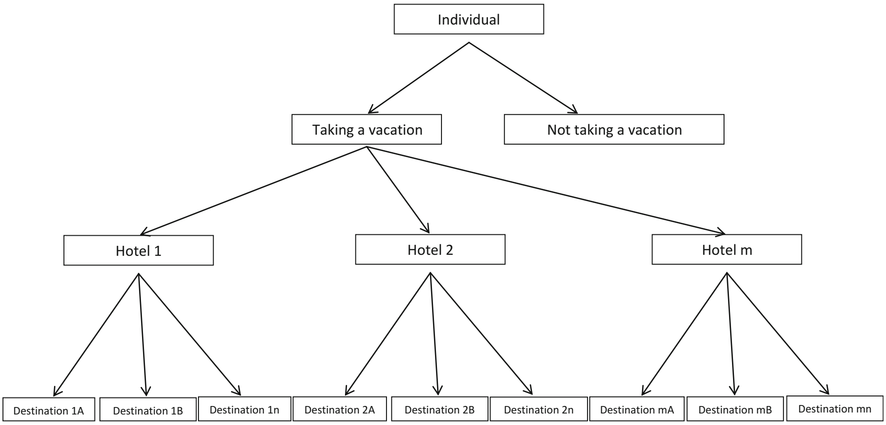


Fig. 4 Hierarchical hotel decision with m hotels and n destinations

3 Research Avenues

This third section explores new avenues for research so that several potential applications are described.

First, knowing the individual utility of a specific tourist gives us information about him or her; information that, he or she himself/herself is not aware that they employed to make the decision. In fact, the estimation of the individual parameters of the utility function of each individual reveals his/her preference structure and allows us to operate with precise information on each individual. At a time when tourists are increasingly demanding and insist on service provision adapted to their specific needs, knowledge of the profile of each tourist allows tourism organizations to offer the most suitable products. Also note that the analysis is based on *real purchase choices* made by individuals (and not on *declarations of intent*), which allows a more accurate representation of the behavior of each tourist.

Second, turning the "market model" into the "click model". The market model is a finance model used to measure the returns of a firm trading on the stock exchange market (for an application in tourism, see Nicolau, 2002). In particular, the rate of returns on the share price of firm i on day t is expressed as:

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it}$$

where R_{it} is the rate of returns on the share price of firm i on day t, R_{mt} is the rate of returns on a market portfolio of stocks on day t. The parameters α_i and β_i are the constant and the systematic risk of stock i, respectively, and ε_{it} is the error term. The analogy would consist of estimating the demand of a product by looking at the number of "clicks" (purchasing clicks, liking clicks, acknowledgment clicks, etc.) where the "clicks portfolio of the market" would be the average number of clicks of the top companies in a industry. Actually, this model would permit the estimation of the expected demand (of clicks) on a specific day. Plus, in the same way that we can estimate the difference between the actual and expected returns by calculating the so-called abnormal returns through the formula:

$$AR_{it} = R_{it} - \left(\hat{\alpha}_i + \hat{\beta}_i R_{mt}\right)$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the estimates obtained from the regression of R_{it} on R_{mt} over an estimation period, we could estimate the difference of the actual number of clicks and the expected amount of clicks. This analysis would give information on the success of a new tourism product or the success of an advertising campaign. For example, if "twitter" were treated as a market where information is exchanged, and the number of "tweets" were considered as a measure of repercussion (or hype), it could be interesting to observe the expectations generated by, say, an innovation announcement on a specific day. Paralleling the market model, it would imply observing whether the amount of "exchanged information" (tweets) derived from a firm's release of news on a given day is abnormally superior to the quantity of "exchanged information" in a normal day, and *whether* and *how many good things* are said.

Third, measuring success of anticipation. The WTTC Report (2014) tells the case of "a match made in heaven": *A passenger boards a transatlantic flight, expecting to plug in the earphones for ten hours straight. But much to her surprise, the passengers on either side of her are also journalists heading to the same conference. Big Data has allowed the airline to engineer the seating arrangement; passengers remember the flight with much more fondness.*

The magazine Hosteltur, in a 2013 article, tells the story of the American writer Janine Driver went to a conference in Nashville and told his audience that the Loews Vanderbilt Hotel where he was staying had visited his profile on Facebook, had downloaded a photo of his newborn son and his older brother, had printed it and left on her bedside table. Driver praised the experience. Both cases are the result of the application of Big Data. However, the following step is to determine how satisfied the people involved are. Not just the specific individuals involved in these two previous examples, but in general terms. Would this strategy be generally favored or would it be considered as interference on one's personal life?

Fourth, in line with Nicolau and Mas (2015), detection of the positioning of both collective and individual brands in people's mind can be done without asking the individuals themselves, just by looking at their decisions and actions. Base on the idea that the meaning of a brand is first individually determined according to people's perceptions; it means that these perceptions will have an influence on the way they will socialize and place their ideas about the brand into social discourse. This social discourse can be examined to discover, not only where they went (so that the analyst can build choice models) but also what they think (so that the analyst can uncover destination positioning strategies).

Fifth, the literature shows that the size of the effect of online reviews depends on whether they are positive or negative, giving rise to asymmetric effects, that is, people perceive extreme ratings (positive or negative) as more useful and enjoyable than moderate ratings (Park & Nicolau, 2015). On account of the importance of online reviews for travel demand, more dimensions can be analyzed and even the ratings of specific attributes of a hotel, airline or destination can be examined.

Sixth, blind booking is a strategy in which an airline offers you different known prices for several unknown destinations. The individual's choice is the "price", not the "destination". This context opens up new research lines as the core element for the tourist's decision, i.e. the destination, is no longer essential. So, people choose prices and their preferences are not based on destination attributes other than price.

Seventh, upselling through auctions. When upselling is the result of auctions, large amounts of data can be obtained that can delve into people's psychology as to the effects of prices.

4 Conclusions

This chapter discusses developments and potential analytic approaches to travel demand modeling with *behavioral Big Data*, with the ultimate goal of generating customer-based knowledge through tourists' feedback and information traces. The advantages linked to the use of Big Data are accompanied with challenges. Accordingly, coordination of the different levels of information is a requisite to properly use this flood of information. This is even more relevant when dealing with destinations as the distinct information is stored by different stakeholders of the destination; thus, heterogeneous data from distinct data sources should be mapped into a homogeneous data format.

Regarding the analysis of Big Data, three empirical problems are to be considered: (1) tourist heterogeneity; (2) the ability to identify all the alternatives available to the tourists when they make their choices; and (3) the inherently hierarchical character of the data at the destination level (e.g. hotels are nested within destinations, destinations within countries). Finally, several new avenues for research are presented. The basic idea is that with the use of Big Data and correctly choosing the analytical tool, we can have a profound understanding of today's travelers' preferences; preferences that they might not even be fully aware that they have.

References

Allenby, G. M., & Rossi, P. E. (1999). Marketing models of consumer heterogeneity. *Journal of Econometrics, 89*, 57–78.

Browstone, D., & Train, K. (1999). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics, 89*, 109–129.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychology Review, 82*(6), 407–428.

Dolnicar, S., & Ring, A. (2014). Tourism marketing research: Past, present and future. *Annals of Tourism Research, 47*, 31–47.

Fuchs, M., Abadzhiev, A., Svensson, B., Höpken, W., & Lexhagen, M. (2013). A knowledge destination framework intelligence application from Sweden. *Tourism, 61*(2), 121–148.

Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2011). The knowledge destination—a customer information-based destination management information system. In R. Law, M. Fuchs, & F. Ricci (Eds.), *Information and communication technologies in tourism 2011* (pp. 417–429). New York: Springer.

Hosteltur. (2013). Las redes sociales, arma para personalizar la experiencia del cliente en hoteles de lujo. http://www.hosteltur.com/118410_redes-sociales-arma-personalizar-experiencia-cliente-hoteles-lujo.html

Invat·tur Report. (2015). Bid Data: retos y oportunidaddes para el turismo. http://invattur.gva.es/estudio/big-data-retos-y-oportunidades-para-el-turismo/

McFadden, D., & Train, K. (2000). Mixed MNL models of discrete response. *Journal of Applied Econometrics, 15*, 447–270.

Mellinas, J. P., Martínez María-Dolores, S.-M., & Bernal García, J. J. (2015). Booking.com: The unexpected scoring system. *Tourism Management, 49*, 72–74.

Nicolau, J. L. (2002). Assessing new hotel openings through an event study. *Tourism Management, 23*(1), 47–54.

Nicolau, J. L., & Mas, F. (2008). Sequential choice behaviour: Going on vacation and type of destination. *Tourism Management, 29*, 1023–1034.

Nicolau, J. L., & Mas, F. J. (2015). Detecting free riders in collective destination brands through a hierarchical choice process. *Journal of Travel Research, 54*, 288–301.

Park, S., & Nicolau, J. L. (2015). Asymmetric effect of online consumer reviews. *Annals of Tourism Research, 50*, 67–83.

Park, S., Nicolau, J. L., & Fesenmaier, D. (2013). Assessing advertising in a hierarchical decision model. *Annals of Tourism Research, 40*, 260–282.

Radojevic, T., Stanisic, N., & Stanic, N. (2015). Solo travellers assign higher ratings than families: Examining customer satisfaction by demographic group. *Tourism Management Perspectives, 16*, 247–258.

Simon, H. (1955). A behavioural model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Steinbruner, J. (2002). *The cybernetic theory of decision*. Princeton, NJ: Princeton University Press.

Train, K. E. (2009). *Discrete choice methods with simulation*. Nueva York: Cambridge University Press.

Vriens, M., & Kidd, P. (2014). What every marketer needs to know about advanced analytics. *Marketing Insights*, 22–29.

WTTC Report. (2014). *WTTC Big Data Report*.http://inspirationhub.imexexhibitions.com/p/4020527602/2014/04/30/wttc-big-data-report

Part II
Analytics in Everyday Life and Travel

Measuring Human Senses and the Touristic Experience: Methods and Applications

Jeongmi (Jamie) Kim and Daniel R. Fesenmaier

1 Introduction

People have senses wherein "our bodily states, situated actions, and mental simulations are used to generate our cognitive activity" such as attitude, behavior, and memory (Krishna, 2012, p. 344). Thus, senses are the central tool of the human body to collect information which is then used as the foundation for understanding or developing meaning. In the context of tourism, people explore a place, they see, hear, smell, touch and taste in combination with their own thought and prior experiences (Csordas, 1999). Thus, it is a traveler's senses which mediate the relationship between the place and meaning (Tuan, 1977). Since our emotional and cognitive responses of places can also be explained by embodied experiences (Gibson, 1966), understanding this process holds the key to 'designing meaningful touristic experience'.

Most research regarding sensing human senses has been conducted either using self-report measures or highly controlled laboratory experiments (Krishna, 2012; Teixeira, Dublon, & Savvides, 2010); however, perceived sensory experiences are not equal to actual sensing experiences. When people perceive external information, they have tuned their own psychological filters such as motivation, past experiences or expectation to particular functional purposes (Sandström, Edvardsson, Kristensson, & Magnusson, 2008). Thus, psychological filters greatly shape 'perceived' sensory experiences and following results.

J. (Jamie) Kim
University of Florida, Gainesville, FL, USA

D.R. Fesenmaier (✉)
National Laboratory for Tourism & eCommerce, Department of Tourism, Recreation and Sport Management, University of Florida, Gainesville, Florida, USA
e-mail: drfez@ufl.edu

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_4

Another important challenge to measuring the traveler experience relates to the fundamental nature of the touristic experience. That is, since the tourism experience is a series of events and not a single event, we need to understand processes of experience creation. Given recent advances in pervasive and ubiquitous technologies, machines can simulate the human senses such as touch, vision, hearing, taste, and smell and capture continuous information that travelers interact, and can make sense of the world around travelers in real-time (Modha et al., 2011). Especially, many wearable and mobile devices are equipped with many sensors perceiving different aspects of the environment (e.g., location, color, sounds, smell, temperature).

The aims of this chapter is describe the role of human senses in the creation of touristic experiences and to introduce a number of approaches for measurement that can be employed in natural environments; as such, the scope of this chapter focuses on measurable and quantifiable human senses and proposes a practical framework which can be used to inform the measurement of traveler's sensory experiences through various sensory modalities. We briefly review senses and related research in consumer behavior and tourism and then discuss how people perceive various environmental which then translates into new technologies that can be used for measuring the nature and extent to which human sense their environment. Last, we discuss the implications of these measurement technologies for tourism research.

2 Senses and Tourism Research

In tourism, the sense or sensory experience has *been* studied using two approaches: the marketing/management approach and the human-geographical/anthropological approach (Agapito, Mendes, & Valle, 2013; Gretzel & Fesenmaier, 2010; Pan & Ryan, 2009). The marketing/management approach to understanding the role of senses and emotions in decision making was introduced by Holbrook and Hirschman (1982) as they stress the experiential nature of consumption and where the sensory experience is understood as outcomes of environmental stimuli. In the more era of experience economy, Schmitt (1999) among others argues that the sensual aspects of experiences are useful ways of differentiate to one products or services Indeed, Pine and Gilmore (1998) argue that the memorable experience needs to engage all five human senses under the conditions of theming. Recent research in consumer behavior also confirms the role of human senses as unconscious triggers which can result in desirable attribute of destination experience (Agapito et al., 2013; Tussyadiah & Zach, 2012) and online tourism marketing (Gretzel & Fesenmaier, 2003; Lee, Gretzel, & Law, 2010).

The human-geographical/anthropological approach explores the embodied sense in tourism (Crouch & Desforges, 2003) where it is argued that embodied sense (or sensation) mediates the relationship between a traveler's body, mind and one's surroundings (Tuan, 1977). Urry also suggests that '[touristic] artifacts are sensed through our bodies'. Hence, a traveler is an active agent who can actively reflect one's own culture, lifestyle, and history into tourism products and services within

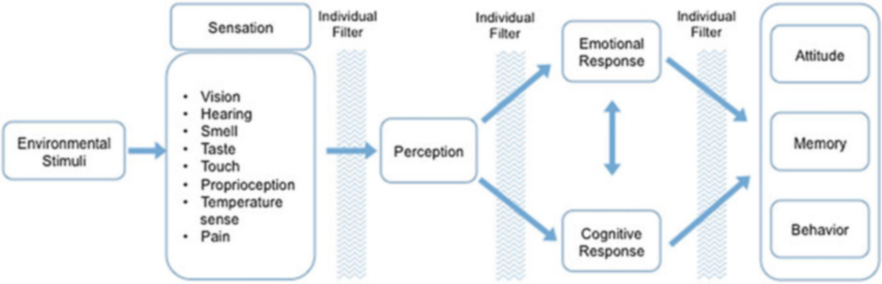


Fig. 1 Framework of touristic experience creation (Adapted from Krishna, 2012)

tourism destinations (Chronis, 2006; Dewey, 1934). Seremetakis pointed out that ‘the interpretation of and through the senses becomes a recovery of truth as collective, material experience’ (1994, p. 6). In this perspective, embodied senses can carry memory and meaning of the place, which often called a ‘sense of place’ (Tuan, 1977) and play an important role in place attachment (Agapito et al., 2013). Thus, the touristic experience can be understood as the process in which various sensory inputs are processed, organized, and interpreted (Larsen, 2007). During this process, a traveler sense and perceive various environmental stimuli, and this will create emotional and cognitive responses influencing one’s attitude, memory, and behavior. However, outcomes of touristic experience process vary based on individual and situational filters such as goals, prior experiences, culture, or travel companions (Sandström et al., 2008). These uncountable dimensions of filter make every touristic experience personal and heterogeneous.

According to Jung, the collective unconscious and complexes of collective unconscious create emotions and meaningful experience (1981). Hence, a big part of human experiences, especially sensory experiences, is not accessible for conscious awareness. Thus, travelers’ perceived senses are not direct records of the world around them. Rather, they are constructed internally along with constraints imposed by the construction of the nervous system and its functional abilities (Gardner & Martin, 2000). Figure 1 provides a conceptual framework for touristic experience creation process. In particular, it describes the elements that play a role in a touristic experience: a traveler and places, services, products, and people that interact with each other. As can be seen the sensory process starts where the environmental stimuli come across the human body’s sense organs which act as the ‘gates’ of emotional and cognitive responses.

3 Psychophysiological Foundations of *Senses*

Before focusing attention on different types of senses and measurements, it is necessary to clarify the terms sense, sensation, and perception. Major schools of thought in psychology consider the sensory experience as systematical process which starts from detecting external stimulus to experiencing and reacting to the stimulus, and then translating knowledge to the perceived situation (Goldstein, 2010; Hekkert, 2006). According to the Oxford Dictionary (www.oxforddictionaries.com), sense is defined as "a faculty by which the body perceives an external stimulus; one of the faculties of sight, smell, hearing, taste, and touch", and "a feeling that something is the case; an awareness or feeling that one is in a specified state". This first definition emphasizing the process of detecting environmental stimulus through the sense organs whereas the second definition focuses more on mental processes; that is, the second definition is about how people 'interpret' the stimulus and 'make meaning' from them. However, the basic processes of detecting environmental stimulus such as light, sound waves and encoding those information into neural energy so that our brains can process is referred to as 'sensation'. As simple example is illustrated by the taste buds in the mouth and olfactory receptor cells which enable people to perceive the texture, temperature, and sweet taste of the dark-colored liquid. This is sensation. However, recognizing 'dark-colored liquid' as a 'hot chocolate' is perception.

Of all kinds of environmental stimuli around us, people record limited information through receptor cells and process these information through brain. Thus, colors, tones, smells, and tastes that we experience are mental creations constructed by the brain out of sensory experience (Gardner & Martin, 2000). As classified by Aristotle, it is generally recognized that humans have five senses such as vision, hearing, smell, touch, and taste. However, recent research suggests that human have more than basic five senses (Gardner & Martin, 2000). Gibson has stated that people have exteroceptive (external) senses and interoceptive (internal). Hence, these studies suggest that senses are physiopsychological systems consisting of a group of sensory cell types that not only respond to an explicit bodily phenomenon but also relate to a specific area in the brain. Therefore, those interoceptive senses also should be recognized as important human senses. Although subject appreciation of some interoceptive senses such as balance, pain, temperatures are below perceptive thresholds, they affect (and are affected by) both emotional states and motivation (Damasio, 1999; James, 1890). Converging evidence from functional imaging studies suggest that recognition of additional human senses and re-mapping the representation of the state of the body could provide the basis for the prelateship between environmental stimulus and feelings, emotions and various activities (Damasio, 1999).

In contrast to the various sensory experiences, all sensory systems convey four basic types of information—modality, location, intensity, and timing—and they are encoded within the human nervous system by specialized subgroups of neurons (Gardner & Martin, 2000). The fact that all sensory systems have the same type of information could be one reason why they could provide objective and context-specific information. Modality defines a general class of stimulus, determined by the type of energy transmitted by the stimulus and the receptors specialized to sense that energy. Each of the sense modalities is characterized by many factors, such as the types of received and accepted data, the sensitivity to the data in terms of temporal and spatial resolutions, the information processing rate, and the availability of the receptors to adapt to the received data. Also each of them induces distinct emotional, cognitive, and behavioral responses. Vision, for example, is more related to operation (Schifferstein & Cleiren, 2005), whereas the sense of smell evokes stronger emotional responses (Schifferstein & Desmet, 2007) and the sense of touch play more significant role in social relationship.

4 Senses and Related Research

The following provides a brief review of basic human senses that are related to touristic experiences based on tourism and marketing research. A short description of innovative measuring techniques for each sense is also provided.

4.1 Vision

Although multisensory stimulus are concurrently available, humans, in general, rely more on vision than on all other forms of sensory modalities (Goldstein, 2010). Visual sensation and perception of an object and its effect on emotional and cognitive responses rely on different visual properties: such as color, size, shape, or movement (King, 2011; Uğur, 2013). For instance, color can trigger certain emotional responses and it plays an important role in perception of other sense. Indeed, Levy (1984) found that different tones of color leads different emotional responses where warm colors raise the level of arousal, whereas cold colors reduce it. More recent research has shown that people reliably correlate specific shape with specific symbolism (Spence & Ngo, 2012a, 2012b) and that circle or curved shapes induce pleasant feelings which, in turn, leads to semantic familiarity of object (Uğur, 2013). Further, Hoegg and Alba (2007) show that color can change the perception of taste; for example simply by changing color of drinks, people perceive the same drinks very differently. In tourism, many studies on traveler's sense of vision has been analyzed from a phenomenological view point—using metaphors "the tourist's gaze". Goodale and Humphrey argued that "The fundamental task of vision is to construct a representation of the three dimensional layout of the world. . .The goal of visual perception is. . . to reconstruct a detailed and accurate model or replica of the three-dimensional world on the basis of the two-dimensional data present at the retinas" (1998, p. 181). Therefore, many researchers have tried to understand how physical environment (e.g., destination, hotel, restaurants, etc.) is conveyed through human visual channels from a holistic viewpoint (Cohen & Cohen, 2012; Larsen, 2001).

Although studies on traveler's visual sense data are scarce, there have been many efforts to develop systems and devices which capture and imitate human visual sense in the fields of media technology and human-computer interaction (HCI). Among vision sensors such as photodetectors, eye-trackers and cameras, the camera is the most widely used sensors because it is affordable, offers high spatial resolution, and rich information (Teixeira et al., 2010). Indeed, recent advances have enabled the creation of wearable camera which allows for recording the sense of visual continuously at the eye through using multiple sensor. For example, Google Glass is equipped with a miniature computer and has many sensors including a front-facing camera to take photographs and a basic eye-tracker which continuously collect visual sensing data of wearer (Ishimaru et al., 2014).

4.2 Hearing

Hearing is one of the important sensory modalities that can elicit emotional responses and sound is stimulus energy. Physical characteristics of sound waves determine psychological dimensions of sound such as loudness, pitch and sharpness and they can result in decrease/increases of sensory pleasantness (Goldstein, 2010). The two most frequently discussed topics related to the experience of hearing in consumer behavior research are perceptions of noise and music. Noisiness refers to particular characteristics of the sound that may or may not induce unpleasant feeling such as annoyance. Any sound may become annoying if it is able to distract listeners from their activities. Interestingly, not all annoying sounds are necessarily noise and any sound may become annoying if it is able to distract listeners from their activities (Fenko, Schifferstein, & Hekkert, 2011). For instance, sound of music or conversation can be experienced as unpleasant sound under certain circumstances, although these sounds do not have a noisy character (Northwood, 1963). Also, low-tone with flat line background noise may reduce tension and surrounding sounds are very indicative of the actions of nearby others taking place at particular moments (Spence & Gallace, 2011). Further, there are certain elements within the music can elicit arousal in the body and induce past memory (Isacsson, Alakoski, & Bäck, 2009; Uğur, 2013); for example, Meyer (1961) demonstrated long ago that the change of rhythm can create expectations about the future development of the music. Also, Webster and Weir (2005) demonstrated that the fast tempo of music is generally associated with happiness and music with slow tempo is associated with sadness.

In tourism, the concept of soundscape has been seen as a useful framework to analyze the touristic experience (Kang & Gretzel, 2012). Travelers sense various sounds during their trip, interpret them and create their own experiences based on them. Similar to the sense of hearing, soundscape in a tourism involves a variety of sounds such as the human voice, natural sounds, media sounds, foreign languages, and even noise (Kang & Gretzel, 2012). For example, in case of outdoor activities, hearing human-caused noise is considered as more annoying and unpleasant than natural sounds. However, these studies measure self-reported 'perceived' sense of sound rather than actual sensing experiences.

The basic physiological principles of human ear have been used to build sound sensors including the microphone. Traditionally, these sound sensors have been fixed a specific location, but now are embedded in many devices including smartphones so that people can easily carry them around and can be equipped with GPS devices so that location be tracked (Kanhere, 2011). SoundSence, for example, collects sound data using microphone of smartphones and can be used to classify 'meaningful' events from them (Lu, Pan, Lane, Choudhury, & Campbell, 2009). Further, Kanhere (2011) and Lu et al. (2009) discuss systems employing crowd-sourcing sound technology so that they can achieve low-cost data collection and analysis through massive coverage in both space and time for observing places and events (Kanhere, 2011).

4.3 Smell

Along with taste, smell is responsible for processing chemicals in our environment into brain (Gardner & Martin, 2000). Smell appears to be the most influential sense to evoke emotional and cognitive responses after sight (Isacsson et al., 2009); also it has the powerful ability to evoke emotional autobiographical memories and the associated emotions (Chu & Downes, 2000, Willander & Larsson, 2006). The strong link between scent, emotion and memory is explained by its direct link with areas of the brain that process emotion, associative learning and memory as olfaction (Herz, 2010). As a consequence, smell is superior (as compared other sensory modalities) in evoking retrieval of autobiographical memories or memories of events that happened before (Goldstein, 2010). Research also show that scents-related inputs decay much slower than other sense-related memory (Isacsson et al., 2009; Zucco, 2003).

The sense of smell in tourism helps to create the character to places and remember (Tuan, 1977). Porteous (1985) examined the power of smell in touristic experience and introduced the concept of 'smellscape' and found that: "The concept of smellscape suggests that, like visual impressions, smells may be spatially ordered or place related...[smellscape is] non-continuous, fragmentary in time" (p. 359). Hence, Dann and Jacobsen (2003) found that smell creates highly emotional responses, which can connect traveler with space and place. Interestingly, however, there is a hue lack of empirical studies examining the role of smell in touristic experience.

4.4 Taste

The sense of taste is closely linked to our perceptions of smell: they both provide about the chemical composition of our surroundings (Goldstein, 2010). Although the sense of smell and the sense of 'flavor' are often used interchangeably, the sense of 'flavor' is actually a combination of true taste and smell. Taste refers the capability to detect the taste of chemical energy through tasting buds in the mouth (Gardner & Martin, 2000). All of our taste sensations can be described as a combination of four basic tastes such as sweet, salty, sour, and bitter (Goldstein, 2010). Several studies have examined the impact of taste on human perception, behavior, and memory (Krishna, 2012). This research indicates that taste is an especially powerful sensory domain for evoking emotional responses, because the consumption of food is directly related to survival and because gustatory stimuli can elicit either positive or negative affective valence (Bartoshuk & Beauchamp, 1994; Goldstein, 2010). Also, the sense of taste is also subject to effects of adaptation (Goldstein, 2010) and Hultén (2011) found that prior experiences and memories influence on the sense of taste. The sense of taste in the tourism context is often regarded as the part of food consumption process (Kim, Eves, & Scarles, 2009). However, most studies in tourism have focused on food itself and overlooked the sensory aspects of taste. As suggested by Cohen and Avieli (2004), tasting during the trip is not just a physiological sensation but also a social, cultural, and symbolic activities. Therefore, most studies have tried to link the physical properties of a food and surrounding environment such as climate and geology with human practices such as traditions, tools, and recipes (Hjalager & Richards, 2002; Kim et al., 2009).

Similar to the electronic nose system, an electronic 'tongue' system which can analyze a number of chemical mixtures (Biswas et al., 2014) where it uses taste sensors to receive information from chemicals and each taste is classified by matching the various chemical compound with taste patterns. Applications of electronic tongues have been found in different areas such as food analysis, environmental monitoring and diagnosis of disease (Biswas et al., 2014).

4.5 Touch

Before discussing the sense of touch, we need to clarify different modalities of somatosensory system, which refers a collective term for sensory signals from the body (Goldstein, 2010). Unlike other sensory modalities, somatosensory systems or the somatic senses are positioned throughout the body rather than in a localized, specific organ (Goldstein, 2010). Therefore, sensitivity of touch differs from area to area (Uğur, 2013). As shown in Table 1, the stimulus of each sensory modality is different (Gardner & Martin, 2000). The sense of touch is one type of somatosensory systems, which is the mechanical distortion of the skin by direct pressure or by bending hairs on the skin (Goldstein, 2010). The sense of touch converts information about an object's weight and location so it plays an important role in object discrimination and manipulation (Gardner & Martin, 2000). The sense of touch comprises two main submodalities: cutaneous and kinesthetic. The cutaneous sense receives sensory inputs throughout the skin of any part of body without motion, and the kinesthetic sense receives sensory inputs from the receptors based on the relative position and movement of body parts (Dahiya, Metta, Valle, & Sandini, 2010). And the haptic sense refers combination of both (Uğur, 2013).

Recent neurological research has found that interpersonal touch triggers direct emotional responses and the importance of touch for humans have been confirmed

Table 1 Sensory modalities and measurement sensors for touristic experience

Sensory system	Sensory modalities	Stimulus energy	Sensor	Sample application
Visual	Vision	Light	Photo detectors Ranging sensors	Camera Glass (e.g., Google Glass)
Auditory	Hearing	Sound	Inertial sensors Vibration sensors	Microphone Ear bud (e.g., SoundApp)
Olfactory	Smell	Chemical	Chemo sensors	Electronic nose
Gustatory	Taste	Chemical	Chemo sensors	Electronic tongue
Somatosensory	Touch	Pressure	Contact sensors Pressure sensors	Silicon fingers
	Proprioception	Displacement	Accelerometer sensors Magnetometer sensors	Wearable clothing Shoes
	Temperature sense	Thermal	Thermal imagers	Thermoelectric bracelet
	Pain	Chemical, thermal, or mechanical	Chemo sensors Temperature sensors Accelerometer sensors Magnetometer sensors	Wearable clothing Shoes Wearable clothing Shoes

Adapted from Gardner and Martin (2000) and Teixeira et al. (2010)

in many empirical studies (Hollins, 2010). Spence and Gallace (2011), for example, found that pleasant feelings induced by touch can modulate a person's overall evaluation of many different products. However, other researchers find that the evaluation and appreciation of product quality and pleasantness of feeling depends on differences in need-for-touch (Krishna, 2012). Due to the nature of somatosensory systems, sense of touch is closely associated with highly subjective, social and intimate emotions (Hertenstein, Verkamp, Kerestes, & Holmes, 2006; Paterson, 2007; Uğur, 2013). Even though engaging various senses is important for creating touristic experience, there is still a lack of empirical studies addressing the sense of touch in tourism (Agapito et al., 2013; Gretzel & Fesenmaier, 2003).

Recently, studies have focused on developed sensors for medical and sport applications that include flexible pressure sensors to measure the sense of touch (Dahiya et al., 2010). In particular, special types of fibers with textile sensors have been used to measure the sense of touch across entire body (Uğur, 2013). For instance, Textrodes were developed by knitting stainless steel fibers in order to monitor the performance of the wearer in sport activities.

4.6 *Other Somatosensory Modalities: Movement, Temperature, and Pain*

As discussed earlier, somatosensory systems are complex sensorial experiences that have various sub-modalities such as proprioception, temperature, and pain (Gardner & Martin, 2000). Firstly, the sense of proprioception refers posture and the movement of parts of the body and it contains the sense of stationary position and movement (Gardner & Martin, 2000). No specific organ works for the sense of proprioception, rather the receptors for this sense are located in the joints, tendons, and muscles of human body (Goldstein, 2010). The sense of proprioception along with other somatosensory modalities provides information about the world around us. This is why it is critical for the ability to move and control the body in space and the physical forces acting upon it (Gardner & Martin, 2000). In tourism, mobility is an important factor in how people experience surrounding world. Also, movements can be regarded as performative actions that evoke various social and personal meaning (Lewis, 2000). It encourages active exploration of external world through traveler's body. This is why the sense of movement often regarded as a 'sixth sense' (Lewis, 2000). When a traveler fully engages in physical activities such as dancing, one's body movement can be used as a means to express self and establish desired social status (Matteucci, 2014).

There is a wide range of products target sports and physical activity based on movement and balance sensors including specialized hardware such as GPS watches and fitness bands as well as apps such as Runkeeper or Endomondo. This technology is mainly devoted to capture those human senses without human efforts to transform human behaviors in desirable ways. For example, RunRight is a system that provided an audio and visual feedback based on running movement (Nylander, Jacobsson, & Tholander, 2014). In order to measure the runner's movement directly from their body they use the chase-belt to capture acceleration in vertical and horizontal direction. The main idea is to configure how to design technology that supports people in learning how a desired movement should feel. Another innovative way to measure the movement was designed by Stienstra, Overbeeke, and Wensveen (2011)). They created a system where skating movements measure the pressures submitted through the blades. These measurements would real-time feed a computer system for further analysis of the balance and movement. As mentioned above, a smartphone itself is widely used as a mobile sensor. GymSkill is one example of the smartphone-embedded balance sensor (Kranz et al., 2013). A basic mechanism of this system is similar to implementing balance and movement sensor: recording accelerometer and magnetometer data. In this system, the smartphone is placed on surface of a balance board, on which user works out. The smartphone interacts with the balance board, store the sensor data, and provide feedback.

Secondly, there is the sense of temperature. It detects increases or decreases in skin temperature using warm and cold thermoreceptors to provide information to maintain the body's temperature (Gardner & Martin, 2000). The sense of temperature can be affected by external temperatures as well as the internal body heat (King, 2011). As the sense of temperature is directly related to physical comfort and personal well-being, rapid changes in skin temperature evoke dynamic emotional responses, both positive and negative ones (Krishna, 2012). Moreover, recent neuroscience research shows that sensations of physical warmth (temperature) affect social warmth such as trust, intimacy, and belongingness. The sense of temperature in tourism has been viewed as a means of securing favorable atmosphere in tourism and hospitality industry (Heide & GrØnhaug, 2006).

Temperature sensors come in a wide variety but all measure temperature by sensing some change in a physical characteristic. In many cases, the temperature sensor is used for the thermic control and environmental monitoring and often integrated with other sensing measure devices such as moisture, gas, and light sensor (Tsow et al., 2009). Recent advances in wearable technology and sensor have allowed not just passively gathering temperature information from environment but also creating pleasant temperature which are optimized to an individual's current needs. Wristify, for example, is a thermoelectric bracelet that monitors air and skin temperature, and sends tailored pulses of hot or cold waveforms to the wrist to help them maintain thermal comfort.

Last is the sense of pain. This may be the most crucial sensory modality for survival by informing the impact of the world on human body (Goldstein, 2010). The sense of pain can be activated by various sources, such as chemical, thermal and mechanical energy (Gardner & Martin, 2000). In general, pain distinctly evoke negative emotional component that interrupts the emotional state (Goldstein, 2010). Interestingly, pain sensations can be moderated by anticipation, prior experience, personal belief, and environmental factors such as companions, pleasant music (Craig, 2009; Gardner & Martin, 2000). While stimulus caused pain is same, response criterion of pain can be increased as time goes by (Gardner & Martin, 2000). This is why, in some case, physical pain can be used as a means of developing personal identity or emotional maturation (Joy & Sherry, 2003). For example, the pains felt in feet after hiking can be perceived as positive and become a happy memory for someone, whilst the other consider it as a stressful event (Crouch & Desforges, 2003). Moreover, the sensing-pain technology is promising to provide many advantages in the field of medical and public health. Pain sensors are expected to revolutionize the remote monitoring of health conditions (Appelboom et al., 2014).

Pain sensing technology is actively utilized in the field of medical and public health (Patel, Park, Bonato, Chan, & Rodgers, 2012). As the sense of pain is a complex and sensitive phenomenon, pain sensors are often comprised of multiple sensors that are typically integrated into a sensor network either exclusive bodyworn sensors or integrating body-worn sensors and other sensors (e.g., chemical, thermal sensor). Therefore, the efforts of developing pain sensors are mostly dependent on incorporating smart wearable sensors embedded in smartphone or wearable devices (Appelboom et al., 2014). An example of such systems is the multi-sensor wearable prototype to monitor low back pain by Chhikara and his colleagues (2008). Their system uses a combination of inertial sensors placed on the lower back and pelvis, a light sensor located on the chest and surface EMG sensors placed on back muscles and buttock muscles, which were expected to be replaced in a lumbar belt or in wearable textiles to facilitate usability in the future study.

5 Capturing Traveler's Senses: Challenges and Possible Solutions

A review of the tourism literature highlights challenges facing touristic experiences research. Especially, capturing traveler's sensory experiences is challenging endeavor for many reasons. The dynamic nature of touristic experiences is one of the main obstacles when capture traveler's sensory experiences. A traveler's emotional and cognitive responses are not simple reactions of specific situation or certain products they encounter, rather "a continuum of sensory experience" (Williams, 1954, p. 98). Therefore, research should try to capture fluctuating moments when these emotional and cognitive changes occur (Nold, 2009). However, due to the methodological and technical challenges, most of studies focus on the specific senses in the specific phases of trips (Agapito et al., 2013). This is particularly evident in the marketing/management literature on the sensory experience of tourism. In line with the foregoing, another issue is dominant research methods, heavy dependence on subjective and context-dependent measure. Currently available biophysiological sensors are only capable to detect specific external sensual energy, but not the perceptive process after sensation (Resch, Summa, Sagl, Zeile, & Exner, 2015). Also unexpected or sudden changes in environmental conditions and may cause some of measurement errors (Teixeira et al., 2010).

To account for this shortcoming, we propose to capture traveler's sensory experiences by measuring multiple sensory modalities through wearable biophysiological sensors. We claim that integrated multiple sensing data including their mobility, opens new outlooks to physical and social dynamics at the traveler-place interaction. Our proposed framework is depicted in Fig. 2 and is based on an extensive review of existing sensing modalities and sensor fusion approaches (Teixeira et al., 2010): (1) Collecting massive amounts of low-cost motion data, (2) Placing a smaller number of fixed camera at key touch-points, and (3) Using mobile technology. Hence, progress in new technologies has given rise to devices and techniques, which allow an objective evaluation of different sensing parameters, resulting in more efficient measurement (Resch et al., 2015). We argue that measuring senses at the lowest level will allow us to empirically gauge the impact of environments on travelers. As shown in Fig. 2, various senses can be measured either single source or multiple different sources (Teixeira et al., 2010). For example, the sense of smell can be captured via chemo sensor, whereas the sense of pain is requiring various sensors.

To further increase our understanding of traveler's sensory experiences beyond the setup described above, we believe more trans-disciplinary approach will

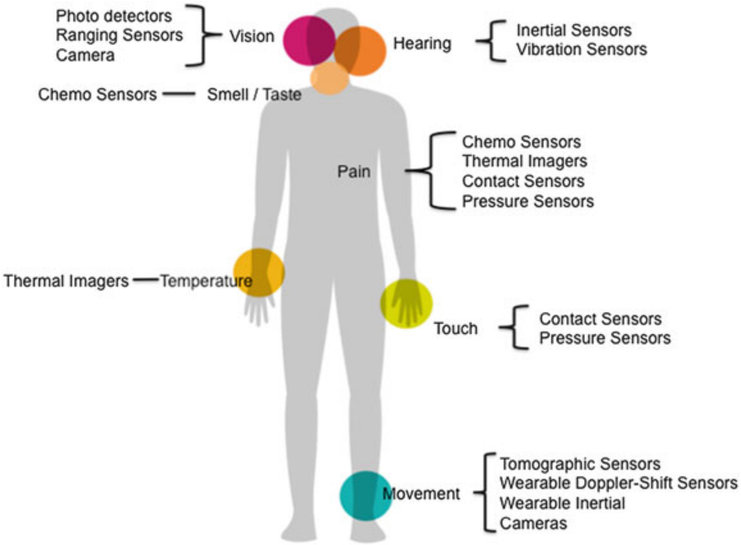


Fig. 2 Biophysiological sensors that may be used to measure traveler's sensory experiences

necessarily take place. First of all, future research can include applying the big data analytics methods. Extracting sensory experience from crowd-sourced data like Twitter or matching both data could provide collective human behavior patterns. It can provide additional insights into the development of both the physical and social structures of inherently complex and dynamic touristic experience (Resch et al., 2015). Also future research needs to map traveler's sensory experience to a series of events (e.g., opening of new attraction, special festival, or natural disasters) to the chronological order. We believe that this effort allows visualization of evolution of traveler's sensory experience by distinct events occurs.

Now we can investigate touristic experiences with enormous sensing data (high volume) using various sensors for different senses (variety) in real time, which shows that the era of big data in tourism research has come. Capturing massive human sensing data and analyzing Big 'human sense' Data have the potential to transform the way tourism researchers use measure traveler's experience and design meaningful touristic experience.

6 Conclusions

In this chapter we have explored the possibility of employing physiological parameters of human senses as the useful channels to understand touristic experiences. Large literatures have shown that the senses are the key to our emotions, the source of our behaviors and long-lasting memories. Therefore, from a psychophysiological perspective and a grounded cognition perspective, sense should be considered as the foundation of how humans interact with environments and make meanings from these interactions.

Understanding how a traveler perceives one's surroundings during the trip can contribute a vital base for tourism experience design and destination development (Gretzel & Fesenmaier, 2003; Tussyadia, 2014). Although each sensory modality provides different information, combining various sensing data together allow us better understanding of how a traveler creates touristic experiences. Thus, it is important for tourism researchers and marketers to recognize how these sensory experiences play their role at different phase of trip as well as how different senses can work together to create more meaningful tourism experiences. Further, measuring those senses has been key goals of engineering, medical and human-computer interaction. Emerging technology to detect changes of environments and developing wearable sensors have changed the way of understanding measuring senses in the field of tourism. As advancement of new technologies has led the development of devices and techniques, there is an increasing opportunity of collecting human-centric data in real-time. Interdisciplinary cooperation with neuroscience, psycho-physiology, and marketing research enables us to understand much better individual's perceptions of the surrounding world (Krishna, 2012). As such, real-time and continuous data collection over different locations and time will help tourism marketers to evaluate their products and services so as to compete more effectively.

References

Agapito, D., Mendes, J., & Valle, P. (2013). Exploring the conceptualization of the sensory dimension of tourist experiences. *Journal of Destination Marketing & Management, 2*(2), 62–73.

Appelboom, G., Camacho, E., Abraham, M. E., Bruce, S. S., Dumont, E. L., Zacharia, B. E., et al. (2014). Smart wearable body sensors for patient self-assessment and monitoring. *Archives of Public Health, 72*(1), 28.

Bartoshuk, L. M., & Beauchamp, G. K. (1994). Chemical senses. *Annual Review of Psychology, 45*, 419–449.

Biswas, P., Chatterjee, S., Kumar, N., Singh, M., Majumder, A. B., & Bera, B. (2014). Integrated determination of tea quality based on taster's evaluation, Biochemical characterization and use of electronics. In *Sensing technology: Current status and future trends II* (pp. 95–117). Springer.

Chhikara, A., Rice, A. S., McGregor, A. H., & Bello, F. (2008). Wearable device for monitoring disability associated with low back pain. *World, 10*, 13.

Chronis, A. (2006). Heritage of the senses collective remembering as an embodied praxis. *Tourist Studies, 6*(3), 267–296.

Chu, S., & Downes, J. J. (2000). Odour-evoked autobiographical memories: Psychological investigations of Proustian phenomena. *Chemical Senses, 25*(1), 111–116.

Cohen, E., & Avieli, N. (2004). Food in tourism: Attraction and impediment. *Annals of Tourism Research, 31*(4), 755–778.

Cohen, E., & Cohen, S. A. (2012). Current sociological theories and issues in tourism. *Annals of Tourism Research, 39*(4), 2177–2202.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*, 59–70.

Crouch, D., & Desforges, L. (2003). The sensuous in the tourist encounter introduction: The power of the body in tourist studies. *Tourist Studies, 3*(1), 5–22.

Csordas, T. J. (1999). Embodiment and cultural phenomenology. In G. Weiss & H. F. Haber (Eds.), *Perspectives on embodiment: The intersections of nature and culture* (pp. 143–163). London: Routledge.

Dahiya, R. S., Metta, G., Valle, M., & Sandini, G. (2010). Tactile sensing—from humans to humanoids. *IEEE Transactions on Robotics, 26*(1), 1–20.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: HarcourtBrace.

Dann, G., & Jacobsen, S. (2003). Tourism smellscape. *Tourism Geographies, 5*(1), 3–25.

Dewey, J. (1934). *Art as experience*. New York: Penguin.

Fenko, A., Schifferstein, H. N., & Hekkert, P. (2011). Noisy products: Does appearance matter. *International Journal of Design, 5*(3), 77–87.

Gardner, E. P., & Martin, J. H. (2000). Coding of sensory information. *Principles of Neural Science, 4*, 411–429.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Goldstein, E. (2010). *Sensation and perception*. Belmont, CA: Cengage Learning.

Goodale, M. A., & Humphrey, G. K. (1998). The objects of action and perception. *Cognition, 67*, 181–207.

Gretzel, U., & Fesenmaier, D. (2003). Experience-based internet marketing: An exploratory study of sensory experiences associated with pleasure travel to the Midwest United States. In A. Frew, M. Hitz, & P. O'Connor (Eds.), *Information and communication technologies in tourism 2003* (pp. 49–57). New York: Springer.

Gretzel, U., & Fesenmaier, D. (2010). Capturing sensory experiences through semi-structured elicitation questions. In M. Morgan, P. Lugosi, & B. Ritchie (Eds.), *The tourism and leisure experience: Consumer and managerial perspectives* (pp. 137–162). Bristol, UK: Channel View Publications.

Heide, M., & GrØnhaug, K. (2006). Atmosphere: Conceptual issues and implications for hospitality management. *Scandinavian Journal of hospitality and Tourism, 6*(4), 271–286.

Hekkert, P. (2006). Design aesthetics: Principles of pleasure in design. *Psychology Science, 48*(2), 157.

Hertenstein, M. J., Verkamp, J. M., Kerestes, A. M., & Holmes, R. M. (2006). The communicative functions of touch in humans, nonhuman primates, and rats: A review and synthesis of the empirical research. *Genetic, Social, and General Psychology Monographs, 132*(1), 5–94.

Herz, R. S. (2010). The emotional, cognitive and biological basics of olfaction: Implications and considerations for scent marketing. In A. Krishns (Ed.), *Sensory marketing: Research on sensuality of products* (pp. 87–107). New York: Routledge.

Hjalager, A., & Richards, G. (Eds.). (2002). *Tourism and gastronomy*. London: Routledge.

Hoegg, J., & Alba, J. W. (2007). Taste perception: More than meets the tongue. *The Journal of Consumer Research, 33*, 490–498.

Holbrook, M., & Hirschman, E. (1982). The experiential aspects of consumption: consumer fantasies, feelings and fun. *Journal of Consumer Research, 9*(2), 132–140.

Hollins, M. (2010). Somesthetic senses. *Annual review of psychology, 61*, 243–271.

Hultén, B. (2011). Sensory marketing: The multi-sensory brand-experience concept. *European Business Review, 23*(3), 256–273.

Isacsson, A., Alakoski, L., & Bäck, A. (2009, November 15). Using multiple senses in tourism marketing: The Helsinki expert, Eckero line and Linnanmaki Amusement Park cases. *TOURISMOS: An International Multidisciplinary Journal of Tourism*, *4*(3), 167–184.

Ishimaru, S., Kunze, K., Kise, K., Weppner, J., Dengel, A., Lukowicz, P., et al. (2014, March). In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with Google Glass. In *Proceedings of the 5th augmented human international conference* (p. 15). ACM.

James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press, 1981. Originally published in 1890.

Joy, A., & Sherry, J. F., Jr. (2003). Speaking of art as embodied imagination: A multisensory approach to understanding aesthetic experience. *Journal of Consumer Research, 30*(2), 259–282.

Jung, C. G. (1981). *The archetypes and the collective unconscious (No. 20)*. Princeton, NJ: Princeton University Press.

Kang, M., & Gretzel, U. (2012). Effects of podcast tours on tourist experiences in a national park. *Tourism Management, 33*(2), 440–455.

Kanhere, S. S. (2011, June). Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In *Mobile Data Management (MDM), 2011 12th IEEE International conference* on (Vol. 2, pp. 3–6). IEEE.

Kim, Y. G., Eves, A., & Scarles, C. (2009). Building a model of local food consumption on trips and holidays: A grounded theory approach. *International Journal of Hospitality Management, 28*(3), 423–431.

King, L. A. (2011). *The science of psychology: An appreciative view* (2nd ed.). New York: McGraw-Hill.

Kranz, M., Möller, A., Hammerla, N., Diewald, S., Plötz, T., Olivier, P., et al. (2013). The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing, 9*(2), 203–215.

Krishna, A. (2012). An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of Consumer Psychology, 22*(3), 332–351.

Larsen, J. (2001). Tourism mobilities and the travel glance: Experiences of being on the move. *Scandinavian Journal of Hospitality and Tourism, 1*(2), 80–98.

Larsen, S. (2007). Aspects of psychology of the tourist experience. *Scandinavian Journal of Hospitality and Tourism, 7*(1), 7–18.

Lee, W., Gretzel, U., & Law, R. (2010). Quasi-trial experiences through sensory information on destination web sites. *Journal of Travel Research, 49*(3), 310–322.

Levy, B. I. (1984). Research into the psychological meaning of color. *American Journal of Art Therapy, 23*, 58–62.

Lewis, N. (2000). The climbing body, nature and the experience of modernity. *Body & Society, 6*(3–4), 58–80.

Lu, H., Pan, W., Lane, N. D., Choudhury, T., & Campbell, A. T. (2009, June). SoundSense: Scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th International conference on mobile systems, applications, and services* (pp. 165–178). ACM.

Matteucci, X. (2014). Forms of body usage in tourists' experiences of flamenco. *Annals of Tourism Research, 46*, 29–43.

Meyer, L. B. (1961). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Modha, D. S., Ananthanarayanan, R., Esser, S. K., Ndirango, A., Sherbondy, A. J., & Singh, R. (2011). Cognitive computing. *Communications of the ACM, 54*(8), 62–71.

Northwood, T. D. (1963). Sound and people. *Canadian Building Digest, 41*, 973–978.

Nylander, S., Jacobsson, M., & Tholander, J. (2014, April). Runright: Real-time visual and audio feedback on running. In *CHI'14 Extended abstracts on human factors in computing systems* (pp. 583–586). ACM.

Pan, S., & Ryan, C. (2009). Tourism sense-making: The role of the senses and travel journalism. *Journal of Travel & Tourism Marketing, 26*(7), 625–639.

Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of Neuroengineering and Rehabilitation, 9*(1), 21.

Paterson, M. (2007). *The senses of touch: Haptics, affects and technologies*. Oxford: Berg.

Pine, J., & Gilmore, J. H. (1998). Welcome to the experience economy. *Harvard Business Review, 76*(4), 97–105.

Porteous, J. D. (1985). Smellscape. *Progress in Physical Geography, 9*, 356–378.

Resch, B., Summa, A., Sagl, G., Zeile, P., & Exner, J. P. (2015). Urban emotions—Geo-semantic emotion extraction from technical sensors, Human sensors and crowdsourced data. In *Progress in location-based services 2014* (pp. 199–212). Springer.

Sandström, S., Edvardsson, B., Kristensson, P., & Magnusson, P. (2008). Value in use through service experience. *Managing Service Quality, 18*(2), 112–126.

Schifferstein, H. N. J., & Cleiren, M. P. H. D. (2005). Capturing product experiences: A split-modality approach. *Acta Psychologica, 118*, 293–318.

Schifferstein, H. N. J., & Desmet, P. M. A. (2007). The effect of sensory impairments on product experience and personal well-being. *Ergonomics, 50*, 2026–2048.

Schmitt, B. (1999). Experiential marketing. *Journal of Marketing Management, 15*(1–3), 53–67.

Seremetakis, N. (1994). *The senses still: Perception and memory as material culture in modernity*. Chicago: Chicago University Press.

Spence, C., & Gallace, A. (2011). Multisensory design: Reaching out to touch the consumer. *Psychology & Marketing, 28*(3), 267–308.

Spence, C., & Ngo, M. K. (2012a). Assessing the shape symbolism of the taste, flavour, and texture of foods and beverages. *Flavour, 1*(1), 1–13.

Spence, C., & Ngo, M. K. (2012b). Assessing the shape symbolism of the taste, flavour, and texture of foods and beverages. *Flavour, 1*(1), 12.

Stienstra, J., Overbeeke, K., & Wensveen, S. (2011, September). Embodying complexity through movement sonification: Case study on empowering the speed-skater. In *Proceedings of the 9th ACM SIGCHI Italian Chapter International conference on computer-human interaction: Facing complexity* (pp. 39–44). ACM.

Teixeira, T., Dublon, G., & Savvides, A. (2010, September). *A survey of human-sensing: Methods for detecting presence, count, location, track, and identity* (Tech. Rep. 09-2010). New Haven, CT: ENALAB, Yale University.

Tsow, F., Forzani, E., Rai, A., Wang, R., Tsui, R., Mastroianni, S., et al. (2009). A wearable and wireless sensor system for real-time monitoring of toxic environmental volatile organic compounds. *IEEE Sensors Journal, 9*(12), 1734–1740.

Tuan, Y. F. (1977). *Space and place: The perspective of experience*. Minneapolis, MN: University of Minnesota Press.

Tussyadiah, I. P., & Zach, F. J. (2012). The role of geo-based technology in place experiences. *Annals of Tourism Research, 39*(2), 780–800.

Uğur, S. (2013). Emotion, design and technology. In *Wearing embodied emotions* (pp. 33–59). Milan: Springer.

Webster, G. D., & Weir, C. G. (2005). Emotional responses to music: Interactive effects of mode, texture, and tempo. *Motivation and Emotion, 29*(1), 19–39.

Willander, J., & Larsson, M. (2006). Smell your way back to childhood: Autobiographical odor memory. *Psychonomic Bulletin & Review, 13*(2), 240–244.

Zucco, G. M. (2003). Anomalies in cognition: Olfactory memory. *European Psychologist, 8*(2), 77–86.

The Quantified Traveler: Implications for Smart Tourism Development

Yeongbae Choe and Daniel R. Fesenmaier

1 Introduction

Imagine a traveler visiting and exploring a destination. With the assistance of GPS-embedded smart shoes, Fred or Sara can easily find his/her way to a particular attraction without gazing at a smartphone screen. At the same time, he/she monitors his/her body temperature and heart beat so as to prevent overheating and other potentially negative health consequences. After returning to the hotel, the room is fully cleaned and set at the desired room temperature. Next morning, Fred's smart watch shows the best route for jogging along a beautiful path near the hotel. During the run, an app on the smartphone senses his emotions and recommends changes in today's trip plan so as include fewer strenuous activities, changes the color and style of the watch and even recommends different music. Further missing his family, Fred simply looks over to a photo of his family placed near the TV wherein it connects to similar photos at home and indicates that they are connected by a glowing frame. Although this vignette by no means represents the current tourism experience and supporting technologies, this scenario describes current technology which is likely to arrive to the tourism industry in the near future.

The continuing evolution of information and communication technology (ICT) has transformed travel in many ways. In particular, smartphones and associated apps have expanded the scope of the tourism experience by enabling travelers to contact and share their experiences with family and friends in different places whenever and wherever they want (Wang, Park, & Fesenmaier, 2012; Wang,

Y. Choe
University of Florida, Gainesville, FL, USA

D.R. Fesenmaier (✉)
National Laboratory for Tourism & eCommerce, Department of Tourism, Recreation and Sport Management, University of Florida, Gainesville, Florida, USA
e-mail: drfez@ufl.edu

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_5

Xiang, & Fesenmaier, 2014a). Parallel to these developments, wearable devices (e.g., Google Glass, Apple iWatch, fitness bands, etc.) have been widely adopted by consumers owing to their advantages of portability and potential usability for travel purposes (Tussyadiah, 2013). These technologies seem to be moving us toward an increasingly data-driven ‘sensor society’ wherein an individual leaves a huge data footprint during the course of his/her everyday life, which creates opportunities for business development (e.g., Andrejevic & Burdon, 2014; Swan, 2012).

Recently, the notion of the “quantified self” has been used to describe the improvement of one’s life through self-knowledge and discovery whereby wearable devices are used to constantly monitor our daily life, to save the collected data future use and, potentially, to share this information with other similarly interested people (Swan, 2012). With wearables devices increasingly used in travel, it is not difficult to imagine that there are many ways to “repurpose” and to ‘extend’ this type of information such that it can be used to provide more details about the traveler including health status, potential considerations for maintaining a diet, possible changes in plans or quality of sleep and emotional status, to communicate with friends and relatives, and even to create a sense of being home. With this background, this chapter discusses the concept of the quantified self in terms of how today’s wearable technologies connect one’s ordinary life and the touristic experience and provides a useful framework (characterized as the quantified traveler) for understanding this technology within the travel context. This chapter then discusses the usefulness of these technologies for smart tourism development.

2 Emergence of the Quantified Traveler and Wearable Technologies

The quantified-self movement is an emerging trend represented by a wide range of technological devices used for self-tracking, life-logging, personal analytics, and personal informatics. The concept of the quantified-self is based upon a new phenomenon wherein people voluntarily monitor their lives to better understand themselves (Lupton, 2014). Indeed, the notion of self-monitoring and tracking has a fairly long history that can be traced back to the 1970s (Kopp, 1988; Marcengo & Rapp, 2014). Since then, the concept of self-monitoring has proven effective in changing people’s attitude and behaviors, which is the goal of an embodied function in the sensing technologies (Choe, Lee, Lee, Pratt, & Kientz, 2014). The motivation behind this movement is to gain self-knowledge by tracking one’s life to “optimize” behavior through the process of quantification (Choe et al., 2014; Marcengo & Rapp, 2014). Having these motivations, quantified-self participants have identified several benefits to this process including acquiring data about their lives, monitors and even challenging themselves, and eventually receiving feedback resulting from comparisons between their actual life activities and goals, and potentially, other similar individuals.

Table 1 Quantified-self categories and measures

Categories	Example of potential measures
Physical states and activities	Body movement, temperature, calories used
Psychological and mental states and traits	Mood, happiness, emotions, self-esteem, thinking patterns, focus, attention, memory, stress, tension
Situation and environmental variables	Location, weather, noise, pollution, context, time of the day, travel, time intervals, places visit, distance traveled
Social variables	Influence, trust, interactions, people you are with, perceived safety

Importantly, the development of wearable devices (e.g., wrist bands and smart watches) which are made possible through relatively inexpensive sensors, easy access to the internet and cloud computing have completely changed the way people track their daily life by lessening the effort and the level of consciousness (Smarr, 2012; Swan, 2012, 2013). The concept of the quantified-self has been applied to a number of the different domains (e.g., health, fitness, and sport) and generates several different types of information about our lives. As can be seen in Table 1, people sometimes are required to have sufficient knowledge and additional effort to manually keep the record of their behaviors and feelings (e.g., steps taken, well-being, happiness, calorie intake, and the number of cups of coffee). However, there are a number of technologies which have the capacity to measure/track people in largely invisible ways (Marcengo & Rapp, 2014; Swan, 2012). These ‘smart’ products and devices now have the capability of somehow capturing or reflecting much of our surroundings and behaviors in real-time unobtrusively and unconsciously and interact with each other so as to gain a general ‘understanding’ of our current circumstances (Lupton, 2014); for example, driving habits and possible drowsiness can be monitored so as to alert drivers to be safe. While self-tracking practices involve a continuous process of recording one’s life, the data does not need to be only quantitative, but can exist in any format such as a picture, video, online social media data, and audio (Augemberg, 2013). Thus, the collected data are both ‘structured’ and ‘unstructured’ depending on the device and method(s) of data capture.

In general, wearable technologies enable us to connect to the Internet, devices, and external environments through digital sensors (Lupton, 2014). Some of these devices can exchange stored information via wireless, NFC, and iBeacon technology so that people can have better conditions, be aware of the environment, and even encourage them to change certain behaviors (Swan, 2012). Smart ‘shoes’, for example, can vibrate so as to point a person in the right direction so that he/she can enjoy the scenery; smart thermometers embedded within clothing can exchange information with other wearable devices in order to adjust the temperature in the room; or similar sensors embedded in a blanket can be used track sleep so as to assess the amount of time and the rhythm of deep sleep one has each night. Figure 1 illustrates some of the applications—wearable devices widely that have been used and which have the potential to measure travelers’ sensory perceptions as well as mediate their travel experiences.

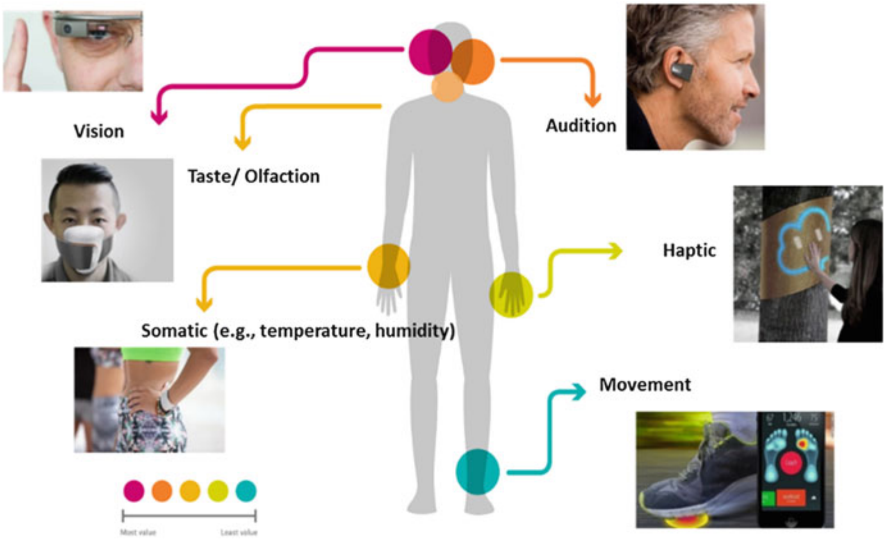


Fig. 1 Applications and wearable devices used for 'Quantified-Self' [adapted from Kim and Fesenmaier (2015)]

Further, the terms 'citizens as sensors', 'people as sensors', and 'collective sensing' have been coined to describe the nature of collective behaviors in terms of understanding context through social media, sensing technologies, and wearable devices (e.g., Goodchild, 2007; Sagl, Resch, & Blaschke, 2015). That is, many people actively use 'sensors' so that they can collect data about their surrounding environment as well as their physical/emotional states (and stored personal historical data) in real time, which in turn, generate huge volumes of data that greatly support individual decisions; for example, many outdoor enthusiasts collect and share information about birds, consistently collect weather information for local reporting, or search the skies of sightings for new phenomena (Goodchild, 2007). Within the context of tourism, managers in a theme park can now easily monitor the flow of incoming visitors at a particular time during the day via the users' location data from the mobile app or RFID tag-embedded ticket. Importantly, these new technologies result in large digital 'footprints' so that destinations 'track' this information in order to build a more comprehensive picture of each visitor as they travel (and make choices) within the destination 'ecosystem.' As such, the new technologies empower both individual travelers and destination management organizations by connecting the real world and the digital world (Sagl et al., 2015).

2.1 The Quantified Traveler and Context-Awareness

As travelers move from one place (or activity) to another along their trip journey (e.g., Gretzel, Fesenmaier, & O'Leary, 2006; Jeng & Fesenmaier, 2002; Yoo, Tussyadiah, Fesenmaier, Saari, & Tjøstheim, 2008), the changing situations and surrounding environments may cause changes in decision-making and behavior (Lamsfus, Wang, Alzua-Sorzabal, & Xiang, 2014). For example, travelers often re-negotiate specific details of a trip when a flight is delayed for many hours; similarly due to physical fatigue, travelers might choose to postpone dinner, a walk through a park or simply going to a museum. Importantly, changes in context and subsequent behavior (in terms of spatial/temporal movements) can transform the way travelers interact and/or experience the destination (Kim & Fesenmaier, 2014; Yoo et al., 2008). As such, wearable devices enable us to track not only those physical behaviors from the external information they provide but also we can guess quite accurately what travelers are thinking and how they are feeling (e.g., emotional state) at a specific moment (Swan, 2012, 2013). Thus, it is argued that context is a fundamental aspect of the tourism experience and knowledge of travelers' context serves as the foundation for tourism design and development and from the destination marketers' perspective, understanding context and mobility empowers them with the ability to influence travelers' decisions in real time (Lamsfus, Martín, Alzua-Sorzabal, & Torres-Manzanera, 2015; Stienmetz & Fesenmaier, 2015). It is, therefore, argued that through the lens of the quantified traveler, there are many opportunities for tourism destinations to capture, understand, and interpret contextual information generated by wearable technologies connected to the Internet.

2.2 *The Quantified Traveler and Ordinary Life*

The data generated during our ordinary life offers huge potential to impact travelers' behaviors at the destination, and consequently, enhance tourism experiences at the destination. In recent years, several papers in the tourism literature (e.g., Gretzel, 2010; Pearce & Gretzel, 2012; Tussyadiah & Fesenmaier, 2009; Wang et al., 2012, 2014a) have shown that technological environments (e.g., smartphone, mobile devices) actively transform the way people travel across all stages of a trip by connecting the moment of tourists (i.e., the tourism journey) to their ordinary life. Further, they argue that the tourism experience is no longer clearly separable and distinguishable from everyday life. Although the basic motivation of travelling is to escape one's ordinary life and seek novelty, many travelers still want to do many of the same things they do in their daily life. For example, if people are on a diet they generally tend to continue within certain diet constraints (e.g., local cuisine, calorie intake, etc.); or if the traveler exercises daily, he/she might want to jog along a walkway, road or beach or workout in hotel's exercise room. In this regard, it is argued that the connection between daily life and tourism experience helps to increase the satisfaction of the tourism experience and, indeed, make their activity even more memorable.

Although the concept of quantified-self emphasizes the individual, it can be easily extended well beyond the scope of individuals to social groups (Swan, 2013). This is because people often share data about their lives (e.g., the level of happiness, walking distance per day) with others with the purpose of collective knowledge development, performance benchmarking and/or participation in social communities. Thus, it is possible that other entities such as actors, agencies, and organizations beyond the personal and private are able to access the information via such communities and/or cloud services and in turn, provide feedback (e.g., a solution, a discount coupon, etc.) in real time (Lupton, 2014). This can be tremendously important and become common in the near future in that recent developments enable us to learn something from others by sharing and comparing how each are doing individually and ultimately discovering the meaningful information and insights from the collective actions (Lupton, 2014; Swan, 2013). With these advantages, businesses may 'repurpose' these data to create commercial value, although the basic data created is based purely on personal activities with a voluntary engagement (Lupton, 2014). As these technologies are being increasingly integrated into everyday life through our phones, clothes and home appliances, travel and tourism can be seen as a field of logical extension of the concept of quantified self, particularly due to its potential applications for smart tourism development (see Fig. 2).

3 The Quantified Traveler and Smart Tourism Development

Smart Tourism refers to the convergence of information technologies, business ecosystems, and tourism experiences (e.g., Gretzel, Koo, Sigala, & Xiang, 2015; Gretzel, Sigala, Xiang, & Koo, 2015). Importantly, Gretzel, Koo, et al. (2015)

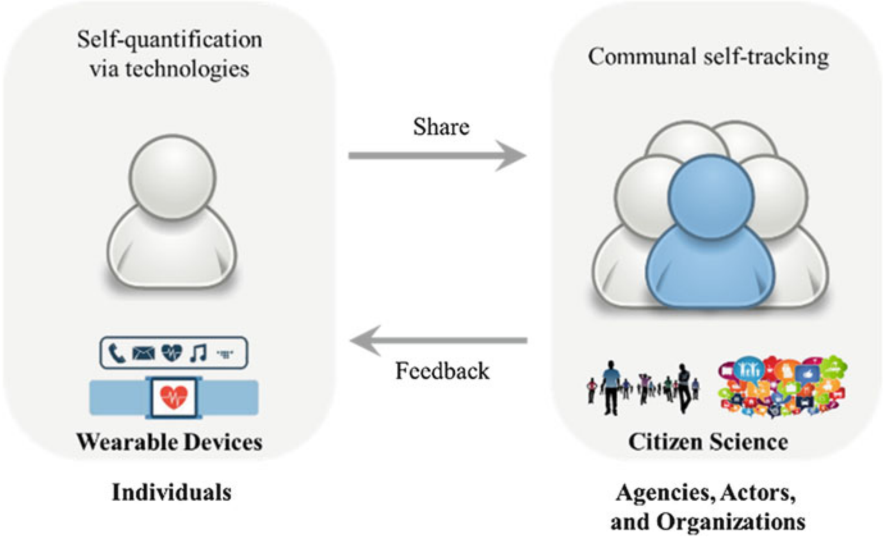


Fig. 2 Data sharing and feedback loop in the 'Quantified-Self' community

argues that the core technology of smart tourism are sensors and mobile devices which enable destinations to create the pervasive technological environments which destination marketers can use to anticipate travelers' needs in real time so as to enhance their experiences and enable the sharing of one's tourism experiences. Thus, they posit that smart tourism development requires destinations and companies to integrate personalization, context-awareness, and real-time monitoring through information collection, ubiquitous connectedness, and real-time synchronization into their management efforts (Gretzel, Sigala, et al., 2015; Neuhofer, Buhalis, & Ladkin, 2015). Within this context, it is further argued that the notion of the quantified traveler holds the key to understanding how today's wearable devices and technologies contribute to the tourist experience and how they can be used to assist smart tourism development. Specifically from a service design and system development point of view the quantified traveler: (1) provides data for context-awareness, (2) connects with one's historical data from everyday life, and therefore, (3) enables us to understand the traveler's interactions with the environment.

Another key to smart tourism development lies in our understanding of how the traveler interacts with and within physical and social environments. Technologies (e.g., wearable devices, sensors, and other agents connected to the internet) have an important but implicit role in facilitating the interaction between travelers and their environment (see Fig. 3). Indeed, technologies have been considered as an effective instrument to create, support, and reinforce tourism experiences by providing information, broadening the choice of traveler behaviors, and enabling travelers

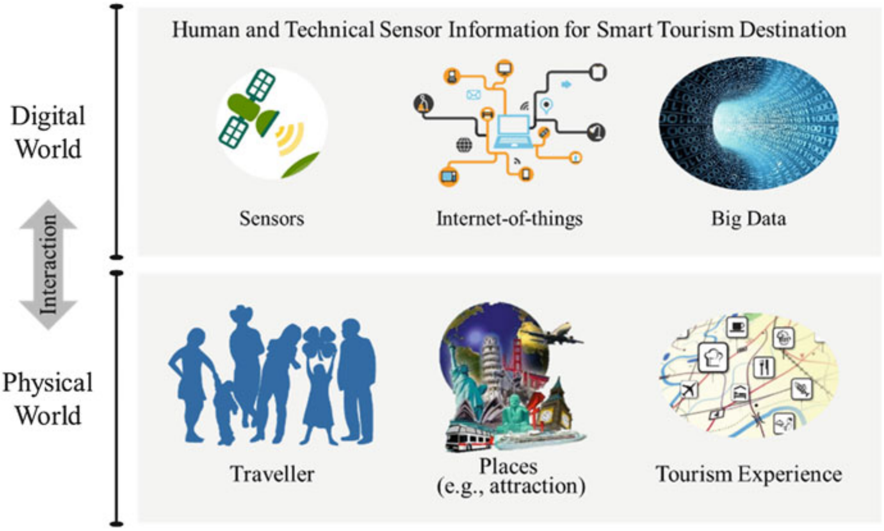


Fig. 3 Context-enriched human and technological sensor information for Smart Tourism Destinations

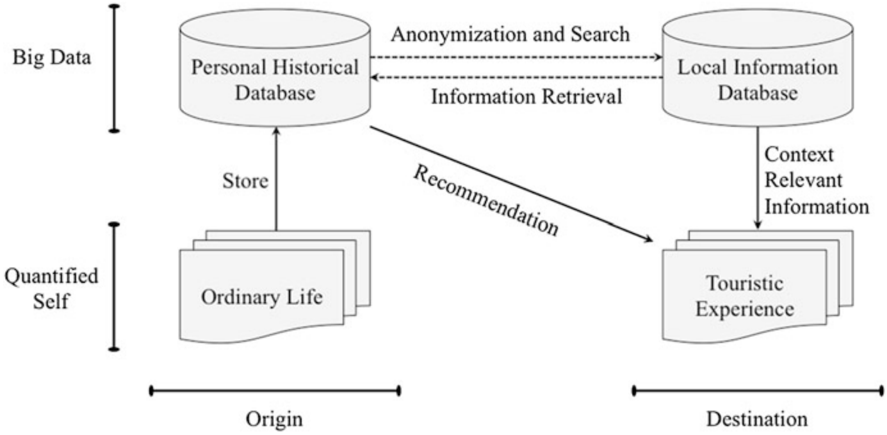


Fig. 4 A basic system integrating the quantified traveler and the touristic experience for SMART tourism destinations

to share their experiences with their family and friends even at the destination (Gretzel, 2010; Tussyadiah & Fesenmaier, 2009; Wang et al., 2012). Among them, information searching and retrieving behaviors are the most vital functions that impact traveler behaviors and experiences (Gretzel, 2010; Wang et al., 2012). Importantly, the use of these technologies affords travelers the ability to create and/or manage their own tourism experiences by not just passively receiving the information from the destination and tourism marketers, but by actively and dynamically engaging in activities within the destination (Zach & Gretzel, 2012).

The quantified traveler provides not only contextual information during travel but also personal historical data generated during ordinary life and the connection of that information to the touristic experience (Wang et al., 2014a, Wang, Xiang, & Fesenmaier, 2014b), which can be used for smart tourism development. An example of a system which uses the data created by monitoring these relationships is illustrated in Fig. 4. That is, by exploiting the increased use of wearable devices and sensors, the physical state (e.g., purchase history, movement, and search history) and the emotional state (e.g., mood, feeling, heartbeat) can be tracked unobtrusively and then stored in real time. Further, this data will be expanded as previously existing data is integrated in the system (Andrejevic & Burdon, 2014).

4 Conclusion

This chapter proposes a framework for assessing the potential use of the concept of quantified-self movement (and wearable devices) by integrating individual travelers' previous behaviors and stored sensor data in their ordinary life into system development during travel. This framework consists of components that systematically encode a disparate sources of heterogeneous personal historical data—individual-level big data—collected from the quantified-self devices and interpret those data to be exploited and explored by a recommender system in conjunction with various contextual information (e.g., local information, weather). This framework considers a wide range of applications and their affordances for contributing to, or enhancing, the touristic experience. As shown in Fig. 5, various affordances of emerging systems can be organized on two axes where they support the individual vs. place and where the various measures are monitored on a daily basis or are only trip-related. For example, the nature of data collected for health occurs on a daily basis and on a personal level; this contrasts to hotel or event reservations which are related to places and are trip specific. Further, Fig. 5 illustrates (see the connected lines) that some of the aspects of daily life such as dinning preferences, communication with family and friends, etc. can easily extend into the travel experience using emerging mobile technologies. Many other connections (and related affordances) can be mapped using this framework.

The following identifies some possible applications of the quantified traveler concept in smart tourism development.

A Persuasive Recommendation System A recommendation system is the most basic but important benefits from the proposed framework and the developed technologies. This system could integrate not only our travel behaviors and stated preferences within the destination but also our historical data (e.g., emotion, habit) and hidden preferences into the suggestions. For example, this system could

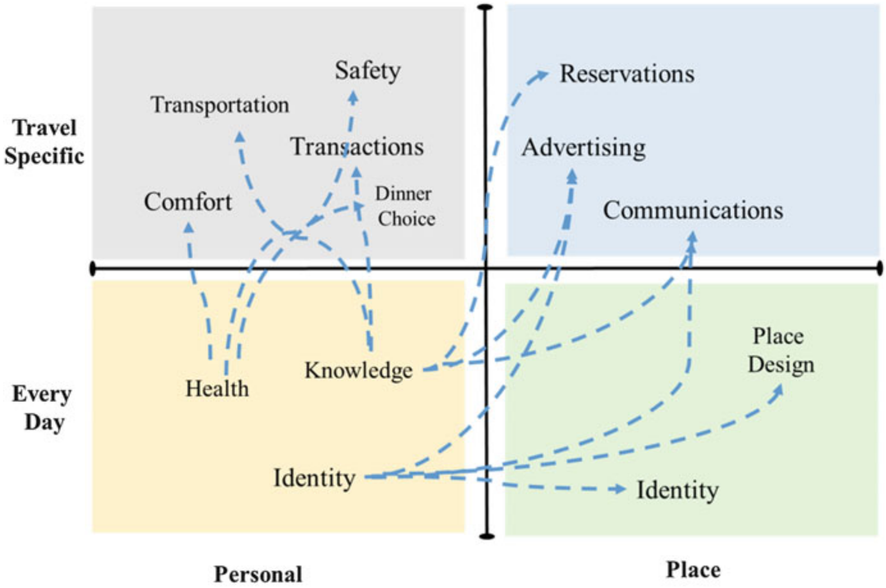


Fig 5 A framework for the quantified traveler

recommend a route for the best jogging route for someone who always run in the morning. In addition, if a traveler walks too much during the trip compared to their original life, this system could suggest him/her to take a little rest at the must-visit restaurant and/or coffee shop depending on their habits.

An Automated and Personalized Hotel Service A wearable device can monitor our body temperature and sleeping habit and then, transfer the information to the sensor installed in a hotel room so that the room environment can be adjusted automatically during the night. Room temperature, light bed, and morning alarm system might be an example for this system.

An Automated Trip Album A device keeps monitoring the entire journey of one's trip to a particular destination. Since the collected data consist of many different types of data format (i.e., photo, video, emotion level, and movement) and are in a huge amount, a device could detect important (and memorable) events based on our saved physical and emotion state automatically as well as manually. This event log will be able to create an automated trip album by an individual traveler and shared with his/her social networks.

A Real-Time Feedback System In order for tourism marketers and destination managers (e.g., theme park, attraction, and hotel operations) to control the quality of their products and services, the devices can keep tracking all the possible situations and provide a continuous but automated feedback to their customers while staying at the hotel and/or enjoying at the tourism attractions. This system could be operated by using a wearable device and/or tag-embedded ticket which can manage waiting time and service failures.

The notion of the quantified traveler provides both opportunities and challenges for the tourism industry. In general, advanced technologies embedded within mobile systems can be used to empower both supply-side (i.e., destinations and tourism businesses) and demand-side (i.e., travelers) to identify, customize, and purchase/produce tourism products (Andrejevic & Burdon, 2014; Sagl et al., 2015; Swan, 2012). Ubiquitous devices and information distributed via these devices can be considered extensions of our five senses, bodies, and minds by repurposing the previously considered role to a more creative facilitator (Kim & Fesenmaier, 2015; Lupton, 2013; Swan, 2012). As such, we can extend or enhance our senses (e.g., voices, gesture, and sight) so as to directly connect to places within the destination or other places and people (e.g., restaurant, shopping, events, and so on). Consequently, travelers are able to access much more diverse information and encounter more possibilities to be creative than ever before, which in turn, enables the traveler to have even more memorable experiences. In this new 'extensive' world, destination and tourism managers can monitor the entire journey from the beginning when a traveler dreams about the destination (and even before actually starting to plan their trip) to the moment that a traveler returns to their daily life and shares their experiences with others.

The way we travel to a destination and the experiences we have at the destination have been constantly evolving because of technology. Now, the advent of the systems supporting the quantified traveler serves as a new generation of tools revolutionizing how people travel. In this new world of the quantified traveler, wearable devices will be used to capture the entire journey (i.e., behavioral outcomes as well as the bodily state) wherein all aspects of the trip can be 'matched' perfectly to the individual traveler in a seamless, unobtrusive fashion. Further, it is argued that these new technologies will induce changes in the value creation process wherein travelers become more creative in designing their trip in a way that closely fits their distinctive travel needs, values, preferences, and so on. Of course, these new devices and new 'informational ecosystems' threaten traditional information channels that simply provide basic destination related information and/or recommendations. As such, the emergence of the quantified traveler requires the destination to develop more dynamic strategies so as to empower each visitor to choose his/her own unique "activated path" depending on his/her needs. Simple examples of these new services include those described as the 'sharing economy' such as Uber or CarShare which are responsive to the immediate needs of the travelers. How to take advantage of these emerging systems, destination managers need to understand better their own products and services within the context of how to design them so as to interact directly with travelers within the destination (Stienmetz & Fesenmaier, 2013, 2015). Several considerations should be emphasized in order for tourism managers to respond to these new innovations. Importantly, gamification, ambient notification, and narrative storytelling should be used to inspire people to adapt those technologies for their own purpose. Additionally, privacy concerns is a very important issue that should be addressed (e.g., Andrejevic & Burdon, 2014) wherein analyses conducted via machine learning are anonymized. Nonetheless, it can be expected that the "big data" generated in travelers' everyday life and during travel as well as the potential business intelligence created based upon these data can serve as the building blocks for the development of smart tourism destinations. With this said, it is argued that the tourism industry is on the verge of a new revolution which will change not only the tools used to plan travel and the way we create travel experiences, but the nature of the tourism industry itself.

References

Andrejevic, M., & Burdon, M. (2014). Defining the sensor society. *Television & New Media*. doi:10.1177/1527476414541552.

Augemberg, K. (2013). *Self-quantification vs self-tracking*. http://measuredme.com/2013/06/self-quantification-vs-self-tracking/

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems—CHI'14* (pp. 1143–1152). New York: ACM Press.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(November), 211–221.

Gretzel, U. (2010). Travel in the network. In M. Levina & G. Kien (Eds.), *Post-global network and everyday life* (pp. 41–58). New York: Peter Lang.

Gretzel, U., Fesenmaier, D. R., & O'Leary, J. T. (2006). The transformation of consumer behaviour. In D. Buhalis & C. Costa (Eds.), *Tourism business frontiers: Consumers, products and industry* (pp. 9–18). Burlington, MA: Elsevier.

Gretzel, U., Koo, C., Sigala, M., & Xiang, Z. (2015). Special issue on smart tourism: Convergence of information technologies, experiences, and theories. *Electronic Markets, 25*(3), 175–177.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*. doi:10.1007/s12525-015-0196-8.

Jeng, J., & Fesenmaier, D. R. (2002). Conceptualizing the travel decision-making hierarchy: A review of recent developments. *Tourism Analysis, 7*(1), 15–32.

Kim, J. J., & Fesenmaier, D. R. (2014). Measuring emotions in real time: Implications for tourism design. *Journal of Travel Research*. doi:10.1177/0047287514550100

Kim, J. J., & Fesenmaier, D. R. (2015). Designing tourism places: Understanding the tourism experience through our senses. In *2015 Tourism Travel and Research Association International Conference*. Potland, Oregon.

Kopp, J. (1988). Self-monitoring: A literature review of research and practice. *Social Work Research and Abstracts, 24*(4), 8–20.

Lamsfus, C., Martín, D., Alzua-Sorzabal, A., & Torres-Manzanera, E. (2015). Smart tourism destinations: An extended conception of smart cities focusing on human mobility. In I. P. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 363–375). Heidelberg: Springer.

Lamsfus, C., Wang, D., Alzua-Sorzabal, A., & Xiang, Z. (2014). Going mobile: Defining context for on-the-go travelers. *Journal of Travel Research*. doi:10.1177/0047287514538839.

Lupton, D. (2013). Quantifying the body: Monitoring and measuring health in the age of mHealth technologies. *Critical Public Health, 23*(4), 393–403.

Lupton, D. (2014). Self-tracking modes: Reflexive self-monitoring and data practices. In *the "Imminent Citizenships: Personhood and Identity Politics in the Informatic Age" workshop*. Canberra, AU. http://ssrn.com/abstract=2483549

Marcengo, A., & Rapp, A. (2014). Visualization of human behavior data. In *Innovative approaches of data visualization and visual analytics* (pp. 236–265). IGI Global.

Neuhofer, B., Buhalis, D., & Ladkin, A. (2015). Smart technologies for personalized experiences: A case study in the hospitality domain. *Electronic Markets, 25*(3), 243–254.

Pearce, P., & Gretzel, U. (2012). Tourism in technology dead zones: Documenting experiential dimensions. *International Journal of Tourism Sciences, 12*(2), 1–20.

Sagl, G., Resch, B., & Blaschke, T. (2015). Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities. *Sensors, 15*(7), 17013–17035.

Smarr, L. (2012). Quantifying your body: A how-to guide from a systems biology perspective. *Biotechnology Journal, 7*(8), 980–991.

Stienmetz, J. L., & Fesenmaier, D. R. (2013). Traveling the network: A proposal for destination performance metrics. *International Journal of Tourism Sciences, 13*(2), 57–75.

Stienmetz, J. L., & Fesenmaier, D. R. (2015). Estimating value in Baltimore, Maryland: An attractions network analysis. *Tourism Management, 50*, 238–252.

Swan, M. (2012). Sensor mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks, 1*(3), 217–253.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1*(2), 85–99.

Tussyadiah, I. P. (2013). Expectation of travel experiences with wearable computing devices. In Z. Xiang & I. P. Tussyadiah (Eds.), *Information and communication technologies in tourism 2014* (pp. 539–552). Cham: Springer.

Tussyadiah, I. P., & Fesenmaier, D. R. (2009). Mediating tourist experiences. *Annals of Tourism Research, 36*(1), 24–40.

Wang, D., Park, S., & Fesenmaier, D. R. (2012). The role of smartphones in mediating the touristic experience. *Journal of Travel Research, 51*(4), 371–387.

Wang, D., Xiang, Z., & Fesenmaier, D. R. (2014a). Adapting to the mobile world: A model of smartphone use. *Annals of Tourism Research, 48*, 11–26.

Wang, D., Xiang, Z., & Fesenmaier, D. R. (2014b). Smartphone use in everyday life and travel. *Journal of Travel Research*. doi:10.1177/0047287514535847.

Yoo, Y., Tussyadiah, I. P., Fesenmaier, D. R., Saari, T., & Tjøstheim, I. (2008). Emergent distributed narratives in spatiotemporal mobility: An exploratory study on mobile 2.0 services. In *Proceedings of the Annual Hawaii International conference on system sciences* (pp. 85–95).

Zach, F., & Gretzel, U. (2012). Tourist-activated networks: Implications for dynamic bundling and en route recommendations. *Information Technology & Tourism, 13*(3), 239–257.

Part III
Tourism Geoanalytics

Geospatial Analytics for Park & Protected Land Visitor Reservation Data

Stacy Supak, Gene Brothers, Ladan Ghahramani, and Derek Van Berkel

1 Introduction

Globally, parks and protected lands (PPL) receive about 8 billion visits annually (Center for Responsible Travel, 2016). The United Nations Environment Programme predicts that ecotourism, nature, heritage, cultural and "soft adventure" tourism will grow rapidly over the next two decades, with global spending on ecotourism expected to increase at a higher rate than the tourism industry as a whole (United Nations Environment Programme, 2011). Therefore, it is increasingly important that PPLs are managed for both the enjoyment of visitors and the protection of natural resources. These sometimes antithetical objectives require managers to strike a balance between visitor impacts and visitor experiences. In this chapter, PPL refers to any natural area that supports recreational tourism on publicly managed land.

Reserving overnight space or access within PPLs helps improve visitor experiences within these natural settings. PPL reservation systems across the globe collect transactional data for reservations at campsites and day-use facilities, as well as permits for the use of trails and backcountry areas. This record of visitation contains information about both destination usage (from the supply side) and visitor behavior (the demand population) (Supak, Brothers, Bohnenstiehl, & Devine, 2015). Geospatial analysis and geovisualizations produced from these data can be used to promote PPL usage and facilitate sustainable visitor experiences; however, PPL reservation databases are rarely leveraged with these goals in mind (Supak et al., 2015).

At the end of the 1990s, the tourism literature started to acknowledge the benefits of using Geographic Information Systems (GIS) to conduct geospatial analytics and

S. Supak (✉) • G. Brothers • L. Ghahramani • D. Van Berkel
North Carolina State University, Raleigh, NC, USA
e-mail: sksupak@ncsu.edu

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_6

create geovisualizations for knowledge generation and decision support related to tourism planning and management (Bahaire & Elliott-White, 1999; McAdam, 1999). At the time, utilization of GIS was rare and awareness of the utility among tourism stakeholders was small. Since then, there has been progress in the use of geospatial analytics to support PPL planning and management, much of which has focused on visitor movement and flow monitoring within PPLs. For example, some studies have utilized GPS visitor tracking to investigate trail impacts (e.g., Beeco, Hallo, English, & Giumetti, 2013), identify spatial demand and competition for the use of acreage within PPLs (e.g., Beeco, Hallo, & Brownlee, 2014), issue visitor advisories (e.g., Chhetri, 2015), and make location-specific recommendations to improve visitor activity experiences (e.g., Wolf, Wohlfart, Brown, & Lasa, 2015). Others have utilized agent-based simulation to better plan for the development of tourist infrastructure such as paths, buildings and viewing platforms (e.g., O'Connor, Zerger, & Itami, 2005) or to explore the influence of alternative management options on recreationist movement, congestion, and crowding (e.g., Bishop & Gimblett, 2000).

Tourism and geography researchers have embraced the geospatial nature of PPL visitation. Studies that model determinants of park visits implicate park services, environmental characteristics like topography, climate, proximity to lakes and rivers, and attractive biomes as well as proximity to attractions and amenities (Hanink & White, 1999; Loomis, 2004; Neuvonen, Pouta, Puustinen, & Sievanen, 2010; Van Berkel, Munroe, & Gallemore, 2014). PPL characteristics as well as the quality of those characteristics have been linked to visitation frequency, with parks assessed as "higher-quality" receiving more visitors from a larger geospatial area than parks deemed "poorer-quality" (Hanink & Stutts, 2002; Hanink & White, 1999). In addition to considering PPL characteristics, visitors must balance the inherent trade-off between the investment of time, money or effort to achieve travel and the time they can spend at the destination (Mckercher & Lew, 2003; Neuvonen et al., 2010). Despite the rich amount of geospatial data collected through reservation transactions (e.g., visitor addresses, destination locations) and the clear influence of geospatial factors on visitation, these data remain largely unexplored.

Within the United States and world-wide, government agencies that manage PPLs can benefit from geospatial analytics of reservation data. Data collected from reservation systems are unlike data collected for traditional scientific inquiry, where collection takes place with specific questions in mind (Miller & Han, 2009). Although these records may not meet the conditions necessary for some types of statistical modeling (e.g., normality or independence), an inductive data mining approach can lead to the discovery of new and unexpected patterns, trends and relationships (Miller & Han, 2009). Geovisualization of these data sets can be a powerful strategy for initial knowledge discovery because the human brain is extremely effective in recognizing patterns, trends and anomalies. Geovisualizations displaying temporally aggregated and spatially summarized reservation data can be used to support decisions related to management objectives. For example, geovisualizations showing visitor attributes that have been summarized by PPL facility can help managers increase knowledge about the usage of a specific facility or group of facilities within a destination region. Examination of the same attributes, summarized and geovisualized by visitor origin can identify which communities are regular users of PPLs, as well as those communities that have historically underutilized recreational opportunities at PPLs. Insight into PPL visitor patterns of demand can help manage, improve and promote visitor experiences for different segments of the population.

Despite the apparent utility of geovisualizing reservation data to address specific PPL challenges, most of these data sets have yet to be leveraged. This is primarily because many organizations find implementing geospatial analytics and geovisualization challenging. The majority of GIS software applications necessary for creating geovisualizations are generic, complicated, and/or expensive; however, a trend is emerging related to the development of innovative and often collaborative, customized web-based mapping applications for creating and sharing geographic information (Haklay, Singleton, & Parker, 2008). Targeted web-mapping applications reduce the complexity and expense associated with investigating tourism reservation data (S. K. Supak, Devine, Brothers, Rozier Rich, & Shen, 2014); however, the only demonstration of a web-mapping application targeted for exploring reservation data is limited to geovisualizing the spatial frequency of visitor origins and it does not help users assess the quality of their data, preprocess their data, generate new attributes (e.g., lead-time or distances traveled), or query specific date ranges, destinations or visitor origins. More comprehensive web-mapping applications are necessary to make geospatial investigation of reservation attributes accessible to interested PPL managers, who may have no data management and GIS training. Until these applications are developed, individuals with data management and GIS skills are likely necessary to help PPL managers get the most out of their reservation data.

This chapter describes the value in utilizing spatiotemporal records of PPL visitation from both the facility and visitor origin perspectives. Section 2 of this chapter examines the nature of reservation data, lessons from private sector tourism, as well as the preprocessing, enrichment, and data mining processes. Further, Sect. 2 describes how Federal and local agencies tasked with tourism and resource management can utilize information generated from geographic data mining to geovisualize both PPL demand populations and destination usage. Section 3 introduces an example PPL reservation data set that includes approximately 12.5 million camping, permitting or ticketing reservations made for U.S. Federal PPL facilities during the years 2007–2015, the majority of which are for overnight stays. Section 4 demonstrates specific management decision support knowledge that can be gained through geospatial analytics of U.S. Federal PPL reservation data. Section 5 discusses the future uses of mined and enriched U.S. Federal PPL reservation data. Section 6 presents a brief review of how geospatial analytics of PPL reservation data can be useful in decision support.

2 Working with PPL Reservation Data Sets

Reservation data in their raw form are not entirely informative and often have accuracy issues that need to be resolved before they are useful for spatiotemporal interpretation by managers. PPL reservation data will at a minimum include destination details, visitor details, order dates, start dates, end dates and administrative information (e.g., fees paid). However, depending on setup of the reservation system, different levels of organization and data quality may exist. Destination information could simply be the name of the reserved facility or it could include details about the facility type and location. In terms of visitor details, names, addresses, party sizes, and credit card information is typically collected and stored, but several of these attributes may not be available for analysis due visitor privacy laws. When full visitor origin addresses are unavailable, typically visitor zip codes are still available for analysis.

The level of geospatial precision recorded in a PPL reservation system may vary for both visitor origins and destinations. At the highest level of geospatial precision, a reservation system would record exact geographic coordinates for PPLs and full home addresses for visitors, which can be converted to exact geographic coordinates through a process called geocoding. If a reservation system lacks these highest levels of precision, it still will likely contain origin and/or destination location information in the form of geographic identifiers such as State, Province, Territories, County, or zip code. Any of these geographic identifiers can be used to geospatially locate reservations by visitor or by destination. Lack of a ‘geospatial strategy’ in record keeping may result in the complete absence of geographic identifiers. If geographic identifiers of visitor origin are not collected within the reservation system, exploration of geospatial demand will not be possible; however, the absence of destination geographic identifiers can be overcome through certain data enrichment processes described later in this chapter.

2.1 Lessons from Private Sector Tourism

Traditional application of geospatial analytics for strategic planning in tourism have been limited to corporate efforts (corporate hotel chains, attractions, and resort destinations) and resource supply-chain flows rather than destinations (Chen, 2007). However, increased use of Internet reservation sites has produced large amounts of empirical data that potentially offer rich insight into tourism behavior (Crnojevac, Gugić, & Karlovčan, 2010). These transactional records are increasingly being used to generate performance metrics. Currently, practitioners in the travel and tourism industry are gaining a better understanding of visitors through the examination of booking rates, stay duration, number in the party and visitor spending.

While some large tourism companies do their data collection and analytics in-house, other companies, destination marketing organizations and tourism agencies contract with Data-services companies. For example, a state office of tourism can purchase a summary of transactions made within thier state from credit cards with billing addresses outside the state. Using the longitudinal data summaries from the transactions, the tourism managers can identify peak visitation periods, determine origin shares within their destination and relative return on marketing efforts for selected origin markets. These data also can be utilized to identify higher-spending visitors and off-season vacancy gaps to fine tune messaging content and timing as well as break into new markets (Lansky, 2016). Knowledge derived from demand population analytics is rapidly replacing the previously relied upon national trend data.

Data-services companies are increasingly providing tourism practitioners with the resources necessary for meaningful spatiotemporal interpretation of visitor data in an effort to create fully data-driven marketing campaigns. These companies are providing destination specific data which include top feeder markets, lengths of stays (duration), booking to arrival time (lead-time), average visitors per booking, search to booking time, visitor spending (from aggregate summaries of credit card transactional information for visitors to their destination) and digital exposure (e.g., media they see before arriving, click through rates of various marketing campaigns) (Lansky, 2016). Many of these companies also break down spending by lifestyle profile. Various data-services companies (e.g., adara.com, airsage.com, buxtonco.com) utilize a mix of Online Travel Agencies, first-party partners, mobile phone data, cookies, and credit card data to deliver their products (Lansky, 2016). As a result, the most comprehensive data-driven support may involve contracting with multiple data service companies.

Methods for gaining knowledge through geospatial analytics and geovisualization of travel reservation data are well established, as the private sector has demonstrated, but these methods are rarely employed for PPL decision support. Without the data that data service companies provide, PPL managers can use in-house collected historical reservation data to identify many of the same destination specific attributes including top feeder markets, lengths of stays (duration), booking to arrival time (lead-time), and average visitors per booking. While corporate decisions aim to increase revenue, these same attributes also can be used to gain knowledge about destination usage as well as geospatial visitor demand, behavior and characteristics. If PPL managers desire this type of knowledge, the raw spatiotemporal reservation data must first be preprocessed, enriched, selected and reduced.

2.2 Preprocessing and Enriching PPL Reservation Data

While geospatial attributes available for analysis are determined by the setup of the reservation system, incomplete values within the records due to human error can present significant data preparation challenges. Common omissions and domain violations in PPL reservation databases could include visitor zip codes not being entered at all or entered incorrectly (e.g., anything other than a five- or nine-digit number). Non-numeric or incomplete transactional records therefore often need to be preprocessed as the first step toward producing new information. Data preprocessing involves "cleaning" data to eliminate duplicate records and determining strategies for handling missing data fields and domain violations (Miller & Han, 2009). For example, nine-digits zip codes may need to be reduced to five-digit zip codes. Data preprocessing can be conducted using a scripting language such as python, which is free, but requires coding ability. Alternatively, there are more user-friendly Graphical User Interface options that are open source (e.g., openrefine.org [formerly Google Refine]) or open source with premium support subscriptions (e.g., datacleaner.org). After preprocessing, data enrichment can provide opportunity for richer geospatial analytics. Avenues for data enrichment are detailed in the following three sections.

Generating new meaningful attributes from attributes that exist within the data is a critical step in extracting information and eventually knowledge from reservation data. The following attributes should be created when possible:

Origin-destination location pairs

- The great circle distance traveled between the visitor and the destination
- The least cost road network travel distance for visitor-origin pairs, when the visitor was likely to travel by motor vehicle (e.g., under 300 great circle miles)
- Total number of alternative PPLs that were closer to the visitor's origin than the reserved PPL
- The number of alternative PPLs within a set travel distance (e.g., 50 miles) of the reserved PPL

The date fields:

- The lead-time: number of days between order date and start date
- The duration: end date—start date
- The person-nights: the number of people participating in the reservation multiplied by the duration. This is a measure of the cumulative human occupancy and subsequent cumulative impact to the natural setting

Customer IDs (when available)

- Number of stays in the last year by customer ID
- Total number of nights stayed by customer ID

Destination Clusters (neighborhood groupings of PPL facilities)

- Number of unique facilities in each cluster
- Number of unique agencies in each cluster
- Total number of reservations for each cluster

2.2.1 Enrichment from Visitor Origin Geography

A 10-year study of outdoor recreation visitor behavior in California shows that visitors' "recreation style," defined as gender, age, ethnicity, spoken language, social status, and socioeconomic status influences recreation choice, and attitudes of visitors toward natural resources (Chavez, 2001). Predicting tourist behavior in order to provide better services and protect natural resources through understanding "recreation style" is a fundamental key to successful management actions (Manning, 2014). When visitor origins are available, external secondary data that includes "recreational style" attributes can be joined to each reservation record based on the origin location. The joining of relevant geodemographic data, which estimate the most probable characteristics of people based on the pooled profile of all people living within the area, can be valuable for decision support. When aiming to identify underserved communities, specific attributes such as median age, median home value, median income, and percent of population over 25 years with a bachelor's degree can help explore socioeconomic disparities. Further, these attributes along with observed visitation numbers could be used to model or predict tourist behavior.

"Recreation style" attributes typically are available at some level of geospatial aggregation (e.g., census tract, zip code, county, state, etc.). Secondary geospatial data is increasingly freely available, which can potentially benefit PPL managers that often have budgetary constraints. In the U.S., demographic characteristics describing the residents of a particular area can be obtained for free from the U.S. Census (http://factfinder.census.gov/). Other population characteristics that may help enrich the data such as consumer spending on travel in the U.S. can be found at the Bureau of Labor Statistics (http://www.bls.gov/data/). Proprietary services such as ESRI's business analyst online (http://www.esri.com/software/businessanalyst/) provide thousands of variables helpful in producing location-driven market insights that may also help PPL managers better understand their demand populations. If the visitor origin and the order date are both considered, location specific historical data also can be joined to the reservation data set (e.g., gas prices, weather, or terror threat index).

The locational precision of visitor origin in the reservation data set will dictate the resolution of the secondary attribute data that can be joined. For example, if full U.S. addresses for visitors are available, the geocoded points representing visitor origins would need to be spatially joined to the census tracts in which it falls, because that is the smallest discretization of space employed by the U.S. census bureau. If visitor origin data is collected only at the zip code level, it should only be joined to other zip code level data or data that has a coarser spatial resolution such as a county or state. Joining zip code level visitor origins to census track level attributes would not be appropriate. It should be mentioned that associating demographic characteristics to populations within zip codes is not without weakness. See Miller (2008) for a detailed discussion of the benefits and limitations of using zip codes and census block groups for demographic proxies.

2.2.2 Enrichment of PPL Destinations Attributes

External secondary data can also enrich a reservation data set to enhance what is known about each specific PPL facility. This can be accomplished in a few ways. First, demographic data previously utilized in the context of visitor origins or other external data (e.g., historic weather data, remoteness index, etc.) can be joined to specific PPL's based on PPL location, as these attributes might influence reservation choice. Second, attributes describing the amenities or activities offered at a specific PPL facility could be joined to the reservation data, if that data exists electronically in a separate database. Another method for obtaining amenity or activity attributes is through the collection of the content that PPL's advertise on the web. For example, the recreation.gov website is a reservation management system that advertises public-resources for tourism, provides location information for campsites, ticketing and permits, describes the amenities available at these destinations, and provides an overview of destination facilities. This amenity data, when available, can be joined to reservation data either geographically or through common identification codes. Importantly, advertised location data found on web pages can be used to create or quality control PPL facility locations stored in reservation systems or supporting databases.

Automated collection of amenity data is possible using web scraping and web crawling techniques. Web scraping uses an algorithm to access and parse content of a website through an automated process which methodically browses a webpage through the website's url links (Olston & Najork, 2010; Thelwall & Stuart, 2006). The consistent structure of website content and organized hierarchies of web pages allows for this systematic information retrieval (Hirschey, 2014). Using such techniques makes it possible to collect large amounts of web content relevant to evaluation of PPL tourism (e.g., amenity details offered by specific campsites).

Implementation of scraping and crawling has become increasingly easy through specialized software. Freely available software (e.g., Kimono, import.io) automatically detect data structure and enable bulk download of website content depending on the web page design. Moreover, scraping software provides a recording interface where the user's actions (e.g., search, page navigation) can be mimicked for complex web searches. While highly user-friendly, such software can be less effective when websites are dynamic (e.g., pop-ups, advertising) or poorly structured. In these cases, coding may be required. High-level programming languages such as Python, Perl and R offer different general and specialized packages that can read and parse text or unicode data for these cases.

2.3 Data Reduction & Geographic Data Mining

Cleaned and enriched reservation attributes are ready to be distilled through data mining practices, so that they may be interpreted within the existing contextual understanding (Miller & Han, 2009). Geographic information is commonly broken into the components of space, time and attribute (Chrisman, 2001). In order to measure one component (time, space, or attribute), one of the other components has to be fixed while the third serves as a control. In classic 2D geovisualizations, time is fixed, space is controlled and attributes are measured. Time is fixed by choosing a specific moment to capture or by aggregating the attribute of interest over some temporal subset of the data. Temporal subsets could be a single day, month, season, or year or the entire time frame over which the data were collected. Selecting temporal periods for aggregation can influence the knowledge discovery process, especially if there are large changes in the attributes over the temporal expanse of a data set. Therefore, management objectives should be considered prior to temporal data subsetting. For example, subsetting the data temporally (e.g., year, season or month) can help identify longitudinal or seasonal changes in visitor behavior. PPL bookings are likely to include temporal trends, such as peaks around holidays, seasonal attractions (e.g., autumn colors of temperate deciduous forests) and 'good' weather that indicate peak tourism demand. Space is then used to control measurement of the attribute for the fixed time frame. PPL reservation data includes spatially explicit origins and destinations, therefore either of these spatial identities could be utilized to control attribute measurement. To measure attributes, an appropriate summarization technique needs to be selected. Standard techniques include reporting for each spatial unit: the count of observations (e.g., total reservation count), the sum of observations (e.g., total number of person-nights) or a descriptive statistics such as median or quantile value (e.g., median distance traveled).

Space, time and attribute decisions for geovisualizations greatly influence what information is communicated. If U.S. zip codes are the most precise geospatial visitor identifier, zip code boundaries that have been drawn to discretize the U.S. will control how the attributes are summarized over chosen time periods. If U.S. counties are the most precise geospatial visitor identifier, county boundaries should be used. If attributes are aggregated or fixed over 1-month time frames, an animation of the individual 1-month geovisualization snapshots could show change over time. Alternatively, including a third dimension in the geovisualization would allow for the temporal aspect of the attributes to be shown; however, depending on the number of records, temporal aggregation may still be necessary, since it is not always possible or useful to render individual tourist activity (Gahegan, 2009). Once space is controlled and time fixed, various attributes can be measured for the entire data set or for a subset of the data. Subsetting the data by some attribute (e.g., visitor origin, PPL facility, PPL facility cluster, lead-time, duration, number in party) can be useful in investigating usage and demand. For example, subsetting the data for a specific PPL facilities or cluster of proximal facilities allows for more precise examination of the geographic dispersion of visitor demand, travel behavior and visitor characteristics for these facilities. Subsetting the data by some property of the PPLs, such as management agency, or specific site type may also provide insight into visitor preference or usage. Regardless of the geovisualization technique employed, distilling data for decision support relies on choosing appropriate time frames to aggregate over, spatial units to control the measurements, and summarization of attributes of interest.

2.4 Utilizing Information Generated from Geographic Data Mining

Data created during the data reduction and geographic data mining process can be stored in a data warehouse, which in contrast to traditional transactional database design should maximize the efficiency of analytical data processing and data examination to support decision making (Miller & Han, 2009). Geographic data warehouses are increasingly necessary for the exchange of information through web-mapping services and location-based mobile applications. A good geographic data warehouse should support online analytical processing tools that provided multidimensional summary views of the data. Further, these warehouses can support data cubes, a powerful and commonly applied online analytical processing tool that creates a set of all possible aggregations based on a particular attribute (e.g., person-nights) and some dimension of interest (e.g., PPL destination or month). These aggregations can be over the zero dimension (e.g., total person-nights), one dimension (e.g., total person-nights by PPL destination or total person-nights by month), two dimensions (e.g., total person-nights by PPL destination and month), continuing on the Nth dimension.

If the infrastructure for data warehousing is not available, the tabular outputs of specific data reduction and mining efforts can be loaded directly into a GIS. If the controlling spatial identifier is descriptive (e.g., zip code, county) and does not contain a coordinate set, data containing the same descriptive attributes as geometric objects (e.g., zip code centroid points) are needed for a tabular join. Some geometric object data is freely available from ESRI (ArcGIS Zip Code Layers, 2016). If coordinates can be associated with the reservation data set, most GIS applications can easily allow for the creation of data layers from coordinates. Regardless of the technique utilized to spatially enable the data, the resulting data layers are typically stored as thematic layers where the geospatial location (i.e. zip code centroid point or zip code polygon) is associated with the summarized attribute information (numeric or string type) for use in a GIS (Gomez, Haesevoets, Kuijpers, & Vaisman, 2009).

Geospatial analysis and geovisualization of PLL reservation data requires strategies for describing visitation trends and tourist behavior that match the needs of administrators. Numerous techniques are available for spatiotemporal data exploration and below we outline several basic, but effective techniques that are viable for PPL management needs. While new sophisticated 3D geovisualization techniques that present the spatial and temporal movements of tourist (Zhao, Forer, Sun, & Simmons, 2013) have been developed, these type of geovisualizations are most appropriate when tracking a visitors movement within a destination. While reservations capture a single tourist's movement from their home to the destination, these paths are usually not complicated, and may not be worth showing.

Basic geovisualizations that presents attribute measurements summarized by PPL facility should display these attributes as points. While basic geovisualizations that presents attribute measurements summarized by visitor origin are most effectively presented as a choropleth map (also called a thematic map) that discretizes the entire area of interest into distinct spatial units (e.g., zip codes). Choropleth maps assign a uniform color or pattern representing a single attribute's value to each spatial unit. The human eye can easily distinguish hot spots and cold spots, so variability in attribute values across geographic regions as well as the level of variability within a region can be easily determined using this technique. When working with U.S. zip codes, a complication arises due to the fact that there are ~40,000 zip codes and zip code centroids, while there are only ~30,000 tabulated zip code areas or polygons. This discrepancy results from the fact that ~10,000 U.S. zip codes have no geospatial footprint because they are post offices or single-site zip codes (e.g., government buildings, universities). To ensure minimal data omission, reservation attributes should be summarized by visitor origin and first joined to zip code centroid coordinates. Then using a GIS, any zip code centroids that falls within a zip code tabulated area can be spatially joined to that area. Through the spatial join process, attributes can be counted, summed or averaged within the tabulated zip code areas.

2.5 *Geovisualization for Pattern Interpretation of PPL Demand Populations*

In the tourism industry, there are well established benefits associated with geospatial analysis of the relationships of various internal and external data sets for the purposes of better market area understanding and customer profiling (Bell & Zabriskie, 1978; Elliott-White & Finn, 1997; Miller, 2008). Since visitation to PPL destinations can be characterized by highly localized utilization, both local and regional utilization and regionally to nationally dispersed utilization with few local residents reserving overnight accommodations (Supak et al., 2015), identifying origin market dispersion is important to characterizing demand populations. Market area definition and subsequent customer profiling of demand populations cannot be accomplished by simply defining distance rings or drive time polygons with respect for each PPL, but rather they should include techniques which account for the geographic dispersion of customers. Examining these dispersions over different fixed time frames allows for investigation of change in demand over time. Alternatively, all records in the enriched data set can be aggregated to identify long-term usage trends for the demand population.

Private sector tourism managers can now pay data services companies to provide historic visitor details including origin markets and other destination specific characteristics for characterizing demand populations. The cost of these services could be prohibitive for PPL managers. Fortunately, much of the data assembled by these companies can also be harvested from an enriched PPL data set (e.g., top feeder markets, lengths of stays, booking to arrival time, average visitors per booking). Geovisualizing the top feeder markets is one way to examine the demand population as well as identify underrepresented communities. Identifying underrepresented communities through market definition is desirable for PPL managers who aim to help engage and create our next generation of PPL visitors, supporters and advocates. For example, the Every Kid in a Park initiative in the U.S. provides an opportunity for every 4th grade student and their family to experience federal public lands in person throughout the 2015–2016 school year. Identifying the underrepresented counties, states and regions through geospatial analytics and geovisualization can help meet the goals of this initiative.

In addition to identifying locations of underserved communities, residents of these locations can be profiled to identify socioeconomic or geographic challenges that might prevent them from accessing public lands for recreational tourism. Tourism marketing has long tracked their customers' characteristics based on their geographic location for the purposes of understanding travel behavior (Miller, 2008; Supak et al., 2014). If visitor demographic and socioeconomic enrichment has occurred, these characteristics can be examined for specific communities as a crude geospatial market profiling technique. There are many examples of more sophisticated market segmentation modeling techniques employed to improve destination marketing such as clustering methods, mixture models, mixture regression models, mixture unfolding models, profiling segments, and dynamic segmentation (Wedel & Kamakura, 2000); however, simple geovisualizations can accomplish a less formal market segmentation analysis (Supak et al., 2015).

For PPLs within large management systems, reservation data can be leveraged to explore and characterize the demand populations for individual PPL facilities, clusters of proximal PPL facilities or at the whole system level. An effective way to allow managers to geovisualize an individual PPL's demand population would be to create a data warehouse, geospatial data cube and corresponding web-map, so that all reservations for a particular PPL could be queried and demand population characteristics summarized by visitor origin over some spatial unit. Data cubes would need to contain aggregate data across desired dimensions, either produced in advance or on the fly, including standard attribute summaries as well as pointers to geospatial objects (Gray et al., 1997; Miller & Han, 2009).

2.6 Geovisualization for Pattern Interpretation of PPL Destinations

Management strategies for effectively balancing visitor impacts and visitor experiences should include knowledge about the spatiotemporal variations of demand population characteristics as well as PPL usage characteristics. When an enriched PPL data set is summarized by PPL facility rather than by visitor origin, geospatial usage patterns of competing use can be explored. Geovisualizations showing aggregated longitudinal destination usage attributes (e.g., person-nights) can provide information to help support management objectives, such as modifying the distribution of visitor impact among facilities within a region. This type of evaluation can be particularly useful when managers want to assess human impact within a region in support of sustainable development and resource allocation decisions. Examining the geospatial relationships of multiple PPL attributes (e.g., total reservation count, lead-time, distances traveled, and duration) for multiple facilities within a specific region can help managers profile the region of facilities as a whole. With this knowledge, availability within the reservation system could be altered, allowing fewer visitors to reserve over-utilized PPLs in favor of neighboring alternative underutilized PPLs. Further, geovisualizations exploring travel distances by PPL can identify PPLs whose demand population is highly localized, both local and regional, or regional to national. If travel distances are small for a specific PPL facility or group of facilities, knowledge of this local utilization can be used for outreach purposes and for improving environmental stewardship.

3 U.S. Federally Managed PPL Reservation Data Set Example

To illustrate the benefits that can come for data analytics, and specifically geospatial data analytics, of outdoor recreation reservation data, we examine 12,473,816 reservation records made at 3272 unique U.S. federally managed PPL facilities. The National Recreation Reservation Service (NRRS) maintains a database that collects reservations open-endedly as part of the U.S. Government's Recreation One-Stop Program. Through collaboration with NCSU researchers, the Recreation One-Stop Program management team has made this historic reservation data set publicly accessible (http://ridb.recreation.gov/). Reservations recorded in this database include those for camping, permitting, or ticketing within PPLs that are managed by seven different U.S. federal partner agencies (e.g., National Park Service, Bureau of Land Management, and Forest Service). These reservations encompass both those made ahead of time by prospective visitors and those made by PPL staff at destinations when visitors arrive without reservations (personal communication DeLappe, 2016). Further, not all PPL recreation destinations managed by the federal government are included in the database; some are truly off the grid and have no Internet access. For these PPLs, some allow reservations through recreation.gov, while others do not.

The preprocessing of these data was accomplished with a python script (Python, 2016) that replaced special characters, stripped nine-digit zip codes down to five-digits, added place-holder end dates where none existed, and replaced empty field values with NaNs. This data set includes records for order dates that range from January 1, 2007 to December 30, 2015, with start dates that range from January 15, 2007 to December 29, 2016. While no records were found to be missing order date or start date values, end date values were missing from 300,804 records. Further, 282,857 records have an order date after the start date. These two values may be related, as many PPLs allow for payment after hours through a drop box and these reservations are then entered into the system after the start date (personal communication DeLappe, 2016). Another 2,563,418 reservation had the same order and start date, many of which were likely booked by PPL employees when visitors arrived at the destination without a reservation.

Figure 1 shows the percent of total reservation records (January 1, 2008–December 31, 2015) that were ordered or started on each day of the year. Data from 2007 was not included in this figure, as many reservations starting in 2007 were ordered in 2006. Order date spikes on the 15th of the month from December to April correspond to Yosemite National Park releasing camping blocks on those days (personal communication DeLappe, 2016). The general trend of the graph shows increasing utilization from the beginning of March to mid July, with decreasing utilization from mid July to the end of October. The U.S. Federal holidays of Memorial Day, Independence Day and Labor Day (the last Monday in May, July 4th, and the first Monday in September, respectively) each capture almost 1 % of the total reservation start dates. The order date dip around the 4th of July Holiday suggests recreation is occurring, keeping prospective visitors away from the reservation system.

Of the 3272 unique facilities, only 31 offer ticketing (e.g., White House Easter Egg Roll), which is likely not for overnight activities. Both overnight and day-use permits are offered at 51 facilities. While ticketing and permitting reservations may not necessarily include overnight stays, these reservations have not been removed from the data set used for this analysis. Given that there are 3190 facilities (97.5 %)

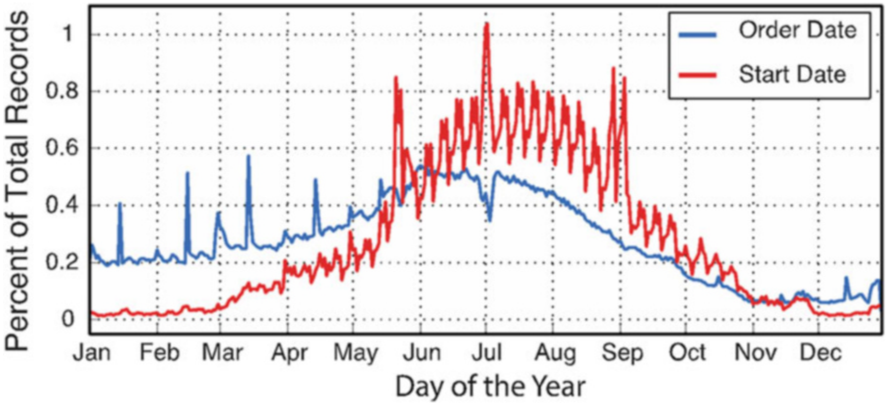


Fig. 1 Percent of total reservations for U.S. Federal PPLs whose order or start date was between January 1, 2008 and December 31, 2015 (n = 12,072,092) binned by day of the year

that provide overnight camping accommodations, it is fair to assume that most reservations represent overnight stays. Of the seven agencies that provide these services, five agencies offer more than two camping destinations and they include the US Forest Service (2233 unique facility names), the US Army Corps of Engineers (697 unique facility names), National Park Service (217 unique facility names), Bureau of Land Management (28 unique facility names), and the Bureau of Reclamation (11 unique facility names).

Fields provided in these historical data that pertain specifically to the destination include facility name, facility, agency, site type, facility zip code and facility coordinates. Of the unique facilities in the data set, many were missing zip code values, but only five were missing coordinates or had discord in coordinate values across the records. These facilities were manually searched and coordinates replaced for 7403 records associated with these five facilities. This led to 100 % geolocation of the facilities; however, inaccurate coordinates may exist within the PPL database. It should also be noted that these facility geolocations are far more precise than previously published facility locations from the same reservation system (Supak et al., 2015), as these geolocations were previously based on facility zip code centroids and not facility coordinates. This increased precision in PPL location allows for examination and comparison of specific facility attributes (e.g., total reservation count, median lead-time, median visitor travel distance) within a geospatial neighborhood, which was previously not possible.

For the demand population, approximately 98 % of reservations included a five-digit zip code of visitor origin. Visitor zip codes were geolocated by joining each record's visitor zip code to the known latitude and longitudes of the zip code centroid. To ensure the highest percentage of geolocation of visitors, a commercial zip code database was used (http://www.zip-codes.com/zip-code-database.asp), which contained 40,145 zip code coordinates. A tubular join of these data led to 96 % of the reservations being successfully geolocated for visitor origin. With both destinations and visitor origins geolocated, enrichment from attributes within the data set commenced. Using the origin/destination location pairs the great circle travel distance between the visitor and the destination were calculated.

In order to better understand the spatio-temporal relationship between distances traveled and seasonality, Figure 2 presents a graph of the distances visitors are willing to travel based on the start date of the reservation. Quantiles show the distribution of travel distances for all records on a particular day. The overall trend of the graph shows that for all start dates, 50 % of the reservations are made for parties traveling less than 200 km for the experience, with virtually no weekend, seasonal or holiday effects. Examining the 75 % quantile shows start dates between April 1 and November 15 have little variability in the distances traveled (between 200 km and 400 km), with a small weekend effects becoming visible. For start dates between mid November and the end of March, 25 % of the reservations (still the 75 % quantile) were for parties that traveled more than 1000 km, with a less regular weekend effect than the rest of the year. This indicates parties are more willing to travel farther distances for recreational tourism in the winter than in other seasons, perhaps for winter specific activities such as back county skiing. When examining the 90 % quantile, the weekend travel distance signal and winter seasonal effect

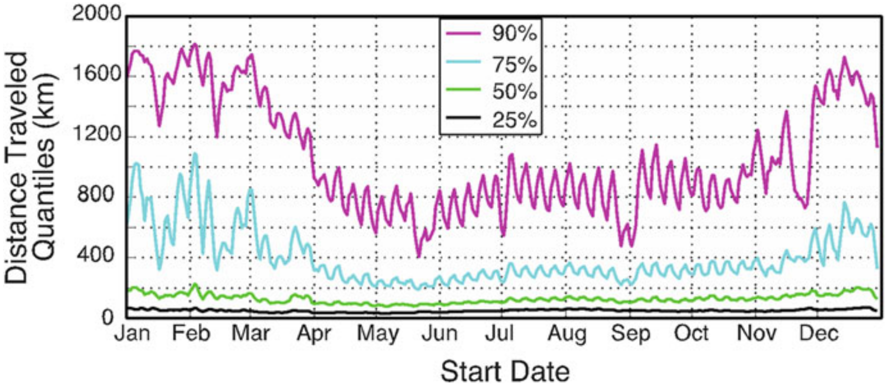


Fig. 2 Great circle distance traveled between visitors and U.S. Federally managed PPL's with start dates between January 1, 2008 and December 31, 2015 (n = 12,072,092) binned by day of the year

become more pronounced. Further at the 90 % quantile, distances traveled drop for Memorial Day, Independence Day, Labor Day and Thanksgiving holidays start dates.

Further enrichment from attributes within the data set produced the attributes of lead-time, duration of stay, and person-nights (a measure of cumulative nights of human occupancy). Enrichment from visitor origin geography and PPL name are not presented in this chapter, but will be considered in future analysis. The lead-time attribute is explored to uncover national trends in temporal reservation making. Figure 3 shows lead-time quantiles for the reservations based on start date. The overall trend of the graph shows that lead-time generally increases in the summer months and for most start dates, 50 % of the reservations are made less than a week prior to the start date. The spikes in the lead-times across all quantiles for the U.S. Federal holidays of Memorial Day, Independence Day, Labor Day and Thanksgiving (fourth Thursday of November) indicate visitors make travel plans for these holidays far more in advance than for other times of the year. These U.S. federal holidays, in addition to Christmas and New Years also are the only start dates for which the 25 % quantile of lead-time is greater than zero. The spacing between the quantile curves within the figure can be used to gauge temporal changes in reservation booking rates. Uniform spacing indicates temporally steady booking rates, while gaps or closures signify non-uniform booking rates. The tighter the curves, the less lead-time variability observed. Notice the tightening of the curve between the 90 % and 75 % quantiles in March. Of reservations made for March start dates, 90 % are booked within 3 months of the start date. This indicates that March visitations are planed less in advance than other months.

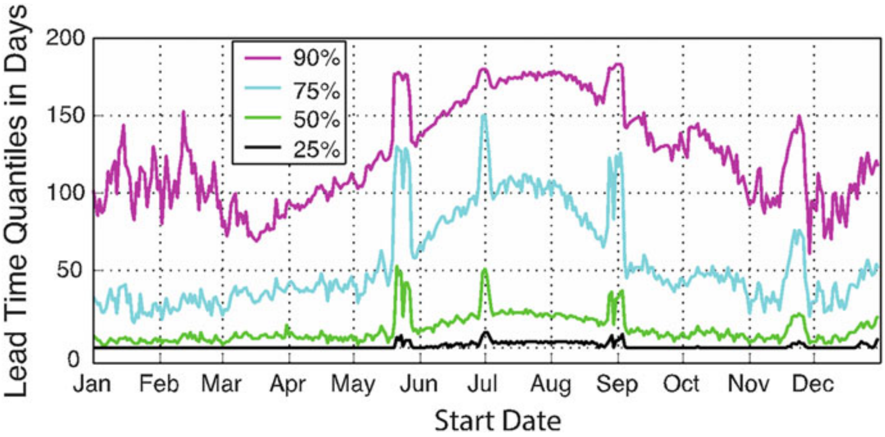


Fig. 3 Lead-time between order date and start date for U.S. Federal PPL reservations with a start dates between January 1, 2008 and December 31, 2015 (n = 12,072,092) binned by day of the year

4 Geovisualizations for U.S. Federally Managed PPLs

The geovisualizations presented in this section examine national trends in the 9-year U.S. Federal PPL reservation data set, which include all records whose origin-destination pairs could be geocoded (n = 12,003,292). While these data could be subset for a particular PPL or group of PPLs, season, or visitor origin region, as described previously in this chapter, the complete data set is considered for this demonstration of the value of geovisualization in PPL decisions support. Figures 4, 5, 6, and 7 present enriched reservation attributes that have been summarized (controlled) by visitor origin zip code. These geovisualizations explore the demand population including an index of national utilization, the median travel lead-time in days, the median travel distance, and the distance difference between the 25th and 75th distance quantiles, which helps describe the distance distribution curve for each origin zip code. Figures 8 and 9 present enriched reservation attributes that have been summarized (controlled) by PPL name. These geovisualizations explore the distribution of total reservation count across the U.S. and a comparison of utilization attributes for distinct facilities located within the Ashley National Forest.

All geovisualizations presented in this section were generated using ESRI's ArcGIS for Desktop10.3.1 (ArcGIS for Desktop, 2016); however, similar figures could be produced with open source GIS applications (e.g., QGIS). The quantile style classification methodology employed (Figs. 4, 5, 6, and 7, 9) aims to put an equal number of unique ZIP codes into each of five categories based on the summarized attribute value. Figure 8 employs a natural breaks classification method because the quantile method was not appropriate due to three reservation count outliers identified in a univariate graph for total reservation count.

In order to better understand national level utilization of federal recreational opportunities, a national index of federal land utilization was generated using the total sum of all person-nights for the 9-year data set (151,655,333). The index simply normalizes the total sum of all reservations person-nights per zip code by ESRI's 2014 total zip code population estimate (Fig. 4). Another way to think about this index is that it captures the cumulative nights of human occupancy and subsequent cumulative impact to the natural setting. A zip code with a Federal Land Utilization index of one would result if each person residing within the zip code made exactly one reservation for one person for one night during the 9-year period. However, it is likely that repeat visitations account for many of the reservations. Without access to customer identification data, the percentage of repeat visitation cannot be determined.

Figure 4 reveals that the south-central Midwest, much of the West coast and pockets along the Appalachian Trail exhibit the highest utilization indices within the contiguous U.S. This means that per capita, residents from these regions make reservations on federal lands more often (red in Fig. 4) than those within other regions. In general, Indiana, Ohio, Eastern New York, costal Carolinas, Louisiana & Western Mississippi underutilize federal lands relative to the national population. This decreased demand may be a function of the lack (or limited availability) of federal PPLs in this region (Figs. 6 and 8); however, alternative reasons such as lack of interest in visiting federal lands, propensity to stay on private or state land for overnight stays, or general disinterest in outdoor recreation require further examination.

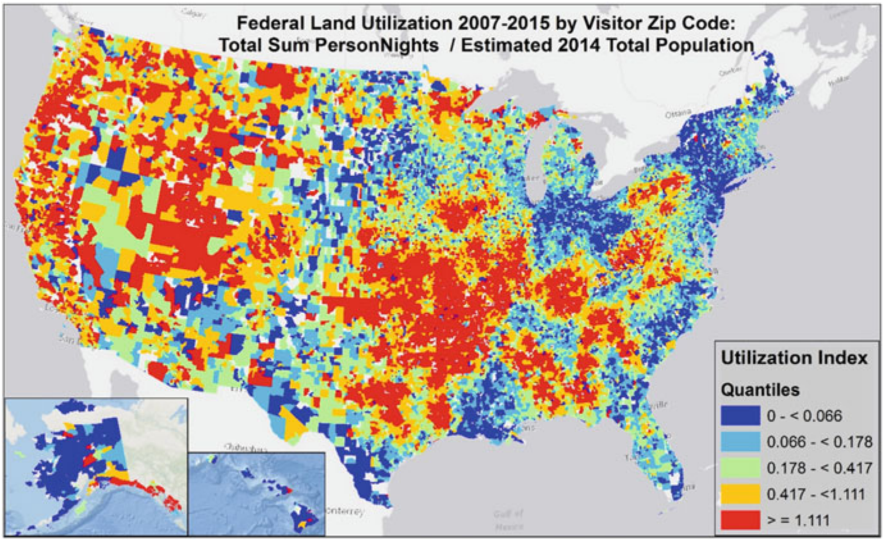

Federal Land Utilization 2007-2015 by Visitor Zip Code:
Total Sum PersonNights / Estimated 2014 Total Population
Utilization Index
Quantiles
0 - < 0.066
0.066 - < 0.178
0.178 - < 0.417
0.417 - <1.111
> = 1.111

Fig. 4 Federal land utilization index by visitor zip code derived from the total sum of person-nights at U.S. federally managed destinations over 9-years (2007–2015) normalized by the estimated 2014 total zip code populations

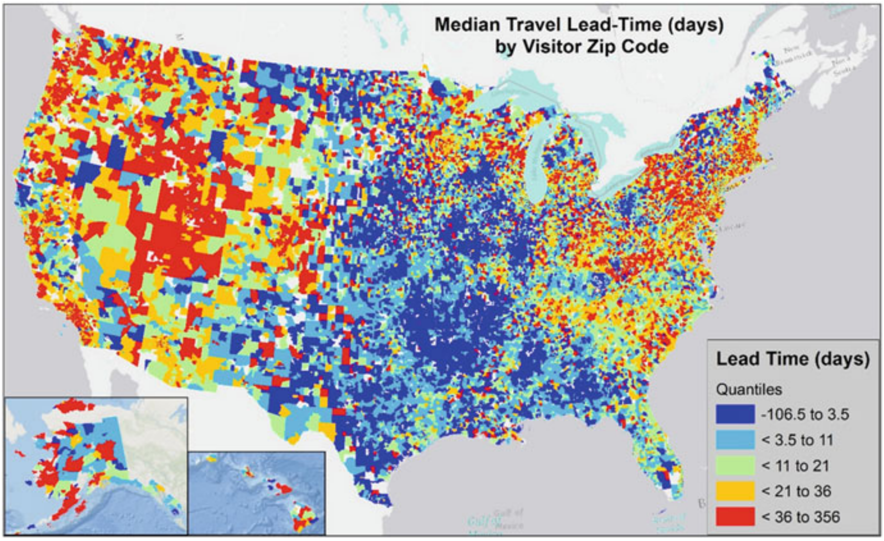


Fig. 5 Median travel lead-time in days when booking U.S. federally managed destinations by visitor zip code, 2007–2015

Next, the median lead-time is investigated (Fig. 5) as the summary statistic rather than the average lead-time because the lead-time distribution for the NRRS data set was not normally distributed. Locations where higher utilization is observed in Fig. 4, largely correspond to locations with short median lead-times. In particular, Arkansas and portions of the states immediately surrounding Arkansas have short median lead-times. Parts of New England and out West, specifically Utah have longer median lead-times. Geovisualization of the 25th and 75th quantiles for travel lead-time (not shown here) also display similar patterns.

Figure 6 presents the median great circle travel distance between visitors and destinations by visitor origin. The median travel distance was selected as the summary statistic rather than the average travel distance because the distance distribution for the NRRS data set was not normally distributed. Visitor travel for recreational tourism on federal PPLs east of the front range of the Rocky Mountains is typified by low median travel distances for zip codes (medium to light blue in Fig. 6) immediately surrounding the facilities. In many cases, a facility or group of facilities become the bull's-eye or locus at the center of concentric circles whose median travel distance increases with increased distance from the facility (dark blue to medium blue to green). This suggests that residents proximal to facilities are in fact reserving local facilities for recreational tourism. Moreover, this pattern of localized utilization is very similar to the utilization pattern shown in Fig. 4. Visitor travel for recreational tourism on federal PPLs west of the front range of the Rocky Mountains is typified by a less concentric bull's-eye pattern, but there is still some local utilization. Regions with few or no federally managed PPLs (e.g., Indiana), not surprisingly have higher median travel distances. However, many PPLs are surrounded by zip codes whose residents travel the farthest, such as in New England, Florida, Southern New Mexico/West Texas, Michigan and Ohio.

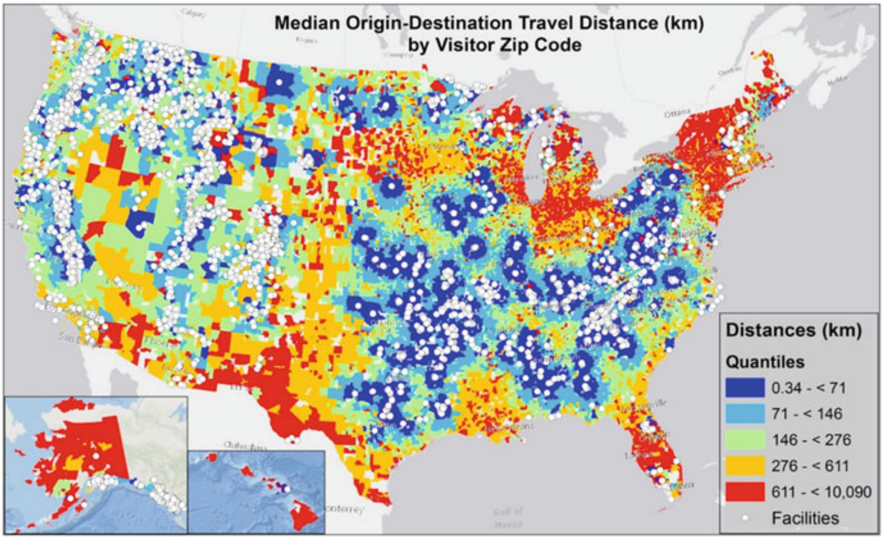


Fig. 6 Median origin–destination travel distance to U.S. federally managed destinations by visitor zip code, 2007–2015

In order to further examine the origin-destination travel distance distribution curve for each origin zip code, for which the median values were presented in Fig. 6, the difference between the 25th travel distance quantile and the 75th travel distance quantile is geovisualized in Fig. 7. If the difference is small for a specific zip code, the origin-destination travel distance distribution curve will be a very tight spike, indicating many visitors (the middle 50 % of visitors) travel relatively similar distances from their origin to their reserved destination. If the difference is large, the origin-destination travel distance distribution curve will be broad and flat, indicating larger variability in the distance visitors are willing to travel to their reserved destinations. The distance differences values measure the range over which visitors from the same origin zip code are wiling to travel, which indicates similarity or dissimilarity in destination preference. For example, if a region is dark blue in both Figs. 6 and 7, then residents of that region make reservations that are highly localized with little variability in the distance visitors are willing to travel to their reserved destination. If a region is light blue in Fig. 6 and red in Fig. 7 (e.g., parts of New England), visitors are utilizing local and regional PPLs, but there is large variability in the distance visitor are willing to travel to their reserved destination and larger distances traveled in general.

Geospatial demand can also be examined from the supply side by looking at enriched reservation attributes summarized by U.S. PPL destination. Figure 8 explores the distribution of total reservation count at federally managed PPL destinations across the U.S. In general, west of the front range of the Rocky Mountains is typified by an abundance of PPLs, many of which have relatively low total reservation counts. Relatively higher total reservation counts can be found

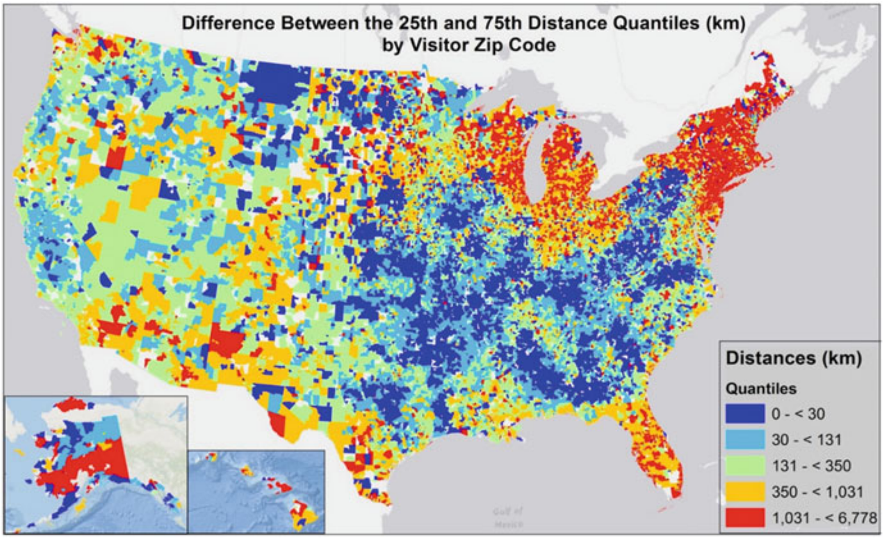


Fig. 7 Difference between the 25th and 75th travel distance quantiles for U.S. federally managed destinations by visitor zip code, 2007–2015

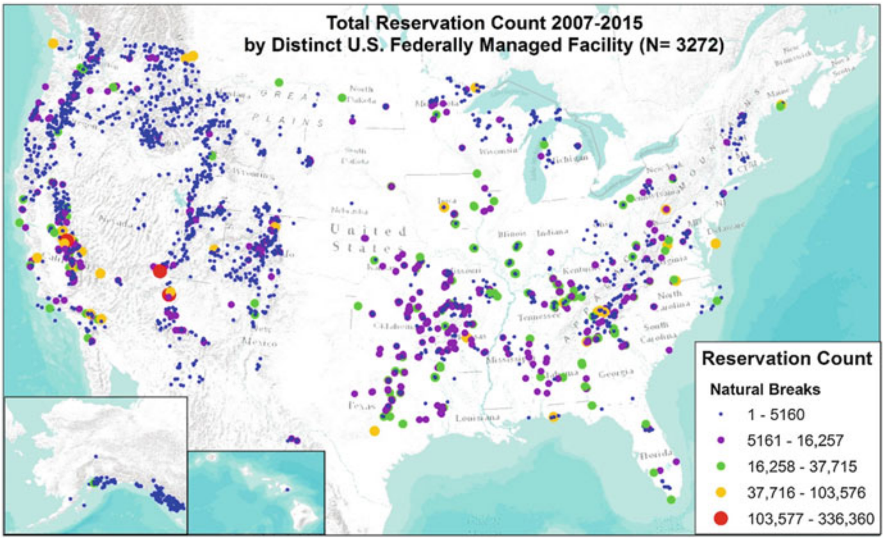


Fig. 8 Total reservation counts for U.S. federally managed facilities, 2007–2015

for PPLs in the eastern U.S., particularly for those PPLs that were at the center of the distance traveled bulls-eyes (Fig. 6) Three unique facilities stand out as having the greatest number of reservations: Mather Campground, Grand Canyon National

Park, 336,360 reservations; Upper Pines, Yosemite National Park, 201,174 reservations; and Watchman Campground, Zion National Park Service, 182,201 reservations.

While the previous national maps provide an opportunity to explore regional distribution patterns, PPL managers also can learn about the utilization of specific facilities within the context of their neighborhood cluster. Figure 9 provides a comparison of demand for all facilities clustered within the Ashley National Forest region near the Green River in northeastern Utah. This figure provides a comparison of the median distance traveled, median lead-time of the reservation, person-nights, and total reservation count for 16 distinct facilities within the reservation system. Median travel distances are quite similar across the 16 facilities and range from 191.8 km to 225.7 km. While one facility does make it into the top national median travel distance quantile, all other facilities in the area fall into the second highest quantile. This likely means that visitors to this destination region are drawn from the same origin markets. The median lead-time for reservations is much more diverse than the travel distance among these facilities, with a high of 360 days lead-time and a low of 19 days. This diversity indicates the varied demand for these facilities (i.e. high demand with high lead-time), which could reflect either or both the attractiveness of the facilities or restricted opportunities (limited availability of reservations).

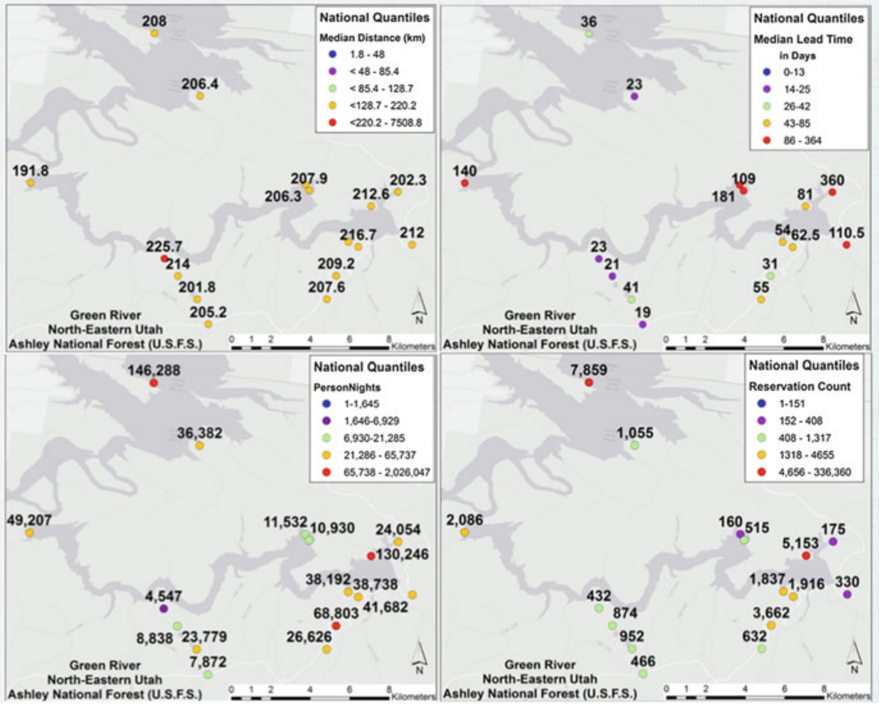


Fig. 9 Reservation attributes for 16 facilities within the Ashley National Forest including distance traveled, lead-time, person-nights, and total reservation count, 2007–2015

Comparing person-nights, a measure of cumulative human occupancy and impact, among the facilities can inform managers about relative demand within the region or for specific facilities. Person-nights for these facilities range from a low of 4547 to a high of 146,288. The facility with the lowest value may have the lowest demand or the site itself may require limited use. This facility had the highest median travel distance for all facilities in the cluster; so, visitors are willing to travel a longer distance for that facility. This observation along with the low lead-time for this specific facility may be related to settings within the reservation system. Additional information about the capacity of this facility and its reservability are necessary to better interpret this geovisualization. Finally, the total reservation count for the region and individual facilities within the region can also inform managers of total visitation demand. The range of the total reservation count for these facilities is from 160 to 7859.

Examining indicators of demand from the PPL perspective (e.g., Fig. 9) can help support management objectives such as increasing knowledge about managed PPLs and supporting sustainable development. Within the Ashley National Forest, the metrics presented indicate that individual facilities have different use load/capacity, as four of the five facilities with highest reservation lead-times (most popular) have the lowest number of reservations (limited or restricted number of opportunities), which creates a high demand. This distribution of demand among facilities could also provide managers knowledge with which they can inform potential visitors of alternative options available or change availability in the reservation system to more equitably distribute use (spatial and temporal) among the facilities within the region.

5 The Past, Present & Future of U.S. Federally Managed PPL Reservation Data

In 1998 the Recreation Information Database (RIDB) was created to provide a web-based resources for citizens interested in accessing supply-side data including recreation area and facility data (Recreation Information Database—RIDB, 2016). In 2006, the newly created Recreation One Stop program integrated RIDB data into an inter-agency recreation reservation website (www.recreation.gov) in an effort to provide visitors a more comprehensive resource. At that time, the RIDB data was still available to third parties, but it lacked site-specific details that were included in the website Recreation.gov. In 2014, the RIDB implemented an XML feed from Recreation.gov to pull in all site-level data as a way to provide a more complete supply-side recreation data set.

With passage of the Digital Accountability and Transparency Act of 2014 (DATA Act), there are now legal requirements to provide open data and many federal agencies are working to overcome the challenges associated with sharing data (*Panel Discussion on Changing the Culture for Open Data*, 2015). In 2015, RIDB developed and deployed an application programming interface (API) to provide data in additional formats (e.g., JSON & CSV) as well as ways to filter or subset recreation supply side data by specific state, activity or organization (Recreation Information Database—RIDB, 2016). On February 10th, 2016 at the requests of NCSU researchers, access to historical reservation data was added to RIDB website as downloadable annual CSV files (2007–2015). Recreation One Stop program staff aim to have the historical data available through the API shortly (personal communication DeLappe, 2016). Providing access to this historical data in open, machine-readable formats, in compliance with President Obama's executive order, is a great example of how government programs can overcome the legal, technological, and cultural barriers that sometimes limit adoption of open data standards.

The geospatial analytics and geovisualization techniques presented in this chapter are one approach for transforming flat CSV files containing historical reservation data into information that supports management decisions. A more effective way to achieve this goal and to promote the use of this open data may be to create a direct link between the reservation system and a historical reservation data warehouse, so that as reservations occur they could automatically be preprocessed, cleaned, and enriched. Geovisualization then can be accomplished through the development of web-based mapping application that utilize simple and customizable APIs, such as Google Maps™ API. One large benefit of utilizing APIs and supporting data infrastructure (e.g., data warehouses and data cubes) for web-mapping application development is that users are provided with accurate real-time data, from which they can extract timey actionable intelligence. These web-maps could be designed and developed for specific management objectives or for perspective-visitor travel planning decision support.

From the management perspective, the U.S. Department of Agriculture's Animal, Plant, Health Inspection Service managers have begun using reservation data to geovisualize the origins of visitors who camp at specific facilities, so that they can track infected firewood dispersion (personal communication DeLappe, 2016). This is just one example of how the geospatial component of reservation data can be used to help managers provide better overall experiences to visitors. Further, historical reservation data can provide decision support knowledge relevant to complex PPL ecosystems that need to be carefully managed with regard to decisions about invasive species, climate change, and development around and within these sites as well as other management issues. For these reasons, agencies tasked with controlling invasive species or the spread of plant and tree diseases, would greatly benefit from the real-time information offered by web-maps.

While this chapter has focused primarily on geospatially examining historical recreational tourism reservation data for improved decision support by managers, it also should be mentioned that allowing the public easy access to historical usage geovisualizations could help support perspective-visitors as they plan travel. In an effort to bring our nation's PPLs closer to the people, to which they collectively belong, the Recreation One Stop program held a developers summit on April 11–12, 2015. The invitation outlined the benefits of bringing "wild areas" within closer reach of the people including improved public health and stress reduction resulting from exercise and time spend outdoors, increased awareness of man's impact on nature and general ecological principles, and increased connections between more citizens and our country's diverse history (myAmerica Developer Summit, 2015).

Among the summit sessions, several could have utilized the geospatial aspects of the historical reservation data. For example, one session described using Census Data to discover communities of interest related to the Every Kid in a Park. These discovered communities could be refined further by eliminating communities that are well represented in the historical reservation record. Another session yielded a prototype that allowed prospective visitors to access spatiotemporal historical reservation data for the Grand Canyon to assist in trip planning, and specifically for identifying the most historically crowded times. This prototype could be extended to allow potential visitors to compare PPLs within desired destination regions or over different date ranges. Prospective visitors could then consider visitation rates as they decide when and where to visit, which could ultimately lead to improved visitor experiences and hopefully a new generation of PPL enthusiasts.

6 Conclusions

Data analytics is booming across domains, industries and data platforms. In the travel and tourism domain, analysis is largely focused on measuring and capturing the tourism experience. To this end, the geospatial component of visitor experience is captured with increasing detail about demand populations (e.g., geo-tagged tweets, Facebook posts, etc.). Increased use of Internet reservation sites has produced large amounts of geospatial data that potentially hold a wealth of information about historic usage, which can be used to the benefit of tourism destinations. The private sector is using these transactional records to generate performance metrics, many of which are highly geospatial such as origin shares within a destination, relative return on marketing effort or geodemographic customer profiling. PPLs can also gain a better understanding of visitors through geospatial examination of metrics such as booking rates, stay durations, lead-times and the number in the party. This information is useful for PPL managers who are tasked with balancing the needs as well as the impact of visitors.

Knowledge discovery from geographic data that is collected without end, such as from a PPL reservation system, is nontrivial and requires distinctive consideration and techniques (Miller & Han, 2009). The inductive approach of mining enriched PPL reservation data sets described in this chapter can serve as a guide for creating historical usage knowledge. To maximize the inductive data mining approach and increase the opportunity of finding relationships, PPL reservation data sets should be enriched from attributes within the data set as well as from relevant external data. Geovisualizations of these mined and enriched data can be an effective means of communication and can deliver insights that might otherwise be lost, as they can present new and unexpected patterns, trends and relationships. These insights can lead to improved management decision support.

Globally, any agency tasked with tourism and resource management at any level of government (i.e. Federal, State, Province, Prefecture, County, etc.) can use this approach to examine patterns and trends within the empirical evidence created though reservation booking. Researchers supporting effective recreational tourism management also can use this approach to modify or formulate new hypotheses, which can in turn help modify existing theoretical frameworks related to tourism and recreation. Specifically, knowledge gained from PPL reservation data can help generate new theories related to distance decay, origin markets, or set choice. These theoretical constructs support more robust data analytics practices such as market segmentation and predictive modeling.

Comprehensively exploring the spatiotemporal aspects of PPL reservation data through geovisualizations can help managers learn about visitor usage. For example, exploring the geospatial demand distribution for groups of facilities within a destination region can help mangers better distribute use among facilities. Further, identifying communities that historically underutilized recreational opportunities at PPLs is the first step necessary in engaging those communities to increase participation. If the data is enriched by visitor origin geography, socioeconomic and geographic challenges that prevent citizens from accessing public lands can be identified. These insights into the spatiotemporal patterns of visitor demand from both the facility and visitor origins perspectives can help managers improve and promote visitor experiences for different segments of the population. These insights could help drive new planning and break down the inequities facing socially and economically underserved communities.

Unlike other countries, the number of visits to PPLs in the U.S. and Japan is decreasing (Balmford et al., 2009). Through geospatial segmentation and community profiling, this trend can be examined across different segments of the population. To support this and various other forms of geospatial analytics within the research community, we aim to publish a data set that includes the complete set of enrichments described in Sect. 2.3 for the U.S. Federal PPL data presented in this chapter. Hopefully providing access to a free cleaned data product will allow other research teams or PPL managers to explore different aspects of this historical record through the creation of geovisualizations, web-maps, models and forecasts. Ideally, this data will be used in support of making open spaces, parks, and public lands easier to reach and explore, so that more citizens can be part of the benefits that these lands afford society (myAmerica Developer Summit, 2015). Finally, web-maps that support visitation planning can help engage citizens and create our next generation of PPL visitors, supporters and advocates.

References

ArcGIS for Desktop. (2016, February). Retrieved March 3, 2016, from http://desktop.arcgis.com/en/

ArcGIS Zip Code Layers. (2016, March 31). Retrieved March 31, 2016, from http://www.arcgis.com/home/search.html?q=zip%20code&restrict=true&focus=layers

Bahaire, T., & Elliott-White, M. (1999). The application of geographical information systems (GIS) in sustainable tourism planning: A review. *Journal of Sustainable Tourism, 7*(2), 159–174.

Balmford, A., Beresford, J., Green, J., Naidoo, R., Walpole, M., & Manica, A. (2009). A global perspective on trends in nature-based tourism. *PLoS Biol, 7*(6), e1000144.

Beeco, J. A., Hallo, J. C., English, W. R., & Giumetti, G. W. (2013). The importance of spatial nested data in understanding the relationship between visitor use and landscape impacts. *Applied Geography, 45*, 147–157. doi:10.1016/j.apgeog.2013.09.001.

Beeco, J. A., Hallo, J. C., & Brownlee, M. T. J. (2014). GPS visitor tracking and recreation suitability mapping: Tools for understanding and managing visitor use. *Landscape and Urban Planning, 127*, 136–145. doi:10.1016/j.landurbplan.2014.04.002.

Bell, R. R., & Zabriskie, N. B. (1978). Assisting marketing decisions by computer mapping: A branch banking application. *Journal of Marketing Research, 15*, 122–128.

Bishop, I. D., & Gimblett, H. R. (2000). Management of recreational areas: GIS, autonomous agents, and virtual reality. *Environment and Planning B: Planning and Design, 27*(3), 423–435.

Center for Responsible Travel. (2016). *The case for responsible travel: Trends & statistics 2015*. Washington, DC. https://ecotourism.app.box.com/s/rxiyp65744sqilmrybfk8mys3qvjbe9g

Chavez, D. J. (2001). *Managing outdoor recreation in California: Visitor contact studies 1989-1998* (No. Gen. Tech. Rep. PSW-GTR-180.) (p. 100). Albany, NY: Pacific Southwest Research Station, Forest Service, U. S. Department of Agriculture.

Chen, R. J. (2007). Geographic information systems (GIS) applications in retail tourism and teaching curriculum. *Journal of Retailing and Consumer Services, 14*(4), 289–295.

Chhetri, P. (2015). A GIS methodology for modelling hiking experiences in the Grampians National Park, Australia. *Tourism Geographies, 17*(5), 795–814. doi:10.1080/14616688.2015.1083609.

Chrisman, N. (2001). Reference systems for measurement. In *Exploring geographical information systems* (2nd ed., pp. 15–35). Wiley. http://www.citeulike.org/group/2170/article/1104881

Crnojevac, I., Gugić, J., & Karlovčan, S. (2010). eTourism: A comparison of online and offline bookings and the importance of hotel attributes. *Journal of Information and Organizational Sciences, 34*(1), 41–54.

DeLappe, R. (2016, February 24). WebEx about data provided by the Recreation One-Stop Program [WebEx].

Elliott-White, M. P., & Finn, M. (1997). Growing in sophistication: The application of geographical information systems in post-modern tourism marketing. *Journal of Travel & Tourism Marketing, 7*(1), 65–84.

Every Kid in a Park. (2015, April 24). Retrieved March 16, 2016, from http://www.nationalparks.org/ook/every-kid-in-a-park

Gahegan, M. (2009). Visual exploration and explanation in Geography analysis with light. In *Geographic data mining and knowledge discovery* (2nd ed., pp. 291–324). Boca Raton, FL: CRC Press.

Gomez, L., Haesevoets, S., Kuijpers, B., & Vaisman, A. A. (2009). Spatial aggregation: Data model and implementation. *Information Systems, 34*(6), 551–576. doi:10.1016/j.is.2009.03.002.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., et al. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery, 1*(1), 29–53. doi:10.1023/A:1009726021843.

Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass, 2*(6), 2011–2039.

Hanink, D. M., & Stutts, M. (2002). Spatial demand for national battlefield parks. *Annals of Tourism Research, 29*(3), 707–719. doi:10.1016/S0160-7383(01)00085-8.

Hanink, D. M., & White, K. (1999). Distance effects in the demand for wildland recreational services: the case of national parks in the United States. *Environment and Planning A, 31*(3), 477–492. doi:10.1068/a310477.

Hirschey, J. (2014). Symbiotic relationships: Pragmatic acceptance of data scraping. *Berkeley Technology Law Journal*, *29*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2419167

Lansky, D. (2016, March 9). *It's time for the Moneyball of destination marketing*. Retrieved March 15, 2016, from https://www.tnooz.com/article/destination-marketing-analytics/

Loomis, J. (2004). How bison and elk populations impact park visitation: A comparison of results from a survey and a historic visitation regression model. *Society & Natural Resources, 17*(10), 941–949. doi:10.1080/08941920490505338.

Manning, R. E. (2014). Research to guide management of outdoor recreation and tourism in parks and protected areas. *Koedoe, 56*(2), 1–7.

McAdam, D. (1999). The value and scope of geographical information systems in tourism management. *Journal of Sustainable Tourism, 7*(1), 77–92.

Mckercher, B., & Lew, A. A. (2003). Distance decay and the impact of effective tourism exclusion zones on international travel flows. *Journal of Travel Research, 42*(2), 159–165.

Miller, F. L. (2008). Using a GIS in market analysis for a tourism-dependent retailer in the Pocono Mountains. *Journal of Travel & Tourism Marketing, 25*(3–4), 325–340. doi:10.1080/10548400802508416.

Miller, H. J., & Han, J. (2009). Geographic data mining and knowledge discovery: An overview. In *Geographic data mining and knowledge discovery* (pp. 1–26). Boca Raton, FL: CRC Press.

myAmerica Developer Summit. (2015, April). Retrieved April 6, 2016, from http://openglobe.github.io/myamerica-devsummit/

Neuvonen, M., Pouta, E., Puustinen, J., & Sievanen, T. (2010). Visits to national parks: Effects of park characteristics and spatial demand. *Journal for Nature Conservation, 18*(3), 224–229. doi:10.1016/j.jnc.2009.10.003.

O'Connor, A., Zerger, A., & Itami, B. (2005). Geo-temporal tracking and analysis of tourist movement. *Mathematics and Computers in Simulation, 69*(1-2), 135–150. doi:10.1016/j.matcom.2005.02.036.

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval, 4*(3), 175–246.

Panel Discussion on Changing the Culture for Open Data, § Government Data Sharing Community of Practice. (2015). The Loft, 600 F Street NW, Washington, DC 2004. http://www.gao.gov/assets/680/674021.pdf

Python. (2016). Retrieved April 4, 2016, from https://www.python.org/

Recreation Information Database—RIDB. (2016). Retrieved April 5, 2016, from https://ridb.recreation.gov/

Supak, S., Brothers, G., Bohnenstiehl, D., & Devine, H. (2015). Geospatial analytics for federally managed tourism destinations and their demand markets. *Journal of Destination Marketing & Management, 4*(3), 173–186.

Supak, S. K., Devine, H. A., Brothers, G. L., Rozier Rich, S., & Shen, W. (2014). An open source web-mapping system for tourism planning and marketing. *Journal of Travel & Tourism Marketing, 31*(7), 835–853.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771–1779.

United Nations Environment Programme. (2011). *Towards a green economy: Pathways to sustainable development and poverty eradication* (p. 419). http://www.unep.org/greeneconomy/Portals/88/documents/ger/ger_final_dec_2011/Green%20EconomyReport_Final_Dec2011.pdf

Van Berkel, D. B., Munroe, D. K., & Gallemore, C. (2014). Spatial analysis of land suitability, hot-tub cabins and forest tourism in Appalachian Ohio. *Applied Geography, 54*, 139–148.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (Vol. 8). Boston: Springer. doi:10.1007/978-1-4615-4651-1.

Wolf, I. D., Wohlfart, T., Brown, G., & Lasa, A. B. (2015). The use of public participation GIS (PPGIS) for park visitor management: A case study of mountain biking. *Tourism Management, 51*, 112–130. doi:10.1016/j.tourman.2015.05.003.

Zhao, J., Forer, P., Sun, Q., & Simmons, D. (2013). Multiple-view strategies for enhanced understanding of dynamic tourist activity through geovisualization at regional and national scales. *Cartography and Geographic Information Science, 40*(4), 349–360. doi:10.1080/15230406.2013.783449.

GIS Monitoring of Traveler Flows Based on Big Data

Dong Li and Yang Yang

1 Introduction

Along with a burst of various types of data sets that store the "digital footprints" of consumers, the concept of "big data" has been introduced to better understand tourist behavior and monitor the tourism demand with timely and precise data sets from a wide variety of sources. Big data analytics have created revolutionary breakthroughs in commerce, science and society by leveraging hybrid methodologies to efficiently extract valuable information from big data sources and transferring them into business insights and solutions. Tourists nowadays adopt many Information & Communication Technology (ICT) tools before, during, and after their trips to improve the overall experiences. As such, many sources of big data that generate tourists' digital footprints becomes of great importance for scholars and practitioners to answer questions that can hardly be handled by conventional ways such as tourist surveys and second-hand statistical data sources.

A major intriguing advantage of big data is that some of them offer precise geo-spatial information of tourist movement. The monitoring and understanding of the spatial-temporal pattern of tourist movement provides crucial insights for destination planning and capacity management (Shoval, Isaacson, & Chhetri, 2013). With the development of reliable and accessible smartphones with built-in GPS antennas, tourists are able to share their user generated contents (UGC) with finer geo-referenced data, such as geo-referenced Twitter sharing (Hawelka et al., 2014) and geo-tagged photos (Vu, Li, Law, & Ye, 2015). Compared to the

D. Li
Beijing Tsinghua Tongheng Urban Planning & Design Institute (THUPDI), Beijing, People's Republic of China

Y. Yang (✉)
Temple University, Philadelphia, PA, USA
e-mail: yangy@temple.edu

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_7

conventional spatial data set, such as GPS data and retrospective travel logs, the geo-tagged/referenced big data contain affluent information over a large representative population covering even a very large area within a coincident period and offer additional temporal information that allows further investigation on temporal spatial patterns.

Several studies have demonstrated how to use different big data sources to predict tourist demand (Yang, Pan, & Song, 2014), explore tourists' experience (Lu & Stepchenkova, 2015), monitor travel routes of tourists (Ahas, Aasa, Roose, Mark, & Silm, 2008), and understand tourist expenditure patterns (Kozak & Sezgin, 2013). In particular, several studies adopted the geo-tagged/referenced big data to investigate tourist spatial behavior at different scales (Hawelka et al., 2014; Shoval & Isaacson, 2007; Vu et al., 2015). With the assistance of proper analytical and visualization tools, big data provides a promising alternative approach to explore and explain tourist flow patterns from both macro and micro perspectives. However, to our knowledge, most of spatial analysis studies on big data are exploratory in nature, without an in-depth analysis on factors shaping the spatial pattern. Moreover, they did not compare the validity of big data with other traditional data sources.

To abridge the research gap, this chapter aims to use the geo-tagged Sina Weibo data to understand the nation-wide Chinese domestic tourist movement patterns during the National Day Golden Week in 2014. The Weibo data are compared with the provincial tourism statistics collected by the government for a validity check. Furthermore, we demonstrate the use of big data as an alternative data source to monitor tourist flows at the national scale, and employ geo-spatial tools, like geographic information system (GIS) and spatial interaction model (SIM), to rigorously explain the pattern of tourist movement. This chapter contributes to the current knowledge of tourist flow analysis by presenting an integrated method to analyze geo-tagged/referenced big data at a large geographic scale.

2 Literature Review

With the development of reliable and accessible smartphones with built-in GPS antennas, tourists can submit the geo-tagged/referenced record through social media, such as Facebook check-ins, geo-referenced Twitter sharing and geo-tagged Instagram photos. Girardin, Fiore, Ratti, and Blat (2008) applied various analytical tools to disclose tourist hotspots and travel routes based on geo-referenced photos from Flickr during their travel. Kádár (2014) highlighted tourist hotspots in Vienna, Prague and Budapest using geo-tagged Flickr photos and found a high level of correlation between this data of tourism statistics. Vu et al. (2015) introduced a Markov chain model for travel pattern mining on the geo-tagged Flickr photos in Hong Kong, and the travel behavior difference between Asian and Western tourists were highlighted. Hawelka et al. (2014) demonstrated the usefulness of geo-located Twitter data to proxy international tourist, and they found the increased mobility of travelers over years and in West European and other developed countries.

Apart from the geo-tagged/referenced social media data, cell-phone roaming data and Bluetooth tracking data can also help understand the pattern of tourist flows at different scales. A group of researchers from Estonia utilized a nationwide roaming mobile dataset of the Estonian GSM network to study the spatial-temporal pattern of inbound tourists to Estonia (Ahas et al., 2008; Nilbe, Ahas, & Silm, 2014; Tiru, Kuusik, Lamp, & Ahas, 2010). Moreover, Bluetooth tracking technology enables researchers to understand tourists' spatial-temporal movement patterns at a small scale (Versichele et al., 2014; Versichele, Neutens, Delafontaine, & Van de Weghe, 2012).

3 Tourist Flow Analysis

The dyadic matrix of tourist flows contains the volume of flows for each origin-destination pair. The spatial pattern of these flows can be examined by a set of indexes. Jansen-Verbeke and Spee (1995) analyzed the inter- and intra- regional tourist flows within Europe and used the Tourist Origin Index (TOI), the Tourist Intensity Index (TII), and the ratio index to study the origin effect, the destination effect, and the balance between the two, respectively. In another research by Li, Meng, and Uysal (2008), they explored the patterns of tourist flows among the Asia-Pacific countries and employed some indexes like the Country Potential Generation Index (CPGI) and the Gross Travel Propensity (GTP). Furthermore, the Relative Acceptance Index (RAI), which measures the difference between actual and expected flows divided by the expected flow, has been applied to the analysis of domestic tourist flows in 31 provinces of Sweden (Pearce, 1996) and the research on second home flows in Sweden (Müller, 2006).

In addition to the description of patterns, the gravity based spatial interaction model (SIM) can be applied to identify factors that facilitate or hinder tourist movement. Based on the estimation results of models, various origin- and destination-specific factors can be analyzed to assess their influence and significance on tourist flows. de Graaff, Boter, and Rouwendal (2009) employed the SIM to evaluated the attractiveness of museums using visitor flow data. Patuelli, Mussoni, and Candela (2013) used the SIM to highlight the importance of World Heritage Sites in attracting cultural tourism flows in Italy. Marrocu and Paci (2013) analyzed the inter-province tourist flows in Italy using the SIM, and underscored several determinants of tourist flows such as income, density, accessibility, and natural, cultural and recreational attractions.

4 GIS Analysis of Tourist Flows

GIS is defined as a computerized system used for the storage, retrieval, mapping, and analysis of geographic data. Farsari and Prastacos (2004) reviewed GIS applications in tourism studies and indicated the great potential of GIS usage. Based on their review, the major uses of GIS include the research of tourism resource inventories/usage, location suitability, tourism impact analysis, and visitor flow management. Along with more user-friendly interface, more available geo-spatial data, and higher computation power of hardware, a myriad of GIS applications have emerged in the field of tourism research (Farsari & Prastacos, 2004; Hall & Page, 2009). As a powerful spatial visualization tool, GIS has been employed to demonstrate and analyze the spatial pattern tourist flows/movements by extracting the spatial information from various sources of tourist survey and second-hand statistical data (Becken, Vuletich, & Campbell, 2007; Beeco et al., 2012; Lau & McKercher, 2006). Wu and Carson (2008) applied GIS techniques to understand the spatial-temporal pattern of multi-destination trips in South Australia by aggregating the tourist survey data. Rather than visualizing the spatial information embedded in the tripographic data directly, the GIS applications are also able to conduct more sophisticate analysis in conjunction with other models. Holyoak, Carson, and Schmallegger (2009) developed a GIS tool to visualize the multi-destination travel route by synchronizing route assignment algorithms and the nation-wide tourist survey data in Australia. Yang and Wong (2012b) applied exploratory spatial data analysis in GIS to highlight the tourism hot-spots among Chinese cities.

5 Methodology

This study investigates inter-province tourist flow pattern in China during the 2014 National Day Golden Week. The Golden Week vacation system was enacted by the Chinese central government in 1999 as a pro-tourism policy to stimulate domestic tourism and related demand (Wu, Xue, Morrison, & Leung, 2012). The Golden Weeks consist of several public holidays, and the National Day week is one them. Due to the lack of paid vacation system in most Chinese companies, the Golden Week vacation becomes a valuable opportunity for Chinese to relax and travel, especially to some long-haul domestic destinations. In 2014, 475 million domestic tourists travelled during the National Day Golden Week, and the overall tourism receipts added up to 245.3 billion RMB, accounting for 8.09 % of the total domestic tourism receipts in that year (CNTA, 2014).

We used the Weibo check-in data to understand the tourist flow pattern. The Weibo from Sina, akin to Twitter, is the most popular Chinese social media, and as of the third quarter of 2014, the number of monthly active users of Weibo came to 167 million. We used the public API interface to acquire check-in data, and geo-coded the data to get the destination information at the city level. Note that only domestic travels are included and those travel records from and/or to places overseas were discarded. The origin information was obtained based on the check-in information in the week ahead of National Day Golden Week. Typically, we regard the check-in record as a tourist's travel record if the origin city is different from the destination city. Because of the Chinese governments' regulation, there should be very few business trips during this legislated holiday, and we regarded these inter-city travels as tourism. After that, we aggregated the data to get a province-to-province tourist flow matrix. In Sect. 4, we will compare this data with the official tourism statistics to check the validity.

In the origin-destination pair-wise tourist flow data, its origin-destination matrix could be of a large number of dimensions depending on the number of origins and/or destinations in the data. In the matrix, each row represents an origin whereas each column represents a destination. Various dimension reduction methods can be used to transform the high-dimensional dimension into fewer dimensions. Common dimension reduction methods, such as multi-dimensional scaling (MDS) and factor analysis (FA) have been introduced to understand the structure of tourist flow matrix (Husbands, 1983; Zhang, Zhang, Li, Liang, & Liu, 2005). In particular, FA is able to extract the spatial structure of the flows by treating each extracted "factor" as a spatial field. Based on the FA results on geographic flows, one can categorize specific flows according to the factor loading and factor score of each origin-destination pair (Lowe & Moryadas, 1975), and therefore, identify the linkage between areas on the basis of spatial interaction structure. More specifically, in the Q-mode FA, the loadings identify destinations sharing similar origins, and factors scores highlight origins connected to these destinations. On the other hand, the R-mode FA helps unveil origins connecting to similar destinations, and factor scores disclose destinations related to these origins (Gober & Mings, 1984).

We adopt a gravity-type spatial interaction model to unveil factors that contribute to inter-province domestic tourism flows in China. A basic SIM equation can be written as

$$Y_{ij} = V_{ij} + \varepsilon_{ij} \tag{1}$$

where Y_{ij} denotes the geo-tagged check-in Weibo counts from origin i to destination j; V_{ij} is the systematic part consists of independent variables to explain the dependent variable; ε_{ij} is the random error term. As suggested by spatial interaction theory, three groups of factors determine the magnitude of spatial interaction, and we re-write V_{ij} as

$$V_{ij} = A_i B_j D_{ij} \tag{2}$$

where A_i represents characteristics related to origin i such that $A_i = \prod_{m=1}^{M} a_{im}^{\alpha_m}$ (a_{im} denotes the mth origin-specific independent variable with a scalar parameter of α_m);

B_j represents characteristics related to destination j such that $B_j = \prod_{k=1}^{K} b_{jk}^{\beta_k}$ (b_{jk} denotes the kth destination-specific independent variable with a scalar parameter of β_k); D_{ij} represents separation characteristics between origin i and destination j such that $D_{ij} = d_{ij}^{\varphi}$ (d_{ij} denotes the *geographic distance between i and j* with a scalar parameter of φ). To estimate the proposed SIM, we use the count data model by assuming that Weibo counts follow certain count data distribution (like Poisson or negative binomial). A Poisson model explains the expected number of counts as a function of independent variables. The Poisson probability function describing the probability of observing the count of y_{ij} is given by

$$f\left(y_{ij}\right) = \frac{V_{ij}^{y_{ij}} e^{-V_{ij}}}{y_{ij}!} \tag{3}$$

where V_{ij} is defined in Eq. (2). The Poisson model can be estimated by the maximum likelihood estimation (Cameron & Trivedi, 2005). A major assumption embedded in the Poisson model is that the conditional mean of dependent variable is equal to its variables. In the context of geographic flows, this assumption is likely to be violated by the unobserved heterogeneity (Scherngell & Barber, 2009). A more flexible count data specification, the negative binomial model, can be used to overcome the problem of unobserved heterogeneity by incorporating a stochastic heterogeneity component $e^{\upsilon_{ij}}$ into V_{ij} (Cameron & Trivedi, 2005) such that

$$V_{ij}^{*} = V_{ij} e^{\upsilon_{ij}} \tag{4}$$

where υ_{ij} follows a gamma distribution with mean 1 and variance δ. The probability function of the negative binomial model becomes

$$f\left(y_{ij}\right) = \frac{\Gamma\left(y_{ij} + \delta^{-1}\right)}{\Gamma\left(y_{ij} + 1\right)\Gamma\left(\delta^{-1}\right)} \left(\frac{\delta^{-1}}{V_{ij} + \delta^{-1}}\right)^{\delta^{-1}} \left(\frac{V_{ij}}{V_{ij} + \delta^{-1}}\right)^{y_{ij}} \tag{5}$$

where $\Gamma(\cdot)$ denotes the gamma function. The negative binomial model collapses to the Poisson model if $\delta = 0$.

6 Data Description

We obtained the geo-tagged Weibo from public APIs, and a total of 1,331,248 records were finally obtained. Note that since our geo-coding algorithm is unable to recognize any intra-city tourist movement, we cannot get the data of intra-province tourist flows in four province-level cities, Beijing, Shanghai, Tianjin and Chongqing. Based on this inter-province tourist flow Weibo data, we calculated the total geo-tagged Weibo counts in each province as destination, and correlated them with the official statistics on the number of domestic tourist arrivals published by each provincial tourism bureau during the National Day Golden Week in 2014. Figure 1 presents the bivariate scatterplot of these two variables. The Pearson correlation coefficient is calculated to be 0.814, indicating a highly positive statistical association between the two variables. A further look at Fig. 1 suggests that many data points above the fitted regression line are associated more developed provinces, suggesting that these provinces with more developed telecommunication infrastructure and concomitant easy access to Internet may over-represent the population of domestic tourists. Hence, we further regressed the officially published number of domestic tourist arrivals during the Golden Week on two variables, the number of geo-tagged Weibo in our data sets and the number of registered cellphone users per capita (capturing the level of telecommunication infrastructure). The estimated coefficient of the former is positive and significant whereas the one of the latter is negative and significant. This result confirms the Weibo data's over-representedness of the tourist population in provinces with easy Internet accesses.

Figure 2 illustrates the top 25 pairs of inter- province tourist flows based on the geo-tagged Weibo data. It demonstrates a large size of tourist movement from Beijing to Hebei, from Shanghai to Jiangsu, from Shanghai to Zhejiang, and from Sichuan to Chongqing. Four clusters are highlighted with strong interactions between provinces in the cluster. They are the North China cluster around Beijing, the Yangtze River Delta cluster around Shanghai, the Pearl River Delta cluster centered in Guangdong, and the Southwest cluster covering Sichuan and Chongqing. Most large tourist flow pairs are found within the cluster with only two

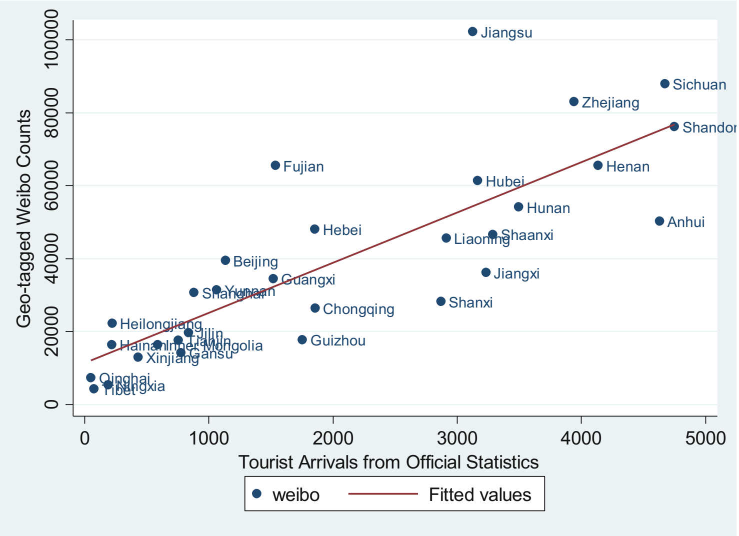


Fig. 1 Scatterplot of geo-tagged Weibo counts and official tourism statistics

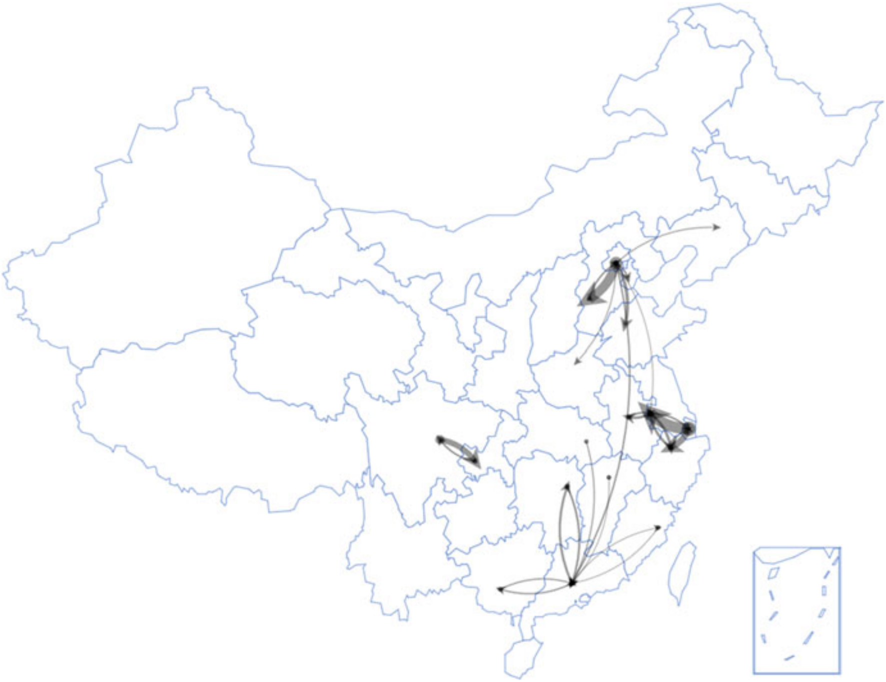


Fig. 2 Major inter-province flows in 2014 National Day Golden Week

exceptions, and they are the flow from Beijing to Jiangsu and that from Beijing to Guangdong.

Guided from the past literature, we incorporate a set of independent variables in the SIM to explain tourist flows. First, lnGDP(O) denotes the log of GDP of origin province (in 10,000 RMB), which is the product of the population size and income per capita of origin. Second, on the destination side, lnGDP(D) denotes the log of GDP of destination province (in 10,000 RMB) (Marrocu & Paci, 2013); lnhotels_pop(D) denotes the log of number of star rated hotels relative to total population (Patuelli et al., 2013); lncell_pop(D) denotes the log of number of registered cellphone users relative to total population; A5(D) denotes the number of AAAAA scenic spots, which is the highest level of scenic spots designated by the national tourism bureau; NP(D) denotes the number of national parks; WHS (D) denotes the number of world heritage sites (Patuelli et al., 2013); Attraction (D) denotes the weighted sum of top-tier tourist attraction, and it is calculated as Attraction(D) = A5(D) + NP(D) + 2*WHS(D); lnCD(D) denotes the competing destination effect of a destination province (Yang, Fik, & Zhang, 2013), and it is calculated as the log of

Table 1 Descriptive statistics of variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Y(OD)	957	1391.064	5652.928	5	113576
lnGDP(O)	957	9.575	0.962	6.694	11.038
lnGDP(D)	957	9.575	0.962	6.694	11.038
lnhotels_pop(D)	957	−2.347	0.481	−3.258	−1.052
lncell_pop(D)	957	4.491	0.227	4.128	5.072
A5(D)	957	5.780	3.284	2	17
NP(D)	957	7.278	5.709	0	19
WHS(D)	957	2.129	1.472	0	6
Attraction(D)	957	17.315	9.258	3	34
lnCD(D)	957	−7.055	0.799	−9.410	−5.709
lndistance(OD)	957	7.016	0.663	4.517	8.176

$$CD_j = \sum_{d_{jp} \le d_{ij}}^{j} \frac{Attraction(D)_p}{distance_{jp}} \tag{6}$$

Lastly, lndistance(OD) denotes the geographic distance (in km) between the capitals of the origin and destination provinces (Yang & Wong, 2012a). Table 1 presents descriptive statistics of dependent and independent variables in the SIM.

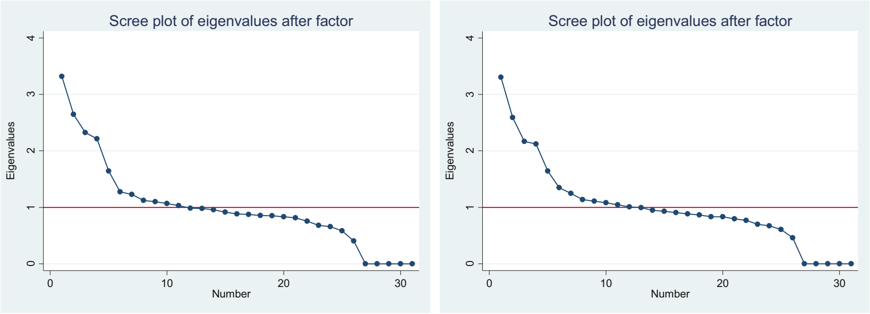
7 Results

Table 2 presents results from R-mode FA. In this analysis, based on the scree plot (Fig. 3a), we extract six factors which together explain 43.28 % of variance. Destination areas are identified with a Varimax rotated factor loading larger than 0.4 in the R-mode FA result. Major origins are provinces with regression-based factor loadings larger than 0.4, and minor origins are those with factor loadings less than 0.4 but larger than 0.3. Figure 4 maps the results from R-mode RA. The map shows six major destination areas for inter-province tourist flows in China. The most dominant one is the destination region around Beijing, the capital of China, which also covers Tianjin and Hebei. The other five destination regions are located in the East, Southwest, Northwest, Northeast, and Central South, respectively. Lastly, the pattern does not highlight any inter-sub-system linkages and therefore, suggests that the nation is composed of spatially independent tourism fields within Golden Week lasting for 7 days.

Table 3 presents results from Q-mode FA. In this analysis, based on the scree plot (Fig. 3b), we extract eight factors which together explain 50.18 % of variance. Origin areas are identified with a Varimax rotated factor loading larger than 0.4 in the Q-mode FA result. Major destinations are provinces with regression-based factor loadings larger than 0.4, and minor destinations are those with factor loadings

Table 2 R-mode FA results of tourist flow matrix

Factor	Eigenvalue	Cumulative variance explained (%)	Destination area	Major origin	Minor origin
Factor1	3.318	9.95	Beijing Hebei Tianjin	Hebei	
Factor2	2.645	18.18	Anhui, Jiangsu, Shanghai, Zhejiang	Jiangsu	Zhejiang
Factor3	2.326	25.52	Sichuan, Tibet, Chongqing	Sichuan	
Factor4	2.214	32.22	Gansu, Ningxia, Qinghai, Shaanxi		Gansu, Shaanxi
Factor5	1.641	37.78	Heilongjiang, Jilin, Liaoning	Jilin, Liaoning	Heilongjiang
Factor6	1.274	43.28	Guangdong, Hubei, Hunan, Jiangxi	Guangdong	Hubei, Hunan, Jiangxi



(a) R-mode FA

(b) Q-mode FA

Fig. 3 Scree plots of R-mode and Q-mode FA

less than 0.4 but larger than 0.3. Figure 5 maps the results from Q-mode FA. The map provides a slightly different results compared to Fig. 4. The three most prominent regions coincide with the findings from R-mode FA. However, Figure 5 highlighted a specific origin region (as Factor 7) covering Ningxia and Inner Mongolia. The Central South region highlighted in Fig. 4 is further disaggregated into two regions in Fig. 5: the one in the central part (as Factor 6) covering Fujian, Hubei and Jiangxi and the one in the South (as Factor 8) covering origins including Guangdong, Guangxi and Hunan.

Table 4 presents the econometric estimation results of SIM. Models 1 and 2 use the Poisson model whereas Models 3 and 4 use the negative binomial model that captures the unobserved heterogeneity with an additional estimate of dispersion parameter δ. Also, we consider different variable sets in measuring the destination attractiveness. In Models 1 and 3, we incorporate three variables to evaluate the

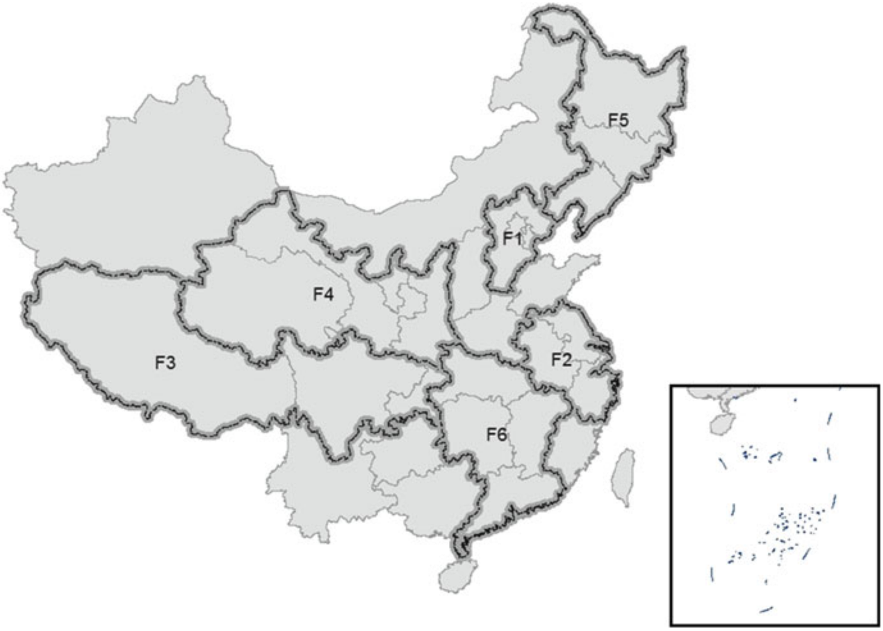


Fig. 4 Clusters of origins with similar destinations from R-mode FA

effects of three different types of top-tier attractions, namely AAAAA scenic spots (A5(D)), national parks (NP(D)) and world heritage sites (WHS(D)), and in Models 2 and 4, we use a single variable—the weighted sum of these top-tier attractions (Attraction(D)). Judging from the goodness-of-fit indexes, such as AIC, BIC and Log likelihood, we find that the negative binomial model fits the data better and the disperse parameter is estimated to be statistically significant at the 0.01 level. Moreover, based on the goodness-of-fit indexes of Models 3 and 4, Model 3 is found to outperform Model 4. Therefore, we choose Model 3 to explain and discuss the SIM estimates.

GDP of both origin and destination (lnGDP(O) and lnGDP(D)) are found to be positively and significantly associated with tourist flows, suggesting large tourist flows are observed between two large economies. The coefficient of lnhotels_pop (D) is estimated to be positive and significant, indicating that provinces with more developed hotel capacity are more likely to receive a large number of domestic tourist arrivals. The variable of lncell_pop(D) is included to correct for the over-respentedness of Weibo data we discovered earlier. However, its coefficient is not statistically significant in the SIM. Moreover, among tourist attraction variables, WHS(D) has the largest positive and significant coefficient, followed by A5(D). However, NP(D) is found to be insignificant. The results show that the both world heritage sites and AAAAA scenic spots are major attractions for incoming domestic

Table 3 Q-mode FA results of tourist flow matrix

Factor	Eigenvalue	Cumulative variance explained (%)	Origin Area	Major destination	Minor destination
Factor1	3.303	10.04	Beijing, Hebei, Shandong, Tianjin	Beijing	
Factor2	2.591	18.09	Anhui, Jiangsu, Shanghai, Zhejiang	Jiangsu, Zhejiang	Anhui
Factor3	2.169	24.80	Sichuan, Chongqing	Sichuan	
Factor4	2.120	30.68	Gansu, Qinghai, Shaanxi		Gansu, Qinghai
Factor5	1.641	36.19	Heilongjiang, Jilin, Liaoning	Jilin, Liaoning	Heilongjiang
Factor6	1.348	40.89	Fujian, Hubei, Jiangxi	Jiangxi	Fujian, Hubei
Factor7	1.248	45.57	Ningxia, Inner Mongolia	Ningxia, Inner Mongolia, Guizhou	Anhui, Guangxi, Hainan, Heilongjiang, Jilin
Factor8	1.136	50.18	Guangdong, Guangxi, Hunan	Guangxi	Guangdong, Hainan

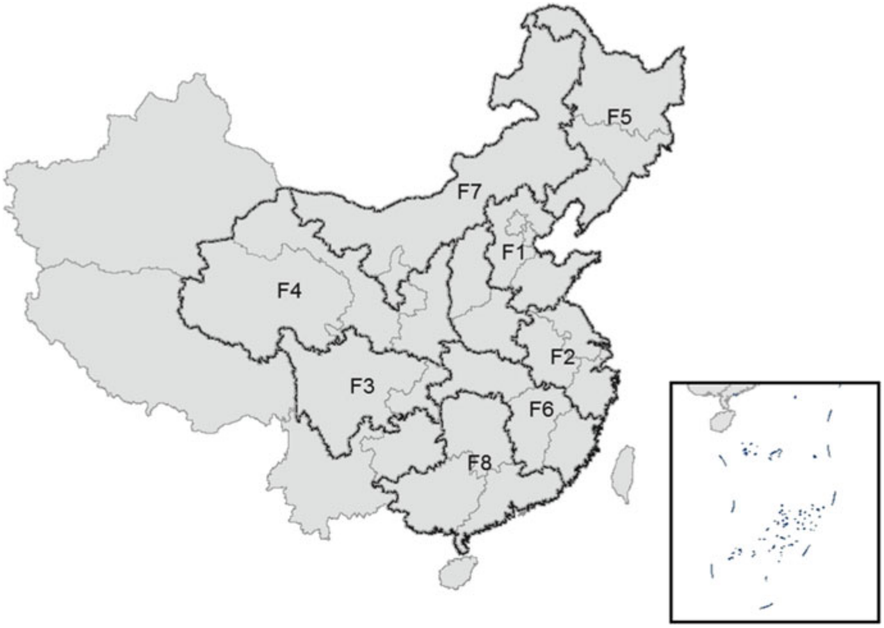


Fig. 5 Clusters of destinations with similar origins from Q-mode FA

Table 4 Estimation results of SIM

	Model 1	Model 2	Model 3	Model 4
Variables	Poisson	Poisson	Negative binomial	Negative binomial
lnGDP(O)	0.591***	0.584***	0.386***	0.385***
	(0.001)	(0.001)	(0.025)	(0.025)
lnGDP(D)	0.458***	0.584***	0.553***	0.610***
	(0.003)	(0.002)	(0.067)	(0.060)
lnhotels_pop(D)	−0.357***	−0.247***	0.297***	0.383***
	(0.004)	(0.004)	(0.115)	(0.106)
lncell_pop(D)	0.149***	0.0397***	0.171	0.134
	(0.007)	(0.006)	(0.182)	(0.182)
A5(D)	0.0493***		0.0221*	
	(0.000)		(0.011)	
NP(D)	0.0122***		−0.00230	
	(0.000)		(0.005)	
WHS(D)	0.0555***		0.0397**	
	(0.001)		(0.019)	
Attraction(D)		0.0193***		0.00674*
		(0.000)		(0.004)
lnCD(D)	−0.857***	−0.815***	−0.220***	−0.214***
	(0.002)	(0.002)	(0.038)	(0.037)
lndistance(OD)	−1.789***	−1.784***	−1.495***	−1.496***
	(0.001)	(0.001)	(0.035)	(0.035)
Constant	0.434***	0.464***	5.877***	5.842***
	(0.030)	(0.030)	(0.848)	(0.853)
δ			0.555***	0.557***
			(0.024)	(0.024)
N	957	957	957	957
AIC	950345.3	958783.3	13638.4	13639.1
BIC	950393.9	958822.2	13691.9	13682.8
Log likelihood	−475163	−479384	−6808.22	−6810.53

Standard error presented in parenthesis

Notes: *** indicates significance at the 0.01 level, ** indicates significance at the 0.05 level, and * indicates significance at the 0.10 level

tourists during the Golden week, and the impact of world heritage sites is larger than AAAA scenic spots. The coefficient of lnCD(D) is significant and negative, and it suggests the intense competition across provinces in attracting domestic tourists. Lastly, the distance decay effect is confirmed by the significant and negative estimate of lndistance(OD) in the model.

8 Conclusions

The study reported in this chapter investigated the inter-province tourist flow patterns using the geo-tagged Weibo data during the 2014 National Day Golden Week. The Weibo data was found to be highly correlated with the official tourism statistics, but moderately over-represented the tourist population in provinces with easy Internet accesses. Factor analysis unveiled several spatially independent fields of tourist flows in China, such as the area around Beijing as well as the regions in the East, Southwest, Northwest, Northeast, and Central South. According to the estimation results from the SIM based on a negative binomial specification, many determinants of tourist flows were highlighted, and they included the distance between the origin and destination, the size of economies in the origin and destination, the hotel infrastructure in the destination. Moreover, we found that world heritage sites play a more important role than AAAAA scenic spots in attracting domestic tourists during the Golden Week.

The results provide several important academic and practical implications. We demonstrated the usefulness of geo-tagged social media data in monitoring the spatial movement of tourists at a large geographic scale. Local tourist authorities and destination marketing organizations can leverage the benefits of geo-tagged/referenced big data sources by integrating different sources of data to present a more comprehensive picture on tourist flow patterns. Moreover, the big data source enables the meso-scale analysis on tourist flows of sub-groups based on the socio-demographic and tripo-graphic information recorded in the big data. Further, our empirical results show that China consists relatively spatially separated tourism fields/regions. Therefore, further management and marketing efforts can be made on strengthening the linkages across different tourism fields/regions.

References

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management, 29*(3), 469–486.

Becken, S., Vuletich, S., & Campbell, S. (2007). Developing a GIS-supported tourist flow model for New Zealand. In J. Tribe & D. Airey (Eds.), *Developments in tourism research* (pp. 107–121). Oxford: Elsevier.

Beeco, J. A., Huang, W.-J., Hallo, J. C., Norman, W. C., McGehee, N. G., McGee, J., et al. (2012). GPS tracking of travel routes of wanderers and planners. *Tourism Geographies, 15*(3), 551–573.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

CNTA. (2014). *Tourism statistical report of national day golden week, 2014*. Beijing.

de Graaff, T., Boter, J., & Rouwendal, J. (2009). On spatial differences in the attractiveness of Dutch museums. *Environment and Planning A, 41*(11), 2778–2797.

Farsari, Y., & Prastacos, P. (2004). GIS applications in the planning and management of tourism. In A. A. Lew, C. M. Hall, & A. M. Williams (Eds.), *A companion to tourism* (pp. 596–608). Oxford: Blackwell.

Girardin, F., Fiore, F. D., Ratti, C., & Blat, J. (2008). Leveraging explicitly disclosed location information to understand tourist dynamics: A case study. *Journal of Location Based Services, 2*(1), 41–56.

Gober, P., & Mings, R. C. (1984). A geography of nonpermanent residence in the U.S. *The Professional Geographer, 36*(2), 164–173.

Hall, C. M., & Page, S. J. (2009). Progress in tourism management: From the geography of tourism to geographies of tourism—A review. *Tourism Management, 30*(1), 3–16.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271.

Holyoak, N., Carson, D., & Schmallegger, D. (2009). VRUMTM: A tool for modelling travel patterns of self-drive tourists. In W. Höpken, U. Gretzel, & R. Law (Eds.), *Information and communication technologies in tourism 2009* (pp. 237–247). Vienna: Springer.

Husbands, W. C. (1983). Tourist space and touristic attraction: An analysis of the destination choices of European travelers. *Leisure Sciences, 5*(4), 289–307.

Jansen-Verbeke, M., & Spee, R. (1995). A regional analysis of tourist flows within Europe. *Tourism Management, 16*(1), 73–80.

Kádár, B. (2014). Measuring tourist activities in cities using geotagged photography. *Tourism Geographies, 16*(1), 88–104.

Kozak, R., & Sezgin, E. (2013). Examining domestic transactions of incoming tourists with credit cards in Turkey. *Management: Journal of Contemporary Management Issues, 18*(1), 23–43.

Lau, G., & McKercher, B. (2006). Understanding tourist movement patterns in a destination: A GIS approach. *Tourism and Hospitality Research, 7*(1), 39–49.

Li, X., Meng, F., & Uysal, M. (2008). Spatial pattern of tourist flows among the Asia-Pacific countries: An examination over a decade. *Asia Pacific Journal of Tourism Research, 13*(3), 229–243.

Lowe, J. C., & Moryadas, S. (1975). *The geography of movement*. Boston: Houghton Mifflin.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management, 24*(2), 119–154.

Marrocu, E., & Paci, R. (2013). Different tourists to different destinations. Evidence from spatial interaction models. *Tourism Management, 39*, 71–83.

Müller, D. K. (2006). The attractiveness of second home areas in Sweden: A quantitative analysis. *Current Issues in Tourism, 9*(4/5), 335–350.

Nilbe, K., Ahas, R., & Silm, S. (2014). Evaluating the travel distances of events visitors and regular visitors using mobile positioning data: The case of Estonia. *Journal of Urban Technology, 21*(2), 91–107.

Patuelli, R., Mussoni, M., & Candela, G. (2013). The effects of World Heritage Sites on domestic tourism: A spatial interaction model for Italy. *Journal of Geographical Systems, 15*(3), 369–402. English.

Pearce, D. G. (1996). Domestic tourist travel in Sweden: A regional analysis. *Geografiska Annaler Series B, Human Geography, 78*(2), 71–84.

Scherngell, T., & Barber, M. J. (2009). Spatial interaction modelling of cross-region R&D collaborations: Empirical evidence from the 5th EU framework programme. *Papers in Regional Science, 88*(3), 531–546.

Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research, 34*(1), 141–159.

Shoval, N., Isaacson, M., & Chhetri, P. (2013). GPS, Smartphones, and the future of tourism research. In A. A. Lew, C. M. Hall, & A. M. Williams (Eds.), *The Wiley Blackwell companion to tourism* (pp. 251–261). Oxford: Blackwell.

Tiru, M., Kuusik, A., Lamp, M.-L., & Ahas, R. (2010). LBS in marketing and tourism management: Measuring destination loyalty with mobile positioning data. *Journal of Location Based Services, 4*(2), 120–140.

Versichele, M., de Groote, L., Claeys Bouuaert, M., Neutens, T., Moerman, I., & Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management, 44*, 67–81.

Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography, 32*(2), 208–220.

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management, 46*, 222–232.

Wu, C.-L., & Carson, D. (2008). Spatial and temporal tourist dispersal analysis in multiple destination travel. *Journal of Travel Research, 46*(3), 311–317.

Wu, B., Xue, L., Morrison, A. M., & Leung, X. Y. (2012). Frame analysis on golden week policy reform in China. *Annals of Tourism Research, 39*(2), 842–862.

Yang, Y., Fik, T., & Zhang, J. (2013). Modeling sequential tourist flows: Where is the next destination? *Annals of Tourism Research, 43*, 297–320.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447.

Yang, Y., & Wong, K. K. F. (2012a). The influence of cultural distance on China inbound tourism flows: A panel data gravity model approach. *Asian Geographer, 29*(1), 21–37.

Yang, Y., & Wong, K. K. F. (2012b). Spatial distribution of tourist flows to China's cities. *Tourism Geographies, 15*(2), 338–363.

Zhang, J., Zhang, J., Li, N., Liang, Y., & Liu, Z. (2005). An analysis on spatial field effect of domestic tourist flows in China. *Geographical Research, 24*(2), 293–303.

Part IV
Web and Social Media Analytics: Concepts and Methods

Sensing the Online Social Sphere Using a Sentiment Analytical Approach

Wolfram Höpken, Matthias Fuchs, Th. Menner, and Maria Lexhagen

1 Introduction

User generated content (UGC), i.e. content generated by Internet users in the form of product reviews, blogs or messages on social media platforms, makes-up an increasing share of overall Internet content (Liu, 2011). In the tourism domain, UGC comes in the form of blogs (like travel diaries or reports), messages on social networks, like Facebook or Twitter, or product reviews on platforms, like TripAdvisor or Booking.com. Especially in tourism, social media gain more and more attention and play an increasingly important role in customers' decision making process (Lexhagen, Kuttainen, Fuchs, & Höpken, 2012). On the one hand, tourists heavily use social media and review platforms in order to express their opinion on tourism products and services after consumption and, on the other hand, to inform themselves about product quality and suitability before consumption takes place.

TripAdvisor, one of the most important travel review platforms, currently, shows over 250 million single reviews for over 5.2 million tourism businesses worldwide (http://www.tripadvisor.com/PressCenter-c6-About_Us.html). While social media tools, like Facebook, are mainly used by tourists to inform friends and social peers about their own holiday experience, review platforms, like TripAdvisor, are used to specifically judge the quality of concrete tourism products or services (Murphy, Gil, & Schegg, 2010). UGC, and especially product reviews, evidently, have a strong influence on tourists' travel decision. 90 % of travelers look at other consumers' comments during trip planning, 87 % of travelers say that

W. Höpken (✉)
Hochshule Ravensburg-Weingarten, Weingarten, Germany
e-mail: wolfram.hoepken@hs-weingarten.de

M. Fuchs • T. Menner • M. Lexhagen
Mid Sweden University, Sundsvall, Sweden

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on
the Verge, DOI 10.1007/978-3-319-44263-1_8

reviews impacted their hotel choice, and 70 % of travelers trust online recommendations (Gretzel, Yoo, & Purifoy, 2007). UGC is, indeed, viewed as more up-to-date, enjoyable and reliable, compared to information from travel service providers.

However, UGC is not only a valuable knowledge source for customers, but also for tourism service providers in order to learn what customers are saying about a service provider and its products. Product reviews, typically, contain concrete and unbiased customer feedback on product quality and suitability for certain customer segments and, thus, constitute a valuable input to product optimization and customer relationship management. Analyzing UGC, in general, means to find out which user expresses which opinion (positive or negative) on which tourism service or topic (e.g. hotel room, service, food, etc.). While the users posting a review on a review platform are typically described by basic demographic characteristics and travel motives in a well-structured format, the opinion itself and the topic have to be extracted from the free text of the customer review. As the number of available customer reviews has increased dramatically, manually analyzing customer reviews is no longer feasible. Thus, automatic methods for extracting knowledge from free text customer reviews (called sentiment analysis or opinion mining) are needed and gained high research attention in recent years. Such approaches, typically, try to automatically extract customer feedback from free text by statistical methods or methods of machine learning. Moreover, these approaches try to find out which topic the feedback is about (topic detection) and the sentiment or polarity of the feedback (sentiment detection) (Liu, 2011).

This chapter provides an overview on different approaches in the area of sentiment analysis, specifically in the tourism domain, and demonstrates their applicability and their usefulness by means of an application case for the leading Swedish mountain destination Åre. The chapter is structured as follows: section two introduces the overall topic of sentiment analysis and briefly presents existing technical approaches and initiatives. The third section discusses different approaches for topic detection, as a subtask of sentiment analysis to identify the topic of the review. The forth section deals with subjectivity detection and sentiment detection, two subtasks aiming to identify the sentiment or polarity of the review. The following section presents an application of the approaches discussed before, and demonstrates their applicability and business benefits in the context of a destination management information system (DMIS). The chapter concludes with a summarization of the most important results and provides an outlook on future research activities and improvements.

2 Sentiment Analytical Approaches in Tourism

UGC, or more precisely, customer feedback in free text form is growing steadily on the Internet. Thus, approaches for analyzing such content, called sentiment analysis, have enjoyed an increasing interest by researchers and practitioners in recent years. The overall task of sentiment analysis can be divided into two subtasks: topic detection and sentiment detection. Topic detection aims at identifying the topic, i.e. the opinion object, which the user is providing feedback for. In the case of a hotel room, for example, typical topics are the room itself (i.e. its size or equipment), food & beverage, service, etc. Sentiment detection, then, deals with the identification of the sentiment or polarity of the feedback, i.e. whether the feedback is a positive or negative statement or a neutral observation. Sentiment detection can be done for a complete user review, a single statement (e.g. sentence) within an overall review, or even a single topic identified by a preliminarily executed topic detection.

Each of the two tasks above can be performed by different approaches. A simple approach, which is nevertheless quite popular, is the dictionary-based approach. Based on a dictionary, i.e. a word-list, with, for example, positive or negative words, this approach can identify whether a text is meant positively or negatively by simply counting the number of positive or negative words (Liu, 2011). Despite their simplicity, dictionary-based approaches reach good results under certain circumstances, whereby their intelligence is based especially in a well-defined and expressive dictionary. Moreover, well-known machine learning approaches are employed in the form of both, supervised and unsupervised learning. Supervised learning is, for example, used to identify the topic of a review by means of a classification, learned by appropriately pre-classified training data. Unsupervised learning approaches solve the same problem by, for example, using a factor or cluster analysis approach identifying often co-occurring words and assigning them, and subsequently the review or statement they are occurring in, to a potential topic. While the approaches above simply treat the underlying text as a “bag of words” (i.e. a word vector), semantic approaches look at the concrete structure of a sentence from which they try to deduce the semantics of single words. Obviously, in order to make use of their respective advantages, concrete applications of sentiment analysis often combine above approaches to so called hybrid approaches.

Based on the above reflections, we adapt the classification proposed by Tsytsarau and Palpanas (2011) and group sentiment analysis approaches into four major categories: (1) Supervised machine learning, (2) dictionary-based, (3) unsupervised machine learning, and (4) semantic approaches, respectively. We are following this categorization when briefly discussing previous research in the field of sentiment analysis within the hospitality and tourism domain.

2.1 Supervised Machine Learning Approaches

The study by Ye, Zhang, and Law (2009) incorporates sentiment classification techniques into the domain of mining reviews from travel blogs. The study compares three supervised machine learning techniques, namely Naïve Bayes, support vector machines and the character-based N-gram model, for sentiment classification of reviews from travel blogs for popular travel destinations in the US and

Europe. Findings indicate that support vector machines and N-gram approaches outperform the Naïve Bayes approach. Interestingly, if training datasets had a relatively large number of reviews, all three approaches reached accuracy levels of at least 80 % (Ye et al., 2009, p. 6527).

The study by Lin and Chao (2010) is focusing on tourism-related opinion mining by utilizing annotated data from blogs of domestic travelers. More precisely, annotators were asked to annotate opinion polarity and the opinion target for every sentence. Subsequently, machine-learning methods are applied to train classifiers. Precision and recall scores of tourism-related sentiment detection amount at 55.98 % and 59.30 %, respectively. In contrast, the scores for target identification (i.e. topic detection) among known tourism-related opinionated sentences stand at 90.06 % and 89.91 %, respectively (Lin & Chao, 2010, p. 37).

Kasper & Vela's (2011) study, first of all, utilizes an already existing (i.e. German) dictionary to initialize sentiment analysis for terms extracted from a hotel review corpora (Waltinger, 2010). In order to achieve the goal of sentiment detection, the extracted 7200 text segments are, subsequently, used to train machine learning-based classifiers (i.e. 4-g with Goodman smoothing) with two polarity classes (i.e. positive/negative). A final cross-validation demonstrates a satisfactory performance in terms of model accuracy (Kasper & Vela, 2011, p. 45).

Similar to the study by Ye et al. (2009), Alves, Baptista, Firmino, de Oliveira, and de Paiva (2014) compare support vector machines with Naïve Bayes classifiers to perform sentiment analysis of tweets (i.e. written in Portuguese) during the 2013 FIFA Confederations Cup. Findings repeatedly indicate that support vector machines outperform the Naïve Bayes technique (Alves et al., 2014, p. 123). Markopoulos, Mikros, Iliadi, and Liontos (2015) create a classifier for sentiment detection by applying the machine learning-based method of support vector machines on hotel reviews written in Modern Greek. Findings are satisfactory after utilizing a unigram language model (Markopoulos et al., 2015, p. 373). Finally, the study by Pablos, Cuadros, and Linaza (2015) introduces the European OpeNER project, a set of free Open Source and ready-to-use text analysis tools (e.g. support vector machines) to perform natural language and text processing tasks, like Named Entity Recognition and Opinion detection. In addition, the paper provides an interesting example of a possible application of OpeNER to the geo-location of hotel reviews (Pablos et al., 2015, p. 125).

2.2 *Dictionary-Based Approaches*

García, Gaines, and Linaza (2012) present a dictionary-based sentiment analysis approach for the tourism domain. The study introduces the use of a lexical database for sentiment analysis of TripAdvisor reviews for the accommodation and food & beverage sectors, respectively. By using the lexicon database with more than 6000 words, a sentiment score, based on negative and positive words appearing in the reviews, is calculated. By using the dictionary, sentences of a review are annotated by its polarity. Finally, the proposed approach also includes a taxonomy to classify fragments by their topic using a list of lemmatized and normalized words, each of them belonging to a different topical category (García et al., 2012, p. 35).

Similarly, a study by Gräbner, Zanker, Fliedl, and Fuchs (2012) proposes a system that performs the classification of customer reviews of hotels. A process is elaborated which extracts a domain-specific lexicon of semantically relevant words based on a given corpus. The resulting lexicon backs the sentiment analysis for generating a classification of the reviews. The evaluation of the classification on test data shows that the proposed system performs better compared to a predefined baseline: if a customer review is classified manually as good or bad, the classification is correct with a probability of about 90 % (Gräbner et al., 2012, p. 460).

2.3 Unsupervised Machine Learning Approaches

A study by Xiang, Schwartz, and Uysal (2015) explores the usefulness of identified guest experience dimensions based upon authentic online customer reviews in order to understand what types of hotels make their guests (un-)happy. Hotels are grouped by experience dimensions and satisfaction ratings using cluster analysis (i.e. unsupervised machine learning). Then, the hotel clusters are examined in relation to topic words in customer reviews with correspondence analysis. Findings show that there are different types of hotels with unique, salient traits that satisfy their customers, while those who failed to do so mostly have issues related to cleanliness and maintenance-related factors. This study points to a promising direction employing authentic consumer experience data to support perceptual mapping and market segmentation for the hospitality industry (Xiang, Schwartz & Uysal, 2015, p. 33).

Rossetti, Stella, Cao, and Zanker (2015) explore different application scenarios to analyze user reviews in tourism with topic models. The method pertains to the statistical approach and is well capable to process textual reviews. Besides contributing with a new model based on the topic model method, this study also includes empirical evidence from experiments on user reviews from the YELP dataset as well as from TripAdvisor (Rossetti et al., 2015, p. 47).

2.4 Semantic Approaches

Kasper and Vela (2012) present a system that automatically monitors user reviews and comments on hotels from various social media sites, making use of semantic techniques. As an important knowledge base for a hotel's quality control, the system provides classified summaries of positive and negative features of a hotel (Kasper & Vela, 2012, p. 471). The study by Xiang, Schwartz, Gerdes, and Uysal (2015) explores the utility of big data analytics to better understand the relationship between hotel guest experience and satisfaction. Their study applies a text analytical approach to a large quantity of consumer reviews extracted from Expedia.com to deconstruct hotel guest experience and to examine its association with satisfaction ratings. Findings reveal several dimensions of guest experience that carried varying weights and, thus, have novel meaningful semantic compositions. The semantic association between guest experience and satisfaction appears strong, suggesting that these two domains of consumer behavior are inherently connected (Xiang, Schwartz, Gerdes, et al., 2015, p. 120).

2.5 Hybrid Approaches

In a final step of the literature review, we additionally present hybrid approaches applied for sentiment analysis. The latter approaches combine supervised machine learning, dictionary-based, unsupervised machine learning and semantic approaches, respectively. A study by Waldhör and Rind (2008) combines a dictionary-based approach for topic detection and a machine learning approach for opinion mining, including semantic (i.e. linguistic) aspects. The proposed semi-automatic software tool for e-blog analysis in the tourism domain includes routines for crawling, sentiment extraction and text categorization, respectively. More precisely, it combines linguistic parsing methodology with information and terminology extraction methods in order to determine polarity and power of expressions. Thus, the proposed approach proves to be especially useful to consider semantic aspects, like negations (i.e. "not") or words which are changing the power (e.g. "very") (Waldhör & Rind, 2008, p. 453).

In a paper by Weichselbraun, Gindl, and Scharl (2013) a hybrid approach that combines a lexical (i.e. dictionary-based) analysis with the flexibility of machine learning to resolve issues of ambiguity and to consider the topical context of sentiment terms, is introduced. The proposed method identifies ambiguous terms that vary in polarity depending on the context and, thus, stores them in contextualized sentiment lexicons. In conjunction with semantic knowledge bases, these lexicons help link ambiguous sentiment terms to concepts that correspond to their polarity. An extensive evaluation applies the method to user reviews across three domains, namely, movies, physical products and hotels (Weichselbraun et al., 2013, p. 39).

A recent study by Schmunk, Höpken, Fuchs, and Lexhagen (2014) presents a hybrid approach for extracting decision-relevant knowledge from UGC and compares different data mining (DM) techniques concerning their accuracy in identifying the polarity of customer opinions and in assigning opinions to topics. More concretely, the study aims at conceptualizing the overall process of information extraction from customer reviews of tourism review platforms, like TripAdvisor and Booking.com, and at comparing different DM techniques (i.e. dictionary-based and machine learning algorithms, like Naïve Bayes, support vector machines and k-nearest neighbor) for identifying both, the topic and the sentiment of the opinion.

The proposed techniques are evaluated in terms of the quality of extracted information, its accuracy and its practical use within a destination management information system (Fuchs, Höpken, & Lexhagen, 2014; Höpken, Fuchs, Keil, & Lexhagen, 2015; Schmunk et al., 2014, p. 254).

Finally, an opinion mining method based on feature-based sentiment classification to extract e-word-of-mouth from weblogs is presented in a recent study by Chiu, Chiu, Sunga, and Hsieh (2015). For opinion extraction, the supervised learning algorithm 'point-wise mutual information' (PMI) is applied to identify words associated with positive or negative paradigms. In addition, a heuristic n-phrase rule is utilized to identify customer opinions about hotel attributes, including hotel image, services, price/value, food and beverage, room, amenities, and location. Findings show that the proposed hybrid approach demonstrates its effectiveness with acceptable classification and forecasting performance, respectively. Finally, a perceptual map based on correspondence analysis visualizes opinion comparisons to provide insight into the hotels' competitive position (Chiu et al., 2015, p. 477).

3 Topic Detection

Topic detection can be executed in a supervised or unsupervised manner. In the supervised case, topics are predefined and, consequently, the number of topics is limited. Examples of such predefined topics in the case of hotel reviews are room, food & beverage, service & personal, location, etc. Supervised topic detection typically takes place on the level of a statement (e.g. a sentence) within a review, as complete reviews tend to deal with more than one topic. Possible approaches to conduct a supervised topic detection are dictionary-based approaches or supervised machine learning techniques (or more concrete classification techniques, like Naive Bayes or Support Vector Machines [SVM]). A clear benefit of supervised topic detection is that topics are fix and, thus, comparable across all reviews and suppliers as well as over time and may serve as valuable input to cross-supplier benchmarking.

In the unsupervised case, topics are not predefined, but any topic customers are talking about can be identified. Consequently, the number of topics is unlimited and typically much higher than the usual number of predefined topics. Unsupervised topic detection typically takes place on the level of single words and, thus, can identify several topics within the same statement or sentence (although it can also be aggregated on the sentence level, which is especially meaningful if the topic detection is to be combined with a sentiment detection, which takes place on the sentence level). Possible approaches for unsupervised topic detection are unsupervised machine learning or statistical techniques, like clustering and factor analysis, or the supervised machine learning technique sequential pattern mining. A clear benefit of unsupervised topic detection is that the most important topics customers are talking about are automatically identified without the need to predefine them in advance. Thus, new topics and a topic shift or trends can be identified.

3.1 Supervised Topic Detection

Supervised topic detection is typically executed by dictionary-based approaches or classification techniques as a type of supervised machine learning. *Dictionary-based topic detections* means that for each class, here topic, a dictionary, i.e. a word-list, is provided, containing a collection of words representing or labelling this class. The topic of a statement is then identified by just counting the number of words of each wordlist and assigning the majority class or topic. Within a prototypical sentiment analysis implementation for the Swedish mountain destination Åre, the dictionary-based approach has been tested on 208 hotel reviews, consisting of 1516 single statements, i.e. sentences, extracted from the platforms TripAdvisor and Booking.com, respectively. Word-lists have been manually defined for the topics food/breakfast, hotel, room, service/personal, location and wellness (containing between three and seven words). The results have been compared to manually classified review statements and the dictionary-based approach reached an accuracy of 71.28 %. On the other hand, *Supervised machine learning approaches* follow the idea of learning how to deduce the class of an observation (i.e. in our case a statement within a review) from its characteristics (i.e. in our case the review text) based on pre-classified training data. The review text is simply represented as a "bag of words", i.e. a word vector based on word occurrences or, more precisely, TF-IDF values (term frequency—invers document frequency), which reflect how specific a word is for a certain document. The learned classifier, then, decides for each review statement, based on the words occurring within the statement, which class (i.e. topic) is the most likely one. As supervised machine learning approaches, we compared support vector machines (SVM), Naïve Bayes (both well known to be suitable especially for text classification), and k-nearest neighbor (k-NN), with the result that SVM, with an accuracy of 72.35 %, outperformed both, Naïve Bayes (49.72 %) and k-NN (57.08 %). In all cases, POS (Part-Of-Speech) tagging has been used to reduce the review text to nouns, which in fact increased accuracy, at least for the SVM approach.

To summarize, SVM as a supervised machine learning approach achieved the best results, especially compared to the other machine learning approaches (Table 1). Although, the dictionary-based approach is slightly inferior compared to SVM, based on its simplicity it is still an option in a practical implementation case.

Table 1 Evaluation of supervised topic detection approaches

Supervised topic detection approach	Accuracy (%)
Dictionary-based	71.28
k-NN	57.08
SVM	72.35
Naïve Bayes	49.72

3.2 Unsupervised Topic Detection

Unsupervised topic detection aims at identifying any topics within review statements without the need to predefine topics. Unsupervised topic detection is typically executed by statistical approaches, like factor analysis, or machine learning techniques, like clustering or sequential pattern mining. *Frequent words* is a simple approach if the problem of topic detection is the identification of *frequent words* as potential topics. The underlying assumption is simply, if several review statements are related to the same product (in our case a hotel), then they will usually mention the same topics or characteristics. Thus, topic-specific words will occur quite frequently. In contrast, non-topic-specific words will show a much higher diversity and, thus, each of them will occur less frequently than topic-specific words. Vice versa, we can conclude that frequently used words (or more concrete nouns), in many cases represent the topics mentioned within a review (Hu & Liu, 2004, p. 168; Liu, 2011, p. 487). When tested on the reviews for hotels in Åre, this approach reached an accuracy of 82.86 %. The accuracy is calculated by comparing the results with manually annotated test data, i.e. reviews with topic-specific words labelled as such. However, a precision of 53.10 % and a recall of 94.20 % for detecting a topic reveal that too many words are identified as topics, constituting a limitation of the approach.

Cluster analysis is used if two review statements deal with the same topic, we assume that they will contain the same (topic-specific) words. If we now combine (i.e. cluster) statements based on their contained words (i.e. each cluster represents a frequently occurring word combination), the clusters can be viewed as latent topics, and each topic is described by the words occurring in the corresponding cluster (Kiran, Shankar, & Pudi, 2010). Analog to the approach above, it is meaningful to restrict the analysis to nouns, and even better to important, i.e. representative, nouns, so called keywords. Nouns are filtered by POS-tagging, important nouns by their TF-IDF value. The keywords, belonging to the identified topic, can then be labelled as topic-specific words within the review statements. The evaluation is again done by comparing the results with manually labelled statements. A k-means clustering with 80 clusters (and the cosine similarity as distance measure) yields an accuracy of 88.45 %.

Latent Semantic Indexing (LSI) is another well-known approach for topic detection (LSI; Miner et al., 2012). Analog to the cluster analysis described above, we assume that reviews dealing with the same topic, will contain the same (topic-specific) words. However, the LSI approach builds on the principle of dimension reduction. The dimensions are the single words occurring in review statements, and words, often co-occurring in the same statements, are grouped together into factors by means of a factor analysis or, more concretely, by Single Value Decomposition (SVD). The resulting factors, then, represent latent topics. We evaluated the approach above on the hotel reviews for hotels in Åre and the LSI with 40 factors reached an accuracy of 88.39 %.

The topic detection approaches above treat the review texts as "bag of words", i.e. the sequence of words and its position within a sentence is not considered. The basic idea of sequential pattern mining for topic detection is to take into account the context of each word, i.e. the words directly before and after. In this case, review statements are no longer represented as word vector (i.e. bag of words) but each single word is stored together with its context, thus, a certain number of words before and after the word, as well as word characteristics, like its position within the sentence, its length, etc. In order to identify which words represent a topic depending on their context and characteristics, *sequential pattern mining* approach needs training data with already labelled topic-specific words, analog to the test data used in the approaches discussed above. Learning sequential patterns, which then enables the identification of topic-specific words, can make use of any kind of classification technique. Although such classification techniques fall into the category of supervised learning, the overall approach still constitutes a case of unsupervised topic detection, as no topics are predefined and the words are not classified as a concrete topic (like room, service/personal, etc.). We applied this approach to the hotel reviews of Åre hotels, using Naïve Bayes as the classification technique and a word context of two preceding and subsequent words, and achieved an accuracy of 92.47 %. Unfortunately, the quite high accuracy has to be attributed to the fact that the same words, labelled as topics within the training data, are just recognized again, if they occur in the test data as well. In order to answer the question, how well topics can be identified which have not been labelled as such within the training data (which is then comparable to the other approaches discussed before), we tested the approach on test data, not containing any pre-labeled topic words, and reached an accuracy of 83.43 %.

To summarize, if we compare the four unsupervised topic detection approaches, presented above, we can conclude that the cluster analysis and latent semantic indexing deliver the best results (Table 2). Besides a lower accuracy, the sequential pattern mining approach, additionally, suffers from the need to provide pre-labeled training data.

4 Sentiment Detection

Sentiment detection deals with the identification of the sentiment or polarity of a complete review or a review statement. This task can be supported by a subjectivity detection as a preparatory step (Liu, 2011). In this case, the subjectivity detection

Table 2 Evaluation of unsupervised topic detection approaches

Topic detection approach	Accuracy (%)
Frequent words	82.86
Cluster analysis	88.45
Latent Semantic Indexing	88.39
Sequential pattern mining	83.43 (92.47)

just identifies whether a review statement is subjective or objective, thus, whether the statement contains an opinion (e.g. the room service is good or bad) or is a neutral observation (e.g. our room is located on the second floor). Then, the sentiment analysis itself only has to deal with subjective statements and, depending on the applied technique, typically achieves better results than a sentiment analysis working on subjective and objective statements at once.

4.1 *Subjectivity Detection*

Subjectivity detection, i.e. classifying whether a review statement is subjective or objective, will in the following be handled by dictionary-based approaches as well as supervised machine learning approaches including k-NN, SVM and Naïve Bayes. For the dictionary-based approach, a word-list containing 6800 positive and negative opinion words is used (Liu, 2011). If a statement contains any opinion words, this statement is assigned to the class “subjective”. Otherwise, the class “objective” is assigned. In our test setting with hotel reviews of hotels in Åre, the dictionary-based approach reached an accuracy of 82.63 %, when compared with pre-classified test data. Stemming, i.e. reducing words to their word stem (e.g. walking to walk), or lemmatization, i.e. mapping inflected forms of words to their lemma or canonical form (e.g. better to good), did not increase the accuracy within our test setting.

The task of subjectivity detection primarily has been conducted using three supervised machine learning approaches, namely k-NN, SVM and Naïve Bayes. Analog to the test data mentioned above, training data for the classification techniques have been created by manually classifying review statements into the classes “subjective” or “objective”. A tenfold cross-validation is used to evaluate all machine learning models and to calculate their accuracy. The best accuracy showed the k-NN (71.50 % with k = 117), followed by SVM (69.70 %) and Naive Bayes (63.00 %). In contrast to the two other approaches, k-NN significantly benefits from using bi-grams (i.e. adding word groups of length two) and filtering nouns and adjectives via POS-tagging.

To summarize, the highest accuracy of 80.37 % for subjectivity detection was achieved by the dictionary-based approach, which is significantly better than the k-NN method (71.50 %) as the best machine learning approach (Table 3). It might be reasonably assumed that the good results of the dictionary-based approach are achieved through the relatively large wordlists comprising more than 6800 words, in comparison to the limited training data set size of 1516 pre-classified statements.

Table 3 Evaluation of subjectivity detection approaches

Subjectivity detection approach	Accuracy (%)
Dictionary-based	80.37
k-NN	71.50
SVM	69.70
Naïve Bayes	63.00

Table 4 Examples for subjectivity detection

Review statement	Detected class	Real class
Hmmm must be a hospital because of that sweet smell of mould and or dead old lady	Subjective	Subjective
Would not recommend unless you have children	Subjective	Subjective
Skiing and staying in Sweden is so different to other European resorts	Objective	Objective
The restaurant is high standard very original and lots of local products	Objective	Subjective
This can be a cost saver for families with children	Subjective	Objective

Table 4 provides some examples for the subjectivity detection. As can be seen by the examples, problems arise if either the statements are ambiguous (e.g. "This can be a cost saver for families with children") or contain a mixture of different opinions (e.g. "The restaurant is high standard very original and lots of local products").

4.2 Sentiment Detection

The sentiment detection builds on the subjectivity detection and classifies subjective statements into positive and negative statements. Analog to the subjectivity detection, the sentiment detection is handled by dictionary-based approaches as well as supervised machine learning approaches, like k-NN, SVM or Naïve Bayes. The dictionary-based sentiment detection makes use of the word-list from Liu (2011), containing about 2000 positive and 4800 negative words. In case of no majority of either positive or negative words, the class *neutral* is assigned (which is not to be confused with the objective statements as result of the subjectivity detection). In our test setting, the dictionary-based sentiment detection reached an accuracy of 71.28 %. Considering negation words (i.e. *not*), which are changing the semantic orientation of a statement, did not increase the accuracy.

Analog to subjectivity detection, the sentiment detection has been executed by the classification techniques k-NN, SVM and Naïve Bayes. Training data has been created by manually classifying review statements into the classes "positive" or "negative". Finally, a tenfold cross-validation is used to evaluate accuracy. The best accuracy showed the SVM approach (76.80 %, using bi-grams), followed by Naive Bayes (69.80 %, using tri-grams) and k-NN (69.60 %, with $k = 8$).

Table 5 Evaluation of sentiment detection approaches

Sentiment detection approach	Accuracy (%)
Dictionary-based	71.28
k-NN	69.60
SVM	76.80
Naïve Bayes	69.80

To summarize, the best result for sentiment detection is gained by the SVM approach (with word bi-grams), showing an accuracy of 76.80 % (Table 5). A likely reason for the somewhat poorer performance of the dictionary-based approach (71.28 %) can be attributed to the fact that, in this case, an additional class "neutral" is considered if opinion words for the classes "positive" and "negative" are equally frequent.

Table 6 shows some examples for the sentiment detection. Here again we can see, that problems occur, if either statements contain multiple opinions with a different sentiment (e.g. "rooms aren't too big but very clean and comfy") or words are used in a misleading way (e.g. "All other guests I would recommend hotel diplomat instead").

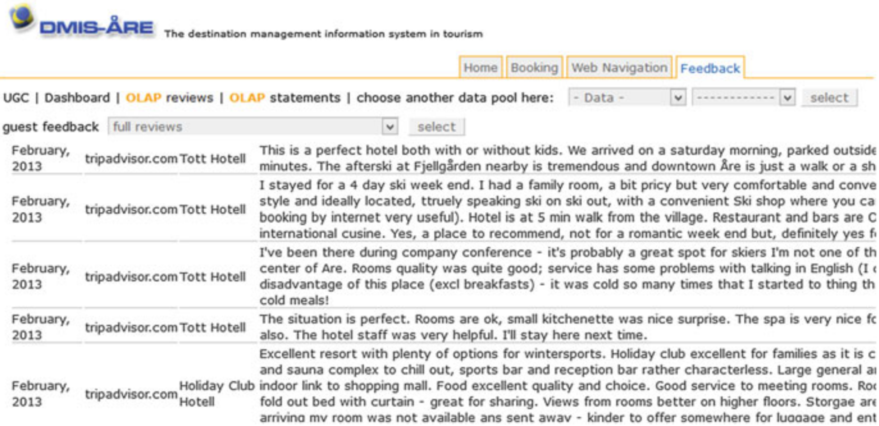
5 UGC as Input to Decision Support: A Case Study

A prototypical destination management information system (DMIS) has been developed as a fully validated and functional prototype for the Swedish mountain destination Åre. The DMIS extracts data from different data sources, like booking systems, webserver log files, customer surveys, etc., stores them in a central data warehouse, and makes them available to destination managers and stakeholders by means of interactive visualizations, e.g. dashboards, and analyses, e.g. OLAP (online analytical processing) analyses (Fuchs et al., 2014; Höpken et al., 2015). The DMIS prototype is implemented based on the business intelligence platform Rapid-Miner®, offering specific support in the area of data integration/preprocessing and data analyses. Besides customer feedback in the form of customer surveys, executed offline and online, UGC in the form of customer online reviews constitutes an important data source in the context of the DMIS described above. Thus, customer reviews, extracted from the online platforms TripAdvisor and Booking.com, have been integrated into the DMIS and its data warehouse.

In the course of preprocessing, (1) html-pages, containing customer reviews, have been fetched from the relevant online platforms by a web crawler, (2) review texts, together with additional information, like date of the review, reviewed hotel, etc., have been extracted from the html-pages, (3) empty or non-English reviews have been removed, and (4) reviews have been split into single statements or sentences. Subsequently, review statements are classified into their topic and their sentiment, by the most appropriate approach, described in the previous sections.

Table 6 Examples for sentiment detection

Review statement	Detected class	Real class
Parts of the hotel seems to be an old hospital	Negative	Negative
All other guests I would recommend hotel diplomat instead	Positive	Negative
The rooms aren't too big but very clean and comfy	Negative	Positive
Good rooms and nicely clean	Positive	Positive
Very nice breakfast room good selection for breakfast	Positive	Positive


DMIS-ÅRE
The destination management information system in tourism
Home
Booking
Web Navigation
Feedback
UGC | Dashboard | OLAP reviews | OLAP statements | choose another data pool here:
- Data -
----
select
guest feedback
full reviews
select

February, 2013	tripadvisor.com	Tott Hotell	This is a perfect hotel both with or without kids. We arrived on a saturday morning, parked outside minutes. The afterski at Fjellgården nearby is tremendous and downtown Åre is just a walk or a sh
February, 2013	tripadvisor.com	Tott Hotell	I stayed for a 4 day ski week end. I had a family room, a bit pricy but very comfortable and conve style and ideally located, ttruely speaking ski on ski out, with a convenient Ski shop where you ca booking by internet very useful). Hotel is at 5 min walk from the village. Restaurant and bars are C international cusine. Yes, a place to recommend, not for a romantic week end but, definitely yes f
February, 2013	tripadvisor.com	Tott Hotell	I've been there during company conference - it's probably a great spot for skiers I'm not one of th center of Are. Rooms quality was quite good; service has some problems with talking in English (I disadvantage of this place (excl breakfasts) - it was cold so many times that I started to thing th cold meals!
February, 2013	tripadvisor.com	Tott Hotell	The situation is perfect. Rooms are ok, small kitchenette was nice surprise. The spa is very nice f also. The hotel staff was very helpful. I'll stay here next time.
February, 2013	tripadvisor.com	Holiday Club Hotell	Excellent resort with plenty of options for wintersports. Holiday club excellent for families as it is c and sauna complex to chill out, sports bar and reception bar rather characterless. Large general a indoor link to shopping mall. Food excellent quality and choice. Good service to meeting rooms. Ro fold out bed with curtain - great for sharing. Views from rooms better on higher floors. Storgae ar arriving my room was not available ans sent away - kinder to offer somewhere for luggage and ent

Fig. 1 Core information extracted from review sites

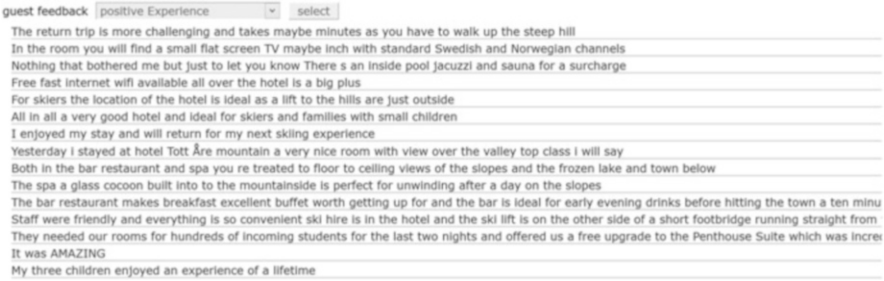
The final outcome of the sentiment analysis provides valuable information on customer reviews and opinions in a structured format. These structured data are stored in the multi-dimensional data structures of the central data warehouse (Höpken, Fuchs, Höll, Keil, & Lexhagen, 2013) and are, thus, available for powerful OLAP analyses and data mining. Figure 1 shows parts of the structured information directly extracted from review sites, namely the date of review, the review site, the hotel name and the full customer review.

As the full reviews shown in Fig. 1 are split into sentences and classified into positive or negative statements, single statements can be filtered according to their sentiment.

Figure 2 shows positive review statement for different hotels in Åre.

Figure 3 shows an OLAP analysis, calculating the average sentiment (over all single sentences) for various accommodation providers. It has to be noted that many hotels have only a few (or even none) reviews and results might, thus, not be representative for the hotel quality.

Figure 4 extends the analysis above by adding the topic as a second dimension, and enables a comparison of average sentiments across accommodation providers and topics, demonstrating powerful benchmarking capabilities.



The return trip is more challenging and takes maybe minutes as you have to walk up the steep hill
In the room you will find a small flat screen TV maybe inch with standard Swedish and Norwegian channels
Nothing that bothered me but just to let you know There s an inside pool jacuzzi and sauna for a surcharge
Free fast internet wifi available all over the hotel is a big plus
For skiers the location of the hotel is ideal as a lift to the hills are just outside
All in all a very good hotel and ideal for skiers and families with small children
I enjoyed my stay and will return for my next skiing experience
Yesterday i stayed at hotel Tott Åre mountain a very nice room with view over the valley top class i will say
Both in the bar restaurant and spa you re treated to floor to ceiling views of the slopes and the frozen lake and town below
The spa a glass cocoon built into to the mountainside is perfect for unwinding after a day on the slopes
The bar restaurant makes breakfast excellent buffet worth getting up for and the bar is ideal for early evening drinks before hitting the town a ten minu
Staff were friendly and everything is so convenient ski hire is in the hotel and the ski lift is on the other side of a short footbridge running straight from
They needed our rooms for hundreds of incoming students for the last two nights and offered us a free upgrade to the Penthouse Suite which was incre
It was AMAZING
My three children enjoyed an experience of a lifetime

Fig. 2 Positive review statements

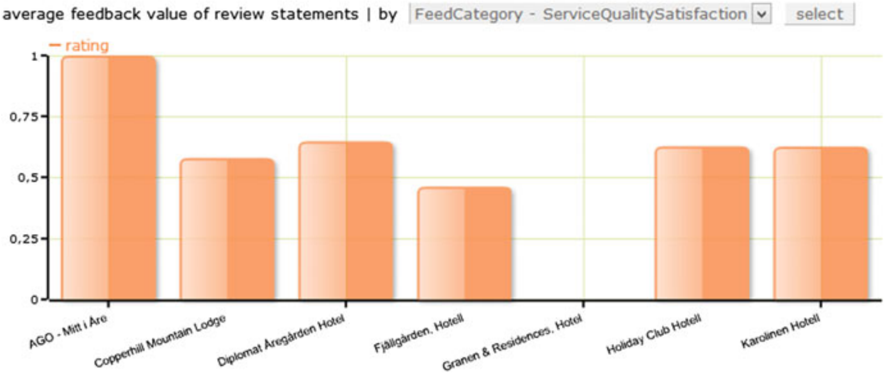


Fig. 3 Average sentiment per accommodation provider

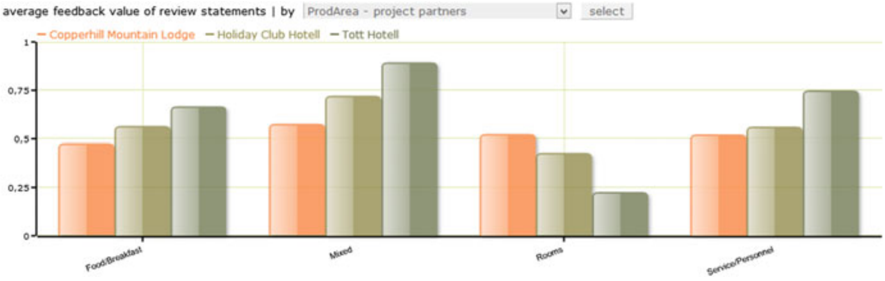


Fig. 4 Average sentiment per topic and accommodation provider

6 Conclusions

Customer online feedback in the form of user-generated content (UGC) increased significantly in recent years. Thus, it has become important for tourism stakeholders and destinations to analyze these reviews on a regular basis. Through the continuous monitoring of customer feedback, companies can gain valuable knowledge as input to product optimization and CRM activities. However, capabilities to manually analyze the huge amount of available reviews are limited. Thus, this chapter presented several different approaches for automatically extracting and analyzing customer reviews from tourism review sites. The task of sentiment analysis has been divided into topic detection, subjectivity detection, and sentiment detection. For each of these tasks, a dictionary-based approach and machine learning approaches, like k-nearest neighbor, support vector machines (SVM) and Naïve Bayes, have been presented. Additionally, for the task of unsupervised topic detection, approaches, like cluster analysis or single value decomposition (SVD), have been discussed. Additionally, optimizations with word n-grams and POS tagging were considered where appropriate.

Supervised topic detection, thus, identifying predefined topics, has been solved best by SVM with an accuracy of 72.35 %. In the case of unsupervised topic detection, cluster analysis and latent semantic indexing reached the best results in identifying manually labeled topic words (both above 88 %). For subjectivity detection, the dictionary-based approach achieved the best accuracy (80.37 %). The sentiment detection was solved best by the SVM approach, showing an accuracy of 76.80 %. Finally, a destination management information system (DMIS) for the leading Swedish mountain destination Åre has been presented, validating the discussed approaches and demonstrating the business benefits of the gained knowledge as input to decision support.

For all subtasks of sentiment analysis, appropriate approaches have been presented, reaching satisfactory results also for a managerial application, like the presented destination management information system. Nevertheless, an important improvement for the future is to apply the presented machine learning approaches to a bigger amount of training data, manually classified by several domain experts independently, as individual classification habits may differ and influence the overall quality. Another vein of future research is a topic-specific sentiment detection. It can be assumed that words representing positive or negative opinions differ depending on the topic the sentiment is about. Thus, executing the discussed machine learning approaches for each single topic separately is expected to further increase the overall accuracy.

References

Alves, A., Baptista, C., Firmino, A., de Oliveira, M., & de Paiva, A. (2014). *Comparison of SVM versus Naive-Bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA Confederations Cup*. Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, Web-Media ‘14 (pp. 123–130). New York.

Chiu, C., Chiu, N.-H., Sunga, R.-J., & Hsieh, P.-Y. (2015). Opinion mining of hotel customer-generated contents in Chinese weblogs. *Current Issues in Tourism, 18*(5), 477–495.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations—A Case from Sweden. *Journal of Destination Marketing and Management, 3*(4), ss. 198–209.

García, A., Gaines, S., & Linaza, M. (2012). A lexicon-based sentiment analysis retrieval system for tourism domain. *e-Review of Tourism Research, 10*, 35–38.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism* (pp. 460–470). Wien: Springer.

Gretzel, U., Yoo, K., & Purifoy, M. (2007). *Online travel review study—role & impact of online travel review*. Hämtat från. http://www.tripadvisor.com/pdfs/OnlineTravelReviewReport.pdf

Höpken, W., Fuchs, M., Höll, G., Keil, D., & Lexhagen, M. (2013). Multi-dimensional data modelling for a tourism destination data warehouse. In L. Cantoni & P. Xiang (Eds.), *Information and communication technologies in tourism* (pp. 157–169). New York: Springer.

Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2015). Business intelligence for cross-process knowledge extraction at tourism destination. *Information Technology & Tourism, 15*(2), 101–130.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *KDD'04*. New York: ACM.

Kasper, W., & Vela, M. (2011*). Sentiment analysis for hotel reviews, Proceedings of the Computational Linguistics-Applications Conference,* (pp. 45–52). Jachranka.

Kasper, W., & Vela, M. (2012). Monitoring and summarization of hotel reviews. In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism* (pp. 471–482). New York: Springer.

Kiran, G., Shankar, R., & Pudi, V. (2010). Frequent itemset based hierarchical document clustering using Wikipedia as external knowledge. In R. Setchi, I. Jordanov, R. Howlett, & L. Jain (Eds.), *Knowledge-based and intelligent information and engineering systems—lecture notes in computer science* (Vol. 6277, pp. 11–20). Berlin: Springer.

Lexhagen, M., Kuttainen, C., Fuchs, M., & Höpken, W. (2012). Destination talk in social media: a content analysis for innovation. In E. Christou, D. Chionis, D. Gursory, & M. Sigala (Eds.), *Advances in hospitality and tourism marketing & management*. Corfu.

Lin, C.-J., & Chao, P. (2010). Tourism-related opinion detection and tourist-attraction target identification. *Computational Linguistics and Chinese Language Processing, 15*(1), 37–60.

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents and usage data* (2nd ed.). Chicago: Springer.

Markopoulos, G., Mikros, G., Iliadi, A., & Liontos, M. (2015). Sentiment analysis of hotel reviews in Greek: A comparison of unigram features of cultural tourism in a digital era. In *Springer proceedings in business and economics 2015* (pp. 373–383). New York: Springer.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Elsevier.

Murphy, H., Gil, E., & Schegg, R. (2010). An investigation of motivation to share online content by young travelers—why and where. In U. Gretzel, R. Law, & M. Fuchs (Eds.), *Information and communication technologies in tourism* (pp. 467–478). Wien: Springer.

Pablos, A., Cuadros, M., & Linaza, M. (2015). OpeNER: Open tools to perform natural language processing on accommodation. In I. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism* (pp. 125–137). New York: Springer.

Rossetti, M., Stella, F., Cao, L., & Zanker, M. (2015). Analysing user reviews in tourism with topic models. In I. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism* (pp. 47–58). New York: Springer.

Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2014). Sentiment analysis—extracting decision-relevant knowledge from UGC. In Z. Xiang & I. Tussyadiah (Eds.), *Information and communication technologies in tourism* (pp. 253–265). Heidelberg: Springer.

Tsytsarau, M., & Palpanas, T. (2011). *Survey on mining subjective data on the web*. Trento: Springer.

Waldhör, K., & Rind, A. (2008). E-BlogAnalysis—Mining virtual communities using statistical and linguistic methods for quality control in tourism. In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), *Information and communication technologies in tourism* (pp. 453–462). New York: Springer.

Waltinger, U. (2010). *Germanpolarityclues: A lexical resource for German sentiment analysis*. Seventh International conference on Language Resources and Evaluation (LREC).

Weichselbraun, A., Gindl, S., & Scharl, A. (2013). Extracting and grounding context-aware. *Sentiment Lexicons, 28*(2), 39–46.

Xiang, Z., Schwartz, Z., & Uysal, M. (2015). What types of hotels make their guests (un-) happy? Text analytics of customer experiences in online reviews. In I. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism* (pp. 33–45). New York: Springer.

Xiang, Z., Schwartz, Z., Gerdes, J., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*(2), 120–130.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527–6535.

Estimating the Effect of Online Consumer Reviews: An Application of Count Data Models

Sangwon Park

1 Introduction

The advent of information technology has resulted in the development of a new form of web communication, known as eWOM (electronic word-of-mouth), operated by consumer participation (Tussyadiah & Fesenmaier, 2009). Online consumer reviews have become one of the vital information sources which allow people to gather sufficient and reliable information about products and services (Liu & Park, 2015). In particular, due to the characteristics of tourism products (e.g. intangibility and perishability), online reviews provide substantial benefits to current travellers, enabling them to obtain authentic and indirect consumption experiences through checking the discourse types of comments (Schuckert, Liu, & Law, 2015). In recognising the importance of online reviews in tourism and hospitality, a number of researchers have investigated the effects of consumer reviews, essentially in terms of product sales (Ye, Law, Gu, & Chen, 2011) and the decision-making process (Sparks, Perkins, & Buckley, 2013). These studies conclude that online reviews have positive influences on increasing revenues and assisting with purchase decisions.

Importantly, easily accessible online reviews facilitate consumers in finding plentiful information (low search costs); however, they also make it difficult for people to determine helpful information (high evaluation costs). Overall, the important question of 'what makes online reviews useful?' still has not been sufficiently discussed. Based on an adaptive decision-making strategy (Payne, Bettman, & Johnson, 1992), consumers are likely to focus on heuristic information cues when the size of information to be evaluated is larger than their cognitive abilities. With regard to the context of online consumer reviews, it has been

S. Park (✉)
University of Surrey, Guildford, UK
e-mail: sangwon.park@surrey.ac.uk

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_9

identified that star rating is a key element of heuristic information, which is regarded as an explanatory variable in this current research.

Therefore, this chapter will examine the relationship between star ratings and perceived usefulness and enjoyment on online reviews. In order to address the research question, over 5000 reviews were collected from Yelp (yelp.com), a well-recognised consumer review website for tourism and hospitality products. This study then employed negative binomial regression, a type of count model (Allison & Waterman, 2002). Analysing secondary data obtained with an unstructured format commonly violates the assumptions of the ordinary least square (OLS) regression, or general count models such as the Poisson regression (Hox & Boeije, 2005). For instance, there can be skewed distribution of the data, zero inflation problems, and overdispersion (where unconditional variance is larger than the mean) (Gurmu & Trivedi, 1996; Jackman, Kleiber, & Zeileis, 2007). Thus, the second aim of this chapter is to discuss count models and, in particular, provide evidence of the usability of negative binomial models in analysing the online review data.

2 Online Consumer Reviews in Tourism and Hospitality

Online travellers like to obtain detailed and up-to-date information and examine indirect experiences of tourism products in order to make a better decision on them (Xiang, Wang, O'Leary, & Fesenmaier, 2015). In this sense, online reviews developed by other consumers have relatively higher reliability and bring about more attention from other consumers. Based on the important role of online reviews in the tourism field, numerous researchers have investigated the effects of online reviews, which can essentially be classified into the three areas of product sales, the decision-making process and evaluation of the information sources (Park & Nicolau, 2015).

Following a statement that the number of consumer reviews written on the social media websites reflects product sales, previous studies have identified a positive relationship between online reviews and revenues in hotels (Xie, Chen, & Wu, 2012) and restaurants (Zhang, Ye, Law, & Li, 2010). For example, Ye et al. (2011) found that a 10 % increase in travel review ratings improves the volume of hotel bookings by more than 5 %. A study conducted by Ogut and Tas (2012) concluded that a 1 % increase in online review ratings leads to increased sales per room by about 2.6 %, depending on destinations. Reviews about the quality and service of restaurants, as well as the volume of reviews, also have positive relationships with restaurant popularity (Zhang et al., 2010). Additionally, high ratings of online reviews tend to generate price premiums (Yacouel & Fleischer, 2012; Zhang, Ye, & Law, 2011). Online reviews, potentially representing service quality, lead consumers to have increased confidence in their decisions. This increase in trustworthiness encourages travellers to pay higher prices when purchasing tourism products.

With regard to the online buying process, Leung, Law, van Hoof, and Buhalis (2013) suggested online consumer contents essentially affect entire phases of the travel planning process, including pre-, during- and post-consumption. Specifically, positive reviews with numerical ratings improve attitudes toward travel products, being associated with the formation of consideration sets (Vermeulen & Seegers, 2009) and purchasing intentions (Sparks & Browning, 2011). Filieri and McLeay (2014) attempted to identify the factors that bring about the adoption of online information by consumers with regard to the elaboration likelihood theory, including the central route (e.g. information accuracy, value-added information, information relevance, information timeliness) and the peripheral route (e.g. product ranking).

Interestingly, several tourism and hospitality researchers have explored travellers' responses to online reviews concerning the trustworthiness, helpfulness and usefulness of the reviews (Racherla & Friske, 2012; Wei, Miao, & Huang, 2013). It has been recognised in this research that positive reviews are likely to be more favourable than negative comments, and heuristic cues of online reviews lead readers to enlarge the perceived helpfulness of the reviews. A recent research by Liu and Park (2015) concluded that the messenger characteristics (e.g. disclosed photo, reviewers' expertise) and message characteristics (number of words, star ratings readability) of the online reviews affect the perceived usefulness of online reviews. When reviewing the literature of online reviews, it was noted that many studies have used a survey method or experimental design approach to estimate the effect of online comments on consumer behaviours (Schuckert et al., 2015). Importantly, however, this study uses data reflecting *actual* user behaviours collected from a real tourism review website. Thus, it is suggested that an alternative method of count models—the negative binomial model—better addresses the research question, as discussed in the following section.

3 Count Models

Count models deal with specific types of data, which are discrete, using a non-negative integer (e.g. 0, 1, 2 …), which stand for counts rather than rankings. In other words, they represent the number of occurrences of an event within a fixed period. Count models aim to identify factors influencing the average number of occurrences of an event. Since count data is distinct from binary data consisting of two values ('0' or '1'), alternative estimations have been suggested for use, such as the Poisson and negative binomial models (Castéran & Roederer, 2013; Czajkowski, Giergiczny, Kronenberg, & Tryjanowski, 2014; Hellerstein & Mendelsohn, 1993). While the linear least square regression coping with continuous variables is applicable, the estimated results can be inefficient, inconsistent and biased (Cameron & Trivedi, 2013). This is because the response variable is categorical or discrete, which often produces skewed distribution of residential errors, as well as making an ineffective approach of a simple transformation.

3.1 Poisson Estimation

The Poisson model is useful when the outcome is count with which the large count becomes rare occurrences (Kutner, Nachtsheim, Neter, & Li, 2004). The Poisson function predicts the number of occurrences of events ($Y = 0, 1, 2 \ldots$) during an interval of time. The Poisson distribution can be expressed as follows:

$$p(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

where Y refers to a Poisson distribution with parameter (or intensity) μ

Therefore it can be said that $\mu = \exp\ (\chi'_{i}\beta)$.

Importantly, one of properties of the Poisson estimation is the equality of mean and variance for $\mu > 0$, known as equidispersion (Cameron & Trivedi, 2013).

$$\mathrm{E}(y|\chi) = var(\mathrm{y}|\chi) = \mu$$

Since the mean is equal to the variance, any factor affecting one element of the equation will simultaneously influence the other.

While the Poisson model is nonlinear, the maximum likelihood estimation facilitates evaluation of the model as a typical count model. Due to the computational convenience of the estimation, a number of researchers in tourism and hospitality have used the Poisson model to understand travel behaviours, including length of stay (Alegre, Mateo, & Pou, 2011), visit frequency to a destination (Castéran & Roederer, 2013) and museums (Bridaa, Meleddub, & Pulinac, 2012), and travel cost analysis (Chae, Wattage, & Pascoe, 2012). However, there is an important limitation in the Poisson model, which may bring about biased and incorrect estimated results (Gurmu & Trivedi, 1996; Zeileis, Kleiber, & Jackman, 2008), denoting overdispersion. The assumption of the Poisson model is the equality of mean and variance. In the context of count data, the conditional variance frequently exceeds the mean. It refers to overdispersion relative to the Poisson model. When the conditional variance is less than the mean, it represents underdispersion. These two cases of over- and underdispersion inhibit the suitability of the Poisson model, resulting from unobserved heterogeneity. In order to manage the restrictions of the Poisson model, this study uses an alternative count model, the negative binomial model, as a type of generalized linear model (Cameron & Trivedi, 2013).

3.2 Negative Binomial Estimation

The negative binomial model is a form of Poisson regression that contains a random component considering the uncertainty about the true values at which events occur for individual cases (Gardner, Mulvey, & Shaw, 1995). In other words, this model addresses the issue of overdispersion by including a dispersion parameter to accommodate the unobserved heterogeneity in the count data. The additional parameter allows the variance to exceed the mean. Hence, the negative binomial estimator can manage 'incidental parameter' bias, and is generally superior to the Poisson estimator (Allison & Waterman, 2002). The negative binomial model can be written as:

$$P(y_t) = \frac{\Gamma(\alpha^{-1} + y_t)}{\Gamma(\alpha^{-1})\Gamma(y_t + 1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + e^{\sum_{k=1}^{K} \beta_k x_{tk}}} \right)^{\alpha^{-1}} \left(\frac{e^{\sum_{k=1}^{K} \beta_k x_{tk}}}{\alpha^{-1} + e^{\sum_{k=1}^{K} \beta_k x_{tk}}} \right)^{y_t} \forall y_t$$

$$= \{0, 1, 2, \ldots\}$$

Where Γ represents the gamma function, x_{tk} the characteristic k of online review t and β_k the parameter which indicates the effect of x_{tk} on *P*(y_t).

The parameter α covers the dispersion of the observations, in such a way that

$$E(y_t) = e^{\sum_{k=1}^{K} \beta_k x_{tk}} = \lambda_t$$

and

$$V(y_t) = e^{\sum_{k=1}^{K} \beta_k x_{tk}} + \alpha \cdot e^{2\sum_{k=1}^{K} \beta_k x_{tk}} = \lambda_t + \alpha \cdot \lambda_t^2$$

One way of verifying the validity of the negative binomial model against the Poisson model is to test the null hypothesis $\alpha = 0$. Note that its acceptance would imply that *E*(y_t) = *V*(y_t), so that the Poisson model is a particular case of the negative binomial when $\alpha = 0$ (Gurmu & Trivedi, 1996).

Due to the benefits of the negative binomial model in managing the restriction of the Poisson model, several tourism scholars have used the estimation in order to understand self-drive trips using the contingency behaviour model (Mahadevan, 2014) to calculate the number of days cars are hired for (Palmer-Tous, Riera-Font, & Rosselló-Nadal, 2007); the length of stays for senior tourists (Alén, Nicolau, Losada, & Domínguez, 2014) and youth travellers (Thrane, 2016); numbers of visitations to a destination (Czajkowski et al., 2014); and number of hotel rooms rented (Yang & Cai, 2016). Thus, this research assesses the appropriateness of models between the Poisson and negative binomial models in understanding the features of the data distribution. Then the effect of online star ratings on information evaluations in terms of perceived consumer usefulness and enjoyment is discussed.

4 Methods

This research collected data on online consumer reviews from Yelp, which constitutes the majority of consumer feedback on restaurants and is regarded as an important travel activity (Park & Fesenmaier, 2014).[1] Consumer reviews were collected relating to restaurants located in two main tourism destinations: London and New York. This approach allowed the researcher to reduce the potential of confounding effects on the estimations with regard to a specific feature of a destination. Other than controlling the location of the restaurants, the researcher took into account the prices and brand familiarity of the restaurants which may affect information search and evaluation (Gursoy & McCleary, 2004). The restaurants were selected according to the classification of price groups and excluding national and local chains. Racherla and Friske (2012) found that a restaurant's position on the website has an influence on users' perception as more attention is drawn to businesses listed in the top places among the reviews. Thus, this study used the collection process in a random manner instead of selecting them in either rankings or alphabetical order. As a result, 45 restaurants in London with 2500 reviews and 10 restaurants in New York with 2590 reviews were chosen for data analysis.

4.1 Model Estimations

This study applied a method to assess the effect of heuristic online reviews (particularly star ratings) on the usefulness of the reviews and the enjoyment of the consumer. The data reflecting the number of votes awarded to individual reviews included features of count data which are nonnegative and occur in integer quantities. According to the integral nature of online review votes, the estimated results using continuous models (e.g., linear regression) that restricts managing censoring (e.g. zeros) brings about biased estimations. Thus, this research used count data models (Hellerstein & Mendelsohn, 1993).

The most well-known approximation is derived from the Poisson distribution $P(\lambda)$, where λ is the average of the random variable, which, in this research, is the number of 'useful' or 'enjoyment' votes awarded to the review in a certain period of

[1] The study uses the same data set as Park and Nicolau's (2015) paper published in the *Annals of Tourism Research*. Detailed descriptions of the data collection and measurements can be found in the article.

time. As discussed above, however, the Poisson model is developed based on the assumption of average-variance equality. It is too restrictive to represent individual behaviours, as it is not able to cope with the heterogeneity of these individuals and creates what is known as the ‘problem of overdispersion’ (Gurmu & Trivedi, 1996). Hence, in order to address the restrictions of the Poisson modelling, this study applied an alternative count model based on a negative binomial distribution (Cameron & Trivedi, 2013).

One way of verifying the validity of the negative binomial model as opposed to the Poisson model is testing the null hypothesis (i.e. dispersion parameter $= 0$ denoting α at the equation discussed in the literature review), reflecting equality of mean and variance $E(y_t) = V(y_t)$. When this hypothesis is rejected (i.e. $\alpha \neq 0$), it can be said that the negative binomial is a more appropriate approach than the Poisson model as it addresses the overdispersion problem (Gurmu & Trivedi, 1996). Furthermore, this approximation copes with the bias problems of regression analysis arising from the discrete character of the dependent variable (Hellerstein & Mendelsohn, 1993).

4.2 Measurement

This research assessed an independent variable—star ratings—that indicates the perceived quality of products and services using five star levels (Chevalier & Mayzlin, 2006; Mudambi & Schuff, 2010; Racherla & Friske, 2012). Given the raw data of the star rating variable, a series of data manipulations were applied. Firstly the data was divided into two categorized variables (i.e. positive and negative reviews) with positive reviews consisting of four and five stars and negative reviews consisting of one and two stars; secondly dummies were given for each star rating. This approach enabled the researcher to investigate the relative influences of reviews on two types of consumer responses (i.e. perceived usefulness and enjoyment) with the medium rating (‘3’) as a reference group. Additionally, these three alternative ways to approach the inclusion of the star rating variable into the model allowed for the identification of the intricacies of different particular effects, as well as confirming robustness in cases where the scores of this variable are highly skewed (mean: 4.28; standard deviation: 0.88). Therefore, examining the variable itself could lead to misleading results, as the mean value could not reflect the whole range of its effect.

There are two dependent variables measured by counting the number of online users who voted that the reviews were useful or pleasurable (Ghose & Ipeirotis, 2011; Van der Heijden, 2004). This research then considered a number of control variables, including identity disclosure (the presence of real names and photos) (Forman, Ghose, & Wiesenfeld, 2008), level of reviewer expertise (the number of previous reviews written by a reviewer) (Chen, Dhanasobhon, & Smith, 2008) and reputation (the number of times that each reviewer achieved the ‘elite’ title) (Gruen, Osmonbekov, & Czaplewski, 2006), review elaborateness (the number of words in each review content) (Shelat & Egger, 2002), and readability[2] (Korfiatis, Garcia-Bariocanal, & Sanchez-Alonso, 2012). These control variables were decided based on the findings of previous studies arguing that the characteristics of messengers and messages affect the perceived evaluations of online consumer reviews. Additionally, the location of the restaurants were added as another control variable so as to test the potential confounding effect on the results (1 = London and 0 = New York).

5 Results

Table 1 presents the results of a linear regression with normally distributed errors. The variables estimated explain 16 % for usefulness and 15 % for enjoyment. In both models, the variable of star rating shows negative relationships while the squared term of star ratings have positive influences on the outcomes. This model, however, is problematic: the main issue is that the data violates the assumption that the variances of the residuals are the same for the original response variable in the regression model (Fox, 1984). To evaluate this property, an approach to testing heteroscedasticity using the White method (Cameron & Trivedi, 2013) was employed. It was identified that the model possesses heteroscedasticity, which potentially results in misrepresenting the estimated variances of the coefficients compared with relevant true variances. Considering count data in which the absolute values of the residuals generally correlate with the explanatory variables, the estimated standard errors of the coefficients are likely to be smaller than their true values (Gardner et al., 1995). The *t*-test results corresponding to the coefficient estimations can be inflated accordingly.

A conventional alternative to responding to heteroscedasticity is transforming the data in order to remove the correlation between the expected counts and residuals. However, the simple transformation approach would not be able to cope with the features of count data generally including many ‘zeros’ (King, 1988). More importantly, the counting numbers are the natural and meaningful values as counts, and thus, the analysis should retain these merits. Therefore, it can be suggested to use certain models dealing with count data.

[2] Readability was examined by automated readability index (ARI) (Zakaluk & Samuels, 1988). This index takes into account the number of words and characters to evaluate the comprehensibility of a text. The estimated value of ARI indicates the educational level required to understand the textual information.

Table 1 The results of OLS regression

	LR[1] Usefulness	LR[1] Enjoyment
Star ratings	-1.642^{***} (0.229)	-0.561^{***} (0.176)
Squared star ratings	0.232^{***} (0.229)	0.100^{***} (0.023)
Exposure name	−0.015 (0.164)	0.047 (0.126)
Exposure photo	0.268^{***} (0.081)	0.168^{***} (0.062)
Reviewer's expertise	0.002^{***} (0.001)	0.001^{***} (0.001)
Reviewer's reputation	0.097^{***} (0.020)	0.097^{***} (0.020)
Information elaborateness	0.155^{***} (0.136)	0.003^{***} (0.001)
Readability (ARI)	0.014 (0.009)	0.001 (0.001)
Location	−0.028 (0.068)	0.008 (0.052)
Constant	2.457*** (0.442)	0.316^{***} (0.341)
R-squared	0.160	0.152
Adjusted R-squared	0.159	0.150
Log likelihood	−11606.26	−10274.5
AIC	4.566	4.043
SIC	4.578	4.056

Note: 1 refers to linear regression
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; numbers in parenthesis refer to standard errors

5.1 *Analysis of Count Models*

The Poisson regression is a more reasonable model to analyse count data than the linear regression model. First, the nature of counts include nonnegative numbers. The Poisson distribution allocates probabilities only to the nonnegative integers of the outcome variable. Second, the variance of the dependent variable increases as a function of mean, referring to equidispersion. Thus, it can be said that the Poisson model has greater validity than the linear regression model (Gardner et al., 1995).

Checking the goodness of fit between models such as LL (log-likelihood), AIC (Akaike information criterion) and SIC (Schwarz criterion or Bayesian information criterion), all of the values for the Poisson model (see Table 3); LL = −8513.1 for PI U and −6480.4 for PI E, AIC = 2.799 and 2.551, and SIC = 2.813 and 2.565) are better than for linear regression (see Table 1); LL = −11606.26 and −10274.5, AIC = 4.566 and 4.043, and SIC = 4.578 and 4.056 for usefulness and enjoyment in linear regression, respectively).

Table 2 The summary of dependent variables

	Observations	Mean	Variance	Min.	Max.
Usefulness	5090	1.22	6.68	0	65
Enjoyment	5089	0.76	3.92	0	55

It is, however, important to consider a critical limitation of the Poisson model, such as over- or underdispersion. When comparing the unconditional mean and variance of the dependent variables (see Table 2), the results do not show equidispersion. That is, the unconditional variances of the outcome variables are much higher than their mean values (variance = 6.68 and 3.92; mean = 1.22 and 0.76 for usefulness and enjoyment respectively). This result provides an indication of an overdispersion problem.

Following the initial assessment, the researcher tested the overdispersion parameter α by applying the negative binomial model. As shown in Table 3, particularly for the models of NB U1 and NB E1, the parameter α is larger than 0 and statistically significant ($p < 0.001$). Furthermore, the models including categorical variables of star ratings (e.g. NB U2, U3, E2 and E3) consistently show the invalidation of the property of mean-variance equality of the Poisson models (Cameron & Trivedi, 1998). This implies the existence of heterogeneity of travel behaviours, which in turn suggests the adoption of a model that manages the variations in order to avoid possible biases in the estimations (Gurmu & Trivedi, 1996). Furthermore, the goodness of fit indexes including AIC and SIC are compared with the Poisson and negative binomial models. It can be confirmed that the indicators related to the negative binomial model are better than the ones associated with the Poisson model. In terms of the explanatory power of the model, statistical evidence including significant likelihood ratio, LR index over 30 % and R-square over 15 % supports the acceptable ability of the negative binomial models to assess the proposed relationships (Hensher & Johnson, 1981; Train, 2009) (see Table 3). Thus, this research uses the negative binomial model as a main data analysis.

5.2 *Assessing the Effect of Star Ratings on Review Evaluations*

The variables of star ratings show a negative linear relationship and a positive curvilinear (U-shaped) relationship with both usefulness ($b = -1.134$ & 0.161, $p < 0.001$) and enjoyment ($b = -0.497$ & 0.100, $p < 0.01$) (see Table 3). The models containing two categorical variables (i.e. positive and negative ratings with a neutral value as a reference) were analysed in order to estimate the relative influences with directional online reviews (see NB U2 and NB E2). Interestingly, only negative reviews are significant in explaining usefulness (NB U2; $b = 0.400$, $p < 0.001$) whereas, in the case of enjoyment, the positive reviews were positively significant (NB E2; $b = 0.474$, $p < 0.001$). This finding implies that online travellers

Table 3 The results of poisson and negative binomial models

		Usefulness				Enjoyment		
	PI[1] U	NB[2] U[3]1	NB U2	NB U3	PI E[4]	NB E1	NB E2	NB[2] E3
Star ratings	−1.277*** (0.075)	−1.134*** (0.140)			−0.780*** (0.123)	−0.497** (0.196)		
Squared star ratings	0.180*** (0.010)	0.161*** (0.019)			0.134*** (0.016)	0.100*** (0.026)		
Positive reviews (4 & 5)			0.081 (0.071)				0.474*** (0.095)	
Negative reviews (1 & 2)			0.400*** (0.111)				0.126 (0.154)	
Positive review (5)				0.225*** (0.073)				0.635*** (0.097)
Positive review (4)				−0.146† (0.076)				0.228* (0.100)
Negative review (2)				0.273* (0.123)				0.167 (0.166)
Negative review (1)				0.733*** (0.178)				0.020 (0.285)
Real name	0.116 (0.081)	0.125 (0.116)	0.095 (0.115)	0.114 (0.116)	0.305 (0.126)	0.258 (0.160)	0.269† (0.160)	0.254 (0.160)
Real photo	0.379*** (0.038)	0.348*** (0.054)	0.350*** (0.054)	0.351*** (0.054)	0.482*** (0.052)	0.480*** (0.070)	0.481*** (0.070)	0.480*** (0.070)
Reviewer’s expertise	0.358*** (0.003)	0.302*** (0.0722)	0.300*** (0.073)	0.316*** (0.073)	0.390*** (0.030)	0.363*** (0.088)	0.355*** (0.089)	0.370*** (0.088)
Reviewer’s reputation	0.113*** (0.008)	0.127*** (0.014)	0.121*** (0.015)	0.126*** (0.014)	0.168*** (0.009)	0.186*** (0.017)	0.181*** (0.017)	0.183*** (0.017)
Review elaborateness	0.003*** (0.001)	0.003*** (0.001)	0.003*** (0.001)	0.003*** (0.001)	0.002*** (0.001)	0.003*** (0.001)	0.003*** (0.001)	0.003*** (0.001)

(continued)

Table 3 (continued)

		Usefulness				Enjoyment		
	PI^1 U	NB^2 U^{3}1	NB U2	NB U3	PI E^4	NB E1	NB E2	NB^2 E3
Readability (ARI)	0.015*** (0.003)	0.012* (0.005)	0.012* (0.001)	0.012* (0.005)	0.004*** (0.004)	0.001 (0.007)	0.001 (0.007)	0.002 (0.007)
Location	0.010 (0.026)	0.081 (0.043)	0.053 (0.043)	0.083 (0.043)	0.048 (0.033)	0.131* (0.054)	0.096 (0.054)	0.134* (0.054)
Constant	0.950*** (0.145)	0.630* (0.266)	−1.181*** (0.136)	−1.217*** (0.136)	−1.089*** (0.252)	−1.741*** (0.387)	−2.348*** (0.1856)	−2.362*** (0.187)
α		0.155*** (0.043)	0.191*** (0.0422)	0.152*** (0.042)		0.521*** (0.050)	0.555*** (0.049)	0.518*** (0.050)
R-squared	0.224	0.214	0.196	0.216	0.190	0.173	0.153	0.175
LR Index	0.162	0.300	0.297	0.301	0.175	0.316	0.313	0.315
LR statistic	3294.8***	6103.4***	6040.9***	6108.0***	2750.5***	4960.4***	4913.5***	4962.1***
Log likelihood	−8513.1	−7108.8	−7140.1	−7106.5	−6480.4	−5375.4	−5398.9	−5374.6
AIC	3.350	2.799	2.811	2.799	2.551	2.118	2.127	2.118
SIC	3.363	2.813	2.825	2.815	2.565	2.132	2.141	2.135

Note: 1 = Poisson model; 2 = Negative Binomial model; 3 = usefulness; 4 = enjoyment
*p < 0.05; **P < 0.01; ***p < 0.001

are more likely to read either positive or negative reviews that enhance the completeness of information, rather than balanced ratings (Cheung, Luo, Sia, & Chen, 2009).

As a way to unravel the asymmetric effects of star ratings on different consumer responses, a more sophisticated analysis composed of binary variables that represent individual star ratings was conducted (see NB U3 and NB E3). In the model estimating usefulness, given middle point as a reference, all variables of each star rating except for 'positive review (4)' are statistically significant at p-value below 5 %. When comparing the relative coefficient values (see NB U3), it was identified that the negative reviews ($b = 0.733$ for rating 1 and 0.273 for rating 2, $p < 0.05$) have higher impacts on review usefulness than positive reviews ($b = 0.225$ for rating 5, $p < 0.001$) (Chevalier & Mayzlin, 2006). Corresponding to NB E2, the findings of NB E3 present the significant effects of positive reviews on enjoyment ($b = 0.635$ for rating 5 and 0.228 for rating 4, $p < 0.05$), but an insignificant result with negative reviews ($b = 0.167$ for rating 2 and 0.273 for rating 1, $p > 0.05$) (Fischer, Schulz-Hardt, & Frey, 2008).

For the control variables, the potential effect of the locations of restaurants (London and New York) was tested with outcome variables (usefulness and enjoyment). Based on the consistent results across OLS regression, the Poisson and the negative binomial models, it is apparent that the variances of dependent variables explained by the different locations are limited. The disclosure of reviewers' information (e.g. photo) and the features of reviewers (e.g. expertise, reputation), as well as the characteristics of the message (e.g. elaborateness), have positive influences on usefulness and enjoyment. Interestingly, review readability seems to be just significant in the aspect of usefulness.

6 Conclusions

Online reviews have become an important and reliable information source to current travellers, which enable them to evaluate the quality of products/services and to have indirect experiences (Liu & Park, 2015). Within the e-WOM strategy, review ratings represent an attempt to quantify service quality perceptions, which is one of the important information elements used by consumers in making a purchasing decision (Ye, Li, Wang, & Law, 2014). This chapter examined potential asymmetries in the effect of online reviews on usefulness and enjoyment, and suggested the use of the negative binomial model as an appropriate method to cope with count data. It was identified that online consumers perceive extreme ratings (positive or negative) as more useful and enjoyable than moderate ratings, illustrating a U-shaped relationship. More specifically, while negative reviews are more useful than positive ones, positive reviews are associated with higher enjoyment. The findings in which the ability to view a real photo, higher levels of reviewer's expertise and reputation, and the review's elaborateness and readability have positive influences on usefulness and/or enjoyment provide important implications. The location of the restaurants has restricted influence on the results, which evidence a limited confounding effect on the estimation.

While there are a number of studies that assess the effect of online reviews on both consumer purchasing behaviours and product sales, the way to address a crucial question of what makes online reviews useful and enjoyable has been restricted. Along with the theory of information diagnosticity, which refers to the extent to which a consumer believes the product information is helpful to understand and evaluate purchase alternatives (Filieri, 2015), online consumers pay greater attention to directional reviews (i.e. positive and negative ratings) to understand the expected advantages and disadvantages derived from the consumption of the product/service.

Specifically, online consumers tend to focus on negative reviews in order to increase the utility of their decisions by reducing the risk of loss (Kahneman & Tversky, 1979). This strongly supports the notion of negativity bias, arguing that rational consumers recognise the purchasing bias, and they compensate for this bias by considering negative reviews more seriously than positive reviews (Hu, Pavlou, & Zhang, 2007). From the enjoyment aspect, the characteristics of tourism products, which refer to experiential (or hedonic) products, suggest that consumers tend to take into account the elements of excitement and pleasure when searching for travel information (Vogt & Fesenmaier, 1998). This could explain the findings of a higher influence of positive reviews on inducing perceived enjoyment than negative reviews. Thus, this chapter elucidated the asymmetric effects of online review as an important information cue on different aspects of information evaluation.

Using secondary data collected from a website with an unstructured format frequently invalidates the properties of using OLS regression or general count models due to non-normal distribution of data (Hox & Boeije, 2005). In particular, considering count data that is discrete, and nonnegative integers, it is important to adopt an alternative method that is suitable for managing the specific features of data (i.e. overdispersion). In this vein, this chapter used the negative binomial model, which allows for addressing those restrictions. Specifically, this research presents a set of procedures to test the appropriateness of the model, including descriptive and analytical estimations, so as to verify the existence of heterogeneity of tourist preferences. Accordingly, it is identified that the negative binomial model not only shows better goodness of fit for the estimated models, but also brings about higher R-square values than the OLS regression and the Poisson model. Thus, the findings obtained from the negative binomial model can avoid possible biases in the estimations.

References

Alegre, J., Mateo, S., & Pou, L. (2011). A latent class approach to tourists' length of stay. *Tourism Management, 32*(3), 555–563.

Alén, E., Nicolau, J. L., Losada, N., & Domínguez, T. (2014). Determinant factors of senior tourists' length of stay. *Annals of Tourism Research, 49*, 19–32.

Allison, P. D., & Waterman, R. P. (2002). Fixed–effects negative binomial regression models. *Sociological Methodology, 32*(1), 247–265.

Bridaa, J. G., Meleddub, M., & Pulinac, M. (2012). Understanding urban tourism attractiveness: The case of the Archaeological Ötzi Museum in Bolzano. *Journal of Travel Research, 51*(6), 730–741.

Castéran, H., & Roederer, C. (2013). Does authenticity really affect behavior? The case of the Strasbourg Christmas Market. *Tourism Management, 36*, 153–163.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge: Cambridge University Press.

Chae, D. R., Wattage, P., & Pascoe, S. (2012). Recreational benefits from a marine protected area: A travel cost analysis of Lundy. *Tourism Management, 33*(4), 971–977.

Chen, P., Dhanasobhon, S., & Smith, M. (2008). *All reviews are not created equal: The disaggregate impact of reviews on sales on Amazon.com*. Working paper, Carnegie Mellon University. Available at SSRN: http://ssrn.com/abstract=918083

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (2009). Credibility of electronic word-of- mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce, 13*(4), 9–38.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book Reviews. *Journal of Marketing Research, 43*, 345–354.

Czajkowski, M., Giergiczny, M., Kronenberg, J., & Tryjanowski, P. (2014). The economic recreational value of a white stork nesting colony: A case of 'stork village' in Poland. *Tourism Management, 40*, 352–360.

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research, 68* (6), 1261–1270.

Filieri, R., & McLeay, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. *Journal of Travel Research, 53*, 44–57.

Fischer, P., Schulz-Hardt, S., & Frey, D. (2008). Selective exposure and information quantity: How different information quantities moderate decision makers' preference for consistent and inconsistent information. *Journal of Personality and Social Psychology, 94*(2), 231–244.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*(3), 291–313.

Fox, J. (1984). *Linear statistical models and related methods*. New York: Wiley.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson and negative binomial models. *Quantitative Methods in Psychology, 118*(3), 392–404.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge Data Engineering, 23*(10), 1498–1512.

Gruen, T., Osmonbekov, T., & Czaplewski, A. (2006). EWOM: The impact of customer-to customer online know-how exchange on customer value and loyalty. *Journal of Business Research, 59*(4), 449–456.

Gurmu, S., & Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *American Statistical Association, 14*(4), 469–477.

Gursoy, D., & McCleary, K. W. (2004). An integrative model of tourists' information search behaviour. *Annals of Tourism Research, 31*(2), 353–373.

Hellerstein, D., & Mendelsohn, R. (1993). A theoretical foundation for count data models. *American Journal of Agricultural Economics, 75*(3), 604–611.

Hensher, D. A., & Johnson, L. W. (1981). *Applied discrete-choice modelling*. New York: Wiley.

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary vs. secondary. *Encyclopedia of Social Measurement, 1*, 593–599.

Hu, N., Pavlou, P. A., & Zhang, J. J. (2007, March 1). *Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication. Overcoming Biases in Online Word-of-Mouth Communication.*

Jackman, S., Kleiber, C., & Zeileis, A. (2007). *Regression models for count data in R* (No. 2007/24).

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science, 32*(3), 838–863.

Korfiatis, N., Garcia-Bariocanal, E., & Sanchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of reviews helpfulness vs. review content. *Electronic Commerce Research and Applications, 11*(3), 205–217.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill/Irwin.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*, 3–22.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management, 47*, 140–151.

Mahadevan, R. (2014). Understanding senior self-drive tourism in Australia using a contingency behavior model. *Journal of Travel Research, 53*(2), 252–259.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185–200.

Ogut, H., & Tas, B. K. O. (2012). The influence of internet customer reviews on online sales and prices in hotel industry. *The Service Industries Journal, 32*(2), 197–214.

Palmer-Tous, T., Riera-Font, A., & Rosselló-Nadal, J. (2007). Taxing tourism: The case of rental cars in Mallorca. *Tourism Management, 28*(1), 271–279.

Park, S., & Fesenmaier, D. R. (2014). Travel decision flexibility. *Tourism Analysis, 19*(1), 35–49.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research, 50*, 67–83.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology, 43*(1), 87–131.

Racherla, P., & Friske, W. (2012). Perceived usefulness of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications, 11*(6), 548–559.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing, 32*(5), 608–621.

Shelat, B., & Egger, F. (2002). *What makes people trust online gambling sites?* Conference on Human factors in computing systems (pp. 852–853).

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323.

Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behaviour. *Tourism Management, 39*, 1–9.

Thrane, C. (2016). Students' summer tourism: Determinants of length of stay. *Tourism Management, 54*, 178–184.

Train, K. E. (2009). *Discrete choice methods with simulation*. New York: Cambridge University Press.

Tussyadiah, I. P., & Fesenmaier, D. R. (2009). Mediating tourists experiences-access to places via shared videos. *Annals of Tourism Research, 36*(1), 24–40.

Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly, 28* (4), 695–704.

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management, 30*(1), 123–127.

Vogt, C. A., & Fesenmaier, D. R. (1998). Expanding the functional information search model. *Annals of Tourism Research, 25*(3), 551–578.

Wei, W., Miao, L., & Huang, Z. (2013). Customer engagement behaviors and hotel responses. *International Journal of Hospitality Management, 33*, 316–330.

Xiang, Z., Wang, D., O'Leary, J. T., & Fesenmaier, D. R. (2015). Adapting to the internet: trends in travelers' use of the web for trip planning. *Journal of Travel Research, 54*(4), 511–527.

Xie, K., Chen, C-C., & Wu, S-Y. (2012). *Leveraging the ranking power of hotels by consumer reviews: Evidence from TripAdvisor.com*. 18th Annual graduate conference proceedings, Washington State University.

Yacouel, N., & Fleischer, A. (2012). The role of cybermediaries in reputation building and price premiums in the online hotel market. *Journal of Travel Research, 51*(2), 219-.

Yang, Z., & Cai, J. (2016). Do regional factors matter? Determinants of hotel industry performance in China. *Tourism Management, 52*, 242–253.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 27*, 634–639.

Ye, Q., Li, H., Wang, Z., & Law, R. (2014). the influence of hotel price on perceived service quality and value in E-tourism an empirical investigation based on online traveler reviews. *Journal of Hospitality & Tourism Research, 38*(1), 23–39.

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability: Its past, present, and future*. Newark: International Reading Association.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software, 27*(8), 1–25.

Zhang, Z., Ye, Q., & Law, R. (2011). Determinants of hotel room price: An exploration of travelers' hierarchy of accommodation needs. *International Journal of Contemporary Hospitality Management, 23*(7), 972–981.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management, 29*(4), 694–670.

Tourism Intelligence and Visual Media Analytics for Destination Management Organizations

Arno Scharl, Lidjia Lalicic, and Irem Önder

1 Introduction

Media coverage is proven to influence international tourism flows (Sealy & Wickens, 2010). Various studies demonstrate the exposure of advertising, media and emerging stories as an influence on tourist behavior, and on attitudes toward a destination. According to Govers, Go, and Kumar (2007), tourists' cultivated images are considered as the first of the travel decision-making process and therefore play a significant role. Unfortunately, the destination itself is often pushed to the background when media cover a breaking event, which leaves only a glance of positive tourism assets (Govers et al., 2007; Sealy & Wickens, 2010; Sönmez & Sirakaya, 2002). As a result, destination managers started to engage in media relations in order to influence how tourists perceive their destination. Furthermore, they started to change their approaches to their branding design. In fact, Destination Management Organizations (DMOs) realize that emotional-based experiences increase tourist satisfaction, as compared to function-oriented approaches (Ekinci, Sirakaya-Turk, & Baloglu, 2007).

According to Aaker (1997) human traits as a part of a branding approach are beneficial for consumers to identify with a brand. Aaker's brand personality scale (BPS) outlines a path to increase a destination's competitive advantages. Significant positive effects between BPS, tourist satisfaction and behavioral intentions are demonstrated (Chen & Phou, 2013; Seljeseth & Korneliussen, 2015). The integration of new branding approaches is necessary for DMOs to remain a stable position in the tourism market. Furthermore, given the large amount of data published through media, DMOs are forced to use new approaches to monitor destination images. Interestingly, only a few studies have attempted to analyze media coverage

A. Scharl (✉) • L. Lalicic • I. Önder
MODUL University Vienna, Vienna, Austria
e-mail: scharl@modul.ac.at

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_10

of tourist destinations as a proxy to all potential tourists' information (i.e., Stepchenkova & Eales, 2011). In other fields such as sports, climate change and politics, media monitoring systems have been design to analyze media streams. However, in tourism such systems are still scarce.

Destinations have to find new ways to leverage big data technologies by monitoring real-time content streams from online media, and incorporate the extracted knowledge into their workflow and decision making processes. This chapter presents a Web intelligence application that addresses this challenge, capturing online media coverage of a tourist destination related to Aakers' brand personality dimensions. The examples in this chapter stem from the webLyzard platform (www.weblyzard.com), which includes a visual dashboard that supports different types of information seeking behavior such as browsing, search, trend monitoring and visual analytics. The dashboard uses real-time synchronization mechanism that helps to analyze and organize the extracted knowledge from published news media, and to navigate the information space along multiple dimensions. It makes use of trend charts and map projections in order to show how often and where relevant information is published, and to provide a real-time account of concepts that stakeholders associate with a topic. Furthermore, the paper supports marketers to approach their branding campaigns from an innovative approach integrating a more emotional-based approach.

2 Media Coverage and Destination Image

The importance of destination image in media is due to its influence on three stakeholder groups: (1) the general public, (2) decision-makers and tourism stakeholders on a national level, and (3) the inhabitants of the destination (Avraham, 2000). The general public is affected by media coverage on issues such as tourism, migrations and investments. For decision-makers it influences decisions regarding revenue grants, capital and resource allocation. Lastly, for the inhabitants it affects the self-image of the inhabitants and their relationships with other destinations' inhabitants (Avraham, 2000).

As a result destination image has been a popular research topic. As Gunn (1972) state, many images are formed before DMOs begin their work. According to Baloglu and McCleary (1999), destinations compete by destination held in (potential) tourists' minds. Bigne et al. (2001) refer to a destination image any idea, belief, feeling or attitude tourists associate with a place evoked by the destination. Beerli and Martin (2004) describe an destination image as an accumulation of consumers' perceptions that result from consumers' decoding, extracting and interpreting the brand signals and associations (Beerli & Martin, 2004). This also implies that a destination image is dominantly based on subjective knowledge which is mediated through information channels, projected image managed by the DMO and actual interaction with the destination (Gunn, 1997).

According to Gunn (1972) image is formed in two different ways: organic and induced images. Organic images are formed from newspaper reports, books, movies, documentaries which are not directly related to tourism. Induced images are formed from marketing promotions and advertisement of destinations. The difference between the two is that the induced images are controlled by the destination, on the other hand organic images are not (Gartner, 1984). This chapter is focusing on the organic images that are formed based on news articles that are related to the destination and published online.

As Gartner (1994) states image formation agents are the forces that produce a specific result and image formation process is a continuum of separate agents. One of these agents is called 'autonomous agents', which include documentaries, movies, and news articles that are independently produced. News articles are seen as unbiased presentation of the situation as a result assumed to have significant impact on destination image formation (Gartner, 1994). In addition, if the event that is reported is major importance then the image can change in a short time. For instance, American tourists were convinced by the North American Press that Jamaica is a dangerous destination to travel in 1970s, when in fact, the unsafe areas were limited (Britton, 1979). On the contrary when foreign travelers in USA were asked about their image of USA, their image was based on news reports portraying violence in the country (United States Travel Service, 1977). However, even if the negative images formed as a result of negative autonomous agents are significant in the short term, it may not be effective in the long term image change (Gartner, 1994). Despite its importance, research is limited in analyzing destination image in media coverage.

Research has suggested implications for managing destination image by integrating the topic of branding. Qu, Kim, and Im (2011) state that branding helps marketers to communicate the expectations of a travel experience as well as differentiate from the competitors. Geuens, Weijters, and De Wulf (2009) refer to consumers have the tendency to select brands that are congruent with their personality characteristics. Aaker (1997) introduced the brand personality concept as a way to design brands based upon human traits and create symbolic meanings. She states that consumers interact and memorize brands in an anthropomorphized way. For example, consumers refer to brands as 'cool', 'exciting' and 'lovely'.

Aaker (1997) developed the Brand Personality Scale (BPS) capturing five dimensions: *competence, excitement, ruggedness, sincerity and sophistication.* This also implies that a brand personality enables the creation of symbolic effects for the consumer: the effective match of brand personality creates a holiday status symbol, and, an expression of a lifestyle (Aaker, 1997). The BPS has been implemented in various research contexts, illustrating the positive effects of a brand personality design on consumers' attachment to the brand and behavioral intentions (Geuens et al., 2009; Selby, 2004; Sirgy, 1982; Sirgy & Su, 2000).

Various studies in tourism research have demonstrated the usefulness of the BPS explaining tourists' satisfaction and behavioral intentions (i.e., Baloglu & Brinberg, 1997; Chen & Phou, 2013; Dickinger & Lalicic, 2016; Ekinci et al., 2007; Hankison, 2004; Morgan & Pritchard, 2004; Murphy, Moscardo, & Benckendorff, 2009; Usakli & Baloglu, 2011). For example, Seljeseth and Korneliussen (2015) demonstrate how a destination personality positively impacts tourists experience value. According to Chen and Phou (2013), destinations that are able to establish instant emotional links with customers can create high levels of loyalty. Furthermore, the higher the match of tourist 'self-concept and a destination, the more likely tourists will have a favorable attitude towards the destination, subsequently leading to intentions to re-visit and word-of-mouth (Murphy et al., 2009; Usakli & Baloglu, 2011). Ekinci et al. (2007) state that through marketing programs such as media construction of a destination, marketer can attribute personality traits to a destination. However, tourism research remains limited on the topic brand personality topic and media coverage.

3 Extracting Tourism Knowledge

Big data refers to datasets in analytical applications that are so large (ranging from terabytes to many exabytes) and complex (e.g. real-time sensor data or discussions on social media platforms) that they require advanced technologies to store, manage, analyze and visualize the data (Chen, Chiang, & Storey, 2012). Some examples of big data include records of credit card transactions, search engine traffic statistics, and user-generated content from social media platforms such as Facebook and Twitter. Big data analysis can reveal trends and complex patterns in such large datasets, and therefore has a variety of applications for business intelligence and decision support.

The webLyzard Web intelligence and visual analytics platform enables such applications. It has been customized to a number of domains including politics (Scharl & Weichselbraun, 2008), climate change (Scharl et al., 2016a), and works of fiction (Scharl et al., 2016b). The environmental domain has been chosen to offer several public showcases of the platform's capabilities:

- The **Media Watch on Climate Change** (www.ecoresearch.net/climate) is a content aggregator on climate change and related environmental issues, currently extended with knowledge co-creation capabilities as part of the DecarboNet research project (www.decarbonet.eu), funded in the European 7th Framework Programme (FP7).
- The **U.S. Climate Resilience Toolkit** (toolkit.climate.gov), hosted by the *National Oceanic and Atmospheric Administration* (NOAA), uses the platform to provide a semantic search function. The toolkit was developed in response to President Obama's *Climate Action Plan* to provide expert knowledge and analytic tools to help communities manage climate-related risks and opportunities.
- **UNEP Live Web Intelligence** (uneplive.unep.org/region/index/EU#web_intelligence) aligns environmental indicators reported to the *United Nations Environment Programme* (UNEP) with content metrics from news and social

media. Its cross-lingual capabilities support English, German, and all UN languages including Arabic, Chinese, French, Spanish and Russian.

The decision support functions presented in this chapter go beyond explorative analyses of unstructured information spaces. They address important questions of decision makers in the tourism domain: What are the driving factors that affect the reputation of a destination among bloggers, journalists and social media users? Are there relevant events that should be tracked, and who are the most influential online voices reporting about these events?

Web intelligence applications help to answer such questions. Having been developed for various domains including sports (Marcus, Bernstein et al., 2011), politics (Diakopoulos, Naaman, & Kivran-Swaine, 2010) and climate change (Scharl, Hubmann-Haidvogel et al., 2013), such Web intelligence applications typically face the following challenges:

- Aggregate large document collections from online sources—heterogeneous in terms of authorship, formatting, style (e.g. news article vs. tweets) and update frequency;
- Extract factual and affective knowledge to automatically annotate and structure the acquired content;
- Compute reliable metrics that reflect the success of communication activities; and,
- Provide visual dashboards to select relevant parts of the online coverage and to analyze trends and relations in the resulting information space.

Contextual information, when properly disambiguated, plays a vital part in addressing these challenges and can improve several steps in the processing pipelines of media analytics platforms. Contextual information can guide content acquisition of tourism-related content via focused crawling (Mangaravite, Assis, & Ferreira, 2012), for example, increase the accuracy of knowledge extraction algorithms tailored to the specifics of user-generated content, or help to understand the role of affective knowledge in the decision-making process (Hoang, Cohen et al., 2013).

3.1 Factual Knowledge

Factual Knowledge includes concepts, instances, and relations among these entities. The tourism intelligence platform presented in this chapter uses the *Recognyze* (Weichselbraun, Gindl, & Scharl, 2014) named entity recognition and resolution component to:

- Identify, classify and disambiguate named entities (people, organizations and locations);
- Align these entities with the corresponding entries of external knowledge repositories such as *DBpedia.org*, *Freebase.com* and *GeoNames.org*; and,

- Create a continuously evolving knowledge repository to better understand the structure of social networks, and the dynamic relations among actors participating in these networks.

3.2 *Affective Knowledge*

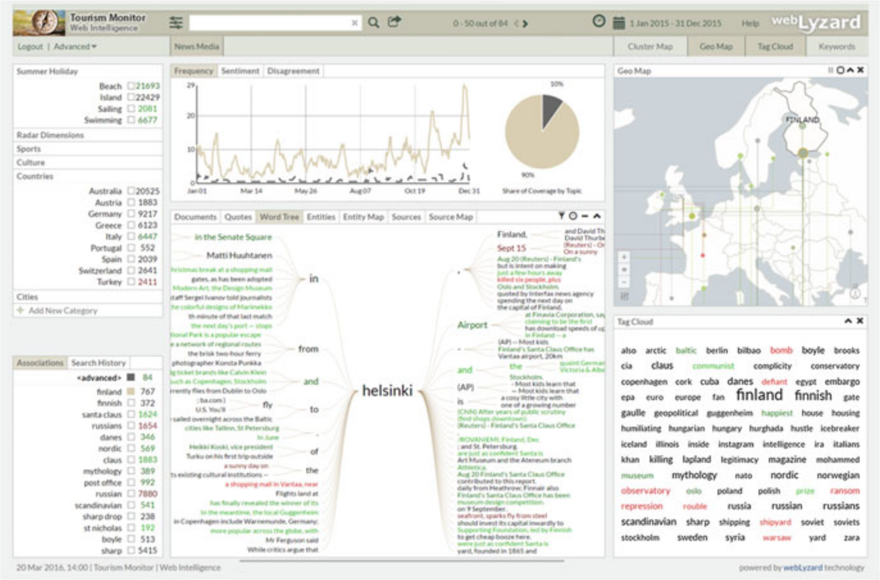
Affective Knowledge includes sentiment and other emotions expressed in a document, which are captured and evaluated by opinion mining algorithms (Weichselbraun et al., 2014; Weichselbraun, Gindl, & Scharl, 2013). Lexical methods rely on sentiment lexicons, which contain known sentiment terms and their respective sentiment values. The ratio of positive and negative terms in a document is a common indicator of overall polarity that is often used for classifiers. Even when considering negations and intensifiers, such methods are computationally inexpensive.

More advanced algorithms rely on dependency parsing or integrate external semantic knowledge bases. This significantly increases the computational demands and calls for more effective approaches to store and analyze data. The factual knowledge extracted by *Recognyze* (see previous section) helps to contextualize the sentiment analysis process, to correctly process ambiguous sentiment terms, and to detect opinion holders and opinion targets.

4 Visual Analytics Dashboard

The visual analytics dashboard shown in Fig. 1 supports tourism managers by identifying trends and topical associations in different online media channels. When applied to user-generated content, the dashboard also reveals what tourists associate with specific destinations, activities or events (traditional surveys help communicators identify value biases in various segments of the public, but do not provide real-time data exploration tools). The visualizations embedded into the dashboard show the geographic distribution of the coverage (for example, destinations most talked about in relation to an activity type), as well as its semantic context (such as the number of documents that report on a specific issue). The dashboard's analytical and visual methods support different types of information-seeking behavior through six main content elements:

- *Sources and settings*. The top menu lets users choose constraints that are relevant for their exploration, including a time interval for accessing longitudinal data, a document source, and a global sentiment filter (unfiltered, positive, or negative). These settings not only affect the trend charts, but also limit search results and dynamic visualizations.


Tourism Monitor
Web Intelligence
Logout | Advanced
News Media
0 - 50 out of 84
1 Jan 2015 - 31 Dec 2015
Help
webLyzard
Cluster Map
Geo Map
Tag Cloud
Keywords
Frequency
Sentiment
Disagreement
Jan 01
Mar 14
May 26
Aug 07
Oct 19
Dec 31
10%
90%
Share of Coverage by Topic
Documents
Quotes
Word Tree
Entities
Entity Map
Sources
Source Map
helsinki
in
from
and
to
of
the
Finland,
Sept 15
Airport
(AP)
is
Geo Map
FINLAND
Tag Cloud
also arctic baltic berlin bilbao bomb boyle brooks
cia claus communist complicity conservatory
copenhagen cork cuba danes defiant egypt embargo
epa euro europe fan finland finnish gate
gaulle geopolitical guggenheim happiest house housing
humiliating hungarian hungary hurghada hustle icebreaker
iceland illinois inside instagram intelligence ira italians
khan killing lapland legitimacy magazine mohammed
museum mythology nato nordic norwegian
observatory oslo poland polish prize ransom
repression rouble russia russian russians
scandinavian sharp shipping shipyard soviet soviets
stockholm sweden syria warsaw yard zara
20 Mar 2016, 14:00 | Tourism Monitor | Web Intelligence
powered by webLyzard technology

Summer Holiday

Beach	□ 21693
Island	□ 22429
Sailing	□ 2081
Swimming	□ 6677

Radar Dimensions

Sports

Culture

Countries

Australia	□ 20525
Austria	□ 1883
Germany	□ 9217
Greece	□ 6123
Italy	□ 6447
Portugal	□ 552
Spain	□ 2039
Switzerland	□ 2641
Turkey	□ 2411

Cities

+ Add New Category

Associations | Search History

<advanced>	■ 84
finland	■ 767
finnish	□ 372
santa claus	□ 1624
russians	□ 1654
danes	□ 346
nordic	□ 569
claus	□ 1883
mythology	□ 389
post office	□ 992
russian	□ 7880
scandinavian	□ 541
sharp drop	□ 238
st nicholas	□ 192
boyle	□ 513
sharp	□ 5415

Fig. 1 Screenshot of the tourism monitor Web intelligence platform, showing a query on "Helsinki" based on news media coverage between January and December 2015

- *Topics*. The left part of the dashboard provides topic management and content navigation. Users can click on a topic to trigger a full-text search; use the topic markers (rectangles) to select which topics are shown in the charts; compute related terms via the "arrow down" symbol; and edit topics or set email alerts via the "settings" symbol.
- *Trend charts*. Interactive charts show weekly frequency, average sentiment, and the level of disagreement regarding selected topics. The sentiment values are based on aggregated polar opinions identified in the document. Disagreement, computed as the standard deviation of sentiment, reflects how contested a particular topic is (references to natural disaster such as "tsunami" or "earthquake", for example, tend to have a low standard deviation because most people agree on their negative connotation). Hovering above a data point displays the associated keywords and daily statistics, whereas a click triggers a search for this topic in the preceding week.
- *Content view*. The content view below the trend charts shows the active document, including its date of publication, keywords, place of publication, and the primary location being referenced.
- *Search results*. The platform's full-text search feature supports wildcard characters, Boolean operators, and regular expressions. The lower third of the dashboard displays the results, including a list of associated terms, and a list of search results with tabs for switching between different views for the document, sentence, and source levels. Each new query also updates the portal's other windows.

- *Visualizations*. To reveal complex and often hidden relations within the document repository, the dashboard rapidly synchronizes a portfolio of visualizations based on multiple coordinated view technology. This portfolio provides insight into the evolution of the underlying document space.

A key strength of the dashboard is its use of multiple coordinated views, also known as *linked* or *tightly coupled views* (Hubmann-Haidvogel, Scharl, & Weichselbraun, 2009), where a change in one view triggers an immediate update of the others. While a user is viewing or editing a new document, for example, the maps pan and zoom to represent its semantic context and offer a holistic, real-time view of the domain. As an alternative to entering query terms to find documents, users can employ the visualizations to retrieve articles related to that particular location, topic, or domain concept. Hovering above a map previews the document closest to the mouse pointer's current position. When previewing documents, the other visualizations automatically adjust to show the previewed documents' immediate context—a crucial feature for supporting the knowledge co-creation process we outline later.

5 Tracking the Brand Reputation of Destinations

The case study presented in this section analyzes content streams from over 150 - English-language news sites and online newspapers (US, CA, UK, AU, NZ), focusing on sentiment expressed in conjunction with Scandinavian capitals. According to Whitelaw, Garg, and Argamon (2005), sentiment analysis is involved with evaluation of a target object as positive or negative. Two things are essential in this process: (1) recognizing how the sentiments are expressed in the texts; and (2) classifying these sentiments as either positive (favorable) or negative (unfavorable) (Nasukawa & Yi, 2003).

In addition to the bipolar classification according to sentiment, the affective knowledge space is analyzed according to Aaker (1997) five-axis structure including *competence, excitement, ruggedness, sincerity* and *sophistication*—with various terms expressing these dimensions, which guarantees a high coverage and ensures the discovery of all relevant concepts. The resulting system provides a comprehensive corpus based on online media coverage for a targeted period. Furthermore, the advanced text mining tools allow an unprecedented level of transparency about emerging trends and the impact of specific events on the public discourse. Figure 1 shows a screenshot of the dashboard. In order to demonstrate the information exploration and retrieval interface ("dashboard") to interactively identify track and analyze coverage about cities, the Scandinavian capitals (Helsinki, Oslo, Stockholm and Copenhagen) are selected. The media coverage is analyzed for the year 2015 divided into four quarters; (1) January–March, (2) April–June, (3) July–September and (4) October–December. The distribution of documents for each quarter is similar for the corresponding destination and the total frequency of

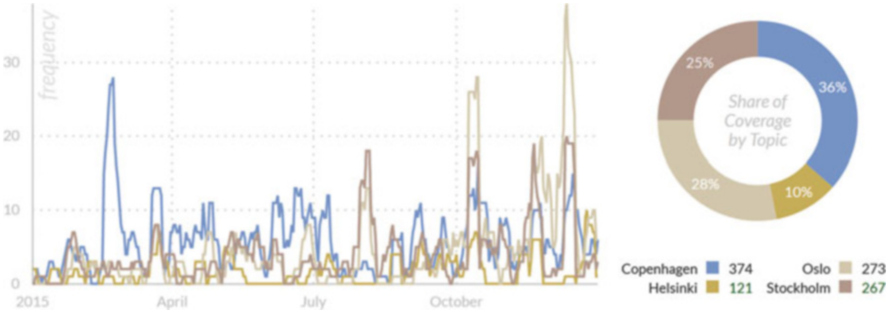


Fig. 2 Weekly frequency of tourism coverage between January and December 2015

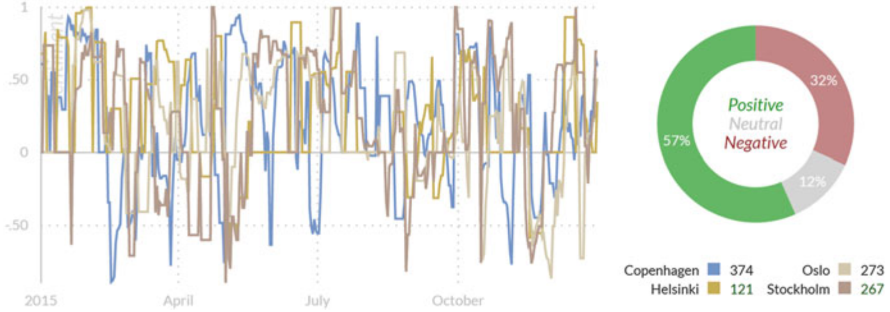


Fig. 3 Sentiment analysis of tourism coverage between January and December 2015

documents are as follows: Oslo (273), Helsinki (121), Stockholm (267) and Copenhagen (374). This shows that Copenhagen is more present in media compared to the other capitals (see Fig. 2).

The sentiments of the documents are then analyzed among the four capitals. The ratio of positive and negative terms found in the surrounding of the target document is used as an indicator of the overall polarity (sentiment) of the document. Through linguistic features (negations and intensifiers) the accuracy of this knowledge extraction process is improved. Sentiment in Fig. 3 is represented by color coding, ranging from red (negative) to grey (neutral) and green (positive). Significant observations from Fig. 3 are the dominant overall representation of positive media coverage in second and third quarters (>80 %). Having a closer look at the distribution per quarter and per capital reveals various outcomes on specific moments. The first quarter, for example, shows a pronounced negative sentiment peak in the second half of February, caused by coverage about the shooting at a free speech debate (BBC, 2015).

Sentiment analyses are a first classification of a destination's representation in user-generated content. However, the dashboard allows further identification of specific affective dimensions (Aakers' five dimensions). Through the use of the radar chart the visualization of the public discourse about destination and Aakers' dimensions is performed. The radar chart is a visual tool that goes beyond sentiment trend charts by profiling a topic across several emotional categories. The radar chart, thus, represents a holistic approach to visualize affective knowledge in the underlying document sources.

Figure 4 illustrates the five multi-dimensional radar charts visualizing media perception of the four capitals based on media coverage in 2015. During the first quarter, the media coverage of Stockholm is dominated by “ruggedness”, Helsinki by “excitement”, and Oslo by “sophistication”. Copenhagen’s media coverage is not dominated by one personality trait but as seen in Fig. 4 leans towards “competence”. During the second quarter, Oslo relates mainly to “sophistication” and “competence”, but also includes “excitement” and “ruggedness”. Copenhagen is more dominant in relation to “sophistication” and “ruggedness” compared to the

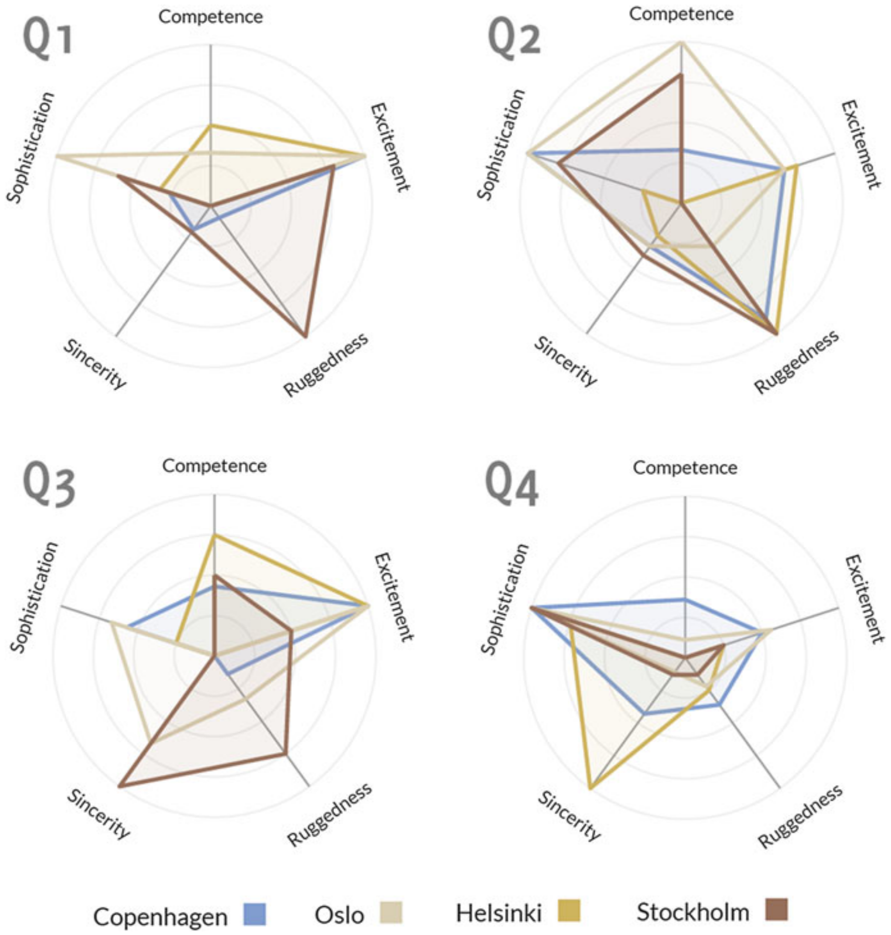


Fig. 4 Quarterly radars charts showing media associations with Scandinavian capitals along the brand personality dimensions of Aaker (1997)

first quarter, whereas the “excitement” trait seems to stay the same as the first quarter. In particular, Helsinki and Stockholm are portrayed by the “ruggedness” trait. In the third quarter, “excitement” is exceptionally related to Oslo, Helsinki and Copenhagen, where as “sincerity” is strongly related to Stockholm. However, in the fourth quarter Helsinki is strongly related to “sincerity” and “sophistication” is more empathized for Stockholm, Oslo, and Copenhagen.

6 Conclusions

Media coverage significantly impacts destination image. Thus, media coverage needs to be continuously monitored and assessed. Given that metadata patterns across various online sources provide novel insights for destination managers and business analysts. These insights will not only yield non-econometric variables to benchmark destinations, but also shed light on emerging discussions of travelers on social media platforms, providing valuable suggestions for operative and strategic improvements. This paper presents a tourism intelligence system for Destination Management Organizations (DMOs) to address the big data challenge. Its dashboard reflects news and social media perceptions along Aakers’ brand personality dimensions, based on comprehensive domain-specific content repositories. The results show the evolution of media coverage on European cities in 2015. This information can be used by DMOs to monitor their destination brands, using visual tools for benchmarking purposes. Destinations should realize the impact of media can have on their tourists’ arrival and react in an accurate manner. The integration of media monitoring systems that processes a large quantity of news media articles allows DMOs to have up-to-date understanding of the image of their destination in the public discourse. The real-time synchronization of the presented dashboard allows DMOs to timely respond to breaking news. Furthermore, the application of various domain-specific topics provides a wealth of information needed to develop appropriate positioning strategies aiming for favorable tourist destination images.

The visual analytics dashboard and the interactive visualizations presented in this chapter support free insight generation without prior modelling of the domain, embracing both unstructured (news media articles, social media postings, etc.) and structured (statistical data, knowledge graphs, etc.) sources. Future work will leverage this flexibility to integrate third-party metrics into the tourism intelligence platform, for example the rich set of survey data contained in TourMIS (www.tourmis.info), an open data platform hosted by *MODUL University Vienna* (Sabou et al., 2013; Brasoveanu et al., 2016). This will enhance the platform’s decision support capabilities since well-informed decisions require not only accurate information about real-world processes such as arrivals per capita and destination-specific metrics, but also on how tourists perceive a destination and its services, and how (and with whom) they communicate about their experiences.

Acknowledgement Some of the visual analytics components presented in this chapter have been developed as part of the ASAP Research Project ("Adaptive Scalable Analytics Platform"), which receives funding from the European Union's *7th Framework Program for Research, Technology Development and Demonstration* under the Grant Agreement No. 619706.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*, 347–356.

Avraham, E. (2000). Cities and their news media images. *Cities, 17*(5), 363–370.

BBC News. (2015). *Denmark Shooting: Gunman targets Copenhagen Free Speech Debate*. Retrieved February 14, 2015, from http://www.bbc.com/news/world-europe-31472741

Baloglu, S., & Brinberg, D. (1997). Affective images of tourism destinations. *Journal of Travel Research, 35*(4), 11–15.

Baloglu, S., & McCleary, K. W. (1999). A model of destination image formation. *Annals of Tourism Research, 26*(4), 868–897.

Beerli, A., & Martin, J. D. (2004). Tourists' characteristics and the perceived image of tourist destinations: A quantitative analysis- a case study of Lanzarote, Spain. *Tourism Management, 25*, 623–636.

Bigne, J. E., Sanchez, M. I., & Sanchez, J. (2001). Tourism image, evaluation variables and after purchase behaviour: Inter-relationship. *Tourism Management, 22*(6), 607–616.

Britton, R. (1979). The image of the third world in tourism marketing. *Annals of Tourism Research, 6*(3), 331–358.

Brasoveanu, A. M. P., Sabou, M., Scharl, A., Hubmann-Haidvogel, A., & Fischl, D. (2016). Visualizing statistical linked knowledge for decision support. Semantic Web Journal, Forthcoming.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Chen, C. F., & Phou, S. (2013). A closer look at destination: Image, personality, relationship and loyalty. *Tourism Management, 36*, 269–278.

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). *Diamonds in the rough: Social media visual analytics for journalistic inquiry*. IEEE Symposium on Visual Analytics Science and Technology (VAST-2010) (pp. 115–122). Salt Lake City: IEEE.

Dickinger, A., & Lalicic, L. (2016). An analysis of destination brand personality and emotions: A comparison study. *Information Technology & Tourism, 15*(4), 317–340.

Ekinci, Y., Sirakaya-Turk, E., & Baloglu, S. (2007). Host image and destination personality. *Tourism Analysis, 12*(5–6), 433–446.

Gartner, W. C. (1994). Image formation process. *Journal of Travel and Tourism Marketing, 2*(2–3), 191–216.

Geuens, M., Weijters, B., & De Wulf, K. (2009). A new measure of brand personality. *International Journal of Research in Marketing, 26*(2), 97–107.

Govers, R., Go, F. M., & Kumar, K. (2007). Promoting tourism destination image. *Journal of Travel Research, 46*(1), 15–23.

Gunn, C. (1972). *Vacationscape*. Austin, TX: Bureau of Business Research, University of Texas.

Gunn, C. A. (1997). *Vacationscape: Developing tourist areas*. Washington, DC: Taylor & Francis.

Hankison, G. (2004). Relational work on brands: Towards a conceptual model of place brands. *Journal of Vacation Marketing, 10*(2), 109–121.

Hoang, T.-A., & Cohen, W.W., et al. (2013). *Politics, sharing and emotion in microblogs*. IEEE/ACM International conference on advances in social networks analysis and mining (pp. 282–289). Niagara Falls, Canada: ACM Press.

Hubmann-Haidvogel, A., Scharl, A., & Weichselbraun, A. (2009). Multiple coordinated views for searching and navigating web content repositories. *Information Sciences, 179*(12), 1813–1821.

Mangaravite, V., Assis, G. T. D., & Ferreira, A. A. (2012). *Improving the efficiency of a genre-aware approach to focused crawling based on link context*. Eighth Latin American Web Congress (LA-WEB 2012) (pp. 17–23). Cartagena de Indias, Colombia: IEEE CPS.

Marcus, A., & Bernstein, M.S., et al. (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration. 2011 Annual conference on human factors in computing systems (CHI-11) (pp. 227–236). Vancouver, Canada: ACM.

Morgan, N., & Pritchard, A. (2004). Meeting the destination branding challenge. *Destination Branding*, 59–79.

Murphy, L., Moscardo, G., & Benckendorff, P. (2009). Linking travel motivation, tourist self-image and destination brand personality. *Journal of Travel & Tourism Marketing, 22*(2), 45–59.

Nasukawa, T., & Yi, J., (2003). *Sentiment analysis: Capturing favourability using natural language processing*. Proceedings of the 2nd International conference on knowledge capture (pp. 70–77). ACM.

Qu, H., Kim, L. H., & Im, H. H. (2011). A model of destination branding; Integrating the concepts of the branding and destination image. *Tourism Management, 32*(3), 465–467.

Sabou, M., Arsal, I., & Brasoveanu, A. M. P. (2013). TourMISLOD: A tourism linked data set. *Semantic Web Journal, 4*(3), 271–276.

Scharl, A., Herring, D., Rafelsberger, W., Hubmann-Haidvogel, A., Kamolov, R., Fischl, D., et al. (2016a). Semantic systems and visual tools to support environmental communication. *IEEE Systems Journal*. Forthcoming. Accepted 31 July 2015.

Scharl, A., Hubmann-Haidvogel, A., Jones, A., Fischl, D., Kamolov, R., Weichselbraun, A., et al. (2016b). Analyzing the public discourse on works of fiction—automatic emotion detection in online media coverage about HBO's Game of Thrones". *Information Processing & Management*, *52*(1), 129–138.

Scharl, A., Hubmann-Haidvogel, A., et al. (2013). From web intelligence to knowledge co-creation—A Platform to analyze and support stakeholder communication. *IEEE Internet Computing, 17*(5), 21–29.

Scharl, A., & Weichselbraun, A. (2008). An automated approach to investigating the online media coverage of US Presidential Elections. *Journal of Information Technology & Politics, 5*(1), 121–132.

Sealy, W., & Wickens, E. (2010). The potential impact of mega sport media on the travel decision-making process and destination choice—the case of Portugal and Euro 2004. *Journal of Travel & Tourism Marketing, 24*(2–3), 127–137.

Selby, M. (2004). Consuming the city: Conceptualizing and researching urban tourist knowledge. *Tourism Geographies, 6*(2), 186–207.

Seljeseth, P. I., & Korneliussen, T. (2015). Experience-based brand personality as a source of value co-creation: The case of Lofoten. *Scandinavian Journal of Hospitality and Tourism, 15* (supp 1), 48–61.

Sirgy, M. J. (1982). Self-concept in consumer behavior: A critical review. *Journal of Consumer Research, 9*(3), 287–300.

Sirgy, M. J., & Su, C. (2000). Destination image, self-congruity, and travel behavior: Toward an integrative model. *Journal of Travel Research, 38*(4), 340–352.

Sönmez, S., & Sirakaya, E. (2002). A distorted destination image? The case of Turkey. *Journal of Travel Research, 41*(2), 185–196.

Stepchenkova, S., & Eales, J. S. (2011). Destination image as quantified media messages: The effect of news on tourism demand. *Journal of Travel Research, 50*(2), 198–212.

United States Travel Service. (1977). *International travel market reviews of selected major tourism generating countries*. Washington, DC: US Department of Commerce.

Usakli, A., & Baloglu, S. (2011). Brand personality of tourist destinations: An application of self-congruity theory. *Tourism Management, 32*(1), 114–127.

Weichselbraun, A., Gindl, S., & Scharl, A. (2013). Extracting and grounding contextualized sentiment lexicons. *IEEE Intelligent Systems, 28*(2), 39–46.

Weichselbraun, A., Gindl, S., & Scharl, A. (2014). Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-Based Systems, 69*, 78–86.

Whitelaw, C., Garg, N., & Argamon, S. (2005, October). *Using appraisal groups for sentiment analysis*. Proceedings of the 14th ACM International conference on information and knowledge management (pp. 625–631). ACM.

Online Travel Reviews: A Massive Paratextual Analysis

Estela Marine-Roig

1 Introduction

In recent years of the knowledge society, there has been a dramatic growth of user-generated content (UGC) in parallel with the rise of the Internet and social media. O'Reilly (2005) talks about the portals facilitating collective work or activity of all web users, and claims that an important part of Web 2.0 is *harnessing collective intelligence*, essentially turning the web into a kind of global brain. Kaplan and Haenlein (2010) define social media as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows for the creation and exchange of user-generated content (p. 61).

This growth of UGC has been widely apparent in the fields of travel, tourism, and hospitality, especially with the exponential increase of online travel reviews (OTRs). For instance, in January 2016, TripAdvisor branded sites made up the largest travel community in the world, reaching more than 320 million reviews and opinions, covering more than 6.2 million attractions, accommodations, and restaurants (TripAdvisor.com, About Us); and Booking claimed to have had more than 75 million verified hotel reviews from real guests (Booking.com, Reviews). There have been many studies on the influence of UGC, and especially OTRs (Schuckert, Liu, & Law, 2015), as types of electronic word-of-mouth (eWOM) marketing of travel-related decisions (Baka, 2016; De Ascaniis & Gretzel, 2013; Fang, Ye, Kucukusta, & Law, 2016; Gretzel & Yoo, 2008; Jalilvand, Samiei, Dini, & Manzari, 2012; Litvin, Goldsmith, & Pan, 2008; Liu & Park, 2015), as well as on the destination image formation (Kladou & Mavragani, 2015; Lai & To, 2015; Li, Lin, Tsai, & Wang, 2015; Marine-Roig & Anton Clave, 2016a, 2016b; Serna, Marchiori, Gerrikagoitia, Alzua-Sorzabal, & Cantoni, 2015). Moreover, to a certain

E. Marine-Roig (✉)
University of Lleida, Catalonia, Spain
e-mail: estela.marine@aegern.udl.cat

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_11

Table 1 Sample of online information about tourist attractions (2016-01-31)

Query	Google indexed pages	TripAdvisor OTRs	TripAdvisor photos
"Central Park" "New York"	58,500,000	61,569	20,825
(Tour OR tower) Eiffel Paris	23,200,000	67,455	30,050
(Basilica OR temple) "Sagrada Familia" Barcelona	10,700,000	65,413	27,639

extent, travel-related writings, as travelogues, travel blogs, and OTRs, can and do function as sources of information for visitors of a destination and can be used in ways similar to conventional travel guidebooks (Peel & Sorensen, 2016, p. 24).

There is also growing the number of tourists who plan and book their trips online (Cao & Yang, 2016; Xiang, Pan, & Fesenmaier, 2014). More than 30,000 European respondents from different social and demographic groups were interviewed and it turned out that Internet websites were the second most-used source of information for making travel plans and by far the most common way to organize a holiday (Eurobarometer, 2015). It is even argued that planning online travel is the most palpable example of how information technologies have changed the domain of travel and tourism (Xiang, Magnini, & Fesenmaier, 2015). Table 1 shows some examples of how much online information can be found about a tourist attraction on Google, the world's leading search engine (Alexa.com, TopSites), and TripAdvisor, the tourism domain's largest user-generated online review site (Baka, 2016). In the case of the Basilica of the Sagrada Familia in Barcelona, Google returns more than 10 million indexed pages; admitting that the results presented by Google represent a minimal part of the indexed pages (Xiang, Wober, & Fesenmaier, 2008), this is a very considerable amount. Moreover, this Catalan landmark has over 65,000 OTRs on TripAdvisor.

According to O'Connor (2010), increased quantities of information can be both a blessing and a curse (p. 756). On the one hand, the availability of a great deal of unbiased, unsolicited, and cost-effective data on a destination is an opportunity for travel-related research to gain insights (Marine-Roig & Anton Clave, 2015), but the study of this vast amount of information requires the use of big data analytic techniques (Krawczyk & Xiang, 2016; Xiang, Schwartz, Gerdes, & Uysal, 2015; Yuan & Ho, 2015). Conversely, this is a serious problem for a vacationer who wants to know relevant opinions of previous visitors of the Sagrada Familia, and finds a hyperlink on TripAdvisor with the following message: "*Read all 65,413 reviews*" (Table 1). Such available information overload prevents consumers from having a comprehensive idea of the attraction and complicates the decision-making process (Fang et al., 2016; O'Connor, 2010).

In this context of inability to read in detail the reviewers' writings (Afzaal & Usman, 2015), the paratextual elements of OTRs, such as review titles, acquire a crucial importance. The term *paratext* was introduced by Gerard Genette in 1987 to define a set of productions (an author's name, a title, a preface, illustrations) accompanying the text of a literary work. One does not always know if one should consider if the *paratext* belongs to the text or not, but in any case they surround it and prolong it, precisely in order to present it, in the usual sense of this verb, but also in its strongest meaning: to make it present, to assure its presence in the world, in addition to its *reception* and its consumption in the form (nowadays at least) of a book (Genette, 1997, p. 1). This French literary theorist divides the *paratext* into *peritext* and *epitext* based on the distance of the elements in relation to the location of the text itself. He devotes a chapter to the publisher's *peritext* to study the whole zone of the *peritext* that is the direct (but not exclusive) responsibility of the publisher, or perhaps, of the publishing house (p. 16). Genette's framework can be used as a language shared by a wide range of disciplines and the paratextual features continue offering a great tool to interpret texts in a digital milieu (Desrochers & Apollon, 2014). In digital media environments, *paratext*s have become an essential part of media consumption (Alacovska, 2015; Gray, 2015). However, in spite of the influence that UGC—such as travel blogs on specialized hosting websites—may exert on destination image formation, little is said about webhost-created content or paratextual information (Azariah, 2011).

Therefore, this paper analyses the paratextual elements of an OTR with particular emphasis on the title and what Genette (1997) names the publisher's *peritext*, which, in this case of writing, refers to the hosted content on a travel-related website that might be called webhost- or webmaster-generated content (WGC), to deduce and distinguish the image perceived by the reviewer as transmitted by the webmaster. For this purpose, both of the most touristic continental regions of the European Union (Eurostat, 2015) are selected: Ile de France, whose capital city is Paris; and Catalonia, whose capital is Barcelona. A random sample of 300,000 OTRs (150,000 for each region) written in English by tourists visiting any of these destinations between 2011 and 2015 is harvested in TripAdvisor. In order to test the effectiveness of the methodology, another random sample of 30,000 titles of OTRs on the Basilica of La Sagrada Familia (Barcelona) written in English is analysed and the results are compared with previous similar studies based on quantitative content analysis of both the title and writing body.

2 Theoretical Framework

Genette (1997) considers that the *paratext* is only an assistant or accessory of the text: if the text without its *paratext* is sometimes like an elephant without a mahout, as in a power disabled, then the *paratext* without its text is a mahout without an elephant, or just a silly show (p. 410). Gray (2015) criticizes this metaphor and concludes that *paratexts* are, in short, part of the text, because they are usually constitutive, central, and absolutely important (p. 230). Paratextual elements are essential for taking advantage of the information contained in the text in light of countless OTRs on an attraction, activity, product, or service. For instance, it is critical to locate a travel blog or review in space and time. Moreover, there are

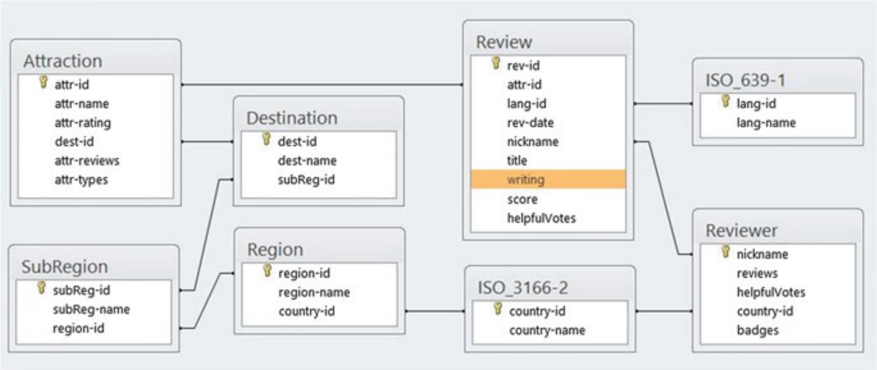


Fig. 1 Relational database of OTR's paratextual elements. *Source*: Author

many thousands of OTRs without text in the writing body with useful information garnered by the title and other paratextual elements that surround it.

Paratextual information influences the evaluation of the nature/genre of a posted text, as it helps the reader understand the content as well as its positioning whilst creating expectations (Azariah, 2011). Following Genette's nomenclature, the OTR *paratext* is divided into OTR *peritext* and OTR *epitext*, and may be UGC, WGC, or a combination of both. The most important elements of OTR *peritext* are title, language, theme or type, date, and geographical location of the destination, followed by the reviewer's profile, number of reviews posted, cities visited, score rating, helpful votes, badges, and even the template provided by the webmaster to write the review. Indeed, webhost paratextual information plays a significant part in positioning a specific UGC post text as a narrative about a particular destination, and has an influence on authorial voice (Azariah, 2011). The OTR *epitext* (related reviews, contextual advertisements, etc.) is not within the scope of this study. Figure 1 shows the relationship between the writing body (selected field) of an OTR and the paratextual elements that surround it.

2.1 OTR Title

An OTR is in itself the combination of a title and a text (Banerjee, Chua, & Kim, 2015; Grabner, Zanker, Fliedl, & Fuchs, 2012). As the communicative purpose of an OTR is the critique (discouragement or recommendation) of a certain travel choice, the narrative component in OTR is not as prominent as in travel diaries and is combined with evaluations and descriptions of the personal travel experience (De Ascaniis & Gretzel, 2013), which renders review titles especially important.

For users, OTR titles are very relevant in a context where there usually is a huge number of OTRs for each tourism product/service (De Ascaniis & Gretzel, 2013).

Information search involves time, effort, and humans who have a limited capacity in processing incoming information (De Ascaniis & Gretzel, 2012, 2013). Users need to find and judge as quickly as possible reviews that meet their needs and, to support this, information retrieval systems' display metadata such as title, date, and URL (De Ascaniis & Gretzel, 2012, 2013). The decision of which source to consult is made by relying on a first impression of search results, based on metadata, titles that serve as the overview and preview of the review, and anticipation, which is absolutely crucial for the online information search (De Ascaniis & Gretzel, 2012, 2013). Hence, titles are fundamental when users have to make a quick first choice to select the reviews that seem most relevant, and they may indeed be the only thing users read of the whole review (De Ascaniis & Gretzel, 2012). Moreover, titles are more recognised by search engines because they have a superior html level. Therefore, results on search engines will be more based on titles than on the review text itself, having a major potential influence.

Titles are interesting because they provide insights into how customers summarize experiences and show the first impressions others may get of a place, product or service. In fact, reviewers are invited to use concise formulations for the title (Grabner et al., 2012). TripAdvisor prompts reviewers to answer the following question when creating titles: "If you could say it in one sentence, what would you say?" as a synthesis of the attraction or experience (De Ascaniis & Gretzel, 2012). In this respect, the role of titles in OTRs can be compared to the role of headlines in newspapers or the role of taglines or slogans for an advertisement (De Ascaniis & Gretzel, 2013). The first impression of a news headline influences people's behaviour and if something catches the person's attention it will be more easily remembered (De Ascaniis & Gretzel, 2012). Furthermore, the linguistic characteristics of titles may enable the development of automated algorithms for the selection and classification of OTRs (De Ascaniis & Gretzel, 2012; Grabner et al., 2012) or even spot market and stock trends based on lexical semantic similarity (Wang & Wu, 2012). Titles are concentrated presentations of the text to come (Wang & Wu, 2012).

The usefulness of titles both for users and researchers is demonstrated by several studies. De Ascaniis and Gretzel (2012) analysed a corpus of 1474 OTR titles about three city destinations published on TripAdvisor in terms of length, informativeness, indication of review orientation, word diversity, and communicative function. These authors found that titles are representative of the review orientation and accomplish the general function of helping readers anticipate what follows in the description. Grabner et al. (2012) conducted a study on hotel reviews from various, large tourism cities, which contained hotel category, overall rating, title, and text, and they automatically extracted text and ratings from TripAdvisor from over 80,000 reviews. Wang and Wu (2012) picked out 2000 blog titles to spot market and stock trends. Banerjee et al. (2015) analysed separately review text and titles to distinguish authentic and fake reviews, and concluded that titles may be a more useful object of analysis because of the greater attention they command. De Ascaniis and Gretzel (2013) focused on the communicative functions of OTR titles and identified the role they play in forming readers' first impression of tourism-related online search results. The authors found that the majority of OTR titles point to the review standpoint, which is to visit the recommendation or to the attraction evaluation argues for in the review.

This is very significant in terms of destination image formation because images are greatly formed before the trip during the search for information, through the influence of various information sources, of which word-of-mouth is one of the most influential (Gartner, 1994). Therefore, this means that review titles can have a key role in forming pre-trip tourist images as they point to the standpoint of the whole text, through the eWOM effect and are highly influential. From a holistic conception of tourist images encompassing both projected and perceived images (Marine-Roig, 2015a), OTR titles become crucial in forming tourist images because they not only show in summary the perceived image of the attraction or place of other tourist-peers, but moreover they are the explicit synthesis of the image or idea of the attraction or place the tourist wants to project or transmit to others, which will likely have more influence on her. It represents the perceived image that she wants others to perceive, written with an audience in mind (De Ascaniis & Gretzel, 2012), where only what is most worth-mentioning is written and that will most strongly impact the user and be remembered. This phenomenon of an elaborate perceived image being transmitted through OTR titles, can be understood in the context of the two-way mutual influence of projected and perceived images (Marine-Roig, 2015a), where tourists reproduce perceived images, by their actions and transmission to others, thus closing the hermeneutic circle of images (Caton & Almeida, 2008).

Moreover, the content of OTR titles seems to be very interesting to analyse tourist images. De Ascaniis and Gretzel (2012) found that OTR titles are especially rich in content items with univocal information (3 out of 5 words), making them especially prone to image content analysis. Titles make strong use of superlatives, slogans, and positive words much more frequently than negative ones (negative ones did not appear in the top keywords list). These results are in the line of Marine-Roig and Anton Clave (2016a) that analysed the affective component of images in OTR (both text and titles) and found that positive adjectives are highly predominant. Besides, many titles try to characterize the destination by highlighting one of its features, which would be the the most representative one for them (De Ascaniis & Gretzel, 2012). In this respect, Marine-Roig and Anton Clave (2015, 2016a) found that the image contained in travel blogs and reviews, in comparison to other types of tourism online sources, is more stereotypical, focused on very specific things (feelings and must-see attractions), and much less diverse. Therefore, it is expected that this tendency is even more accentuated in titles, seen as the synthesis of the image to be transmitted to others. The analysis of destination images through OTR titles would therefore enable the reader to spot the “tip of the iceberg” of the destination image, its synthesis, which is the visible part of the perceived image that becomes transmitted and mostly seen by others.

However, it is important to note that in OTR websites, titles are part of the paratextual elements and review webhosts also add information to the same titles, so this should also be considered in terms of destination image formation. As

Azariah (2011) points out, those who analyse travel blogs to study destination image must recognize contribution of the webhost to the positioning of the blog as a travel narrative. In travel blog hosting sites, similarly to OTR websites, the content provided by the webhost coexists and competes for space with titles created by the users in a manner that can influence the positioning of the text (Azariah, 2011). Although titles are the first main element to examine for textual and authorial identity matters in UGC posts, webhosts introduce other information such as the location (country and destination) of the post in the title (Azariah, 2011). In fact, host-generated content can take precedence over personal discourse as represented by user-generated content, especially in terms of the identification of the post with an author (Azariah, 2011). The same could be said in the case of OTR. In this online context, destination image construction is also influenced by the image transmitted by the webhost in browsers through paratextual elements and in search engines through *meta tags*, which will enable the user to find a specific review, give it a specific positioning, and thus have more potential influence on other users.

2.2 Another OTR peritext

Most authors who have analysed OTRs have taken into account the language, topic, date and/or geographical location of the destination, such as: Dickinger and Lalicic (2016) on destination brand personality and emotions; Fang et al. (2016) on perceived value of OTRs; Johnson, Sieber, Magnien, and Ariwi (2012) on web harvesting; Liu and Park (2015) on review usefulness; Marine-Roig (2015b) about feelings and religiosity; Schmunk, Hopken, Fuchs, and Lexhagen (2014) on sentiment analysis; and Wang, Chan, Ngai, and Leong (2013) on reviewer credibility. Further to cope with the information overload mentioned in the introduction, some authors have delved into the analysis of the other elements of OTR *peritext* to deduce aspects such as readability, reliability or, in short, the usefulness of a review for other users who are planning a trip. For instance, Liu and Park (2015) point out that many review websites have designed peer reviewing systems where users vote to assess the usefulness of a review in their decision-making. For example, Amazon provides a service that displays the top two most helpful, favourable, and critical reviews posted by online users in order to help its customers evaluate each displayed product easily. In this respect, Wang et al. (2013) proposed an impact index to compute the reviewer's credibility, which evaluated both expertise and trustworthiness, based on the number of reviews posted by the reviewer and the number of helpful votes received by the reviews. In their index, the more reviews, the higher the expertise of the reviewer and thus her impact index. Similarly, the more helpful votes, the higher the trustworthiness of the reviewer.

In terms of the helpfulness of reviews, Fang et al. (2016) found that the readability of a review text is correlated with its perceived helpfulness. Reviews with precise details that are easily understandable will receive more helpfulness votes. Moreover, the perceived helpfulness of a user's reviews will be influenced and can be inferred by her historical rating distribution. Specifically, the mean rating of the historical ratings of an author can be used to infer the starting point attitude towards travelling reviews, either positive or negative. Usually, positive reviewers, with higher means will receive more helpfulness votes. Further, Johnson et al. (2012) argue that, from a tourism research perspective, UGC posts are especially suitable to obtain information on niche tourism or 'off the beaten track' tourism amenities. OTRs offer several possibilities to harvest specific types of information. For instance, the authors harvested from TravelReview the quantitative overall star rating out of five, plus the amenity-type specific ratings out of five (such as cleanliness and service for accommodations). Moreover, using web harvesting, it was possible to extract star ratings for each amenity reviewed. These authors found that star rating for Nova Scotia were high, with 75 % of accommodations, 79 % of attractions, and 69 % of restaurants receiving a four- or five-star rating. However, the authors point out that star rating data is insufficient to understand the experience of tourists and it should be combined with the analysis of the review description (text and title). Therefore, OTR *peritext* elements such as review ratings and helpfulness votes should be taken into account as influential for review positioning and potential influence in the destination image formation of users.

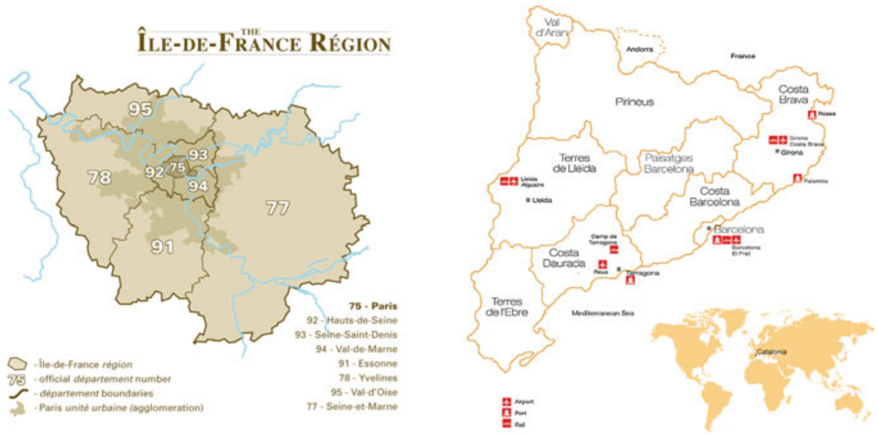
3 Methodology

The methodology used to achieve the objectives of this chapter is an adaptation of the methodology to analyse massive UGC data, as defined in Marine-Roig and Anton Clave (2015), and detailed in Marine-Roig and Anton Clave (2016b). This method is divided into five stages: destination choice; webhost selection; data collection; pre-processing; and analytics.

3.1 Destination Choice

Given the scanty amount of text in the titles, it is interesting to have many OTRs increase the reliability of the results. That is why we have chosen the two most touristic regions of the European Union by overnight stays (Eurostat, 2015): Ile de France, whose capital city is Paris; and Catalonia, whose capital is Barcelona. There is another European region with more tourists, the Canary Islands, but it is not located on the European continent and is specialized in nature tourism and in the tourism of sun, sea, and sand for its year-round mild climate.

Ile de France and Catalonia have similar characteristics that make them comparable. Both regions have a big capital city surrounded by subregions that complement the tourist offer (Fig. 2). With regard to the hotel business in 2015, Ile de



Author: J. M. Schomburg (WikiMedia) Author: Official work (CTB, 2016)

Fig. 2 Ile de France and Catalonia European regions

France recorded 32.4 million travellers who spent 66.3 million overnight stays (CRT, 2016) and Catalonia recorded 17.6 travellers who spent 52.0 million respectively (IDESCAT, 2016). In 2015, Paris represented about 50 % of the hotel activity in the region and Barcelona about 40 %.

3.2 Webhost Selection

The analysis of websites hosting OTRs used in previous works (Marine-Roig, 2015b; Marine-Roig & Anton Clave, 2015) has verified that TripAdvisor (TA) is the most suitable source for the case study by far if compared to other websites. For example, compared to VirtualTourist (VT), the second most important site in January 2016, VT had less than 600 reviews on the most important landmarks of the two regions (Eiffel Tower and Basilica of La Sagrada Familia) while TA had over 65,000 OTRs of each (Table 1). Therefore, it is not considered necessary to include reviews of the other websites in the data set because their corresponding weight would be negligible.

3.3 Data Collection

Since the analysis is intended to infer the image perceived by the reviewer, only OTRs on “things to do” in the destination are downloaded, excluding the hotel and restaurant reviews for its high specialization and because they are the subject of other types of studies such as those carried out by Krawczyk and Xiang (2016), O’Connor (2010), and Xiang, Schwartz, et al. (2015). Once the filters are established (Marine-Roig & Anton Clave, 2016b), the OTRs on both regions are downloaded by means of a web copier, Offline Explorer Enterprise (OEE). OEE is a scalable solution supporting massive downloads and fast data processing. It includes the ability to download up to 100 million URLs (Uniform Resource Locator) per project and the fastest-possible multi-threaded processing of downloaded files by using all CPU cores (MetaProducts.com). OEE saves the OTR contents and records related attractions on the local hard disk, keeping information on its hyperlinks in the filenames.

3.4 Web Data Mining

Considering that the aim of this study is to analyse the image perceived by the reviewer and transmitted by the webmaster from the title and other paratextual elements of an OTR, in this stage it is necessary to extract from the downloaded web pages the data that appear in the fields of the Review, Reviewer, and Attraction tables (Fig. 1). With regard to the titles, one works with the information that the user visualizes in web browsers. Then, the participation of the reviewer and webmaster in their generation of content is as follows:

- Heading title (UGC). This is the title written by the reviewer according to the webmaster template (WGC) and one sees in the OTR page formatted by the CSS (Cascading Style Sheets) style rules of the web site (WGC). It consists only of literary text and is essential to analyse the image perceived by the reviewer.
- Page title (UGC + WGC). This is the reviewer’s title (UGC) and further information added by the webmaster (WGC). It is between HTML tags (<title > and </title>) and can be seen in the title bar of the web browser. The WGC data is valuable for knowing the attraction and destination associated to the OTR.
- Title hyperlink (WGC). This is the URL pointing to the OTR page. It goes between HTML tags (<a href = and </a>). That is, the heading title is the anchor text associated with this hyperlink. Hyperlinks are very important for search and they play a central role in search ranking algorithms (Liu, 2011). They can be seen in the address bar of the web browser. This hyperlink contains important data (destination, attraction, and OTR codes; and attraction, destination, and region names) that allows classifying the reviews and setting up relationships between the various items in the database (Fig. 1).

Titles have been extracted from the files using a search-and-replace utility, Replace Studio Pro (RSP). RSP is a *grep* (Global Regular Expression Print) utility that has an extensive repertoire of *regular expression* functions that let you search or replace using wildcard operators (Funduc.com). For example, for the titles of 100,000 HTML pages, <title> is inserted into the search box and RSP generates a plain-text file with 200,000 lines. The odd lines have the file name and full path that includes the hyperlink, and the even the title of the page bordered by the HTML tags<title> and </title>. The remaining *peritext* fields of the Review, Reviewer, and Attraction tables (Fig. 1) are obtained in a similar way by introducing *regular expressions* in the search box of the RSP utility, but before, to prevent RSP from collecting redundant information, related reviews (*epitext*) added by the webmaster on the page of each item must be removed.

3.5 Data Arranging

The data collected in the previous subsection are organized according to the basics of relational databases (Fig. 1). In order to save space and processing time, tables are stored in CSV format (Comma Separated Values). The CSV files contain only data in text format and allow separate records of tables in Fig. 1 with a line feed and fields with a semicolon and can therefore be handled with any word processor. Moreover, CVS files are compatible with spreadsheet applications and database management systems (DBMS). The relationships between the items of Fig. 1 allow for the exchanging of information without having to duplicate data; for example, in Table 2, an attraction is related to its subregion, just as each of the nine Catalan brands are related to all the attractions of its territory and, through them, to all of reviews about these attractions. On the other hand, the collected *peritext* allows for conducting multiple classifications of OTRs, for example, in the Tables 3 and 4 there is a ranking, per years and subregions, of the 300,000 reviews sample.

We used the UltraEdit Pro (UEP) programme to work with the large, raw-data files. As stated in its slogan, UEP is the multi-purpose text editor that loves multi-gigabyte files (UltraEdit.com) and can handle and edit files in excess of 4 gigabytes. UEP also has full *regular expressions* support to the find and replace function. Continuing with the above example file with 200,000 lines, the so-called special characters—those that are not in the English alphabet—should first be replaced by a character ISO 8859-1 (Latin 1). For example, *Î}\^{I}{le de France* has an uppercase ‘i’ with a circumflex accent that the web server can write variously as UTF-8 (*Ã*Ž), HTML number (&#206;), or HTML name (&Icirc;). Semicolons with the remaining comma should be replaced to avoid altering the structure of the CVS files. UEP then replaces the newline and the first HTML tag (<title>) by a semicolon, and deletes another tag (</title>). In a few seconds, the text file of 200,000 lines becomes a CSV file with 100,000 records with two fields (filename with full path and heading title).

Tables 3 and 4 show the trends of the OTRs in space and time. There is an evidential increase in the number of reviews over the years and metropolises (Paris -75, and Barcelona -Barna) have a much higher weight than the other subregions.

Table 2 Example results of database queries (Fig. 1) in the case of Catalan destinations

attr-id	dest-id	dest-name	subReg-id	subReg-name
d3932772	g665816	Badalona	Barna	Barcelona
d1755008	g187502	Sitges	cBarc	Costa Barcelona
d668671	g494960	Lloret de Mar	cBrav	Costa Brava
d667082	g562814	Salou	cDaur	Costa Daurada
d191040	g187501	Montserrat	pBarc	Paisatges Barcelona
d2025081	g1072494	Sort	Pyren	Pirineus
d615292	g1916989	Miravet	tEbre	Terres de l'Ebre
d2334329	g187500	Lleida	tLlei	Terres de Lleida
d3929158	g664637	Baqueira	vAran	Val d'Aran

Table 3 Random sample of 150,000 OTRs on Ile de France per district and year

	75	77	78	91	92	93	94	95
2011	5284	464	208	0	22	19	4	5
2012	22339	1627	691	1	37	54	18	13
2013	23592	1735	857	6	59	84	17	21
2014	29365	1984	1266	12	68	125	32	38
2015	54618	3091	1858	20	123	119	44	80

Table 4 Random sample of 150,000 OTRs on Catalonia per brand and year

	Barna	cBarc	cBrav	cDaur	pBarc	Pyren	tEbre	tLlei	vAran
2011	3730	48	173	502	43	5	2	13	1
2012	17,783	229	927	1861	292	43	6	11	3
2013	22,910	376	1120	1907	658	86	10	12	2
2014	30,953	616	1499	2342	373	133	20	43	6
2015	50,930	1331	4372	3569	734	218	46	48	14

3.6 Parser Settings

The ad hoc parser used in this work divides textual input into words and counts them. For this purpose, it needs the following information:

- Composite words. List that contains groups of words together have a different meaning than each of them separately as "must see" or "must-see" and compound nouns like "Eiffel Tower" or "Tour Eiffel".
- Black list. This contains the so-called stop words such as adverbs, determiners, and prepositions, which are not considered useful for quantitative analysis. They have also been ignored for parsing words with less than three letters.
- Word delimiters. These are characters that separate one word from the next. In this case study we have used all characters that are not letters in Catalan, English, French, and Spanish languages: NOT [a-zA-ZáàâæéèëêíïîóòôœúùüûçñÿÁÀÂÆÉÈËÊÍÏÎ}\^{I}{ÓÒÔŒÚÙÜÛÇÑŸ].

The parser creates a unique-words set. Each word or compound word of the set is associated with its frequency. The parser algorithm follows these steps: (1) Loads input text and list files, and converts text to lowercase; (2) Reads compound words, removes them from the input text, and adds them to the set with its frequency if greater than zero; (3) Divides the remaining text into words according to the delimiters; (4) For each word it increments the counter of whole words, and if it has a length greater than two characters and is not in the blacklist, it adds it to the set (if the word already was in the set, it increases its frequency); and (5) Creates a CSV file with three columns: whole word; frequency; and percentage related to all the words of the input text including stop words.

3.7 Categorisation

As Stemler (2001) asserts, text content analysis is a technique that compresses many words into fewer content categories. Categories are groups of words with similar meaning or connotations and must be independent, mutually exclusive, and exhaustive. In quantitative content analysis there are two main approaches in creating categories: (1) A priori categorization (Stemler, 2001) in which categories are pre-established by the researcher and obtained deductively. Usually, category word frequency counts are obtained, organized in a matrix, and associations/ correlations between categories can be calculated (Stepchenkova, Kirilenko, & Morrison, 2009). (2) Emergent categorization or a posteriori (Stemler, 2001), in which categories are created from the data themselves in an inductive way. This correlational model determines categories from the text analysed, extracting themes from the matrix of word frequencies by means of different techniques (Stepchenkova et al., 2009).

In this case study, due to the exploratory nature of the content of OTR paratextual elements, emergent categorisation has been used. This is in the line with findings of Dann (2014), who argues for an emergent categorisation based on tourists' contents. In this context, "when the data are content analysed, categories emerge that are uniquely founded on the *ipsissima verba* of the subjects (p. 49)".

After a preliminary word frequency analysis, some categories emerged from the 25 most frequent words in both UGC and WGC together and UGC alone. These categories were words referring to: destinations, positive feelings, and attractions. It should be noted that in the case of UGC titles, users usually do not mention again the name of the attraction they are reviewing, but conversely in the case of WGC the name of the attraction and the destination or location are central elements introduced. This emergent categorisation entails a relationship with the different components of tourist image: mainly cognitive and affective (Beerli & Martin, 2004).

4 Results

Content analysis has been conducted first to the whole OTR title (including both UGC and WGC) and then only to the UGC title (only to what the user has written). This allows the comparison between the global image transmitted to other users by the titles of the reviews (which includes both the webhost contents and the titles explicitly written by the tourist) with the image closest to the perceived image of tourists, where only what has been purposefully written by the user is taken into account. This will give us a double vision of the transmitted image of OTR titles vs. the perceived image present in OTR titles.

It is also important to note that the percentages seen in Tables 5 and 6 and Fig. 3 have been obtained from the total of words, also including the stop words in the black list. Although both samples are highly comparable, percentages are used to be more precise and to avoid absolute numbers. Concerning the 25 top keywords in OTR titles from Ile-de-France (Table 5) and Catalonia (Table 6), we observe very significant parallelisms, in the composition of words belonging to UGC + WGC and only UGC. The first similarity is that the most frequent word in both UGC + WGC and UGC are the capitals of the regions: Paris in the case of Ile-de-France and Barcelona (1st in UGC + WGC and 2nd in UGC) in the case of Catalonia. In both case studies, UGC + WGC titles mention much more the destination or location of the post than UGC. This can be seen with the UGC + WGC mentions of “Paris”, “France”, “Marne-la-Valee” and “Versailles” in Ile-de-France and “Barcelona”, “Spain”, and “Salou” in the case of Catalonia. However, in the case of UGC only “Paris” and “Barcelona” are mentioned among the top words and with a much weaker percentage of presence.

Moreover, in terms of the specific attractions that are mentioned, only titles including UGC + WGC include precise references to attractions among its top words (e.g., Eiffel Tower and Musee d’Orsay in Ile-de-France or Sagrada Familia and Park Guell in the case of Catalonia). In the case of UGC only generic words such as museum and art appear in the case of IdF and of city and park in Cat. This can be explained by the fact that when reviewers are within the template to review a specific attraction they do not usually mention again the name of the attraction in the title, whilst this information is one of the most added by webhosts. However it is remarkable that in the case of Cat there is a word that exceedingly relates to the major attractions of the region and to its tourist identity, which is the architect Gaudi. This could give us indications on the review attractions (e.g., Gaudi masterpieces, such as Sagrada Familia, Park Guell, and Casa Batllo).

Another important difference between UGC + WGC information and UGC alone concerns the mentioning of positive feelings. Very remarkably, UGC titles mainly consist of positive feelings and attributes (great, beautiful, amazing, best, worth, nice, good, wonderful, fun, must see, excellent, fantastic, and interesting), which mostly coincide with the two case studies. These positive words represent about half of the 25 top posts. Although in UGC + WGC titles some of the top words also refer to positive attributes (great, beautiful, best, amazing, must-see, nice, and

Table 5 Twenty-five top keywords in both (UGC and WGC) OTR titles

Ile de France (IdF)	UGC + WGC		Ile de France (IdF)	UGC	
Whole words	Unique		Whole words	Unique	
2,997,044	18,432	Percent	915,512	17,543	Percent
paris	167736	5.59671	paris	17345	1.89457
france	151063	5.04040	great	12389	1.35323
review	150158	5.01020	beautiful	8520	0.93063
tripadvisor	150003	5.00503	tour	7463	0.81517
tours	13923	0.46456	museum	5243	0.57269
great	12403	0.41384	amazing	5036	0.55007
eiffel tower	11729	0.39135	best	4943	0.53992
musee d’orsay	11068	0.36930	place	4634	0.50616
tour	9931	0.33136	visit	4303	0.47001
notre dame cathedral	8873	0.29606	worth	4110	0.44893
beautiful	8520	0.28428	see	3967	0.43331
marne-la-vallee	7845	0.26176	way	3686	0.40262
musee du louvre	7802	0.26032	experience	3521	0.38459
museum	6341	0.21158	nice	3443	0.37607
versailles	6188	0.20647	very	3427	0.37433
place	5906	0.19706	good	3226	0.35237
arc de triomphe	5643	0.18829	wonderful	3104	0.33905
best	5079	0.16947	view	3090	0.33752
luxembourg gardens	5066	0.16903	fun	3060	0.33424
amazing	5036	0.16803	day	2944	0.32157
musee	4953	0.16526	must see	2933	0.32037
disneyland park	4910	0.16383	art	2606	0.28465
river seine	4743	0.15826	time	2600	0.28399
bike	4702	0.15689	excellent	2389	0.26095
visit	4303	0.14357	fantastic	2240	0.24467

Source: 150,000 OTR titles on Ile de France

worth), these are much less mentioned as a whole and if compared to UGC titles, they have a much lower percentage of mentions. It is important to note that no negative feelings appear among the top words, confirming previous results, which pointed out that UGC posts, reviews, and review titles are eminently positive (De Ascaniis & Gretzel, 2012; Marine-Roig & Anton Clave, 2016a).

In this respect, it should be said that while seeing the words provided by UGC + WGC, it would be easily deduced that these words refer to Ile-de-France/Paris, or to Catalonia/Barcelona. It should be expected that among the top words would appear some identity elements of the destinations. However, this is not the case for the top words in the titles. One could not easily say (apart from the two words Paris or Barcelona) what destination these words are related to. Moreover, this could confirm the tendency that tourists only mention the most essential or synthesized places such as Paris or Barcelona, and no other destination appears

Table 6 Twenty-five top keywords in both (UGC and WGC) OTR titles

Catalonia (Cat)	UGC + WGC		Catalonia (Cat)	UGC	
Whole words	Unique		Whole words	Unique	
2,978,651	16,959	Percent	868,547	15,643	Percent
barcelona	159936	5.36941	great	15229	1.75339
spain	150627	5.05689	barcelona	11898	1.36987
review	150108	5.03946	tour	8680	0.99937
tripadvisor	150006	5.03604	amazing	7927	0.91267
sagrada familia	22337	0.74990	beautiful	6903	0.79478
great	15246	0.51184	place	5201	0.59882
tours	14779	0.49616	best	5071	0.58385
tour	12585	0.42251	nice	4815	0.55437
salou	10082	0.33848	worth	4567	0.52582
park guell	9915	0.33287	gaudi	4471	0.51477
day	7962	0.26730	visit	4461	0.51362
amazing	7927	0.26613	experience	4324	0.49784
beautiful	6904	0.23178	very	4252	0.48955
casa batllo	6368	0.21379	day	4174	0.48057
gothic quarter	5961	0.20012	good	4152	0.47804
barri gotic	5710	0.19170	fun	3859	0.44431
place	5208	0.17484	must see	3692	0.42508
best	5108	0.17149	way	3597	0.41414
camp nou	5093	0.17098	excellent	3509	0.40401
gaudi	4881	0.16387	see	3492	0.40205
nice	4815	0.16165	fantastic	3302	0.38018
experience	4672	0.15685	interesting	2703	0.31121
worth	4567	0.15332	wonderful	2639	0.30384
visit	4462	0.14980	city	2593	0.29854
good	4322	0.14510	park	2331	0.26838

Source: 150,000 OTR titles on Catalonia

among the top words. It is remarkable that in both case studies, in UGC + WGC titles, the name of the host (TripAdvisor) appears fourth in the list. It is also remarkable that in UGC + WGC the theme park DisneyLand and Marne-la-Vallee (where it is situated) are mentioned, and in the case of Catalonia, Salou is mentioned, which is the place associated to the theme park PortAventura (although this is not among the top 25 words).

Figure 3 shows in more detail the distribution of the broad categories (destinations, attractions, and good feelings), which emerged from the analysis of the 25 top posts in both case studies and in both UGC + WGC and only UGC titles. In the case of UGC + WGC in both IdF and Cat, the weight of the destination names or the locations represents about 5.8-5.7 of the total weight accounted by the 25 top words, demonstrating that webhosts contribute to the title's positioning by adding multiple information, especially destinations. Then in both case studies about 20 % of the

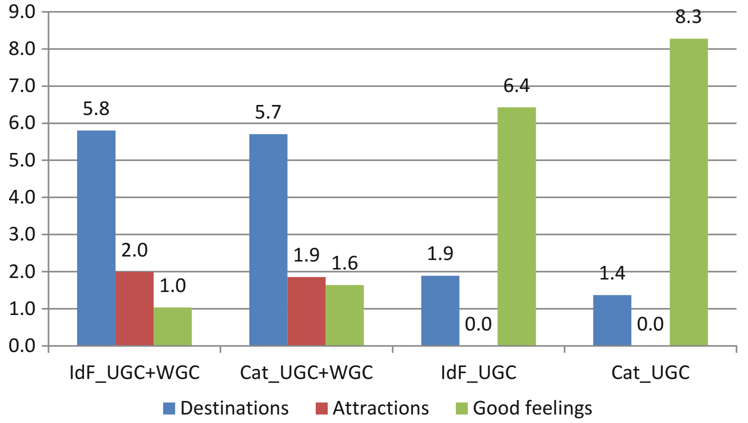


Fig. 3 Categories weight (%) in the whole-words set. *Source*: Twenty-five top keywords (Tables 5 and 6)

total weight of words corresponds to attractions. In both case studies, good feelings come third, but in the case of Cat these have a much higher weight (1.6) than in IdF (1.0).

Conversely, results are very different in the case of UGC. In both IdF and Cat, good feelings are by far the largest category, followed by the mentioning of a destination. However, no specific attraction is mentioned within the 25 top words, showing that the more functional aspect of tourist image, more related to attraction factors (Marine-Roig & Anton Clave, 2016a) is not importantly present among the main elements of titles written by users (although there are some generic elements: museum, art, city or specific:). This is a particularity of reviews as most probably users do not write again the name of the attraction when the template they write in is already about the said attraction, and they assume this paratextual information will be added by the webhost. This can be a main difference from individual travel blogs, for example, or other types of online information. Moreover, destinations are mentioned, but only the capital cities, with a weight of 1.9 in IdF and 1.4 in Cat. In the case of UGC titles, it is interesting that good feelings in Cat are 2 points above good feelings in IdF, which may give indications on the possible comparative orientation of review titles, texts, and user satisfaction. From these results, the affective component of image is more importantly present in the case of Cat and in a positive sense. This may indicate that the affective image component perception, and later transmission, of visitors to Cat is more positive and more related to positive feelings and attributes than in the case of IdF.

The fact that good feelings are so important in the case of the image transmitted by the tourists, confirms what De Ascaniis and Gretzel (2012) pointed out: that review titles are helpful and influential because they are opinionated information. Good feelings clearly denote a judgement or an opinion (a majority positive opinion) about the destination or attraction. These positive feelings and attributes may not only give information about the general orientation of the review to come,

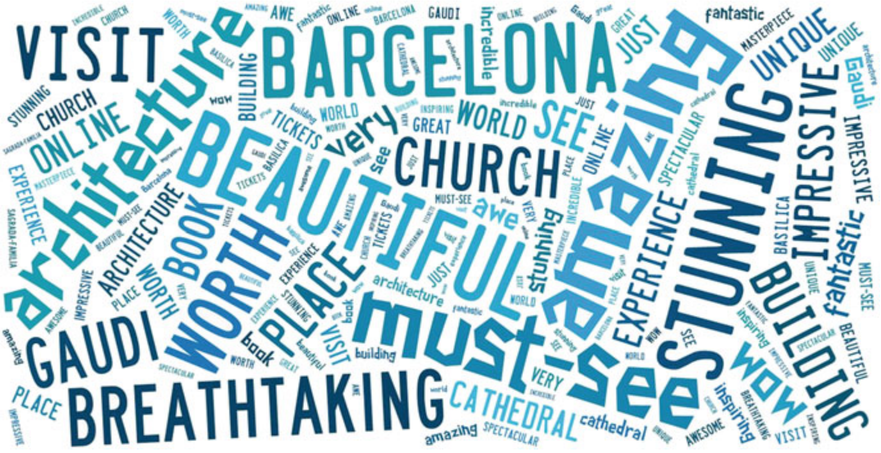


Fig. 4 Hundredth of the 30 most frequent keywords about La Sagrada Familia. *Source*: Random sample of 30,000 OTR titles written in English between 2011 and 2015

but also on the nature of the attraction: e.g., fun vs. beautiful vs. interesting, denote very different characteristics or attributes of the attraction of place, which shows both an elaborate perceived image of the tourist that will probably contribute in forming tourist images. In this respect, in UGC titles, the affective component of image (Marine-Roig & Anton Clave, 2016a) is much more present than in UGC + WGC.

Finally, it is relevant to mention that among the 25 top keywords in the case of UGC most of them were significant or univocal, supporting previous results (De Ascaniis & Gretzel, 2012). This significance of words can be related to the three categories that emerged. Among the 25 top words, most UGC + WGC words could be classified as 8.8 (IdF) and 9.2 (Cat), and in the case of UGC alone, 8.3 (IdF) and 9.7 (Cat). This means that more than 80 % of both UGC + WGC and UGC title words could be classified into destinations (this is more than 4 in 5 words), attractions, and good feelings. This aspect should be further researched in order to understand better the content of the review and perfect review analyses.

Concerning the specific study of the most popular attraction in Catalonia's OTRs, Fig. 4 shows a hundredth part of the 30 most frequent keywords in a random sample of 30,000 titles of OTRs on the Basilica of La Sagrada Familia (Barcelona) written in English between 2011 and 2015 and represents it graphically. These results of analysis of titles are very useful for analysing the affective component of the perceived image (Marine-Roig & Anton Clave, 2016a) because they prominently reflect the feelings of visitors. Results are also consistent with previous similar studies based on content analysis of the entire OTR (title and body of the writing), such as a quantitative content analysis of a sample of 18.884 OTRs from TripAdvisor and VirtualTourist written in English (Marine-Roig, 2015b) and a sample of 7481 OTRs written during 2014 (Marine-Roig & Anton Clave, 2015).

In all three case studies three groups of related keywords can be highlighted: Feelings (positive adjectives and other good feelings), material structure (building, cathedral, church, Gaudi, architecture, . . .) and access (book, tickets, online, . . .). The main difference is that the keywords related to good feelings acquire more weight in the analysis of the titles, whereas the issues related to the purchase of tickets and access to the basilica (Marine-Roig & Anton Clave, 2015) go on to fill third place.

5 Conclusions

Within the context of the growing amount of online information and increasing OTR creation and use, paratextual elements become crucial to: help with the information search; position an OTR; assess tourism decision-making; help to understand and take advantage of the information contained in the text; and evaluate and create expectations. Tourist UGC have been found to be seen as trustworthy and influence destination image formation through the eWOM effect. In this context, OTR paratextual elements can be considered influential elements for destinations image formation. In framed contexts such as review hosting websites, the paratextual information the reader sees, which will influence their pre-trip image formation, is not only what has been written by the user but also what has been added by the webhost, which also influences the positioning of the post. Hence, this study has contributed to adding value to the WGC, distinguish it from UGC, and advocates taking it into account for OTR and online image studies. This research found that the UGC information created by the user is different and should be distinguished from the global result the final reader sees, which includes WGC. In the case of the top word analysis of UGC + WGC, some relevant destinations of the region appear (especially the capital city) and in the case of UGC, the only destination name mentioned is the capital city of the region. In terms of the mentioning of specific attractions, this could only be seen in the case of UGC + WGC, deducing that it is the webhost who adds the major part of information concerning the attraction and its specific location.

Both UGC and UGC + WGC contained positive feelings among the most mentioned words, confirming previous results (De Ascaniis & Gretzel, 2012; Marine-Roig, 2016a) in that most reviews are positively oriented. However, in the case of UGC the presence of good feelings was much more prominent.

In this respect the nature of OTR may be eminently different from the content of its text. Although, in line with Marine-Roig and Anton Clave (2015, 2016a) and De Ascaniis and Gretzel (2012) this study's results show that OTR paratextual content shows and focuses only on very specific features—the most representative, the feelings, and must-see attractions—,a great difference was found in this respect between UGC and UGC + WGC. While UGC + WGC really do contain the must-see attractions, feelings, and destinations in a prominent place (among the top words), UGC is mainly focused only on feelings. This trend was confirmed with the specific analysis of the Sagrada Familia titles, which eminently contained positive feelings or attributes. Therefore, in terms of the analysis of OTRs as sources for destination image formation, this study suggests distinguishing between the UGC paratextual information, which can be considered and understood as an elaborately perceived image of the tourist experience and satisfaction, and the final review title with which the reader is in contact with. This includes both UGC and WGC, and can be considered the final transmitted image of the destination that other tourists will encounter when looking for travel information, and that as we have seen, has different characteristics from UGC.

The fact that UGC titles written by users most strongly contain feelings (positive feelings), may be relevant for destination image studies, as UGC titles in OTR will be useful to assess the affective perceived image component, but may not be so useful in order to assess the functional or conative components. Conversely, the analysis of UGC + WGC may give more insights into the spatial image component, as well as to the functional image component (with attraction factors) to a greater extent, as well as to the affective image component. Moreover, this study confirms previous results (De Ascaniis & Gretzel, 2012) in that most words in review titles are significant, even to a greater extent (4 out of 5), which reaffirms their interest as objects of study. When studies analyse OTRs or review titles, it is often to assess the influence they may exert on other users and their decision-making, due to the greater attention they command. Thus, this study suggests that in the study of online UGC framed sources such as OTR, WGC must be taken into account and given the role it corresponds as part of the final review that the user reads that will influence search results and the final tourist decision-making. This study should be combined with other relevant information such as the assessed review helpfulness and readability (Fang et al., 2016). Moreover, WGC may influence the way reviewers write the OTR titles, for example, without mentioning the name of the attraction or the location in the title because the webhost is assumed to do so. Therefore, to understand UGC posts and how users express their perceived images in titles and even inside the review itself, the specific webhost framework and the WGC should be thoroughly studied. Future studies should determine the actual influence of WGC paratextual information in the way that tourists express themselves and in the contents they post in OTR titles and text.

This research contributes tour understanding of assessing how the paratextual contents of OTR are similar across different destinations. With a massive analysis of OTR of the two most touristic regions in continental Europe, this study shows that a certain pattern of contents emerges in UGC and UGC + WGC in titles, at least as it pertains to the words that have a major presence and weight in attraction factors, destinations, or good feelings. This framework consisting of *Destination choice, Webhost selection, Data collection, Web data mining, Data arranging, Parser settings, and Categorisation* has been suitable and effective for the purposes of this study and allows multiple classifications and organization of data. Specifically, it has enabled us to analyse, distinguish, and categorise the content of the Heading title (UGC), the Page title (UGC + WGC), and the Title hyperlink (WGC) information of the OTR in a considerable way. The results obtained on the perceived and transmitted image of the destination can be of great interest to national tourism (NTO) and destination marketing (DMO) organisations to improve the tourism supply chain.

So far, researchers have analysed specifically textual and paratextual elements of OTRs. In the theoretical framework proposed in Sect. 2, it is observed that the review forms a whole with the paratextual elements that surround it. That is, you cannot place or even understand content written by the reviewer (UGC) without reference to other webpage elements (WGC). This work has been limited to paratextual elements visible by Internet users, but all of them are reproduced in the HTML meta-tags to be read by search engines and there is no need to insist on the great importance of Internet searching in travel planning. This study used two random samples of a given size (150,000 OTRs) for comparative analysis. The randomised algorithm used is highly reliable because it generates decimal numbers between zero and one with 15 decimal places, which makes it almost impossible for any number to be repeated. But the randomness does not guarantee that there is a proportional population representation (reviewers, dates, destinations, attractions, etc.) in the sample.

Acknowledgements This work was supported by the Spanish Ministry of Economy and Competitiveness [Grant id.: MOVETUR CSO2014-51785-R].

References

Afzaal, M. & Usman, M. (2015). *A novel framework for aspect-based opinion classification for tourist places*. Proceedings of Tenth international conference on digital information management (pp. 1–9). doi:10.1109/ICDIM.2015.7381850

Alacovska, A. (2015). Legitimacy, self-interpretation and genre in media industries: A paratextual analysis of travel guidebook publishing. *European Journal of Cultural Studies, 18*(6), 601–619. doi:10.1177/1367549415572318.

Azariah, D. R. (2011). Whose blog is it anyway? Seeking the author in the formal features of travel blogs. In S. Adams et al. (Eds.), *Proceedings of eleventh humanities graduate research conference*. Perth, WA: Curtin University, Faculty of Humanities.

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management, 53*, 148–162. doi:10.1016/j.tourman.2015.09.004.

Banerjee, S., Chua, A. Y. K., & Kim, J. J. (2015). *Using supervised learning to classify authentic and fake online reviews*. Proceedings of the 9th International conference on ubiquitous information management and communication (article 88). New York: ACM Digital Library. doi:10.1145/2701126.2701130

Beerli, A., & Martin, J. D. (2004). Factors influencing destination image. *Annals of Tourism Research, 31*(3), 657–681. doi:10.1016/j.annals.2004.01.010.

Cao, K. & Yang, Z. (2016). A study of e-commerce adoption by tourism websites in China. *Journal of Destination Marketing & Management*. In Press. doi:10.1016/j.jdmm.2016.01.005

Caton, K., & Almeida, C. (2008). Closing the hermeneutic circle? Photographic encounters with the other. *Annals of Tourism Research, 35*(1), 7–26. doi:10.1016/j.annals.2007.03.014.

CRT (2016). *Bilan de l'activite touristique de l'annee 2015* [Balance of tourist activity in 2015]. Comite Regional du Tourisme Paris Ile-de-France. Retrieved February 29, 2016, from http://pro.visitparisregion.com

CTB. (2016). *Press Pack '16*. Catalan Tourist Board. Retrieved February 29, 2016, from http://www.act.cat/press-pack

Dann, G. M. S. (2014). Why, oh why, oh why, do people travel abroad? In N. K. Prebensen, J. S. Chen, & M. S. Uysal (Eds.), *Creating experience value in tourism* (pp. 48–62). Oxfordshire: CABI.

De Ascaniis, S., & Gretzel, U. (2012). What's in a travel review title? In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism 2012* (pp. 460–470). Wien: Springer. doi:10.1007/978-3-7091-1142-0_43.

De Ascaniis, S., & Gretzel, U. (2013). Communicative functions of online travel review titles: A pragmatic and linguistic investigation of destination and attraction OTR titles. *Studies in Communication Sciences, 13*(2), 156–165. doi:10.1016/j.scoms.2013.11.001.

Desrochers, N., & Apollon, D. (Eds.). (2014). *Examining paratextual theory and its applications in digital culture*. Hershey, PA: IGI Global.

Dickinger, A., & Lalicic, L. (2016). An analysis of destination brand personality and emotions: A comparison study. *Information Technology & Tourism, 15*(4), 317–340. doi:10.1007/s40558-015-0040-x.

Eurobarometer. (2015). *Flash Eurobarometer 414: Preferences of Europeans towards tourism*. Brussels: European Commission.

Eurostat. (2015). *Eurostat Regional yearbook 2015*. Luxembourg: Publications Office of the European Union.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management, 52*, 498–506. doi:10.1016/j.tourman.2015.07.018.

Gartner, W. C. (1994). Image formation process. *Journal of Travel and Tourism Marketing, 2* (2–3), 191–216. doi:10.1300/J073v02n02_12.

Genette, G. (1997). *Paratexts: Thresholds of interpretation*. New York: Cambridge University Press.

Grabner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism 2012* (pp. 460–470). Wien: Springer. doi:10.1007/9783709111420_40.

Gray, J. (2015). Afterword: studying media with and without paratexts. In L. Geraghty (Ed.), *Popular media cultures: Fans, audiences and paratexts* (pp. 230–237). Basingstoke: Palgrave Macmillan.

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In P. O'Connor, W. Hopken, & U. Gretzel (Eds.), *Information and communication technologies in tourism 2008* (pp. 35–46). Wien: Springer. doi:10.1007/978-3-211-77280-5_4.

IDESCAT. (2016). *Statistical yearbook of Catalonia 2015*. Statistical Institute of Catalonia. Retrieved February 29, 2016, from http://www.idescat.cat/en/

Jalilvand, M. R., Samiei, N., Dini, B., & Manzari, P. Y. (2012). Examining the structural relationships of electronic word of mouth, destination image, tourist attitude toward destination and travel intention: An integrated approach. *Journal of Destination Marketing & Management, 1* (1–2), 134–143. doi:10.1016/j.jdmm.2012.10.001.

Johnson, P. A., Sieber, R. E., Magnien, N., & Ariwi, J. (2012). Automated web harvesting to collect and analyse user-generated content for tourism. *Current Issues in Tourism, 15*(3), 293–299. doi:10.1080/13683500.2011.555528.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. doi:10.1016/j.bushor.2009.09.003.

Kladou, S., & Mavragani, E. (2015). Assessing destination image: An online marketing approach and the case of TripAdvisor. *Journal of Destination Marketing & Management, 4*(3), 187–193. doi:10.1016/j.jdmm.2015.04.003.

Krawczyk, M. & Xiang, Z. (2016). Perceptual mapping of hotel brands using online reviews: a text analytics approach. *Information Technology & Tourism, 16*(1), 23–43. doi:10.1007/s40558-015-0033-0

Lai, L. S. L., & To, W. M. (2015). Content analysis of social media: a grounded theory approach. *Journal of Electronic Commerce Research*, *16*(2), 138–152. Retrieved February 29, 2016, from http://www.jecr.org/node/466

Li, Y. R., Lin, Y. C., Tsai, P. H., & Wang, Y. Y. (2015). Traveller-generated contents for destination image formation: Mainland China travellers to Taiwan as a case study. *Journal of Travel & Tourism Marketing, 32*(5), 518–533. doi:10.1080/10548408.2014.918924.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*, 458–468. doi:10.1016/j.tourman.2007.05.011.

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin: Springer.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management, 47*, 140–151. doi:10.1016/j.tourman.2014.09.020.

Marine-Roig, E. (2015a). Identity and authenticity in destination image construction. *Anatolia—An International Journal of Tourism and Hospitality Research, 26*(4), 574–587. doi:10.1080/13032917.2015.1040814.

Marine-Roig, E. (2015b). Religious tourism versus secular pilgrimage: The basilica of La Sagrada Familia. *International Journal of Religious Tourism and Pilgrimage, 3*(1), 25–37.

Marine-Roig, E., & Anton Clave, S. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management, 4*(3), 162–172. doi:10.1016/j.jdmm.2015.06.004.

Marine-Roig, E., & Anton Clave, S. (2016a). Affective component of the destination image: A computerised analysis. In M. Kozak & N. Kozak (Eds.), *Destination Marketing: An international perspective* (pp. 49–58). New York: Routledge.

Marine-Roig, E., & Anton Clave, S. (2016b). A detailed method for destination image analysis using user-generated content. *Information Technology & Tourism, 15*(4), 341–364. doi:10.1007/s40558-015-0040-1.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management, 19*(7), 754–772. doi:10.1080/19368623.2010.508007.

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software*. Sebastopol, CA: O'Reilly Media.

Peel, V., & Sorensen, A. (2016). *Exploring the use and impact of travel guidebooks*. Bristol: Channel View Publications.

Serna, A., Marchiori, E., Gerrikagoitia, J. K., Alzua-Sorzabal, A., & Cantoni, L. (2015). An auto-coding process for testing the cognitive-affective and conative model of destination image. In I. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 111–122). Cham: Springer. doi:10.1007/978-3-319-14343-9_9.

Schmunk, S., Hopken, W., Fuchs, M., & Lexhagen, M. (2014). Sentiment analysis: Extracting decision-relevant knowledge from UGC. In Z. Xiang & L. Tussyadiah (Eds.), *Information and communication technologies in tourism 2014* (pp. 253–265). Cham: Springer. doi:10.1007/978-3-319-03973-2_19.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing, 32*(5), 608–621. doi:10.1080/10548408.2014.933154.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation, 7*(17). Retrieved February 29, 2016, from http://PAREonline.net/getvn.asp?v=7&n=17

Stepchenkova, S., Kirilenko, A. P., & Morrison, A. M. (2009). Facilitating content analysis in tourism research. *Journal of Travel Research, 47*(4), 454–469. doi:10.1177/0047287508326509.

Wang, F., & Wu, Y. (2012). Mining market trend from blog titles based on lexical semantic similarity. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 261–273). Berlin: Springer. doi:10.1007/978-3-642-28601-8_22.

Wang, Y., Chan, S. C., Ngai, G., & Leong, H. V. (2013). Quantifying reviewer credibility in online tourism. In H. Decker et al. (Eds.), *DEXCA 2013* (pp. 381–395). Berlin: Springer. doi:10.1007/978-3-642-40285-2_33.

Xiang, Z., Magnini, V. P., & Fesenmaier, D. R. (2015). Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet. *Journal of Retailing and Consumer Services, 22*, 244–249. doi:10.1016/j.jretconser.2014.08.005.

Xiang, Z., Pan, B., & Fesenmaier, D. R. (2014). Foundations of search engine marketing for tourist destinations. In S. McCabe (Ed.), *The Routledge handbook of tourism marketing* (pp. 505–519). New York: Routledge.

Xiang, Z., Schwartz, Z., Gerdes, J. H., Jr., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130. doi:10.1016/j.ijhm.2014.10.013.

Xiang, Z., Wober, K., & Fesenmaier, D. R. (2008). Representation of the online tourism domain in search engines. *Journal of Travel Research, 47*(2), 137–150. doi:10.1177/0047287508321193.

Yuan, Y. L., & Ho, C. I. (2015). *Rethinking the destination marketing organization management in the big data era*. Proceedings of the ASE BigData & Social Informatics 2015 (article 60).

Conceptualizing and Measuring Online Behavior Through Social Media Metrics

Bing Pan and Ya You

1 Introduction

Social media is almost ubiquitous these days. As of 2015, 70 % of the U.S. population has at least one social networking profile; the number of worldwide social media users is predicted to grow to 2.5 billion by 2018 (Statista, 2015). Businesses have realized the power of social media and considered social media marketing as an important part of their integrated marketing communication effort. According to a recent Forrester research, the expenditure in social media marketing would increase from $7.52 billion in 2014 to $17.34 billion in 2019 in the U.S. (Forrester Research, 2014). With social media, brands are able to increase brand recognition, enhance brand loyalty, and improve sales conversion rate. In particular, fans and followers would become voluntary advocates for the brand they love on their social networks. Undoubtedly, social media will continue to have a significant impact on businesses with the ability to reach out and communicate with their target customers on a personal level and on a daily basis.

However, understanding how to measure social media effectiveness is far behind the exploding speed of social media usage by marketers. For example, the 2014 CMO Survey reveals that only 15 % of the responded marketers could quantify the impact of social media on their business (CMO Survey, 2014). Moreover, the 2015 Social Media Marketing Industry Report shows only 42 % of over 3700 surveyed marketers agreed they are able to measure the return on investment of their social media activities (Stelzner, 2015). These survey results are not surprising given that the lack of categorization and standardization of metrics across social media

B. Pan (✉)
Penn State University, State College, Pennsylvania, USA
e-mail: bingpan@psu.edu

Y. You
College of Charleston, Charleston, USA

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_12

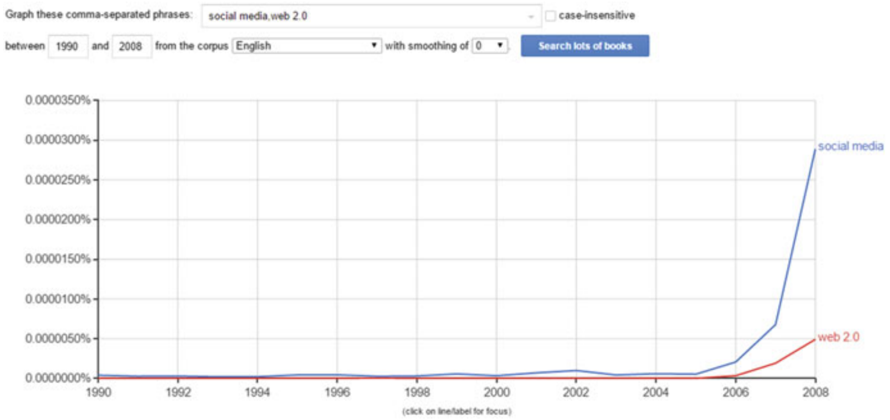
platforms limits managers' ability to truly measure their value. Social media includes not only the comments, shares, and links posted on diverse platforms such as Facebook, YouTube, or blogs, but also organizations and people who follow or subscribe the online community. Therefore, an effective and integrated framework of social media metrics is important to understand how social media really works in creating values for businesses. Furthermore, the social media metrics can be used to develop dashboard-like interfaces and create key metrics, which reflects the key drivers and outcomes within the organization, diagnoses excellent or poor performance, and facilitates managerial decision-making.

2 The Evolution of Social Media

The need for measurements is coincided with the commercialization of social media. In the earliest days, users logged in interesting websites they found on the web and recorded their own lives and musing, on a web page (Oxford English Dictionary, 2015). Apparently the need to capture how many people read the web page and for how long was insignificant. When businesses adopted social media to maximize their revenue and profit, the measurement starts to gain significance.

Social media has gone through many stages of evolution. The birth of social media probably happens in 1997, when the first social site sixdegrees.com was established (Boyd & Ellison, 2007); some web users started first weblogs in the same year (Oxford English Dictionary, 2015) and the word "social media" appeared (Bercovici, 2010). However, not until 2003–2005, did the concept of "social media" start to gain popularity (Fig. 1). The first academic report on social media probably appeared in 2003 (Kline & Arlidge, 2003) and the first conference on social media was held in 2004 (Thompson, 2004). It's really 2005 when the concept started to take off in books and reports (Fig. 1). The concept of "Web 2.0" gained popularity after 2006, with the wide adoption of social media platforms and the production of great amount of social information (Fig. 1). Especially in the sharing economy today, when new businesses thrive on the number of users and the way they connected with each other (Cannon & Summers, 2014), the measurement of sharing activities might be the single determining factor in the livelihood of those companies. Thus, the measurement of social media becomes more crucial than ever.

The forms of social media are also evolving. The early date of text-based blogs was slowly expanded to images and videos; the platforms transformed from the sharing of text and pictures to social networks where users got connected. As a result, the way of measuring social media is also evolving: from the early days of the number of views and reads to later time of network measurement in terms of friends, fans, and connections. However, no matter how much the realm of social media has changed, we argue that the underlying structure is still the connections between users and information artifacts (Pan & Crotts, 2012). In this chapter, we intend to propose a conceptual framework for social media and couch the



Graph these comma-separated phrases: social media,web 2.0 □ case-insensitive

between 1990 and 2008 from the corpus English with smoothing of 0

Fig. 1 The concept of social media and Web 2.0 on Google Books. *Source: Google Books NGram viewer, retrieved January 6, 2016*

measurements of social media in this integrated framework. We argue that various measurements are the representations of different aspects of a multidimensional network and they can be potentially measured and combined to produce metrics for monitoring business performance and optimizing operations.

3 A Conceptual Framework

Many behavioral and psychological theories can explain and abstract the complex phenomenon of social media. The multi-dimensional social network theory can capture the nature of the interaction between users and information artifacts and is the inspiration for our social media framework (Contractor, 2009). In this section, we outline a multi-dimensional network to represent social media content and actors. Different from other mass media, social media is based on the underlying structure of the social networks composed of users. It involves multiple relationships, including relationships between actors (users) and actors (users), information artifacts with other information artifacts, and actors and information artifacts. A multi-dimensional network framework can represent the nature of this communication platform (Contractor, 2009). This network is composed of nodes and links. The nodes are of two types which we term **actors** (a Facebook account, a twitter account, etc.) and **memes** (a text post, a tweet, a video, a picture, etc.). There are multiple types of relationships among these (Fig. 2).

Figure 2 shows one example of a multi-dimensional social media network, using Twitter platform as an example. The actors are four different twitter accounts: John, Steven, Tom, and Company A; three memes are a video, a tweet, and a retweet. First, there are relationships between actors: John and Steven are following Company A as followers; Tom is following Steven. There are relationships between

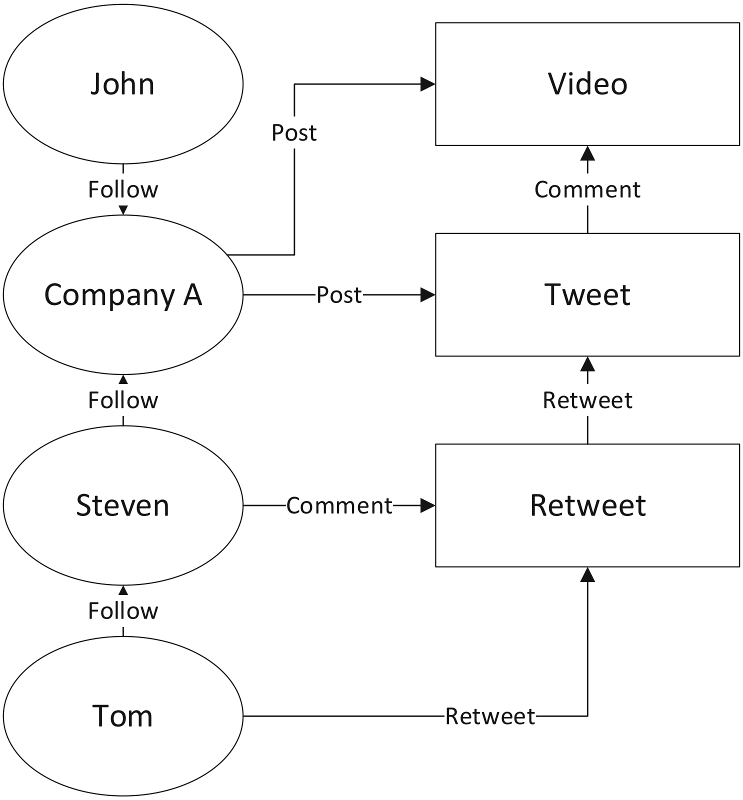


Fig. 2 Example of multi-dimensional network framework

actors and memes: Company A posted a video on Twitter and later made a tweet about it. The video was retweeted by Steven later; Tom further retweeted that retweet made by Steven. So in this way, Tom is indirectly connected Company A with a retweet. There are also relationships between memes: the tweet is on the video, and the retweet is about video indirectly. Thus, almost all the relationships, from following to retweeting, are different types of directional connections. With this model, various measurements of social media are actually different representations of this nodes-links network with different levels of aggregation. For example, one can calculate the following metrics for Company A: the number of followers; the number of memes; the number of tweets and retweets generated by the video; the number of interactions of the followers with the content Company A produced. Thus, this multi-dimensional network provides a valid way to abstract all the interactions in the social media sphere.

4 Social Media Measurement

In this framework, we can calculate all the web-based and marketing-oriented measurements. Web-based measurements refer to those behavioral indices on the level of user interaction with web content; marketing-oriented measurements connect those web-based measurements in relation to business performance. They can refer to the same numbers but contain different meanings. Table 1 describes the different components of this multi-dimensional network, different types of relationship and representations, as well as commonly used web-based and marketing-oriented measurements. For example, actor-actor connections are the relationship on “friends with” or “following”; meme-meme relationships are “hyperlinking”, or “commenting on”; actor-meme connections are “tweeting”, “posting”, or “tagging”. Thus, reach can be measured by the number of friends or followers; audience engagement can be calculated by the proportion of users who comment on a meme to the total number of views on the same meme. Response rate is the percentage of inquiries or complaints responded by company’s official account.

4.1 Connections with Other Social Media Measurements

There are several measurement frameworks proposed in recent years. In this section, we use the aforementioned framework to explain and connect with these.

Table 1 Social media measurements and multi-dimensional network

Network	Type	Representation	Web-based measurements	Marketing-oriented measurements
Nodes	Actors	Facebook account, Twitter account, Instagram account		Social media presence
	Memes	A text post, a picture, a video, a place	Posts; sentiments; articles;	Social media involvement
Connections	Actor—Actor	Friend with; follow	Followers; Likes; friends; fans; group members;	Awareness; Reach; Lead generation; Brand liking
	Meme—Meme	Hyperlink to; explaining	Inbound links, comments	Response rate
	Actor—Meme	Tweeting; retweeting; posting; commented on	Shares; hides; retweets; tweets; check-ins; views; impressions; redemptions; bookmarkings; response time; clickthroughs	Recommendations; Virality; Audience engagement

Table 2 Social media framework by Kaushik (2011)

Metrics	Facebook measurement	Multi-dimensional network measurements
Conversion rates	The number of audience comment per post	Commenting connections between memes and memes
Amplication rate	The number of shares per post	The average number of sharing connections between actors and memes
Applause rate	The number of likes per post	The average number of like connections per post meme

We argue that they are all the different representations of the same multi-dimensional network.

4.2 Kaushik (2011) Social Media Framework

Avinash Kaushik has written a few books on web analytics and is an influential figure in the arena of online marketing (Kaushik, 2007, 2009). In his influential blog site named Occam's razor, Kaushik proposed a few more meaningful measurements such as conversion rates, amplication rate, applause rate, and economic value (Kaushik, 2011). Table 2 shows the social media framework developed by Kaushik in connection with our framework. Using the conceptual framework proposed above, one can easily measure these metrics using the numerical values in the multi-dimensional network. However, economic value is hard to measure with our conceptual model since it is to do with the assigned value of each meme.

4.3 Zarrella's Social Media Metric Framework

Dan Zarrella, a web developer and a social media marketing author, has probably written the first textbook on social media (Zarrella, 2009). In the book, he argues that a marketer should create measureable objectives first and then map them to the metrics second. Those business objectives are exposure, engagement, influence, impact, and advocacy. Table 3 reproduces those objectives with related metrics. For example, exposure could be Total Followers. In Fig. 2, we can easily calculate it by counting the number of inbound links to Company A; if visit is one type of linkage, Page Visits can be calculated by counting the number of links. On the other hand, one metric for engagement is clickthroughs. This can be measured by a meta-metric—the ratio of the number of clicks on one specific meme to the number of views. However, there are a few metrics which can be hard to obtain by the calculation of the metrics through the multi-dimensional network, which were highlighted by italic in Table 3: for example, purchase consideration, likelihood to recommend, number of sales leads, conversion rates, sales, and repeat sales.

Table 3 Social media metrics framework by Zarrella (2009)

Business objective	Metrics	Multi-dimensional network
Exposure	Page Visits; Visitors; Unique Visitors; Visits per Channel (Source); Reach; Total Followers (Audience Count); Opportunity-to-See; CPM (Cost per Thousand Exposures)	The count of behavior on official actors
Engagement	Repeat Visits; Time Spent on Site; Total Interactions on Post/Page; Likes; Shares; Comments; +1 s; Clickthroughs; Number of Followers; Mentions; People talking about Brand	The count or meta-count of behavior on actors
Influence	Links; Association with Brand Attributes; Sentiment (Positive, Neutral, Negative); *Purchase Consideration*[a]*; Likelihood to Recommend; Net Promoter Score; Klout Score*	The count of connections between memes
Impact	New Subscribers; Number of Referrals to Website; Downloads; Number of App Downloads; *Abandoned Shopping Carts; Number of Sales Leads; Conversion Rates; Sales; Repeat Sales*	The number of actors and the number of behavior on Memes; the meta-score of behavior
Advocacy	Ratio Mentions to Recommendations; Number of Brand Fans/Advocates; Content of Ratings/Reviews; Employee Ambassadors; *Online Ratings*	The meta-score of connections between memes and actors, memes and memes, and actors and actors

[a]*Note*: metrics that can be hard to obtain by calculation of the metrics were highlighted by *italic*.

Currently these metrics are likely to be off-line and disconnected with proposed network.

4.4 Connecting with Business Performance

Though the aforementioned frameworks are different in the number of terms and measurement, the fundamental question is: how do we measure the economic values of social media? We argue that the total numerical matrix of all these connections and actors are the total social value of the network, including the number of followers, the number of posts, retweets, shares, etc. By measuring these numerical values in time as well as tracking the revenue and profits of the company, one will be able to build the economic values of these through econometric methods such as Granger modeling techniques (Granger, 1969) or graph structure learning (Heckerman, Geiger, & Chickering, 1995).

One example can demonstrate the measurement of economic value of a social media content in a viral video case. For example, Disney Cruise Line produced a viral video and posted it on its YouTube channel. This video will be liked, shared, tweeted, blogged, discussed, and even modified and reposted. The numbers of those connections to this meme, the numbers of likes, comments, and followers, are different measurements of the reach, involvement, and amplification of this video meme. In addition, the subscribers to DisneyCruiseLineNews increase by thousands of fans. All these will lead to a change of the ego-network structure of the social media platform. Utilizing historic time series models, one can validate that this video generates 2 % of cruise customers in the following year with a monetary value of $2 million; the profit generated is $200,000 and the Return on Investment is 150 %.

5 Conclusions

In this chapter, we proposed a conceptual framework of social media and couched the social media metrics under this multi-dimensional network. We argue that social networks are multi-dimensional networks of actors and memes. All the different measurements of social media could be calculated based on the numerical values of these nodes or connections. Furthermore, more meaningful measurements, such as conversion, amplication, and applause, could be measured through this multi-dimensional network. The amalgamate of all these measurements of this network could be tracked along a time scale and analyzed with the revenue and profit of a business's performance. This way, the economic values of different social media could be quantified.

The arena of social media metrics is as complex, dynamic, and evolving as any other Internet phenomena, if not more. However, we believe the proposed framework could explain and simplify these metrics in an automatic process of calculation. However, the next step for this line of research is to construct a multi-dimensional network with a sample social network and test these assumptions. This require a tremendous amount of hardware, software, and programming support. Nonetheless, the potential of its application in the marketing field will be tremendous.

References

Bercovici, J. (2010). Who coined "social media"? Web pioneers compete for credit. *Forbes. Disponível.*

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210–230.

Cannon, S., & Summers, L. H. (2014). How Uber and the sharing economy can win over regulators. *Harvard Business Review*, 13.

CMO Survey. (2014). CMO Survey report: Highlights and insights. Retrieved from https://cmosurvey.org/results/survey-results-february-2014/

Contractor, N. (2009). The emergence of multidimensional networks. *Journal of Computer Mediated Communication, 14*(3), 743–747.

Forrester Research. (2014). Forrester Research Social Media Forecast, 2014 To 2019 (US), Q3 2014 Update. Retrieved from https://www.forrester.com/report/Forrester+Research+Social+Media+Forecast+2014+To+2019+US+Q3+2014+Update/-/E-RES119281

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society, 37*, 424–438.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197–243.

Kaushik, A. (2007). *Web analytics: An hour a day (W/Cd)*. New Delhi: Wiley-India.

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability & science of customer centricity*. Indianapolis, IN: Sybex.

Kaushik, A. (2011). Best social media metrics: Conversation, amplification, applause, economic value. Retrieved from http://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/

Kline, S., & Arlidge, A. (2003). *Online gaming as emergent social media: A surve*y. Media Analysis Laboratory, Simon Fraser University. Online: http://www. sfu. ca/medialab/onlinegaming/report.html

Oxford English Dictionary. (2015). *weblog, n.* Oxford University Press.

Pan, B., & Crotts, J. C. (2012). Theoretical models of social media, marketing implications, and future research directions. In E. C. M. Sigala & U. Gretzel (Eds.), *Social media in travel, Tourism and hospitality: Theory, practice and cases* (pp. 73–83). Burlington, VT: Ashgate.

Statista. (2015). Statistics and facts about Social Networks. Retrieved from http://www.statista.com/topics/1164/social-networks/

Stelzner, Michael A. (2015). 2015 social media industry report. Retrieved from https://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2015.pdf

Thompson, C. (2004). Chris Shipley announces BlogOn 2004: The business of social media conference to explore rising business opportunities in blogging and social networking [Press release]. http://www.prnewswire.com/news-releases/chris-shipley-announces-blogon-2004-the-business-of-social-media-conference-to-explore-rising-business-opportunities-in-blogging-and-social-networking-74420252.html

Zarrella, D. (2009). *The social media marketing book*. Sebastopol, CA: O'Reilly Media.

Part V
Case Studies in Web and Social Media Analytics

Sochi Olympics on Twitter: Topics, Geographical Landscape, and Temporal Dynamics

Andrei P. Kirilenko and Svetlana O. Stepchenkova

1 Introduction

A successful destination must be favorably positioned in the public mind. Mega-sporting events are an integral part of destination marketing because they attract worldwide visibility and publicity, provide a nucleus for a positively framed public discourse, and have the potential to improve attitudes toward countries and the destinations within them. At the same time, a host country's domestic and international policies can negatively affect destination perceptions because these policies are the subjects of representation, discussion, and interpretation, not only in the traditional mass media but also in social networks and user-generated media, such as Twitter.

In 2012, Russia entered the list of the world's top ten tourism destinations for the first time, with 26 million international arrivals, but discrepancies between inbound and outbound tourism expenditures indicate that the tourism industry in Russia remains in the developing stages (United Nations World Tourism Organization, 2013). With the Sochi Olympics in 2014 and the FIFA World Cup in 2018, the Russian government has been investing in tourism infrastructure and marketing communications to improve the country's economic situation and image (Hilton Worldwide, 2014).

The Sochi Olympics were held on February 7–23, 2014, with the opening rounds of certain events held on February 6. The dates for the Paralympics were March 7–16, 2014. In preparation for the 2014 Sochi Winter Games, organizers focused on modernizing the telecommunications, electric power, and transportation infrastructures in the region. Though it was estimated at USD 12 billion in the original Olympic bid, the budget for the Sochi Olympics increased to more than USD

A.P. Kirilenko (✉) • S.O. Stepchenkova
University of Florida, Gainesville, FL, USA
e-mail: andrei.kirilenko@ufl.edu

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_13

51 billion, ultimately surpassing the cost of the 2008 Summer Olympics in Beijing (USD 44 billion) as the most expensive Olympics in history. Considering the greater number of sports event at the Beijing Olympics, this difference is even more striking, with USD 520 million per event spent at Sochi vs. USD 132 million at Beijing. The cost of the Olympic Games themselves was estimated USD 6.5 billion; the remainder consisted of the costs of Sochi infrastructural projects (Oliphant, 2013).

Four main factors contributed to such a colossal deviation from the original budget: the geographic location of Sochi, with its sub-tropical summer climate; a hefty amount of construction, which was required to conduct the Games in the area (e.g., the 31-mile mountain road between the Olympic venues); the security required to suppress terrorist activity in the region; and corruption (Taylor, 2014). With regard to the scale of the last factor, in 2013, Russia ranked 127th of 175 countries on the Transparency International Corruption Perception Index with a score of 28 out of 100, where 100 indicated "no corruption" (Transparency International, 2013).

Russia bet on a number of benefits from hosting the Games, such as developed infrastructure, an improved reputation, and an enhanced image. With regard to the country's image, releasing "high profile" dissidents Mikhail Khodorkovsky (a -Russian oligarch and the former owner of the Yukos oil company) and Nadezhda Tolokonnikova and Maria Alyokhina (members of the Pussy Riot punk rock group) prior to the Games was considered an attempt by the Russian government to control the narrative about the Sochi Olympics in a situation in which the relationships between Russia and the West had been noticeably "cooling off", which started with the Magnitsky Act and the Dima Yakovlev Law in 2012. A controversy over gay rights in Russia, threats of terrorism, and the unfolding events in Ukraine also adversely affected the image of the country. Thus, the Sochi Olympics were considered an image builder and a catalyst for attracting other major cultural and sporting events to Sochi, e.g., the Formula 1 Grand Prix series and the 2018 FIFA World Cup.

The Sochi Olympics greatly boosted national pride and inspired volunteering movements. Russia promoted the Olympics through social networks, such as Facebook, YouTube, Twitter, Вконтакте, and Flickr, with several PR campaigns starting as early as 2005 (Losevskaya, 2013). For example, an official Twitter project oriented primarily toward international users had 36,370 followers and accumulated approximately 4000 tweets. The official page for Russian-language users[1] had 109,000 readers, with content consisting of approximately 3000 messages (Losevskaya, 2013). Despite the broad promotional outreach for the Sochi Olympics in social media, however, the "new media" were used in the "old way", i.e., as a channel for transmitting information with minimal interactive content (Losevskaya, 2013).

[1] http://twitter.com/sochi2014_ru

This study focuses on Twitter as a new medium and aims to investigate how the Sochi Olympics were portrayed on Twitter by hosts and guests. Twitter is an Internet service that allows users to post brief online messages that are visible to their social networks. "The simplicity of publishing such short updates in various situations ... makes microblogging an innovative communication method that can be seen as a hybrid of blogging, instant messaging, social networking and status notifications" (Ross, Terras, Warwick, & Welsh, 2011). Discussions on Twitter provide vast amounts of data about various topics of social importance. A recent meta-analysis of 575 peer-reviewed publications that used Twitter data identified the wide range of domains to which these studies belonged, including geography, marketing, natural disaster management, linguistics, politics, and many others (Williams, Terras, & Warwick, 2013).

Twitter as a source of data can also be useful for tourism research in the areas of destination marketing, imaging and branding, crisis management, and risk analysis. Notably, a search in the Science Direct database of social sciences and sports and recreation academic journals with the keywords "Twitter AND tourism" and "Twitter AND Olympic" in the abstract returned only three articles (in *Public Relations Review*, *Sport Management Review*, and *Annals of Tourism Research*). Thus, this study demonstrates an approach to extracting topical, spatial, and temporal information from Twitter messages to answer the following questions: What is the geographic landscape of Twitter messages about the Sochi Olympics? What issues were the most salient before, during, and after the Games? What are the temporal dynamics of issues concerning the Sochi Olympics, as reflected on Twitter? With regard to the last question, we intended to determine which of these issues were pertinent to any mega-sporting events and which were country-related and influenced by political events.

We approached answering these questions by collecting more than 7 million Twitter messages about the Sochi Olympics for a period of 1 year starting on October 31, 2013, in the most popular languages. The "people-as-sensors" (Goodchild, 2007) geo-distributional nature of Twitter messages, their internal structure, and the attached auxiliary information are conducive to creating geographic and temporal distributions of issues and attitudes toward the Sochi Olympics and toward the hosting country and destination. Twitter messages are particularly suited for computer-assisted content analysis because they have already undergone the process of data reduction, i.e., standardizing the size of the messages, "shedding off" of irrelevant content, and incorporating conventions and shortcuts, the "meaning" of which is shared by all Twitter users (Kirilenko & Stepchenkova, 2014). Sentiment analysis algorithms allow for the monitoring of public attitudes with respect to the temporal and geo-locational dimensions.

The paper proceeds as follows. Section 2 covers the data collection process, including descriptions of the procedures, key words, languages, and data "cleaning". It also addresses the methods for assigning geo-locational coordinates to the collected tweets. Section 3 explains rationale for data reduction using hash tags and aggregation algorithms. The section also reports the results. Finally, the

Conclusion section summarizes the analyses with respect to the stated research questions, discusses the limitations, and outlines directions for future research.

2 Method and Data

2.1 Data Collection

For 1 year, from November 1, 2013, to October 31, 2014, we systematically collected Twitter data related to the Sochi Olympics by performing a Twitter search with adaptive frequency, ranging from once every 3 min during the Olympic Games to 12 times per day during the pre- and post-Games period. The searches were performed with a Python code that used Twitter REST API, version 1.1. In total, 7.8 million tweets were collected using the keywords *sochi*, *olympics*, and *paralympics* in various languages. A time interval from several weeks prior to the Games to immediately after the Games generated the majority of all of the collected tweets. The number of tweets has remained relatively small since March 2014. Figure 1 illustrates the temporal frequency distribution of the collected tweets.

To ensure the quality of the collected sample, the data were filtered to remove those tweets that were unrelated to the Sochi Olympics. First, we selected a random sample of 20 tweets per month in the two most frequently used languages (English and Russian) and manually classified them into two categories: related and unrelated to the Sochi Olympics. Over time, the percentage of irrelevant tweets changed dramatically, varying between 0 % and 5 % from November 2013 to March 2014 and then increasing sharply to 90 % irrelevant tweets in October

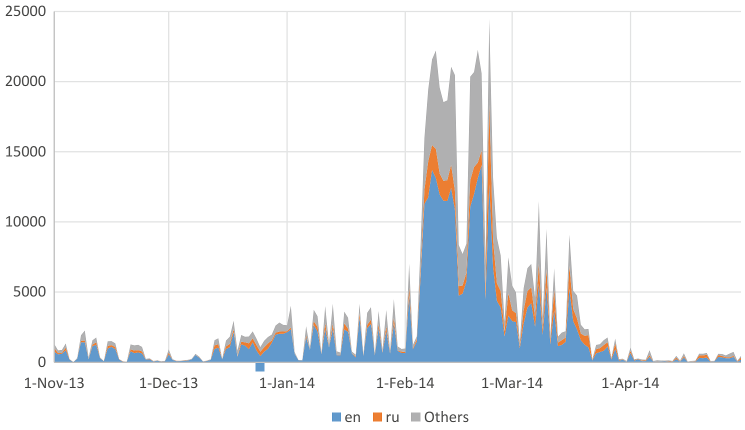


Fig. 1 Distribution of collected daily number of tweets in two the most prevalent languages, English (en), Russian (ru), and in other languages (Others)

2014. Based on this analysis, we restricted the database to the period from November 1, 2013, to March 31, 2014.

Similar to time period filtering, the most frequent hash tags were extracted from the collected tweets, and a sample of 100 tweets per each hash tag was manually classified into two categories: related and unrelated to the Sochi Olympics. Only the hash tags that had at least 90 % relevant tweets were left in the dataset, i.e., #sochi2014, #сочи2014, #olympics, #олимпиада, #teamusa, #sochiproblems, #temacanada, #wearewinter, #winterolympics, #paralympics, and #roadtosochi. In total, 616,333 tweets spanning the period between November 1, 2013, and March 31, 2014, were selected for further analysis.

2.2 Geolocation

The term *geolocation* relates to the process of identification of the geographic location of an object or person. The majority of spatially explicit studies that employ Twitter data utilize the geographic coordinates reported by the GPS-enabled user's device, such as his or her cell phone (further geo-coordinates). This method has two critical disadvantages. First, only a small number of tweets (typically 1 % or less of the dataset (Graham, Hale, & Gaffney, 2014)) carry these data (are *geotagged*), which results the majority of the information being discarded. Second, it is unknown whether and how the content of the retained geotagged tweets, which are presumably produced by the most technically savvy users (Graham et al., 2014), differs from the entire dataset. Alternatively, multiple researchers have attempted to extract locational data from the texts of tweets. The methods rely on searching for geographical information contained in tweets, augmented by linguistic geographical matching (e.g., matching the greeting "howdy" to Texas residents) (Cheng, Caverlee, & Lee, 2010; Eisenstein, O'Connor, Smith, & Xing, 2010; Gelernter & Mushegian, 2011; Ghahremanlou, Sherchan, & Thom, 2014). These methods, however, suffer from low reliability; e.g., Cheng et al. (2010) were able to resolve only half of a sample of tweets with an accuracy of within 100 miles.

We opted for a mixed approach and assigned each tweet the coordinates that were estimated from the user self-defined location (SDL). The important difference from the aforementioned methods of tweet geotagging is that this approach attempts to define the person's permanent place of residence, rather than his or her current location. Similar to other methods, the SDL approach is not free of ambiguities: the following problems must be resolved.

(1) Ambiguity of toponyms. The geographical information in tweets is represented by means of toponyms, which in many cases are ambiguous. For example, a person listing London as his or her place of residence might be referring to several different cities;
(2) Usage of nicknames, e.g., "City of Angels" for Los Angeles, USA;
(3) Alternative names and spellings, e.g., Saint-Petersburg, Sankt Peterburg, or St. Petersburg;

Table 1 Haversine distances (km) between true tweet locations and coordinates estimated from SDLs

		Percentile			
N	Mean	50	75	90	95
3445	534.8	9.3	39.3	1362.5	2757.1

(4) Errors in location names, e.g., “San Fransisco”; and,
(5) Nonsensical locations, e.g., “Milky Way”

We applied the Python code, which utilized the GeoNames online geographical database[2] to resolve the geographic locations of tweets from SDLs. The database contains more than 7 million geographical features together with their alternative names and nicknames. We resolved geolocation ambiguity by assigning the *largest* of possible populated places as the place of residence; e.g., we resolved “London” to “London, UK” and “London, CA” to “London, Ontario, Canada”. Then, we selected the top user-defined locations and manually parsed the geolocation results, resolving the remaining alternative names and errors. We defined the top locations as those found in at least 20 records; of 49,890 distinct SDLs, 2137 locations belonged to this category; manual disambiguation corrected the locations of 4.6 % of the records with the top locations. In total, the locations were defined for 215,840 tweets.

The geolocation results were validated against the true locations of Twitter users specified in 13,996 geotagged tweets. The haversine distance between estimated SDL coordinates and GPS coordinates provided with tweets was used to quantify the precision of the adopted geolocation algorithm. When multiple tweets coming from the same Twitter user were available, differences in GPS coordinates were assumed to reflect the user’s travelling pattern; accordingly, the minimum haversine distance was used. Finally, only tweets for which the geolocation was resolved with a county or a more precise location were used (e.g., the tweets with location defined as “Arizona” were excluded from validation). Table 1 shows the validation results. The geographic coordinates of 75 % of the tweets were defined within a 40-km radius of the true location; based on this result, the geolocation algorithm was presumed to resolve the points of origin of Twitter users acceptably.

3 Results

3.1 Distribution of Tweets Across Languages

The three most frequent languages in our sample—English, Russian, and Japanese—together account for 75 % of the 616,333 collected relevant tweets. The distribution of tweets across languages differs significantly from the distribution in a random sample of all tweets (Table 2). This random sample of 3,275,263 tweets

[2] http://www.geonames.org

Table 2 Distribution of tweets across languages responsible for at least 1 % of all tweets

Code	Language	Count	Olympic Tweets (%)	Random sample (%)
en	English	372,221	60.4	33.8
ru	Russian	66,885	10.9	0.9
ja	Japanese	28,857	4.7	14.8
fr	French	25,782	4.2	2.2
es	Spanish	22,632	3.7	11.9
de	German	18,162	2.9	0.9
tl	Tagalog (Philippines)	11,317	1.8	1.9
nl	Dutch	11,023	1.8	1.2
it	Italian	9463	1.5	0.9
bg	Bulgarian	9074	1.5	0.0
id	Indonesian	6187	1.0	12.2

Language codes (Code) are explained according to the IANA language subtag registry

was considered a tweeting "baseline" and was collected from June 13 to July 3, 2013. Compared with this baseline, the differences in language share are especially significant for English (60.4 % in the Olympics sample vs. 33.8 % in the baseline sample) and Russian (10.9 % vs. just 0.9 %) languages. For such languages as Japanese, Spanish, and Indonesian, the percentage of Olympic tweets is considerably lower than expected, based on Twitter popularity (baseline sample). This difference can be partially explained by the method of sample selection, but it is also believed to reflect the differences in the relative interest in the Winter Olympics in different countries. The percentage of different languages present in the sample also changes over time, with the relative number of tweets in Russian gradually increasing after the end of the Paralympics, thus reflecting the decrease in interest in the Games in countries other than Russia (Fig. 2).

3.2 *Content Analysis of Hash Tags*

Content analysis of Twitter messages can be based on hash tags, which are user-provided metadata that serve to group similarly tagged tweets. In total, there are 54,756 hash tags in the database for 616,333 tweets; however, only six hash tags (#Sochi2014, #Сочи2014, #Olympics, #SochiProblems, #RoadToSochi, and #TeamUSA) account for more than half of all hash tag usage. We removed the top two hash tags, #Sochi2014 (41.4 %) and #Сочи2014 [#Sochi2014][3] (2.9 %), from consideration because the same keywords were used to select the sample. Of the remaining hash tags, the 75 most frequently used hash tags were used in more than 50 % of tweets, 867 hash tags were responsible for 75 % of hash tag usage, and 7107 hash tags made up more than 90 %. The top 75 hash tags, listed in order of

[3] Brackets designate translations from other languages; parentheses designate explanations of hash tags.

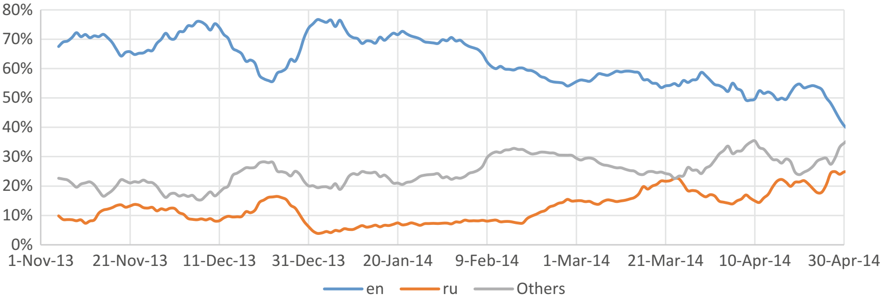


Fig. 2 Distribution of tweets over the languages (percentage, decade running mean)

descending frequency, were as follows: Olympics, SochiProblems, RoadToSochi, TeamUSA, Olympics2014, Paralympics, Sochi, Олимпиада [Olympics], TeamCanada, WeAreWinter, Russia, GoCanadaGo, Paralympic, BBCSochi, Olympic, FigureSkating, Сочи, Canada, USA, WinterOlympics, CanadaProud, オリンピック [Olympics], hockey, GoTeamUSA, GoldForCanada, Curling, Хоккей [hockey], ClosingCeremony, CBCOlympics, Figureskate, ソチ [Sochi], OpeningCeremony, Россия, Lgbt, Putin, Everywhere, Gold, Ukraine, NHK, JO2014, Новости, Олимпиада2014, フィギュアスケート [Figure skating], CanVsUSA, SochiFail, Biathlon, YunaKim, SeeYouInSochi, Slopestyle, Паралимпиада [Paralympics], БолеемЗаНаших [Cheering for Russia], TeamGB, Chatwing, Whatsthere, パラリンピック [Paralympics], Visaソチ [Visa Sochi], ソチ五輪 [Sochi Olympic], IceHockey, スカパー [SKY, Japan satellite broadcasting], News, SkyOlimpiadi [Italy sport broadcasting], TeamVisa (sponsoring young athletes to qualify for Olympics), 羽生 [Yuzuru Hanyu, 2014 Olympic figure skating champion from Japan], Биатлон [biathlon], Euromaidan (2013 Ukrainian protest movement), Live, TorchRelay, Create, CANvsSWE, Halfpipe, キムヨナ [Yuna Kim], JamaicaBobSled, USAHockey, LoveCurling, and OS2014 [Olympics2014]. To this list of the 75 most frequent hash tags, we added hash tags that were in the "top 10" on a certain day or in the "top 25" during a certain week, so we would not miss any topic of popular interest that was "localized" in time. Then, synonyms for the items on this expanded list were merged under the most frequent hash tag for every set of synonyms. For example: hockey, хоккей, icehockey → Hockey. Finally, the merged hash tags were classified into eight broad themes/categories described below. The numbers in parentheses refer to the percentages of all hash tags that were classified using the procedure described above.

1. ***Events (20.8 %)***: three types of hash tags, related to the opening ceremony, closing ceremony, and Paralympics. The first two tags were tightly distributed within two or three days after the corresponding event. Dissimilar to the Olympic Games tags, which uniformly appeared throughout the data collection period, the Paralympics tag did not appear outside an approximately three-week period centered on the Paralympics week.
2. ***News (14.9 %)***: the hash tags of general media, broadcasting companies (e.g., #bbcsochi), and generic news hash tags (e.g., #news)

3. ***Sports (12.1 %)***: the hash tags related to specific sports (e.g., #hockey), sports in general (e.g., #sport), sporting organizations (e.g., #ioc—International Olympic Committee), particular games (e.g., #canvsusa—Canada vs. USA) and epithets used with particular sports (e.g., #icequeen with figure skating). There were spikes in this category during the pre-Olympic trials (e.g., January 12–18) and during key events in the most popular sports, such as ice hockey. Notably, there was a distinct interest in figure skating from Japan and the Republic of Korea that contributed a significant number of tweets to this category. There was a drastic reduction in the number of tweets in this category during the Paralympics.
4. ***Anticipation (11.8 %)***: hash tags related to awareness campaigns (e.g., #roadToSochi), countdowns (e.g., #100days), and the Olympic torch relay (e.g., #torch); the last category also included the names of the cities and places on the torch route (e.g., #iss—International Space Station). *Anticipation* is the strongest category throughout almost the entire pre-game period, with high variability related to a particular event (e.g., #iss torch travel to the International Space Station between November 6 and 12), to countdowns (e.g., "#50days"), or to awareness campaigns (e.g., #roadToSochi campaign of support and raising awareness of the games by Team GB).
5. ***Cheering (5.5 %)***: the hash tags related to specific countries and their teams (#sweden, #teamusa, #gocanada) and Olympic cheering campaigns (#teamvisa). This category had two distinct spikes during the first few and final few days of the Games. Similar to the *Sports* theme, there were few tweets in this category during the Paralympics. Two "country" hash tags, #russia and #ukraine, were not classified under the *Cheering* theme. Hash tag #russia was assigned to the theme *Other* because of a large number of generic tweets related, for example, to Russia's role as the host of the Games. Hash tag #ukraine was classified under the theme *Problems and Politics* because the political events in Ukraine were the predominant topic under this tag.
6. ***Problems and Politics (1.8 %)***: the hash tags directly related to problems during the Olympics, e.g., #sochiproblem, #ソチ四輪 (an Olympic Ring did not open) and hash tags related to political issues surrounding the games, e.g., #putin, #lgbt, and #ukraine. Several major topics drove this theme. The most significant political issue by frequency of mentions present in tweets under this theme was LGBT discrimination in Russia and related concerns about athletes' safety. Related hash tags included not only those directly related to the issue (e.g., #lgbt) but also the #cheersToSochi campaign by Coca-Cola and McDonald's, which were initially conceived to inspire athletes but were overtaken by LGBT activists. The second significant political issue was the protests over the Games' location allegedly selected at the site of the genocide of the Circassian people. Finally, the #sochiProblems and similar hash tags indicated problems with Olympics infrastructure and organization. This sub-category was responsible for more than half of *all* Olympics-related tweets during the single day immediately preceding the opening ceremony, when large numbers of Games guests and participants arrived in Sochi. After the first few days of the Games, the number of tweets in this category gradually decreased, but it did not reach zero until the end of the Paralympics.

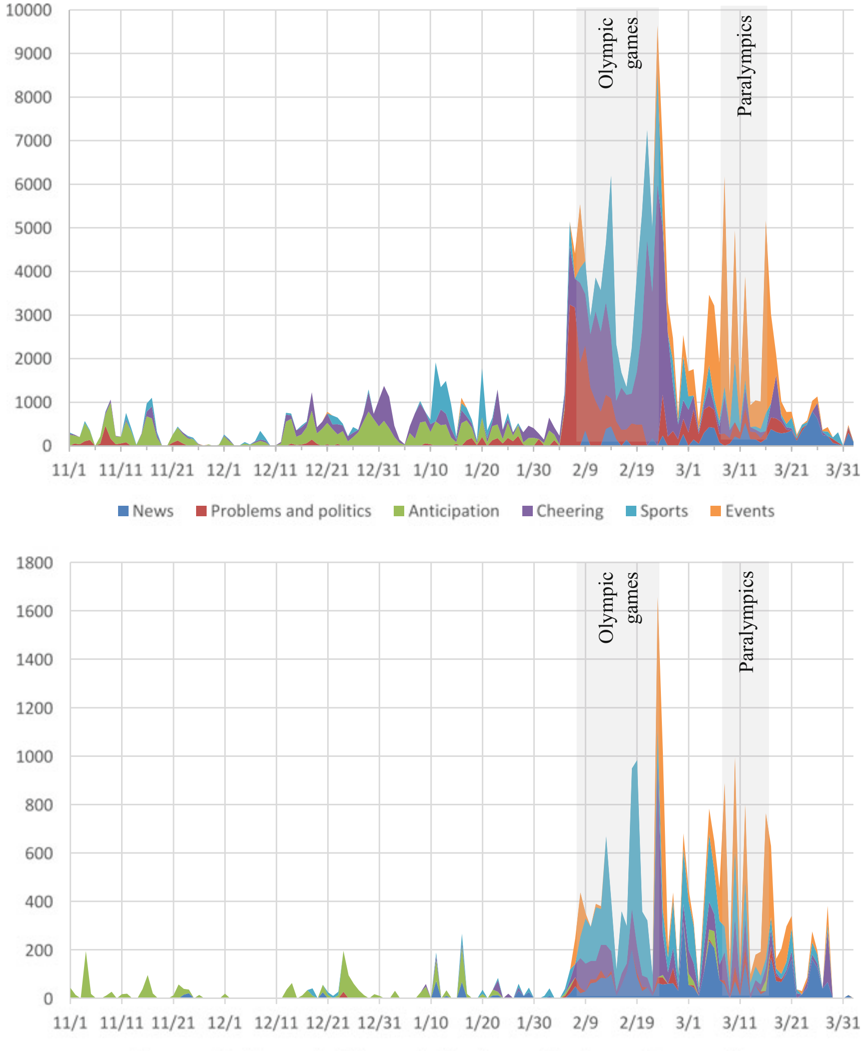


Fig. 3 Dynamics of tweets over categories: overall (*top*) and Russia (*bottom*). The categories "Others" and "Volunteering" are not shown

7. ***Volunteering (0.4 %)***: the hash tags related to volunteering at the Olympic Games. Due to a relatively small number of tweets in this category, it is included only in the overall hash tag analysis.
8. ***Other (32.7 %)***: generic hash tags and tags that did not fit into any of the above categories (e.g., #sochi, #юмор (jokes)).

Figure 3 illustrates the changes in the numbers of tweets in the eight categories over time in the English and Russian languages.

The overall numbers of tweets in each of these categories are shown in Table 3 for the English, Russian, and Japanese languages. The most noticeable difference among the languages is that significantly higher numbers of tweets about Games' problems were tweeted in English (15.3 %) than in Russian (4.1 %) and Japanese (1.7 %). The most frequently mentioned problems in English tweets were issues with Olympic objects not being ready for the Games (38.2 %) and with human rights in Russia (20.4 %). A relatively small percentage of hash tags in English were devoted to specific sports disciplines (17.4 % vs. 49.1 % in Japanese), but a much higher percentage of tweets consisted of cheering for the national teams (32.0 % vs. 1.5 % in Japanese). The percentage of tweets about specific sports disciplines varies significantly among the languages, e.g., for the popular ice hockey and figure skating:

- English: hockey—33.1 %, figure skating—12.4 %;
- Russian: hockey—39.8 %, figure skating—48.3 %; and
- Japanese: hockey—0.0 %, figure skating—85.7 %.

Not surprisingly, there were no tweets in the *Volunteering* category in languages other than Russian. Finally, in the *Other* category, the top place was universally occupied by the tags #sochi and/or #russia. Aside from these two hash tags, for both the English and Russian languages, promotion tags were prevalent. In English, the third and fourth most common tags in *Other* were #create and #chatwing tags, advertising the online chat service ChatWing. In Russian, the third place, after #sochi and #russia, was taken by status-inflating hash tags promising to reciprocate followers (those who subscribe to read your tweets), e.g., #ru_ff.

3.3 Large-Scale Geography of Tweeting

The geography of tweeting about the Sochi Olympic Games is analyzed using the geolocational algorithm described in Sect. 2.1. Additionally, to analyze the pattern of tweeting about the Olympics at the city level, we identified tweeting locations with a precision of a county or better and assigned all of these locations to the nearest urban area. In effect, this process distributed 161,823 tweets with a county or better precision over 798 identified urban areas (Fig. 4a). The most represented countries tweeting about the Sochi Olympics are the USA (25.9 % of all tweets with identified locations), Canada (23.0 %), and Russia (16.4 %), followed by the UK (6.7 %) and Japan (4.2 %). The most tweeting cities are also located in these countries; they are Toronto (6.2 % of all tweets), Moscow (4.8 %), London (3.7 %), New York (3.2 %) and Vancouver (2.7 %). The host city of the Olympics, Sochi, is in 6th place (2.6 % of all tweets).

This geographical distribution of tweets, however, does not reflect the interest of country populations in the Olympic Games due to differences in Twitter penetration between countries. To account for regional variations in Twitter usage, we collected a sample of all geotagged tweets using Twitter streaming API over one week,

Table 3 Distribution of the most frequent hash tags in the categories (percentages) in three most frequent tweeting languages of the sample

Language	N	News	Problems and politics	Volunteer	Anticipation	Cheering	Sports	Events	Others
English	762,475	5.6	15.3	0.0	9.2	32.0	17.4	10.6	9.8
Russian	165,425	10.3	4.1	1.2	5.9	12.8	23.3	12.4	30.0
Japanese	75,025	15.3	1.7	0.0	6.8	1.5	49.1	10.9	14.8

Note that the total number of hash tags N is greater than the total number of tweets in the sample because one tweet might contain multiple hash tags

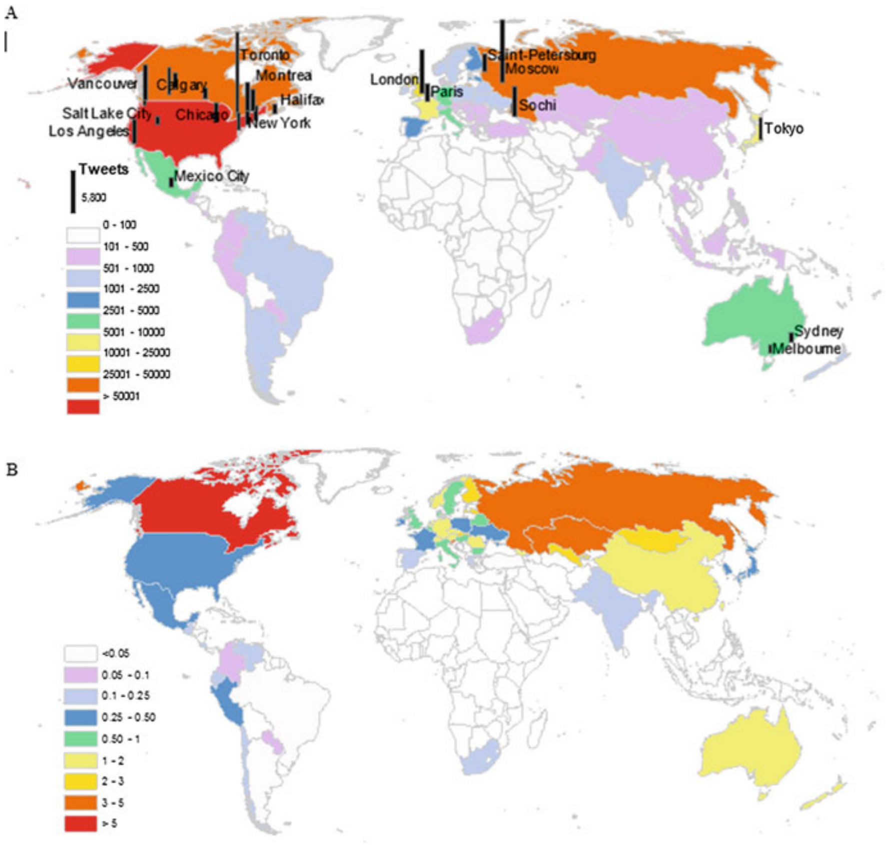


Fig. 4 Distribution of tweets between the countries: (**a**): without correction; (**b**): corrected for Twitter penetration

accumulating 2,015,000 tweets over the period from February 13 to 19, 2015. These tweets were then assigned to the countries based on their geographical coordinates. The relative number of tweets adjusted for Twitter penetration T^a was then calculated as

$$T_i^a = \frac{T_i}{\overline{T}} / \frac{S_i}{\overline{S}}$$

where i designates a country, T_i and S_i are numbers of tweets in the Olympics sample used in this study (616,333 tweets) and the overall random sample (2,015,000 tweets), respectively, for a particular country, and $\overline{T}$ and $\overline{S}$ are the average numbers of tweets per country in the corresponding samples. We excluded countries with small numbers of Olympic tweets (fewer than 100) from this computation. The geographical distribution of the relative number of tweets adjusted for Twitter penetration (Fig. 4b) was radically different from the distribution of the raw number of tweets. Whereas the USA was the leading country in terms of the absolute number of collected tweets, its relative tweeting intensity was less than average. In contrast, Canada, which trailed the USA in absolute numbers, was 8.7 times more active that the USA in discussing the Sochi Olympics on Twitter after adjusting its overall tweeting activity. The host of the Games, Russia, had the second highest relative tweeting activity ($T_i = 3.2$).

3.4 Fine-Scale Geography of Tweeting

Fine-scale analysis of the tweeting pattern was possible using the geotagged tweets. Although half of the tweets contained identifiable SDLs, the percentage of geotagged tweets was much smaller (Table 4). In our sample, the geographic latitude and longitude of a twitterer's location were present in 13,996 records (2.3 %) of the sample, which was slightly greater than the percentage of geotagged tweets in our random sample from 2013 (1.9 %). Compared with the percentage with SDLs, which remained roughly the same for tweets in different languages, the percentage of geotagged tweets varied from 10.0 % for Bulgarian to 0.9 % for Portuguese (only languages with at least 5000 tweets were included in the analysis). Notably, compared with the tweets with SDLs, the percentage of geo-tagged tweets was noticeably greater (2.1 % vs. 2.9 %); we hypothesize that Twitter users who were inclined to broadcast their place of residence were also more likely to use GPS devices to broadcast their immediate location.

Figure 5 shows the spatial travel patterns of tweeters who visited the Sochi Olympics. Note that only the tweets from the Olympic sample are shown; consequently, the maps illustrate patterns of travel for people who started tweeting about the Olympics prior to the Games and then travelled to Sochi. This pattern of travel shows is quite homogeneous: the majority of foreign travelers came to Sochi from the USA and Canada (Fig. 5a). The domestic travel shows even more homogeneous, as the majority of domestic travelers came to Sochi from just one city: Moscow (Fig. 5b).

The Sochi Olympic Games were concentrated in two tourist clusters, coastal and mountain clusters, with the venues in the coastal cluster located within walking distance from one another. Sochi International Airport (AER), which served as the major gateway for the Games, is located only 5 km south of the coastal cluster

Table 4 Percentages of tweets with self-defined locations (SDLs) and geotagged tweets in the two most common languages of the Games: English and Russian

Language	N tweets	SDL percentage	Geotagged percentage
English	372,221	51.4	2.1
Russian	66,885	49.8	4.2
Others	177,227	43.7	2.9
Overall	616,333	49.0	2.6

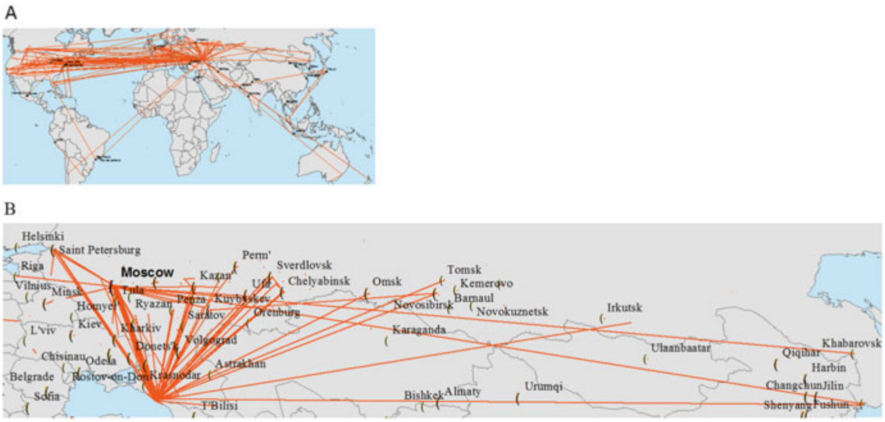


Fig. 5 Travel pattern of Sochi Olympics attendants worldwide (**a**) and inside Russia (**b**)

venues. The triangular pattern of travel (Fig. 6a) reflects the concentration of the Games, with the majority of Twitter users moving between two Olympics venues and the nearest city, Sochi. Very few tweets came from the outside this triangle, which was not the long-distance travel to the region reflected in Fig. 5 and may reflect little visitation of the Games by local people aside from people from Sochi. This pattern of local travel was very different between the English language users, who were used as proxy for foreign travelers, and the Russian language users, who were used as a proxy for domestic travelers. Over the 5 months of data collection, the domestic travelers moved among all three hubs of the Games (Fig. 6c); this pattern remained the same during the Games (Fig. 6d). At the same time, foreign travelers moved almost exclusively between the tourist clusters of the Games and the airport; very few tracks included the city of Sochi.

3.5 *Authoritative Internet Sources of Information About the Sochi Olympics*

The Sochi Olympics tweets frequently referenced external resources. Of the entire database, 28 % of the tweets (174,428) contained references to at least one resource on the Internet. In total, these URLs pointed to 11,203 different domains; however, 50 % of the references pointed to only 43 (0.39 %) of the most popular domains (Table 5). By far, the most frequently referenced domains (28 %) belonged to photography and video sharing services Instagram and YouTube. Among the traditional news outlets, the largest number of tweets referred to the BBC and Vesti.ru (a Russian-language news channel) and to dedicated sports channels, such

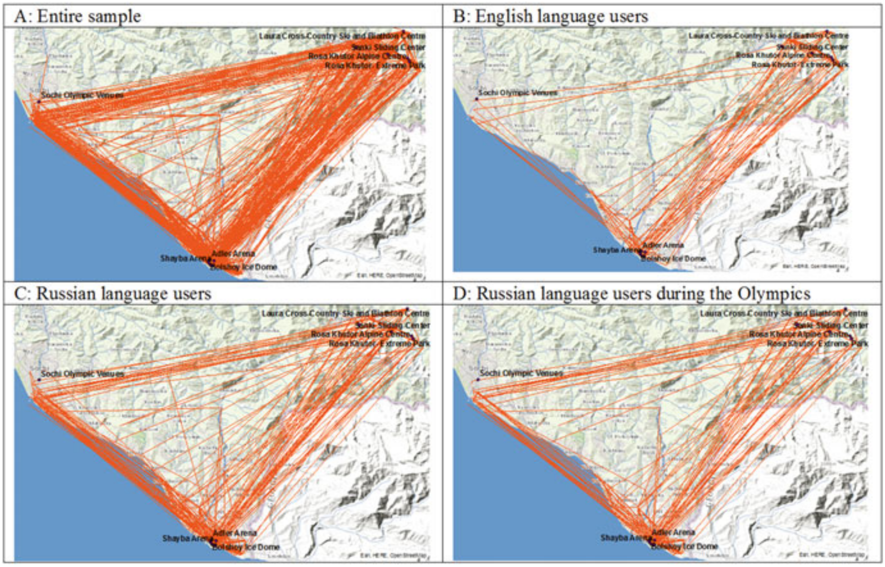

A: Entire sample
B: English language users
C: Russian language users
D: Russian language users during the Olympics
Laura Cross-Country Ski and Biathlon Centre
Sanki Sliding Center
Rosa Khutor Alpine Centre
Rosa Khutor Extreme Park
Sochi Olympic Venues
Adler Arena
Shayba Arena
Bolshoy Ice Dome

Fig. 6 Travel pattern of Twitter users in the Olympic Games area entire sample; English language users separately; Russian language users, and Russian language users travelling within the Games period. Note that there were very few English language tracks outside the Games period

as NBC Olympics (the holder of the USA broadcast rights), sportbox.ru, and other Russian-language sports outlets. A considerable number of tweets referenced the domains belonging to national committees and to the International Olympic Committees—e.g., Olympic.org, Olympic.ca—and to other sports organizations, such as the National Hockey League. Among the non-traditional news sources, the Facebook, Tumblr, and Worldpress blogging and social network domains, together with the Russian language social networks V Kontakte (vk.com) and LiveJournal, were the most popular. Finally, a considerable number of tweets referred to other domains, such as the currently defunct domain sochiproblems.com, the Web site of the President of Russia, LGBT rights Web sites, and an auction site for collectables.

These frequently referenced Internet resources differed among languages (Table 6), with little intersection but with similar distributions among traditional news, social networks, and dedicated sports outlets, including the sites of national Olympics organizations. In English, the BBC, Facebook, and the Canadian Olympic Team were the most popular domains. In Russian, the most popular were the TV channel Vesti.ru, the social network V Kontakte, and the sports outlet sportbox.ru. In Japanese, 13.6 % of all URLs referenced the site of Japanese National Olympic Committee, 4.9 % referenced the BBC, and 2.6 % pointed to the Infoseek news portal.

Table 5 The most frequently referenced domains in tweets about the Sochi Olympics (percentage of the total number of tweets with references greater than 0.3 %)

	Domain	%	Explanation
General news	bbc.co.uk	2.5	BBC
	vesti.ru	1.3	Russian news
	tvmix.com	1.0	Online TV
	ap.org	1.0	Associated Press
	cbc.ca	0.9	CBC
	wsj.com	0.6	Wall Street Journal
	theglobeandmail.com	0.6	Globe and Mail
	rt.com	0.5	Russian TV foreign edition
	russia.tv	0.5	Russian TV Channel 1
	nbcnews.com	0.4	NBC
	washingtonpost.com	0.4	Washington Post
	rg.ru	0.4	Russian government daily newspaper
	theguardian.com	0.3	The Guardian
	cnn.com	0.3	CNN
	thestar.com	0.3	Toronto Star newspaper
News aggregators	yahoo.com	0.9	Yahoo portal
	go.com	0.7	Disney news landing domain
	mashable.com	0.6	Online news media
	tunein.com	0.4	Online radio
	buzzfeed.com	0.3	Online news media
Sports news	nbcsports.com	1.3	NBC
	sportbox.ru	0.7	Russian online sports media portal
	nbcolympics.com	0.7	NBC
	sports.ru	0.5	Russian online sports media portal
	insidethegames.biz	0.4	Online sports media portal
	sovsport.ru	0.3	Russian sport newspaper
Organizations	olympic.ca	1.8	Canadian Olympic Team
	olympic.org	1.0	International Olympic Committee
	teamusa.org	1.0	Olympic committee—USA
	paralympic.org	1.0	Paralympic Committee
	joc.or.jp	0.8	Olympic Committee—Japan
	nhl.com	0.6	National Hockey League
	hockeycanada.ca	0.4	Hockey Canada
	paralympic.ca	0.4	Paralympic committee—Canada
	teamgb.com	0.4	Olympic committee—GB
	iihf.com	0.2	International Ice Hockey Federation
Blogs	facebook.com	2.6	Facebook
	vk.com	1.1	Russian social network
	tumblr.com	0.5	Microblogging
	livejournal.com	0.4	Russian social network
	wordpress.com	0.4	Blogging

(continued)

Table 5 (continued)

	Domain	%	Explanation
Other	sochiproblems.com	1.3	Currently defunct web site
	allout.org	0.5	LGBT rights campaign
	kremlin.ru	0.4	President of Russia
	legendsdepot.com	0.3	Auctions and collectors

In total, these domains are responsible for 33.0 % of all references (50.8 % with instagram.com and youtube.com). Photo and video sharing services and redirecting services are not included in the table but were considered in computing the total percentages

Table 6 The most frequently referenced resources in tweets about the Sochi Olympics in different languages

English	%	Russian	%	Japanese	%
bbc.co.uk	3.0	vesti.ru	7.6	joc.or.jp	13.6
facebook.com	2.7	vk.com	5.8	bbc.co.uk	4.9
olympic.ca	2.6	sportbox.ru	4.0	infoseek.co.jp	2.6
sochiproblems.com	2.0	russia.tv	3.0	olympic.org	2.1
nbcsports.com	1.8	sports.ru	2.9	yahoo.co.jp	2.0
tvmix.com	1.5	rg.ru	2.3	skyperfectv.co.jp	1.7
teamusa.org	1.5	livejournal.com	1.8	fc2.com	1.4
ap.org	1.4	sovsport.ru	1.6	nbcsports.com	1.1
cbc.ca	1.4	facebook.com	1.5	twipple.jp	1.0
yahoo.com	1.2	olympic.org	1.5	pixiv.net	0.9

Photo and video sharing sites, such as Instagram.com, are not included

4 Conclusions

This study examined Twitter messages pertaining to the Sochi Olympics to demonstrate an approach for extracting topical, spatial, and temporal information from Twitter messages and to answer the following questions: What is the geographical landscape of Twitter messages about the Sochi Olympics? What issues were the most salient before, during, and after the Games? What are the temporal dynamics of issues concerning the Sochi Olympics, as reflected by Twitter? The following paragraphs briefly summarize the findings related to the stated research questions.

The answer to the question about the geographic landscape of Twitter messages is illustrated by Fig. 4a and b. If one supposes that the number of tweets indicates general interest in the Sochi Olympics, it seems that interest in the Games varied among countries and in comparison to their “tweeting baseline” (Figs. 1 and 2). The people who were most interested in the Games were those from “winter sports countries” and from countries that had hosted Olympic Games in the past. In addition, the adopted algorithm of establishing the geo-location of tweets and the geographic landscape of Twitter messages allowed for the obtaining of insights into the travel patterns of guests to the Games. The heaviest travel to the Sochi Olympics was from North America, Europe, and elsewhere in Russia. The travel patterns of foreigners in Russia were primarily contained in two Sochi tourist clusters: coastal and mountain (Fig. 6). The method used to collect the data included only messages that contained the pre-selected key words, thus limiting the generalizability of travel pattern analysis because the long-distance travel patterns of the people not tweeting about the Games outside the Games' venues were not represented. Extending the data collection to all of the tweets from a person tweeting about the Games at least once might have removed this limitation.

The answer to the question about the most salient issues discussed on Twitter about the Sochi Olympics and their temporal dynamics is illustrated by Figs. 1 and 3 a & b. The tweet volume starts with an anticipation phase 3–4 months prior to the Games, which is highly concentrated around the Games period and decreases quickly after the Games are over. The Paralympics generate noticeably much less excitement than the Olympics events. The main topics of discussion on Twitter were the opening and closing ceremonies, news and updates about the Games tweeted by various media outlets and the general public, discussions of particular sports, such as ice hockey, figure skating, and others, cheering for national teams and following the route of the torch relay.

Most of the issues tweeted about the Sochi Olympic Games (those classified under the topics of Anticipation, Cheering, News, Sports, Events, and Volunteering) could be viewed as pertaining to mega-sporting events in general. Country-specific problems and political issues surrounding the Games had a 1.8 % share of all hash tags that accounted for approximately 15 % of all of the tweeted messages in three languages. Among the most prominent topics that cast an unfavorable light on Russia were Sochi's infrastructure problems and the situation of LGBT rights in Russia. Several times during the pre-Games period, LGBT issues surged in the whole volume of Twitter messages, especially relative to the anti-LGBT laws ("homosexual propaganda laws") passed in Russia. Notably, this topic was pronounced in English but not in either the Russian of the Japanese segments of the collected data.

Interestingly, the issues of terrorism, corruption, and budget overspending, ecological problems pertaining to infrastructure development, and discussions of Russia's international politics were not very prominent in the collected tweets, which might indicate that although those tweeting about the Olympic Games did reference various media resources, including political news agencies (see Tables 5 and 6), mega-sporting events are viewed by the general public as predominantly non-political events. Different topical patterns of hosts and guests were registered: Russian tweets practically did not mention problems or politics, and Russians were more active in discussing various Olympics events than were guests at the Games.

Even though this study uses Twitter to investigate how the Sochi Olympics were reflected in user-generated content, the approach is expandable to user generated content from a wider set of social networks and other Web 2.0 applications. Overall, we demonstrated that despite a virtual absence of research into mega-sporting events based on the analysis of social networks in the scientific literature, social networks provide ample data about the major topics of interest and the political issues surrounding events, including their spatial and temporal dynamics. The ability of assigning location to user generated messages and content analysis allow for investigation of differences between the countries and regions in relation to a specific sporting event. Additionally, messages with geographical locations provided by users allow for the analysis of travel patterns of the events' visitors, as well as sentiment analysis. Finally, although it was unexplored in this paper, an important opportunity of mapping the connections inside the collected data has the potential to provide insights into information flows between people and groups discussing events.

References

Cheng, Z., Caverlee, J., & Lee, K. (2010) *You are where you tweet: A content-based approach to geo-locating Twitter users*. Proceedings of the nineteenth ACM Conference on Information and Knowledge Management (CIKM '10), Toronto, Ontario.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010) *A latent variable model for geographic lexical variation*. Proceedings of the 2010 conference on empirical methods in natural language processing, Cambridge, Massachusetts.

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS, 15*(6), 753–773.

Ghahremanlou, L., Sherchan, W., & Thom, J. A. (2014). Geotagging Twitter messages in crisis management. *The Computer Journal*, doi: 10.1093/comjnl/bxu034.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *Professional Geographer, 66*(4), 568–578.

Hilton Worldwide. (2014). *Balancing Russia's tourism deficit: A report of the future of the industry*. Retrieved 23 July, 2015, from http://news.hiltonworldwide.com/assets/HWW/docs/2012/2863HWRussianReportEng.pdf

Kirilenko, A. P., & Stepchenkova, S. O. (2014). Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change, 26*, 171–182.

Losevskaya, E. (2013). Роль «новых медиа» в информационной кампании «Сочи-2014» (The role of new media in the Sochi-2014 informational campaign). Журнал социологии и социальной антропологии (The Journal of Sociology and Social Anthropology), *16*(5), 203–220.

Oliphant, R. (2013). *Sochi: chaos behind the scenes of world's most expensive Winter Olympics*. London: The Telegraph.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation, 67*(2), 214–237.

Taylor, A. (2014). Why Sochi is by far the most expensive Olympics ever. *Business Insider*.

Transparency International. (2013). *Corruption perceptions index 2013*. Retrieved 24 July, 2015, from https://www.transparency.org/cpi2013/results

United Nations World Tourism Organization. (2013). *UNWTO tourism highlights*. Retrieved 24 July, 2015, from http://www.e-unwto.org/doi/pdf/10.18111/9789284415427

Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation, 69*(3), 384–410.

Leveraging Online Reviews in the Hotel Industry

Selina Wan and Rob Law

1 Introduction

The advent of Web 2.0 has changed the manner by which hotel guests solicit information. They no longer merely rely on communication messages from hotels but also read or share hotel experiences on different social media websites. Past studies showed that the average traveler browses 38 websites prior to purchasing vacation packages from online travel agencies (Schaal, 2013). Hotel websites account for 4.1 % of the online sources used, whereas review websites comprise nearly 6.8 % of this total (Schaal, 2013). Approximately 60 % of European consumers actively engage in social computing activities, such as reading and writing online reviews (Forrester Research, 2007). As one of the largest travel review websites, TripAdvisor.com has received over 200 million online reviews worldwide (TripAdvisor, 2014). Approximately 70 % of global customers rate online reviews as one of the most trustworthy sources of information (NielsenWire, 2009). Thus, exposure to online hotel reviews (either positive or negative) may also increase the likelihood of including a particular hotel in a decision choice set (Vermeulen & Seegers, 2009).

Consumers are evidently becoming powerful, knowledgeable, and sophisticated as a result of the rapid development of the Internet and mobile technologies; thus, they are increasingly difficult to please (Buhalis & Law, 2008). A consumer review is "a mixture of fact and opinion, impression and sentiment, found and unfound tidbits, experiences, and even rumor" (Blackshaw & Nazzaro, 2006, p. 4);

S. Wan (✉)
City University of Hong Kong, Kowloon Tong, Hong Kong
e-mail: mkselina@cityu.edu.hk

R. Law
The Hong Kong Polytechnic University, Hung Hom, Hong Kong
e-mail: rob.law@polyu.edu.hk

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_14

therefore, such information is further referred to as "big data" (Lu & Stepchenkova, 2015). Thus, the emerging patterns in hotel preferences must be determined from a large-scale user-generated content (UGC). The consumer insights solicited by online reviews are expected to create value and provide evidence, thereby inspiring hotel managers to enhance the competitiveness and reputations of the hotels they are working for in the market.

The objective of this chapter is fivefold: discuss the emergence of the online review movement in the hotel industry, describe the nature and characteristics of online reviews, illustrate the influence of online reviews on the hotel industry and on consumer behavior, demonstrate how academic scholars and hoteliers solicit and leverage customer intelligence from online reviews, and present both the successful and poor responses of hotel management to online reviews to highlight the best practices in reputation management. An analysis of the theoretical framework of academic studies and the current practices of hoteliers enables this chapter to collate and clarify the issues related to online reviews, as well as their influence on hotel performance.

2 Emergence of Online Reviews in the Hotel Industry

The rapid development of the Internet and Web 2.0 has empowered people to disseminate, collaborate, and exchange travel experiences online (e.g., O'Connor, 2010; Ye, Law, Gu, & Chen, 2011). People cannot evaluate hotel products and services before availing of them; thus, such individuals may conduct extensive research on a hotel using multiple sources of information to limit perceived risks (e.g., Lewis & Chambers, 2000; Lin, Jones, & Westwood, 2009). Online reviews are among the emergent forms of UGC and electronic word-of-mouth, and such reviews play a pivotal role in consumer decision making (Litvin, Goldsmith, & Pan, 2008). Forester Research explained that in 2012, more than 50 % of travelers would not reserve a hotel that does not have online reviews (McEvilly, 2015). Therefore, online reviews inevitably influence the hotel industry.

Numerous types of review websites are available online. In particular, a few stand-alone review websites, such as TripAdvisor.com and Yelp.com, are written by actual hotel customers without solicitation and booking functions; hence, the reviews posted on these sites are perceived as credible sources of information (Law, 2006; O'Connor, 2010; Travel Media Group, 2015). Google+, Yahoo! Travel, and other search engines allow people to write hotel reviews on the corresponding websites by registering their e-mail accounts. Expedia.com, Hotels.com, Orbitz.com, Travelocity.com, Booking.com, and other online travel agencies (OTAs) promote verified reviews. However, only customers who book stay through the corresponding websites are invited by e-mail to post hotel reviews.

Despite a period of tension and the lawsuits filed by hotels against review websites regarding the publication of fake and misleading reviews in 2011, the relationship between these parties has improved since 2012 (McEvilly, 2015). Several hotel chains have partnered with review websites to leverage this online trend (Farley, 2012). The Four Seasons group enables potential customers to click through travel reviews via the group's hotel website (Short, 2013). Each of the property pages of the Four Seasons hotels includes online reviews derived directly from TripAdvisor, Twitter, and Facebook (Four Seasons, 2012). Wyndham and Accor encourage guests to write reviews on TripAdvisor after their stay; the total number of reviews, average ratings, and recent review content on this website are then displayed on each property page. Eric Danziger, CEO of Wyndham Hotel Group, explained that "travelers have an insatiable appetite for online reviews; reviews can influence hotels to correct problems… providing easy access to reviews has already translated into greater bookings." Consequently, hotel bookings at Wyndham increased by 30 % with the addition of the TripAdvisor logo, star ratings, and review links to the hotel website (USA Today, 2012).

With the increasing popularity of online reviews, several hotel companies have considered integrating consumer review functions into their websites. Starwood is the first global hotel chain to invite Starwood Preferred Guest members to share their thoughts and opinions on the websites of the individual hotels where they were accommodated; these members are guests who stay at any hotel property for more than 50 nights annually. This program was implemented at the end of 2011 (Miller, 2011), and has been proven to be successful as evidenced by the publication of 528,930 new and unedited reviews on all property websites from January to mid-September 2013. Furthermore, management response rate was significantly higher (48 %) than the industry average (25 %) within the same year (Ady, 2013). Marriott follows a similar practice by embracing the presentation of online reviews on its Marriott Rewards Insiders website (Farley, 2012). Nonetheless, only knowledgeable Marriott club members who are frequent guests are eligible to post comments although these reviews are open to the public. Third-party travel review websites, such as TripAdvisor, cannot verify whether or not a reviewer has actually stayed at a hotel (O'Connor, 2010); thus, various hotel websites have launched their own guest review programs, thereby enhancing the authenticity of reviews because reviewers' identities require verification by hotel staff. More importantly, no incentive is provided to reviewers, and posts are left unedited. This strategy not only prevents hotel companies from paying considerable commission fees to OTAs, but also encourages potential customers to browse and to reserve rooms via the hotel website (Levere, 2014). In the process, the online presence of these hotels increases.

3 The Nature and Characteristics of Online Hotel Reviews

As a "mega-trend" influencing the hospitality and tourism industry, online reviews (a popular form of social media) have become an important topic that has motivated academic scholars to aim at generating knowledge and a theoretical framework for a discipline (Leung, Law, van Hoof, & Buhalis, 2013). Previous studies have examined the nature and characteristics of online hotel reviews and have identified two interesting phenomena. First, online hotel reviews are heavily skewed toward positive ratings (Racherla, Connolly, & Christodoulidou, 2013). For example, more than 70 % of the 1.28 million hotel reviews on TripAdvisor were positive (4 or above on a 5-point rating scale), whereas merely 15 % were negative (under 3) (Melián-González, Bulchand-Gidumal, & López-Valcárcel, 2013). Second, the valence becomes more balanced when the volume of reviews increases, thereby mitigating the negative effects of hotel reviews (Melián-González et al., 2013). Therefore, the damaging influence of negative reviews can be diluted if hotel managers can encourage more guests to post reviews on websites.

Another stream of research explores the motivation for writing and reading online reviews. Yoo and Gretzel (2008) determined that customers are generally motivated to write reviews as to satisfy the need for enjoyment and for positive self-enhancement, fulfill the desire to help a travel service provider, and to show concern for other consumers. The act of venting negative feelings is not as salient as expected by the general public. Öğüta and Cezara (2012) further investigated the motivations to write reviews in the hotel context, and the empirical results of this study showed that customers who are satisfied with a hotel's price and overall hotel performance are more willing to post online reviews on hotel websites compared with dissatisfied customers. Even though a hotel's star score and star rating do not significantly influence the intention to write reviews, extreme satisfaction or dissatisfaction with a hotel does not affect customers' writing motivation either. Nonetheless, this study did not explain the reasons behind the findings.

Conversely, several researchers aimed to clarify why people read online reviews. Three motivating factors were identified, namely, convenience and quality (e.g., a fast and efficient method of obtaining information, best value, and reduced hotel prices), risk reduction (e.g., for making the right buying decisions), and social reassurance (e.g., to determine if others share their feelings regarding a hotel, to be part of a community, to compare a personal evaluation with others) (Kim, Mattila, & Baloglu, 2011). The survey results reported by the aforementioned academic scholars indicated that women generally rely on online hotel reviews to minimize risks and to seek convenience and quality. By contrast, men depend solely on their personal levels of expertise in online booking when reading hotel reviews. When perusing online reviews, the top five specific hotel attributes customers consider are room cleanliness, location and accessibility to other points of interests, value for money, customer service, and safety (Ong, 2012). Ong's study revealed that the two important attributes for review readers are the "reviewers' profile" and "the date of their hotel stay or date the review was posted." Vague opinions included in the hotel's publicity materials and travel guidebooks (e.g., hotel staff members' behavior and attitude) are the major reasons that account for the interest of people in reading online reviews (Williams, van der Wiele, van Iwaarden, & Eldridge, 2010).

Understanding why people read or write online reviews is increasingly important to hoteliers because several ranking systems in review websites are based on online review activity. For example, TripAdvisor's Popularity Index is determined by the quantity, quality, and recency of online reviews (Short, 2013). The content and activity of the Google + review website affect its search engine optimization ranking as well. More, better, and up-to-date reviews may enhance the position of a property hotel on the TripAdvisor Popularity Index or in the Google search engine (Travel Media Group, 2015). Therefore, such strategy may enhance the online reputation of and induce additional bookings at a particular hotel.

4 Influence of Online Reviews on the Hotel Industry

A typical hotel review provides both statistical evidence (e.g., the total number of reviews received by a hotel, traveler ratings, and ratings for different hotel attributes, including location, rooms, service, value, and helpfulness of a review) and narrative evidence (e.g., a reviewer's opinion of a hotel, such as "This is the best hotel I've ever stayed at," "It is modern and conveniently located," and "I highly recommend this hotel to everyone," as well as hotel tips, such as "Try to get a harbor view room reservation" and "Must visit their pool") (Hong & Park, 2012). Reviewers and hoteliers may even post photographs on the review website. Several websites provide Q&A function to facilitate online conversations between potential customers and hotel managers. The abundant numerical, textual, and visual information that can be obtained from online review websites has piqued the interest of academic scholars in determining how online reviews affect the hotel business.

Several studies have empirically tested the influence of online reviews on hotel performance. The pioneer study conducted by Ye et al. (2011) employed the log-linear regression model in predicting the online sales of hotel rooms in China. The results showed that the variance in the valence of rating scores across reviews does not significantly influence online sales. Nonetheless, a 10 % increase in a review rating leads to a 5 % increase in online bookings. Although a high hotel room rate may reduce the number of online bookings, the convenient locations of hotels in large cities help boost online sales. Ye et al. (2011) did not obtain the actual sales figures for hotel rooms in China; accordingly, these researchers applied the number of published reviews on a travel review website as a proxy for online hotel room sales.

Industry practitioners evaluate hotel performance according to actual hotel occupancy, average daily rate (ADR), and revenue per available room (RevPAR). In this regard, academic scholars have attempted to partner with syndicated research companies or online travel agencies (e.g., ReviewPro, Smith Travel Research, comScore, Travelocity, and TripAdvisor) to obtain actual data for empirical modeling. Anderson (2012) investigated the effects of social media on a hotel's return on investment (ROI) via logistic regression. This study estimated that hotel price (as measured based on a hotel's ADR) increases by 11.2 % with a 1-point increment in review rating (e.g., from 3.5 to 4.5 on a 5-point scale on the Travelocity review website) when the occupancy or market share of a single hotel is maintained. Meanwhile, a 1 % increase in a hotel's online reputation (as measured by ReviewPro's Global Review Index) results in a 1.42 % increment in RevPAR. This effect is considerably stronger on a mid-scale hotel than on a luxury hotel. The use of the model for the dynamic generalized method of moments (GMM) enabled Duverger (2013) to determine that short positive reviews (i.e., the number of words in a review) reduce the negative effect on a hotel's market share. The increase in the length of a review with either positive or negative tone increases the negative effect on market share as well. Blal and Sturman (2014) determined the differential effects of review volume and valence on hotel RevPAR and showed that the sales performance of low-tier hotels increases if the number of reviews increases. By contrast, hoteliers from luxury properties should concentrate on increasing review ratings rather than aiming only to boost the number of reviews. Exceeding customer expectations by providing excellent services is particularly important in driving sales and revenue from bookings at luxury hotels. Using the new methodological approach of Artificial Neural Network (ANN) enabled Phillips, Zigan, Silva, and Schegg (2015) to determine that regional room star rating enhances a Swiss hotel's RevPAR whereas room quality, positive regional review, and regional hotel reputation negatively affect hotel performance.

5 Influence of Online Reviews on Consumer Behavior

Online reviews can be examined from different dimensions, such as linguistic style, emotional expressions, helpfulness, framing, reviewers' identities, credibility, trust, review valence, review length, and volume; thus, previous studies determined how different review dimensions interact and perhaps affect consumer behavior (e.g., Kusumasondjaja, Shanka, & Marchegiani, 2012; Lee, Law, & Murphy, 2011; Noone & McGuire, 2014; Sparks & Browning, 2011; Sparks, Perkins, & Buckley, 2013; Tsao, Hsieh, Shih, & Lin, 2015; Vermeulen & Seegers, 2009; Xie, Miao, Kuo, & Lee, 2011).

Vermeulen and Seegers (2009) learned that positive online reviews improve consumers' attitudes toward a hotel. This effect is particularly significant for hotels with low brand awareness. The laboratory experiment of Sparks and Browning (2011) explored the role of online hotel reviews in perceived trust and in the intention to reserve a booking at a hotel. These authors selected four key dimensions of review as independent variables, namely, the target of the review (i.e., whether or not the content of the message is related to the core features of a hotel, such as the cleanliness of a guest room; or the services of a hotel, e.g., the friendliness of hotel staff), overall review valence (i.e., whether positive or negative comments were posted by the guests), the framing of the reviews (i.e., whether the initial review is positive or negative), and the presence or absence of a numerical rating. The aforementioned study determined that consumers generally adopt the easy-to-process approach when evaluating a hotel. The persuasive influence of an online review is magnified when the overall review content is negative and is negatively framed. Interestingly, review rating alone does not increase the number of hotel bookings nor enhance perceived trust. When a review set is positively framed based on customer service standards and is supported by a review rating, consumers tend to believe in a hotel. Therefore, the intention to reserve a hotel room increases. By considering an eco-resort hotel as the experimental context, Sparks et al. (2013) tested the influence of information source (i.e., whether a review was posted by a consumer or the management team of the resort hotel), content style (i.e., whether the information related to the sustainable tourism practices and facilities of the resort is vague or specific), and peripheral cues for credibility (i.e., the presence or absence of eco-certification logos in a review) on purchase intention and hotel guests' beliefs in the utility of reviews, trustworthiness, and the corporate social responsibility (CSR) of the hotel in question. The aforementioned authors determined that the inclusion of both UGC and firm-generated information with specific content in online promotions facilitates the usefulness and effectiveness of a communication message. Awards, logos, or credentials further increase the perceived trust and CSR held by the hotel; trust is particularly important to hotel attitude formation and purchase intention.

6 Customer Intelligence from Online Reviews

Hotel managers have long recognized the importance of guests' opinions and comments in customer buying decisions. A particular concern of hoteliers is the type of customer intelligence that should be extracted and the process of extracting the information (e.g., demographic and psychological profiles for existing and potential customers, primary interest, desired hotel services, and preferred facilities) from different sources of data to enhance customer satisfaction and hotel performance (Lau, Lee, & Ho, 2005). The main traditional customer intelligence methods employed by academic scholars or hoteliers are surveys (e.g., Shanka & Taylor, 2004), opinion polls (Li, Law, Vu, Rong, & Zhao, 2015), focus groups, "mystery guest" evaluations, and managers' breakfasts (Withiam, 1995). However, the major disadvantage of these research methods is the use of sample data instead of data from the entire population within the study period (Lau et al., 2005). The use of travel reviews as a source of data for analysis can minimize the limitation of small sample sizes because large amounts of data are readily available on review websites. The information on these websites is based on actual travelers' post-purchase experiences (Li et al., 2015). Owing to the advantages of simplicity, low cost, quickness, and a non-intrusive nature of soliciting guests' opinions, online reviews have emerged as a source of data for investigating customer insights over recent years (Li et al., 2015; Lu & Stepchenkova, 2015).

Previous studies have adopted different approaches to analyze textual comments from reviewers. In particular, content analysis (e.g., Au, Buhalis, & Law, 2014; Barreda & Bilgihan, 2013; Memarzadeh & Chang, 2015; O'Connor, 2010; Tong, Lee, Tse, & Law, 2014; Zhou, Ye, Pearce, & Wu, 2014) and text mining (Berezina, Bilgihan, Cobanoglu, & Okumus, 2015; Li, Ye, & Law, 2013) are used extensively to solicit customer intelligence from online travel reviews.

Table 1 Twenty-three hotel attributes that influence customer satisfaction

Attribute category	Detailed attributes
Physical setting (Room)	Room/bathroom amenities, room size and layout, room cleanliness, additional welcome facilities
Physical setting (Hotel)	Availability of Wi-Fi, public facilities (lounge, lobby, pool, and fitting center), dated level (old/new), noise level, entertainment facilities
Physical setting (Food)	Food variety (including Western food), food quality, dining environment, availability of special food services (room service; vegetarian and gluten-free options)
Value	Room, food and beverage, and other prices
Location	Nearness to attractions, city center, airport/railway stations; accessibility
Staff	Friendliness of staff members, language skills of staff members, efficiency of staff members in solving problems

Source: Zhou et al. (2014)

By applying content analysis, O'Connor (2010) identified the common themes of guest satisfaction and dissatisfaction in the London market as reflected on TripAdvisor. Hotel location, room size, good staff service, cleanliness, good breakfast options, in-room facilities, comfort, temperature, maintenance, and noise are the top 10 concerns of the respondents. Barreda and Bilgihan (2013) identified the elements of hotel experiences (either positive or negative) as motivating factors for customers to evaluate a hotel and post a review on TripAdvisor. This study noted that service experiences, bedroom and bathroom interiors, location and cleanliness, sleep quality and value, the physical attributes and ambiance of the hotel, as well as the amenities and complementary services are the main hotel experience themes mentioned in travel review websites. In negative reviews, travelers tend to relate concerns regarding cleanliness to a major hotel experience element. A convenient location is the common factor appraised in most positive reviews. To enhance brand image, hotel managers should focus on service quality delivery because this concern is a major theme that inspires travelers to write positive online travel reviews. By analyzing 1345 hotel-related reviews from Agoda.com in the emerging tourist city of Hangzhou in China, Zhou et al. (2014) identified 23 attributes that influence hotel customer satisfaction (Table 1). Among these attributes, good public hotel facilities can enhance customer satisfaction. The language skills of hotel staff members, especially English fluency, are the major concern of international travelers who stay at hotels in Hangzhou.

Established as a teaching and research hotel by the Hong Kong Polytechnic University, Hotel ICON is devoted to applying new concepts. Three "prototype" guest rooms with a theme of well-being, technology, and sustainability were purposely designed for testing the new concepts in hotel management (Tse, 2012). To investigate guests' attitudes toward the application of new technology at Hotel ICON, Tong et al. (2014) analyzed both English and Chinese reviews from several hotel review websites between May and August 2013. Among the 3088 reviews, 12 technology-related keywords were identified (Table 2). This study

Table 2 Twelve technology-related keywords

Blu-ray disc player	Printer
Bose iPod dock/speaker	Technology
Electronic door locks	Television (TV)
Electronic lighting system	VoIP Phone
iPad	Website
iPhone app	Wi-Fi/Internet

Source: Tong et al. (2014)

determined that the two most important benefits for travelers when selecting a hotel are whether or not guest rooms have Internet connections, and whether or not free Wi-Fi is available because Wi-Fi/Internet was frequently mentioned in the reviews. Interestingly, the findings also suggest that travelers are more concerned with in-room technology applications than with those applied to a hotel's service delivery processes, such as check-in and check-out services at the front desk, the food delivery service at hotel restaurants, and catering functions. Although technology-related reviews merely account for 20 % of total reviews, Tong et al. (2014) reported that innovative technology improves the brand image of a hotel and, more importantly, enhances the guests' overall lodging experience and satisfaction. Au, Buhalis, & Law (2014) compared the complaint behaviors of Chinese and non-Chinese guests at hotels in Mainland China. Although service quality accounts for the majority of online complaints by all customers, the posted negative reviews showed non-Chinese guests focus specifically on the practical use of hotel facilities. Chinese guests rarely complain regarding hotel prices, a tendency that can be explained by the cultural value of saving face. Memarzadeh and Chang (2015) focused on negative word-of-mouth diffusion regarding Southeast Asian luxury hotels. Three types of consumer complaints were identified through the content analysis of online reviews: the inferior quality of hotel facilities, inattention to guests' orders, and inappropriate attitudes of hotel staff members. Instead of complaining directly to hotel managers, hotel guests express their anger toward such issues through travel review websites as a major channel.

As previously mentioned, text mining is increasingly used to analyze online reviews. The use of search engines under text mining software (e.g., IBM's intelligent Miner for Text and SAS Text Mine) minimizes the human effort exerted in exploring the patterns and rules of voluminous textual data (Lau et al., 2005). To some extent, the number of human errors logged in the data tabulation process is reduced. Li et al. (2013) studied the determinants of customer satisfaction with hotels in Beijing by using the text-mining technique. The convenience of transportation, food and beverage, accessibility to tourist destinations, and value for money are the prevalent factors that motivate travelers to reserve hotels in China. Although beds, front-desk services, guest room size, and decoration are important in selecting a hotel, most travelers are disappointed with these attributes. Hotel lobbies and sound insulation are also important to international travelers who are booking luxury hotels rather than budget hotels in this country. Berezina et al. (2015)

determined through text-link analysis that hotel services satisfy customers, whereas furnishings (e.g., beds, carpets, and towels) and financial issues (i.e., money, charges, credit, and cost) motivate dissatisfied customers to post negative reviews online.

Content analysis and text-mining focus on identifying one-shot hotel demand attributes; thus, these methods cannot detect changes in the underlying factors over time. In a recent study, Li et al. (2015) suggested a new research method called the emerging pattern mining (EPM) technique to identify emergent demands for hotel attributes that are important to travelers' decision to book a hotel. This study considered extensive, text-based online reviews posted from 2009 to 2013 (i.e., a total of 118,000 pieces of data over 5 years), and identified the emergent hotel features of interest to travelers. Rooms, staff members, location, breakfast options, service, cleanliness, food, pool, floors, and views were determined as the 10 most popular hotel features based on the online reviews. Among these features, hotel facilities (e.g., clubs, lounges, and pools), services, and food have been and continue to be important hotel preferences, as reflected in online reviews. By adopting EPM, hotel managers can customize their market offerings and maximize ROI by refining their resources to meet future customer needs and demands.

7 Responses to Online Reviews

Positive online reviews generally add value to the hotel industry because the practice of reviewing effectively generates free publicity (Stagg, 2011). Negative feedback damages corporate reputations if the hotels in question do not respond appropriately to outrageous claims. Nonetheless, not all hotel managers are aware of the severe influence of negative word-of-mouth on consumers; a few operators even underestimate the damage caused by this negative action to corporate reputation (Mauri & Minazzi, 2013). Even though several hotel managers realize the increasing importance of using online reviews to address customer concerns, not all hotels respond to such reviews. Only a few hotels actively manage their goodwill on online review websites (O'Connor, 2010). Less than 0.5 % of reviews have received responses from the management even though review websites, such as TripAdvisor, have provided a mechanism that allows hotel managers to deliver immediate feedback to negative comments (O'Connor, 2010). By interviewing the hotel managers responsible for responding to online reviews, Park and Allen (2013) determined that managers who respond frequently to online reviews perceive such reviews as an honest reflection of customer sentiment. By contrast, managers who never respond believe that such posts are merely exceptionally positive or negative. Several managers even feel threatened by customers who write poor reviews on purpose to obtain freebies (BBC, 2014). Consequently, managers hesitate to acknowledge guest complaints. The following section relates an experience by a hotel:

To prevent "customers from defaming" the business, the Broadway Hotel, a three-star, family-run budget hotel in Blackpool, England incorporated a "no bad review policy" into its terms and conditions. Guests were required to sign the booking document that contains the following statement:

"Despite the fact that repeat customers and couples love our hotel, your friends and family may not. For every bad review left on any website, the group organizer will be charged a maximum £100 per review."

A retired van driver Tony Jenkinson, 63, and his 64-year-old wife Jan, who had stayed a night at the Broadway Hotel in August 2014 were disappointed with the hotel. They even described the place as a "filthy, dirty rotten stinking hovel run by Muppets" on TripAdvisor and posted three shocking photos showing peeled wallpaper, cracked plaster, and dirty shower. The couple wrote ". . . the wallpaper was peeling off the walls, the carpet was thin, dirty, and stained. The bed was something else, it must have come out of the ark, the base was all scuffed and the springs in the mattress attacked you in the night. . .I don't know if they are ever inspected, but if so, I don't know how this place has passed!"

A few days later, the couple was fined £100 by the hotel due to the negative review, although the room rate charged to the couple was only £36. The hotel explained in the statement that "we exercised this policy with Mr. and Mrs. Jenkinson as we felt extremely upset by their actions and insulting comments towards our staff. . .we agree there is room for improvement at our establishment and we desperately want to turn things around."

The incident made international news and generated mass media coverage. Upon receiving a warning from the authorities under the Blackpool Trading Standards, the hotel management cancelled the policy, which is deemed an unfair trading practice, and refunded the extra charge to the couple (BBC, 2014; Cockroft, 2014; Pidd, 2014; Quinn, 2014).

Thus, the imposition of heavy penalties and threats by hotels to stop aggrieved customers from spreading negative word-of-mouth information online may result in a poor outcome. More importantly, this practice could definitely affect the hotel image. The responses of hotel management to online reviews are therefore increasingly important to hoteliers in managing customer relationships and corporate reputations. Zhang and Vásquez (2014) analyzed 80 responses by management to negative online reviews and noted that one-third of such responses are non-specific; that is, the hotel managers from the same chain responded to multiple reviews. Management responses to two different types of customer complaints were identical in a few extreme cases (e.g., one review was on beds, whereas the other review was on the overall service). Only 24 % of management responses provided detailed explanations or specific steps of actions to solve the problems raised. This finding is consistent with that of the recent study conducted by Sparks and Bradley (2014), which postulates that not all hotel responses to negative reviews include any explanation. In fact, 33 % of responses do not even account for service failures.

In the event of negative online reviews, the perceived level of trust and communication quality of specific hotel responses are higher than those of generic responses (Wei, Miao, & Huang, 2013). Hotel responses that contain empathetic statements (e.g., "We know that it does not feel good to wait and we know that such a situation is frustrating") and paraphrases of the cited problems inspire potential guests to evaluate such responses favorably (Min, Lim, & Magnini, 2015). The promptness of a response does not significantly influence customer satisfaction (Min et al., 2015). This finding contradicts the common belief that response promptness is critical in the service recovery process. A standard hotel response reduces the credibility of its message because people perceive such responses as commercial communications (Mauri & Minazzi, 2013). Hotel responses moderate hotel performance and ratings for several service attributes in combination with review volume and variation (Xie, Zhang, & Zhang, 2014). By conducting a panel data analysis of reviews of 843 hotels on a review website, Xie et al. (2014) noted that articulate management responses to the geographic condition of a hotel may positively influence hotel sales and location ratings. By contrast, excessive responses by management regarding the hygiene condition of a hotel may negatively affect hotel sales and cleanliness ratings. Therefore, hotel managers should respond to customer complaints differently with reference to the services offered to avoid perceptions that their replies are too general or defensive (Chen & Xie, 2008).

How does an hotelier leverage the response strategy to manage customer relationships and enhance the online reputation of the hotel? The following section presents the successful story of a hotel.

Upon realizing that the TripAdvisor popularity ranking of the Four Seasons Hotel in Austin had dropped from 20th to 27th, General Manager Rob Hagelberg developed an effective response strategy to engage reviewers and to restore the hotel's reputation. Four best practices were implemented by the hotel.

(1) **The monitoring of social activity to detect new customer reviews**. By subscribing to social review monitoring tools (e.g., Revinate), a hotel can receive notifications regarding new reviews on various social media websites. Therefore, they can respond immediately to negative reviews or citations.

(2) **The setting of deadlines for staff members to respond to negative reviews**. The hotel policy stipulates that negative reviews (three stars or below) should be responded to within 24 h. By examining the daily review reports produced by social review monitoring tools, the concerned department may investigate the incident immediately and inform all employees if necessary. The manager in charge then replies to reviewers and devises a way to avoid similar problems in the future.

(3) **The use of a standard template to respond to all negative reviews**. Hagelberg developed a specific structure for responding to negative online reviews. Five elements are included in the template to ensure consistency in responses, although each response should be customized for individual situations. The five elements are as follows:

1. Thanking the customer for taking the time to write a review;
2. Acknowledging any positive comment;
3. Apologizing for the specific complaint or issue;
4. Explaining a specific, forward-looking plan that will address the problem; and
5. Inviting the customer to return.

(4) **The posting of personalized messages to thank people for positive reviews**. Hagelberg responds to positive reviews even though this action is not part of the corporate policy. The manager thanks the reviewers, reiterates the value he finds in a guest's appraisal, and invites the guest to stay at the hotel again. Finally, a personal touch (e.g., a belated birthday wish) is incorporated to enhance the overall hotel experience.

By implementing the aforementioned practices, the Four Seasons Hotel has become the highest-rated hotel in Austin for a period of less than three years (Rajan, 2013; Sanchez, 2013; Short, 2013).

Finally, hotel managers should carefully identify the bottom line when designing response strategies; that is, whether the response simply aims to solve problems in the short term or to improve the hotel in the long run. Consumer reviews can be regarded as a component of the elements for strategic planning; such action can potentially enhance the operational efficiency and effectiveness of a hotel, engage customers, and develop innovative service offerings because the action can enhance the availability of dedicated resources and the generation of a highly collaborative internal environment (Park & Allen, 2013). Although eliminating all negative reviews is impossible, good service quality control, along with complaint resolution and post-service relationship management mechanisms, is crucial in mitigating the damaging effects of such ratings (Duverger, 2013). To illustrate, Hotel ICON has successfully leveraged online reviews to provide the best possible experience to guests.

> Wholly-owned by the Hong Kong Polytechnic University (PolyU) and an extension of its School of Hotel and Tourism Management (SHTM), Hotel ICON is positioned as a world-leading teaching and research hotel. The hotel aims to "train next generation managers while evolving the ideal hotel environment and guest experience." An independent hotel without the support from a chain, Hotel ICON relies significantly on customer feedback to understand guests' needs. From day one, a research was conducted by a PolyU professor to find out how many guests preferred to stay at Hotel ICON because of the reviews they had read before. Furthermore, the top 10 features that guests liked most about the hotel were also identified from online reviews. This study found that nearly 70 % of respondents stay at Hotel ICON after reading online reviews. The location was the only negative factor as per feedback provided by the reviewers. Realizing the disadvantage of hotel's inconvenient location, the hotel has set up a free shuttle bus service for the guests, which cost the hotel approximately US$10,000 per month. This direct response strategy to customers' feedback proved to be effective as few comments about the location was found from the travel review websites afterward. The daily monitoring system from travel review websites such as TripAdvisor, and the posting of professional and immediate responses by management (usually within five days) has become a major practice for the hotel. This response strategy was also complimented by guests, thereby reinforcing the hotel's reputation.
>
> According to Richer Hatter, the General Manager of Hotel ICON: "…from day one, we rely entirely on people's endorsements and recommendations… through the TripAdvisor reviews, we have been able to react and to build customer loyalty and to start engaging our customers."
>
> Hotel ICON was recognized as one of the "Top 25 Hotels in China" and one of the "Top 25 Hotels for Service in China" in the 2016 TripAdvisor Traveler's Choice Awards. The hotel also received a United Nations World Tourism Organization Award for Innovation in Enterprises in early 2014 and was listed as a Forbes Travel Guide Four-Star hotel, the only teaching and research hotel in the world to be so named. (Hotel ICON, 2015, 2016; Hotel Online, 2010; PATA, 2015; Tse, 2013)

8 Conclusions

Given the readily available and affordable wireless mobile devices in the market, such as smartphones and tablets, consumers can easily search for and disseminate word-of-mouth information on the Internet through social media websites. Online reviews exert a considerable influence on consumers' choices and purchasing behavior. Hotel companies also consider all aspects of leveraging online reviews to understand, communicate with, delight, and engage customers. Through partnerships with review websites or the incorporation of review functions into hotel websites, hotel companies seek to drive bookings to their personal proprietary websites and to manage their online reputations. As a component of "big data," online reviews provide valuable information for hoteliers that could facilitate the identification of customer intelligence. Hotel managers must keep pace with the rapid changes in customer preferences and embrace the most recent technology tools to monitor online reviews. Therefore, the response policy of a well-established hotel is expected to reinforce its corporate image and retain customer loyalty.

References

Ady, M. (2013). *From beds to being on top of guest reviews, Starwood is brilliant.* [Online]. Retrieved February 23, 2016, from http://www.trustyou.com/miscellaneous-de/beds-top-guest-reviews-starwood-is-brilliant-3?lang=de

Anderson, C. (2012). The impact of social media on lodging performance. *Cornell Hospitality Reports*, *12*(15), 6–11. [Online]. Retrieved April 12, 2015, from http://scholarship.sha.cornell.edu/chrpubs/5/?utm_source=scholarship.sha.cornell.edu%2Fchrpubs%2F5&utm_medium=PDF&utm_campaign=PDFCoverPages

Au, N., Buhalis, D., & Law, R. (2014). Online complaining behaviour in Mainland China hotels: The perception of Chinese and Non-Chinese customers. *International Journal of Hospitality & Tourism Administration, 15*(3), 248–274.

Barreda, A., & Bilgihan, A. (2013). An analysis of user-generated content for hotel experiences. *Journal of Hospitality and Tourism, 4*(3), 263–280.

BBC. (2014, November 19). *Trip Advisor bad review 'fine' to be refunded by Blackpool hotel.* [Online]. Retrieved April 12, 2015, from http://www.bbc.com/news/uk-england-30111525

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding satisfied and dissatisfied hotel customers: Texting mining of online hotel reviews. *Journal of Hospitality Marketing & Management, 25*(1), 1–24.

Blackshaw, P., & Nazzaro, M. (2006). *Consumer-generated media (CGM) 101: Word-of-mouth in the age of the web-fortified consumer*. New York: Nielsen.

Blal, I., & Sturman, M. C. (2014). The differential effects of the quality and quantity of online reviews on hotel room sales. *Cornell Hospitality Quarterly, 55*(4), 365–375.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism Management, 29*(4), 609–623.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*(3), 477–491.

Cockroft, S. (2014, November 19). *Pictured: Inside the 'filthy stinking hovel' of a hotel which tried to fine couple £100 for leaving a negative review on TripAdvisor.* [Online]. Retrieved April 12, 2015, from http://www.dailymail.co.uk/travel/travel_news/article-2840131/Couple-fined-100-hotel-called-hovel-review-Punishment-TripAdvisor-comments-illegal.html#comments

Duverger, P. (2013). Curvilinear effects of user-generated content of hotels' market share: A dynamic panel-data analysis. *Journal of Travel Research, 54*(4), 465–478.

Farley, A. (2012, May 4). *How hotels are embracing the online customer review*. [Online]. Retrieved July 2, 2015, from http://www.travelandleisure.com/articles/how-hotels-are-embracing-the-online-customer-review

Forrester Research. (2007). *60 percent of Europeans have adopted social computing*. [Online]. Retrieved April 3, 2015, from https://www.forrester.com/Europeans+Have+Adopted+Social+Computing+Differently/fulltext/-/E-res42156

Four Seasons. (2012). *The luxury consumer in the new digital world: Then & now*. Four Seasons Group. [Online]. Retrieved July 2, 2015, from http://www.fourseasons.com/content/dam/fourseasons/web/pdfs/landing_page_pdfs/2012_TRD_Report_final.pdf

Hong, S., & Park, H. S. (2012). Computer-mediated persuasion in online reviews: Statistical versus narrative evidence. *Computers in Human Behavior, 28*(3), 906–919.

Hotel ICON. (2015). *The Hotel ICON Story by TripAdvisor*. [Online]. Retrieved April 25, 2015, from https://www.youtube.com/watch?v=ZUgOSaFALrg

Hotel ICON. (2016). *Awards*. [Online]. Retrieved February 12, 2016, from http://www.hotel-icon.com/about-the-hotel.aspx#/award

Hotel Online. (2010). *PolyU unveils name of its teaching and research hotel: Hotel ICON redefines standards in hospitality and hotel management*. [Online]. Retrieved February 13, 2016, from http://www.hotel-online.com/News/PR2010_1st/Jan10_PolyUIcon.html

Kim, E. E. K., Mattila, A. S., & Baloglu, S. (2011). Effects of gender and expertise on consumers' motivation to read online hotel reviews. *Cornell Hospitality Quarterly, 52*(4), 399–406.

Kusumasondjaja, S., Shanka, T., & Marchegiani, C. (2012). Credibility of online reviews and initial trust: The roles of reviewer's identity and review valence. *Journal of Vacation Marketing, 18*(3), 185–195.

Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. *Cornell Hotel and Restaurant Administration Quarterly, 46*(3), 344–362.

Law, R. (2006). Internet and tourism—Part XXI. *Journal of Travel & Tourism Marketing, 20*(1), 75–77.

Lee, H., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing, 28*(7), 675–688.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22.

Levere, J. L. (2014, January 15). Wider reach for hotel reviews: Business travelers rely increasingly on websites in planning their trips. *International New York Times*. [Online]. Retrieved July 2, 2015, from https://www.questia.com/newspaper/1P2-36314093/wider-reach-for-hotel-reviews-business-travelers

Lewis, R. C., & Chambers, R. E. (2000). *Marketing leadership in hospitality, foundations and practices* (3rd ed.). New York: Wiley.

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management, 46*, 311–321.

Li, H., Ye, Q., & Law, R. (2013). Determinants of customer satisfaction in the hotel industry: An application of online review analysis. *Asia Pacific Journal of Tourism Research, 18*(7), 784–802.

Lin, P. J., Jones, E., & Westwood, S. (2009). Perceived risk and risk-relievers in online travel purchase intentions. *Journal of Hospitality Marketing & Management, 18*(8), 782–810.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458–468.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality and Marketing, 24*(2), 119–154.

Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management, 34*, 99–107.

McEvilly, B. (2015, July 8). How online review sites are affecting your hotel. *HospitalityNet* [Online]. Retrieved July 2, 2015, from http://www.hospitalitynet.org/news/4070901.html

Melián-González, S., Bulchand-Gidumal, J., & López-Valcárcel, B. G. (2013). Online customer reviews of hotels: As participation increases, better evaluation is obtained. *Cornell Hospitality Quarterly, 54*(3), 274–283.

Memarzadeh, F., & Chang, H. J. (2015). Online consumer complaints about Southeast Asian luxury hotel. *Journal of Hospitality Marketing & Management, 24*(1), 76–98.

Miller, M. J. (2011, October 21). *Starwood, bravely, posts hotel guests' reviews online*. [Online]. Retrieved July 2, 2015, from http://brandchannel.com/2011/10/21/starwood-bravely-posts-hotel-guests-reviews-online/

Min, H., Lim, Y., & Magnini, V. P. (2015). Factors affecting customer satisfaction in responses to negative online hotel reviews: The impact of empathy, paraphrasing, and speed. *Cornell Hospitality Quarterly, 56*(2), 223–231.

NielsenWire. (2009, July 7). *Global advertising: Consumers trust real friends and virtual strangers the most*. [Online]. Retrieved November 27, 2014, from http://www.nielsen.com/us/en/insights/news/2009/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most.html

Noone, B. M., & McGuire, K. A. (2014). Effects of price and user-generated content on consumers' pre purchase evaluations of variably priced services. *Journal of Hospitality & Tourism Research, 38*(4), 562–581.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing and Management, 19*(7), 754–772.

Öğüta, H., & Cezara, A. (2012). The factors affecting writing reviews in hotel websites. *Procedia-Social and Behavioral Sciences, 58*, 980–986.

Ong, B. S. (2012). The perceived influence of user reviews in the hospitality industry. *Journal of Hospitality Marketing & Management, 21*(5), 463–485.

Park, S. Y., & Allen, J. P. (2013). Responding to online reviews: Problem solving and engagement in hotels. *Cornell Hospitality Quarterly, 54*(1), 64–73.

PATA. (2015). *A 'Smart Hotel' redefines hospitality*. [Online]. Retrieved April 26, 2015, from http://www.pataconversations.com/richard-hatter-hotel-icon/

Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management, 50*, 130–141.

Pidd, H. (2014, November 21). Please don't fine me, I'm a journalist: A night in Blackpool's Broadway hotel. *The Guardian*. [Online]. Retrieved April 26, 2015, from http://www.theguardian.com/travel/2014/nov/21/night-blackpool-broadway-hotel-tripadvisor

Quinn, B. (2014, November 19). TripAdvisor couple fined £100 by hotel for bad review. *The Guardian*. [Online]. Retrieved April 26, 2015, from http://www.theguardian.com/uk-news/2014/nov/19/tripadvisor-couple-bad-hotel-review-charged-blackpool-broadway

Racherla, P., Connolly, D. J., & Christodoulidou, N. (2013). What determines consumers' ratings of service providers? An exploratory study of online traveler reviews. *Journal of Hospitality Marketing & Management, 22*, 135–161.

Rajan, R. (2013). *How did the Four Seasons in Austin jump 26 spots on TripAdvisor*. RethinkHotels. [Online]. Retrieved June 1, 2015, from http://rethinkhotels.com/four-seasons-austin-maintain-top-tripadvisor-rankings/

Sanchez, A. (2013, November 13). *How the Four Seasons become the highest-rated hotel in Austin*. E-Marketing Associates. [Online]. Retrieved June 1, 2015, from http://www.e-marketingassociates.com/how-the-four-seasons/

Schaal, D. (2013, August 26). *Travelers visit 38 sites before booking a vacation, study says*. [Online]. Retrieved April 6, 2015, from http://skift.com/2013/08/26/travelers-visit-38-sites-before-booking-a-vacation-study-says/

Shanka, T., & Taylor, R. (2004). An investigation into the perceived importance of service and facility attributes to hotel satisfaction. *Journal of Quality Assurance in Hospitality & Tourism, 4*(3–4), 119–134.

Short, T. (2013, October 31). *How the four Seasons Hotel maintains top TripAdvisor ratings*. [Online]. Retrieved May 1, 2015, from http://overnight-success.softwareadvice.com/how-the-austin-four-seasons-hotel-maintains-top-tripadvisor-ratings-1013/

Sparks, B. A., & Bradley, G. L. (2014). A "Triple A" typology of responding to negative consumer-generated online reviews. *Journal of Hospitality & Tourism Research*. doi:10.1177/1096348014538052

Sparks, B., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323.

Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management, 39*, 1–9.

Stagg, J. (2011). Online review sites must ensure rules are policed. *Caterer & Hotelkeeper, 201* (4689), 5.

Tong, K., Lee, A., Tse, T. & Law, R. (2014). An analysis of online reviews on technology application: A case of Hotel ICON [Online]. In P. M. Chien (Ed.), *CAUTHE 2014: Tourism and hospitality in the contemporary world: Trends, changes and complexity* (pp. 613–621). Brisbane: School of Tourism, The University of Queensland.

Travel Media Group. (2015, March). *Hoteliers' guide to responding online reviews*. [Online]. Retrieved April 6, 2015, from http://travelmediagroup.com/wp-content/uploads/2015/03/TMG-How-To-Respond-To-Online-Reviews-Whitepaper.pdf

TripAdvisor. (2014). *Fact Sheet*. TripAdvisor. [Online] Retrieved November 23, 2007, from http://www.tripadvisor.com.tw/PressCenter-c4-Fact_Sheet.html

Tsao, W. C., Hsieh, M. T., Shih, L. W., & Lin, T. M. Y. (2015). Compliance with eWOM: The influence of hotel reviews on booking intention from the perspective of consumer conformity. *International Journal of Hospitality Management, 46*, 99–111.

Tse, T. (2012). The experience of creating a teaching hotel: A case study of Hotel ICON in Hong Kong. *Journal of Hospitality & Tourism Education, 24*(1), 17–25.

Tse, T. (2013). The marketing role of the Internet in launching a hotel: The case of Hotel ICON. *Journal of Hospitality Marketing & Management, 22*(8), 895–908.

USA Today. (2012, October 9). *Wyndham to invite guests to review its hotels on TripAdvisor*. [Online]. Retrieved July 2, 2015, from http://www.usatoday.com/story/hotelcheckin/2012/10/09/wyndham-tripadvisor-consumer-reviews/1622111/

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management, 30*(1), 123–127.

Wei, W., Miao, L., & Huang, Z. J. (2013). Customer engagement behaviors and hotel responses. *International Journal of Hospitality Management, 33*, 316–330.

Williams, R., van der Wiele, T., van Iwaarden, J., & Eldridge, S. (2010). The importance of user-generated content: The case of hotels. *TQM Journal, 22*(2), 117–128.

Withiam, G. (1995). Measuring guest perceptions: Combine measurement tools. *Cornell Hotel and Restaurant Administration Quarterly, 36*(6), 16.

Xie, H., Miao, L., Kuo, P. J., & Lee, B. Y. (2011). Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *International Journal of Hospitality Management, 30*(1), 178–183.

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management, 43*, 1–12.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 27*(2), 634–639.

Yoo, K. H., & Gretzel, U. (2008). What motivates consumers to write online travel reviews? *Information Technology & Tourism, 10*(4), 283–295.

Zhang, Y., & Vásquez, C. (2014). Hotel's responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context and Media, 6*, 54–64.

Zhou, L., Ye, S., Pearce, P. L., & Wu, M. Y. (2014). Refreshing hotel satisfaction studies by reconfiguring customer review data. *International Journal of Hospitality Management, 38*, 1–10.

Evaluating Destination Communications on the Internet

Elena Marchiori and Lorenzo Cantoni

1 Introduction

Destination Management Organizations (DMOs) represent organizations within the tourism industry in charge of the promotion and marketing of a tourism destination, and can be categorized according to the geographical and political level at which they operate. The role of a DMO is crucial in the tourism industry, as it represents a key success factor for a country as a whole, as well as for regions and cities, because of its efforts to reach global audience (Buhalis, 2003). DMOs are primarily marketing organizations, in particular dedicated to the development of a destination's image, and to coordinating internal stakeholders to provide tourism products and services to visitors (Govers & Go, 2009; Gretzel, 2006). According to Gretzel (2006), a DMO's main activities can be summarized as follows:

- Coordination of shareholders (including the political and business industry representatives);
- Leadership role and advocacy for tourism within the local community, in order to create awareness among the residents on the relevance of the tourism industry;
- Support for the development of tourism facilities and attractiveness;
- Information supporting tourists before and during their visit; and,
- Assistance to third parties such as tour operators and travel agents.

Therefore, one of the main focus of a DMO is to manage the place branding of a destination, which is related to the process of destination image communication to specific audiences (Govers & Go, 2009). A destination can communicate with prospective travelers using the Internet, and in particular managing its own official website, which is considered the main online official communication channel for

E. Marchiori • L. Cantoni (✉)
USI - Università della Svizzera italiana, Lugano, Switzerland
e-mail: lorenzo.cantoni@usi.ch

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_15

DMO. Moreover, website content is of primary importance for destination managers since it affects the perceived image of the destination. As one of the main purposes of a DMO's website is to attract and increase visits to the destination (Qi et al., 2007), the quality of its online communication should be very high; otherwise, as explained by some empirical studies, there might be the risk that a huge amount of users leave the website because of usability problems. Moreover, as stated by the World Tourism Organization, DMO websites can promote destination products and act as a bridge in promoting destination's services and products and communicating with the market (Pike, 2005). Furthermore, a system, which hosts other services or products (or in other words, gives visibility to third party websites like a DMO website for the hospitality players within the destination), should be well-designed and should have great performances in order to satisfy both investors and end-users. In this context, good website usability normally leads to a good website performance; therefore usability performance is a key success factor for a website (Douglas & Mills, 2004; Nielsen, 2006). However, besides the website, there are also many other activities for improving online communication and web marketing for a DMO such as: search engine promotion and marketing, web usages statistics, usability studies, mobile applications, augmented and virtual reality, social media and e-word-of-mouth (Cantoni & Ceriani, 2007; Marchiori, Pavese, & Cantoni, 2012). In particular, online conversations taking place on social media platforms are generally outside the control of whoever has the responsibility to communicate and promote the DMO. While DMOs are realizing the importance of online communication for the accomplishment of their mission, they still need a general framework to analyze their actual performances and systematically improve them. Therefore, a more systemic approach is needed, which is able to detect problems from different data sources about the users and the design, and to prioritize them so that DMO website managers could take informed decisions, better invest their budget and resources, and finally ensure that these efforts are aligned with their business goals.

In this chapter, a combination of a destination website analysis and an online reputation analysis is conducted, showing how they can complement each other, and how they can provide DMO managers with instruments useful to evaluate a destination's online performance. The combined approach is based on two models elaborated on the authors' previous research. The first model is the UsERA (User Experience Risk Assessment) model, which describes the interplay between usability and usages analysis formalizing the interaction between the two approaches and leading to more structured information for website destinations' managers. The second model proposed is DORM (Destination Online Reputation Model), which provides a framework to map online narratives about a destination. The following sections will introduce the theoretical background of this study corresponding to research on destination communications on the Internet. The two models for the evaluation of destination communications on the Internet are then introduced, together with their study methods and case studies. Theoretical and practical implications are finally discussed in the conclusions.

2 Destination Online Communications

Contents published online, whether from official (e.g. a DMO official website), or unofficial sources (e.g. online conversations on social media platforms) can become an object of analysis in order to better investigate:

- From a tourism industry perspective: what the prospective travelers can perceive from online contents will help tourism managers to understand what travelers experienced at the destination; what future travelers may need/search/visit, and ultimately which kind of topics they can encounter online, topics that might influence the decision to visit the destination.
- From a tourist perspective: what they are going to choose as an investment for their future trip, what to expect from a destination, getting ideas, forming their opinions about the place.

For a destination manager, dealing with the DMO website can be considered subject to some degree of risk: actual users are often unknown (although possibly predicted during design), the actual behaviors of the users on the site are often unknown from the outset, and the actual effect or outcome of the experience with the site on the user is difficult to predict. Most importantly, the complexity of the design features of large web applications (and their emergent properties due to their interconnectedness) poses additional levels of unpredictability to such factors, augmenting the risk of negative user experiences. A proper analysis of the user experience risk should inform project managers, communication and web designers in making decisions concerning questions such as: which parts of the application require immediate attention for re-design or improvement? Are my users exposed to potentially negative experiences? How can I optimize the good experiences on my site? Our innovative contribution is the elaboration of basic constructs to analyze and characterize such hurdle of risk issues by holistically leveraging current approaches to usability analysis and usage studies.

Managing online reputation is essential, as shown in Marchiori and Cantoni (2012: 147), in that "the online environment matters for reputation either because it provides published opinions (= proxies of reputation) or because it provides individual opinions (= instances of public opinion). The latter are published and accessible, which makes a dramatic difference in comparison with the pre-online situation. There, mass media provided published opinions, while individual opinions were accessible only through surveys and were not able to extremely influence others' opinions. Online, the same item can be treated as individual instances of public opinion (person X has opinion Y about object Z), and at the same time as published opinion, due to its accessibility to others; even more, it may become highly influential because it is accessed by a number of people; for instance, because that individual user-generated content is well ranked on a search engine".

Online reputation analysis allows tourism operators such as hotel managers and/or destination managers to: monitor mentions on social media platforms; categorize reviews based on topics, and also on importance or urgency; review past and current trends; compare own online presence with the one from competitors, manage customer feedback such as sending immediate responses to reviews. In online reputation analysis contents are generally clustered based on the type of the topic expressed, and two main directions for an online content classification are:

- A bottom-up/inductive approach, which foresees a creation of topic categories after the content analysis. That is, following a saturation approach, every time a new topic is recognized, it is classified in a new category, until no new categories are found. This approach allows to create custom analysis and to have a precise map of the topics associated with the destination.
- A top-down deductive approach: topic categories are created upfront, using a pre-established model, which allows for a systematic analysis, and to perform comparison and ranking between and among a destinations.

The second approach saw an increase in the creation of professional tools for data harvesting and data classification, so-called "tools for monitoring the online reputation". Methods used by those tools are mainly frequency analysis of keywords, sentiment analysis of the online mentions, and topic association with brand values, and/or with predefined topic categories. The rise of those numerous tools, the related online reputation indexes, and content analysis processes rise a methodological issue: in fact, there are not yet standard models and procedures. Therefore, in this context, the DORM model is proposed in order to perform an online reputation analysis for a tourism destination, it implies a human-coding procedure that allows an in-depth analysis of online conversations about a tourism destination.

3 The OCM Model as a Map to Evaluate Destination Online Communication

Cantoni and Tardini (2006; Tardini & Cantoni, 2015) developed a valuable model, named Online Communication Model (OCM), to describe the components, actors involved, and relationships. According to this model, the online presence is a mixture of five components with four pillars related to online presence managed by the owner, while the fifth element representing the communication market in which communication takes place. OCM can be applied to a destination's online communication in order to identify its main components, and in turn being able to evaluate its performance. Applying to a DMO and its official website, it is possible to identify the following components of its online presence applying the OCM framework (see Fig. 1):

- *Pillar I*: *A cluster of contents and functionalities*. It includes all the contents, news, navigation options/functionalities which compose the official website of a tourism destination. E.g. home page contents, photo gallery, possibility to leave a comment, eCommerce, connection with social media, virtual tour, etc. At this level, information and quality of the contents are of the utmost importance. With

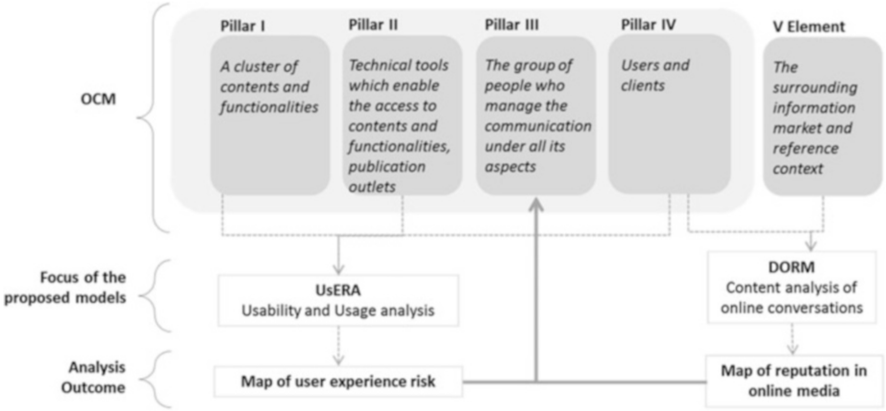

OCM
Pillar I
A cluster of contents and functionalities
Pillar II
Technical tools which enable the access to contents and functionalities, publication outlets
Pillar III
The group of people who manage the communication under all its aspects
Pillar IV
Users and clients
V Element
The surrounding information market and reference context
Focus of the proposed models
UsERA
Usability and Usage analysis
DORM
Content analysis of online conversations
Analysis Outcome
Map of user experience risk
Map of reputation in online media

Fig. 1 Application of OCM to describe the focus of UsERA and DORM models

the advent of social media, a DMO can have an official account on several social media platforms. The contents produced by a DMO on social media platforms are considered as new contents which co-create the online narratives of a destination projected by a DMO.

- *Pillar II*: *Technical tools, which enable the access to contents and functionalities*. This aspect deals with, more specifically, hardware, software, Human Computer Interface, input and output instruments, and the information architecture of the website as a whole. This pillar includes not only owned media but also all other publication outlets: the earned and paid ones (e.g. Social Media channels).
- *Pillar III: The group of people* who *manage communication under all its aspects*. It comprises people who project, design, develop, produce, maintain, promote, evaluate the online communication, and interact with users.
- *Pillar IV: Users and clients*. They are the purpose of the online communication itself, who uses and enjoy the contents and functionalities during the three phases of their tourism experience: before visiting the destination, during their trip, and afterwards.
- *V Element: The surrounding information market and reference context*. This aspect refers to the overall information market around the destination. Due to the uncountable presence of websites, easiness to reach online data and replicate online contents, the online competition is getting heavier for a single website. Moreover, the presence of new publication platforms—such as social media platforms like YouTube or Facebook—and the success of mobile connectivity, makes it possible to remain connected everywhere, and the internet has become the largest, and crowded, public square available.

Figure 1 depicts the connections of the proposed models with the OCM's pillars. The outcome of the proposed analytical models implies suggestions for destination managers in order to balance/modify/improve existing communication, and/or plan new ones for a better and useful online presence. Three main analytical methods are proposed and refers to OCM:

Investigating the Adequacy of Contents/Functionalities and Accessibility Tools Used in a Website

This analysis refers to pillars I, II and IV, and it allows to identify the potential usability risks of an in-house communication activity represented by the contents and functionalities created by a DMO online outlets. The method proposed is a usability analysis, intended as "the adequacy of contents/functionalities (pillar I), and accessibility tools (pillar II), between themselves and with respect to the users (pillar IV) and the relevant context (world). Moreover, this adequacy has to be measured by taking into consideration the goals of people who commission, design, develop, promote and run the website (pillar III)" (Cantoni & Tardini, 2006: 129–130).

Usability studies have typically focused on the empirical evaluation of the efficiency and effectiveness of the website to support user goals and tasks, with the aim of improving the quality of the design (Au Yeung & Law, 2003; Brinck, Gergle, & Wood, 2002; Nielsen & Mack, 1994), as per the definition of ISO—the International Organization for Standardization: "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO, 1998: n. 11). Regarding the evaluation presence on social media platforms: only the contents can be managed by a destination manager, while the functionalities and the overall structure of a social media platform are generally own and managed by a third party, and thus a usability analysis is not generally performed on those platforms. Therefore, the proposed model UsERA uses the usability analysis as a first step for the analysis of a website user experience risks in a website or mobile app.

Investigating the Performance of Contents/Functionalities Presented in a Website/Mobile App

This analysis refers to pillar IV, and it allows to identify which contents and functionalities are generally accessed, and overall which are the navigation. The method proposed is a usages analysis (through a third-party service, like Google Analytics, or through log files analysis). From a more technical viewpoint, log files are the traces left by the user while visiting the web site. This specific group of files record users' activities while interacting with the server. The study of log files is not only an engineering activity: log files analysis can give interesting information at a communicative level (Cantoni & Ceriani, 2007) such as the study of the users' paths along the website (Pitkow, 1997), by which it is possible to optimize the communication flow of the application. The study of usages has mainly addressed the analysis of traffic, aggregated user's paths and ecological factors (e.g.: referrals, keywords used on search engines prior to access, etc.), with the purpose of informing marketing actions and visibility (Atterer, Wnuk, & Schmidt, 2006). Thus, this activity allows the understanding of the audience of the websites: who are the users? Where do they come from? When? Through which links? Having searched which kind of keywords on search engines? Time of visiting? How often?

After a promotional campaign or another event? Which contents/functionalities do they access? etc. When it comes to analyze the usages of an account within a social media platform, in general it is a common practice to refer to the statistics provided by the social media platform itself. Therefore, the proposed model UsERA uses the usage analysis as a second step for the analysis of a website user experience risk.

Investigating What Is Said Online by Travelers About a Given Destination This analysis refers to pillar V (and partially IV), and it allows to identify the pros and cons related to the online communication of a tourism destination. The method proposed is an online reputation analysis in the form of content analysis, intended as the analysis of the topic(s) expressed in the online contents and their related sentiment. In tourism, the reputation of a destination is important as prospective travelers who do not have previous experience with a destination encounter several risks/limitations during their decision making and, therefore, use the reputation of a place to guide their travel decisions. Recently, several researchers (Gretzel, 2006; Passow, Fehlmann, & Grahlow, 2005; Tussyadiah & Fesenmaier, 2008; Yang, Shin, Lee, & Wrigley, 2008) have noted that the role of recommendations from several second-hand sources, which act as reputation mediators, is crucial in this decision-making. Moreover, as noted from authors' previous research (Marchiori & Cantoni, 2012), in the online domain, word-of-mouth comments are generally found on social media websites and can be considered to be proxies of readers' perceived reputation, and of a dominant public opinion (reputation). Besides, Tussyadiah and Fesenmaier (2008), found that the narrative reasoning people possess and with which they can retrieve information is more effectively presented through stories, particularly if users can identify themselves with the story's characters. Thus, a strict connection between the online messages (where narratives/opinions are expressed) and the concept of reputation is underlined by the fact that the perception of stories in a place may be due to the act of mentally summarizing what has been learned from online content exposure. Therefore, the proposed model DORM uses a content analysis of online conversations as a procedure to analyze the online reputation of a tourism destination. The following paragraphs present and discuss two proposals, i.e., UsERA—User Experience Risk Assessment Model, which refers to the OCM's pillars I, II, and IV, and DORM—Destination Online Reputation Model is presented, which refers to the OCM's V element.

4 Evaluating Destination Website Usability and Usage: The UsERA Model

The proposed User Experience Risk Assessment model (UsERA) refers to previous research partially conducted by the authors (Adukaite, Inversini, & Cantoni, 2013; Inversini & Cantoni, 2009; Inversini, Cantoni, & Bolchini, 2011; Marchiori & Cantoni, 2013; Marcus, Schieder, & Cantoni, 2013). The UsERA model has been developed in order to integrate usability evaluation and analysis of usages for DMO websites in an innovative and holistic framework. As usage analysis gives merely inferential indications about users' behavior on the website (Cantoni & Ceriani, 2007), with the UsERA Model it was possible to map usability shortcomings inherent to design (threats), the actual usages of the website or the exposure to usability problems (vulnerability), and the ability of the users to overcome the usability problems (resilience). The model emphasizes the relationship among usability and usages in order to provide indications to website's managers. UsERA model results are generally compared with the objectives and goals of the website managers (as it is stated in the usability definition by Cantoni & Tardini, 2006; Tardini, Adukaite, & Cantoni, 2014), to formulate appropriate design interventions.

Usability problems of a website and/or a digital application can be identified though usability methods, which generally aim to identify significant risk factors for a detrimental user experience. Usage analysis of a website and/or a digital application can identify the probability for users to be actually exposed to the usability problems. Based on this theoretical elaboration, a proper analysis of the user experience risk would inform project managers, communication and web designers in making decisions concerning questions such as: what parts of the digital application require immediate attention for re-design or improvement? Are my users exposed to potentially negative experiences? How can I optimize the good experiences on my site? Therefore, UsERA provides constructs and procedures to analyze and characterize such potential usability obstacles, and the related user experience risks by leveraging current approaches to usability analysis and usage studies.

5 Components of the UsERA Model

The UsERA model proposes to treat the user experience risk as composed of three main elements (see Fig. 2), which are explained in the following paragraph, and are: (1) threats, as usability problems inherent to the design; (2) vulnerability, as the exposure to usability problems, and (3) resilience, as the users' ability to overcome usability problems.

(a) "Threats" as Usability problems inherent to the design
The design complexity of large destination websites is often prone to usability problems. A usability problem is defined as a design defect that is a potential threat to an optimal user experience. A long standing tradition of web usability analysis and web engineering acknowledges that usability problems of varying severity are typically inherent to how the application has been designed and, therefore, eventually lie at one or more of the following seven design dimensions (Triacca, Inversini, & Bolchini, 2005):

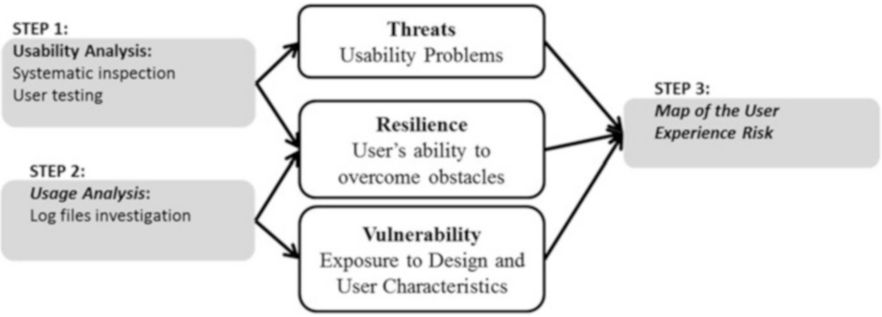


Fig. 2 Components of the UsERA model

- Content: the core information messages of the websites, from text, to multimedia. An example of potential threat, or usability problem, at this level is the presence of obsolete content, or the absence of contact information.
- Information architecture: the overall organization of the content in chunks and sections. An example of potential threat at this level is the classification of the content using a limited set of criteria (e.g. only by geographical location), which do not correspond to the user's natural reasoning in exploring information (e.g. I want to go skiing, no matter in which specific location).
- Navigation and interaction: the strategies by which users can use and move around the information architecture through links and interact with content. An example of threat at this level is the lack of intuitive mechanisms to use an interactive map, to navigate to in-depth information from it, and to print the desired information.
- Services & transactions: the strategies by which specific operations and services are organized, structured and made accomplishable by the user.
- Search functionality: the way an internal search engine supports accurate and efficient retrieval of information.
- Labeling and interface semiotics: the way in which all the above mentioned aspects are conveyed at the interface level through naming conventions, layout strategies, metaphors and labels.

This analytical modeling of the threats reveals critical areas of the site that do not necessarily determine a risk of negative experiences, but need to be carefully and jointly considered with respect to the vulnerability that our users may manifest.

(b) "Vulnerability" as exposure to usability problems

An area of a DMO website (e.g.: a list of hotels) with severe usability problems (out-of-date or missing contact information) may not be considered a too dangerous threat if, for instance, no user ever bumped into it. The fact that the list of hotels is very difficult to find mitigates the potentially destructive effect of the threat because the actual exposure of our users to it can be considered very low or null. This example shows that user's vulnerability to a threat can be defined as the exposure of the users to it, identified in terms of actual traffic or potentially accessible pathways. However, the fact that few users access the list of hotels has to be carefully analyzed: is that in line with the overall website goals or that should be the most important website area? Are there usability problems (most probably in the navigation layer) that prevent users from reaching it? Are promotional activities bringing the most appropriate publics to the website? In addition, vulnerability also depends on the specific characteristics of our users, which may be more or less sensitive to a threat. For example, web-savvy users may find no problem in downloading an additional Flash player to enjoy a video. Senior users new to web technology, on the contrary, may find it difficult to install a plug-in.

(c) "Resilience" as the users' ability to overcome usability problems

Resilience is defined as the users' ability to overcome obstacles/threats. Whereas vulnerability identifies the danger of a potential or actual exposure to a threat, risk can be highly mitigated by considering how and whether users actually overcome a threat. Let's assume that users often visit the section to subscribe to the newsletters, and they are exposed to a set of poorly organized pages to create an account, necessary to subscribe to the newsletter. The fact that 90 % of the people accessing the newsletter section are eventually able to complete the task is a clear sign of high resilience, and this mitigates the overall risk of negative user experiences. As the user population changes, however, this resilience may vary, causing a high level of risk.

5.1 Components of Usability Analysis

The usability analysis can be performed through (1) expert inspection, and/or (2) user testing. On the one hand, (1) experts are able to identify usability issues, and can compare many websites/mobile apps belonging to the same domain; on the other hand, (2) "naive" users are able to spot unforeseen problems, which might be overlooked by the experts (Inversini & Cantoni, 2009). Table 1 presents a simplified map of different approaches that can be taken by usability experts and/or users:

Heuristics are guidelines to be applied in order to find possible design problems: the term itself comes from the ancient Greek verb that signifies "to find". Among the most known and prolific authors who have proposed usability heuristics, we mention here Jakob Nielsen (see the most popular online resource of usability heuristics: www.useit.com. In 1995, Nielsen proposed "10 Usability Heuristics for User Interface Design", later on integrated with hundreds of additional items, linked to specific domains or devices. The most popular heuristics proposed by

Table 1 Components of usability analysis

Expert inspection	Heuristics
User testing	User scenarios

Nielsen is: “Visibility of system status”, and it appears relevant as it is paradigmatic from a communication viewpoint. That is, in the act of interacting with a website / mobile app, a system should provide constant and meaningful feedback to users, in order to keep them informed about their navigation. This is done through icons (e.g.: a moving arrow, a completing bar, a turning hourglass), through anticipation of steps (e.g.: in a booking funnel we are informed about the steps we have already taken, and the ones that remain), through signaling of turning points (e.g.: “by clicking this, you accept/your credit card will be charged”), through confirming statements (e.g.: “you have just booked seat X, on flight Y”). Therefore, every usability expert can build up her/his own set of heuristics, linked to specific application domains, which usually encompass content, navigation, interface, and technology.

In ***user testing,*** it is necessary to recruit common users belonging to the same groups for which the application has been developed. It is important to consider that for the selection of the sample, it is not possible to have a statistical representativeness (which cannot be reached at all), but it is important to reach a saturation point. Every user can find new issues during the usability test, which might partially overlap with those already found by others. Indeed, after a certain number of users (this occurs with about 8/12 users) it is very unlikely that new major issues will be found (hence saturation is reached). For a user test it is possible to ask users to freely navigate an application/website, however, the most common strategy is to define suitable user scenarios. User scenarios are also extensively used by experts to guide their own inspections.

A ***user scenario*** is a vivid story about a successful interaction with the website, it describes an experience that is desirable for both end-users and publishers. They are defined in collaboration with different stakeholders, including the publisher, so to make sure that they are in line with their communication and business goals. A user scenario presents three main components:

- user profile;
- overall goal;
- specific tasks: activities a user should be able to execute on the interface.

Below, there is an example of a user scenario that can be applied to a cruise company’s website:

User Profile Maria—62 years old—has not any previous cruise experiences. She accesses the website from home and office; her connection speed is good (high speed ADSL). She is familiar with the computer and the internet. She does not have any domain knowledge. She is used to surfing to find news about luxury exotic vacations. She is a “freedom boomer”, affluent empty nesters with grown up children and passion for travel. She loves to travel with her husband and couple of friends on luxury land tours or boutique resorts. She found the website of cruising company X and is sufficiently motivated to find out about the offer for luxury cruise vacations.

Goals To see destinations, ships and onboard activities. She also wants to understand what company X's lifestyle is.

Tasks Five tasks are proposed as following:
T1. Find X's Unique Selling Points / Peculiarities
T2. Find covered destinations
T3. Explore the vessels
T4. See proposed accommodation
T5. Find information about the life onboard

In a user testing session, after a briefing, the user is invited to perform the scenario(s), while the screen is recorded (also her/his face and voice might be recorded). The person managing the user testing follows the navigation, taking notes of specific problems/issues, which are then discussed with the tester at the end of the session. To get a better understanding of the stream of thoughts of the user-tester (and of the reasons why user-testers choose specific options), many times they are requested to verbalize their thoughts (thinking aloud). Eye-tracking features might be also added, so to follow eye movements of testers, and better understand how they visually process the interface (in those cases retrospective thinking aloud is used: the navigation is re-played and users comments about their eye movements).

Indeed, not all usability problems are likely to have the same impact on the overall experience of the website/mobile app, and on its effectiveness. They can be interpreted as threats to a satisfying user experience. Corresponding risks are to be considered as the relationship between such threats and the vulnerability of the users. In order to assess how many users are exposed to such threats, and how many are actually vulnerable to them, web analytics can be fruitfully combined with usability, as it has been proposed by the UsERA model (Inversini et al., 2011; Inversini, Cantoni, & Bolchini, 2010; Tardini et al., 2014).

5.2 *Applying UsERA to Analyze Destination Website Usability and Usages*

The following paragraphs explain the application of the UsERA approach to the evaluation of the user experience risk of a destination website. The case study of an Italian destination website: www.turismo.ravenna.it, presented in Inversini, Cantoni, and Bolchini (2010), is used to show how the UsERA model can be translated into an analytical instrumentation composed by three main steps to inform new discoveries in the study of the user experience and to enable better digital designs.

5.2.1 Step 1: Usability Analysis

The usability analysis of the DMO website revealed the following usability issues:

- *Accuracy of the information* (e.g. information not precise, and poorly structured. Hyperlinks pointing to external resources sometimes broken).
- *Second level menus position consistency* also affects different tasks (e.g.: the navigation menu placed above the text and not visible by users using screens with a resolution less than 1324 × 768. In this issue two different usability problems can be highlighted:
  - *Layout conventions:* it is very uncommon to have second level navigation menus positioned above the main content. Existing patterns in web page design follow different convention.
  - *Orientation*: the awkward position of the menu affects the sense of overall orientation in the navigation architecture (Where I am and where I can go).
- *Segmentation of the information:* Information segmentation refers to editorial decision of the website designers and content managers to actually segment the information within different pages. Information should be well divided and organized in the whole website in order to let the user easily access each piece of information.

A section affected by the usability issue related to the information segmentation was the accommodation section, and predominantly its booking system. In this case, the main usability issues were related to:

- *In-depth anticipation*: the user is not aware of the path and steps s/he is supposed to do inside the application to accomplish the task s/he has in mind; the user should click several times before arriving to the dedicated hotel page—choosing an hotel according to the availability—and discover that s/he needs to ask (trough hotel website or as in most of the cases via email) the room availability.
- *Icons predictability:* icons are explained above the table of the hotel availability, but there is no clue about the meaning of the number and the letter in the cells.
- *Labeling consistency:* two labels identify accommodation sections: "ospitalità" (i.e. hospitality) and "disponibilità alberghiere" (i.e. accommodation availability); information is not consistent and these two different buttons lead to different pages/sections.

5.2.2 Step 2: Usage Analysis

Usage analysis describes the end-users traffic volume on the website outlining the most viewed sections in the website and analytically describing the different sections viewed. The method used for studying the usages is a log files analysis. Log files can be analyzed using ad hoc software (otherwise a third-party service can be used, such as Google Analytics). For this case, a log files analysis was conducted in order to assess vulnerability and resilience of the real users on the website over 1 year timeframe (1st of October 2007—1st of October 2008). The results show that the DMO website received 29,637,297 hits in 461,980 visitor sessions, for a total number of 289,714 unique visitors. The most visited page was the home page, which collected 69.1 % (354,925) of the total hits. Then different pages about the destination events and initiatives such as: "notte d'oro (= a night event)" received 34,283 hits, and "mare d'inverno" (= "winter sea" event) received 9,834 hits. One unexpected popular session also dealt with the bus and cycling paths download (4,099 sessions). Most of the single sessions were just initiated and terminated in the home page (116,193 sessions).

5.2.3 Step 3: Map of the User Experience Risk

Usability results and log files results were then displayed on a website map obtained with a reverse engineering exercise: different colors highlighted different website sections where risks of negative user experience could be found. Usability evaluated threats (i.e. usability problems inherent to the design), while log files analysis helped in evaluating resilience as users' ability to overcome the obstacle (i.e. accomplish tasks) and vulnerability as the degree of exposure to design (i.e. general visits). In general terms, it was possible to claim that:

- Events section was the most popular section within the website (1,331,800 hits and 334,683 user sessions); users could find and download online guides, maps and brochures.
- Accommodation section presented few accesses and was further investigated: total hits count for the whole accommodation section over the given period was 901 (i.e. 0.003 %). Among these, 297 hits were on the home page of the accommodation section (32 % of 901) but only few user sessions stopped on the accommodation home page (0.002 %): anyway due to the website structure it was not possible to follow all the user paths because some contents/functionalities had been hosted on a different web server (different server log files were not available for analysis).

Log file data showed two important results: on one hand, most of the users of Ravenna DMO website overcame usability issues (i.e. threats) to reach the given piece of content (e.g. news/events) demonstrating a high resilience in a high visited section (high vulnerability); risk was not high and usability problems seemed not to strongly influence the user experience. On the other hand, threats (i.e. usability problems) within the accommodation section were quite high; resilience was high because few paths stopped into the accommodation home page (i.e. 0.002 %) so that users overcome the obstacle; risk is low, mainly due to the low users' exposure.

Once the data are known, with the caveat that they are always approximations, it is important to understand and evaluate what kind of implications it has on our business. That is, it is necessary to make hypotheses and inferences, which can then guide managerial decisions. Here below, based on the OCM, are presented three major strategies on how to perform data investigation:

- Operate on the content (pillar I): remove pages/sections never or poorly accessed, optimize content to make it more suitable for human readers and for search engines (SEO: Search Engine Optimization). For example, let us consider the following case:
  - *Data*: the section with the Russian translation of a destination's website is almost never visited
  - *Hypothesis*: people are not interested
  - *Decision*: it is discontinued to avoid useless translation costs
- Operate on the structure or on the publication outlet (II pillar): distribute contents on different publication channels, reorganize the navigational structure to ensure more internal visibility to under-used sections, or to remove obstacles against the completion of relevant processes (e.g.: booking funnel):
  - *Data*: same as above
  - *Hypothesis*: people landing on pages other than the home, do not realize that Russian translation is available
  - *Decision*: the website is re-engineered in a way so that in every single page it is possible to access a different language
- Operate on the users themselves (pillar IV), putting in place adequate promotional activities (online marketing, online PR, SEM: Search Engine Marketing) to invite the right users:
  - *Data*: same as above
  - *Hypothesis*: Russian-speaking people do not know about the existence of the website
  - *Decision*: several promotional activities, both offline and online, are done in order to make the website known to the Russian market

Finally, it is important to consider that who is managing a website/mobile app, should be constantly aware of the fact that the competitors are just a click away, and that if we fail to deliver a high quality online experience, users can bounce out.

6 Evaluating Destination Online Reputation: The DORM Model

The proposed Destination Online Reputation Model (DORM) refers to previous research conducted by the authors (Inversini et al., 2009; Inversini, Marchiori, et al., 2010; Marchiori & Cantoni, 2012; Marchiori & Cantoni, 2015; Marchiori et al., 2012; Marchiori et al., 2010, 2011). The research perspective of those studies is rooted in the media effects, social psychology, linguistic, and organizational reputation studies. Valuable contributions to reputation studies have been made in particular in the field of organizational reputation (Berens & van Riel, 2004; Money & Hillenbrand, 2006), where scholars provided a map of the current reputation measurements for the investigation of instruments that allow for understanding the value of reputation for a business. In these studies, authors used the Walsh and Wiedmann (2004) theoretical causal framework of reputation, which sees the reputation construct composed of antecedents and consequences. Money and Hillenbrand (2006) referred to the Fishbein and Ajzen (1975) causal framework for the investigation of the perception components to study within reputation research: experiences, beliefs, attitudes, intentions, and behaviours with respect to a given object, where:

- Experiences are considered as information elements, which concur in the creation of beliefs;
- Beliefs are considered elements that determine people's attitudes toward an object; and,
- Attitudes toward an object are related to people's intention to perform certain behaviors with respect to the object, and each intention is related to the corresponding behavior.

Within this stream of reputation studies, the proposed evaluation approach to destination reputation presence focuses on the intangible assets of a complex object such as a tourism destination (i.e.: belief about thematic dimensions and attitudes expressed as emotional appeal) expressed online in the form of online conversations. The consequences level is also investigated and is represented by the potential change confirmation/disconfirmation of prior beliefs, which in turn might generate intention-behaviors towards a destination. As mentioned before, as a tourism destination is a complex and (partially) unstructured organizational network, in order to proceed with a systematic analysis of the dimensions of a destination, which are the objects of the online conversations, and are perceived as dominant, it is necessary to break a tourism destination down into measurable dimensions (multidimensional traits). Previous research has shown applications of the organizational reputation principle to the tourism-related domain, such as the Country Reputation Index (Passow et al., 2005), and its revised version by Yang et al. (2008). The development of this index followed a similar process as that seen in the Reputation Quotient model (Fombrun & Shanley, 1990) used for corporate reputation analysis, where a set of appeal dimensions is used to capture stakeholders' perceptions (beliefs and attitudes) related to a specific object. Findings from these studies underline how the analysis of reputation in tourism-related studies using the causal reputation framework allows for systematic analysis.

In this stream of research, a deductive approach has been used in order to develop the theoretical classification system for tourism destination-related online conversations, named DORM: Destination Online Reputation Model. DORM contributes to the online content analysis studies in tourism by introducing a top-down deductive perspective. It provides pre-established topic categories about the reputation dimensions, which allow for a systematic content classification and a comparison among similar objects, such as tourism destinations.

6.1 Components of the DORM Model

DORM considers the specific characteristics of a tourism destination as a unique and complex organizational unit of the tourism industry. Researchers used the Reputation Quotient (RQ) and the adapted version RepTrak (2006) presented by the Reputation Institute, which are based on 23 drivers that work as predictors of reputation (Vidaver-Cohen, 2007). Using these two models (RQ and RepTrak) as a base, the authors were able to adapt the core dimensions and reputation drivers to the reputation of a tourist destinations considering the peculiar characteristics of the tourism industry. The framework was created and adapted thanks to an extensive literature review and it was validated through semi structured interviews with domain experts in order to collect the interviewees' perception on how the elements of the proposed model relate and influence the perception of reputation in regards of a tourism destination. During the semi structured interviews, domain experts were asked to rank the importance of each of the core dimensions emerged from the literature, and to add any additional element perceived as having an influence upon the overall reputation of a destination and which was not previously considered.

DORM (see Table 2) consists of five dimensions (Products and Services, Society, Governance, Environment, and Performance), nine sub-categories for the dimension Products and Services (Accommodation, Food & Beverage, Site attractions, Outdoor activities, Events, Entertainment, Transportation and accessibility, Infrastructure and facilities, and Other), and 14 drivers (four for the dimension Products and Services, two for the dimensions Society and Governance, three for the dimensions Environment and Performance).

DORM was then tested through a tourists' survey held in two Italian airports (July and August 2010) with the main objective to define which online topics were relevant or missing by tourists regarding their decision-making process to visit a destination. This study highlighted that travelers were aware of the existence of online content produced by other tourists, and they believed that they were influenced by them in their decision-making process. In particular, five main topic dimensions emerged as the most relevant topics in the tourist information seeking process. These are online content information regarding: the tourism destination products and services that are good value for money; the local cultures and traditions at the destination; the tourism experience at the destination; the safety of the environment at the destination; and the weather at the destination. This framework was created and tested with online case studies (Inversini, Marchiori, et al., 2010; Marchiori, Inversini, Cantoni, & Dedekind, 2010). Guidelines for online content interpretation specific to online conversations (can be communicated by text, image, video, or other symbol) have been described in Table 2. A coder can use those guidelines to classify the main relevant topic expressed according to the given reputation drivers and indicate the sentiment expressed using a 5-point Likert Scale.

Table 2 DORM core reputational dimensions and related drivers

Core dimensions	Drivers	Examples of topic expressed (can be communicated by text, image, video, or other symbol)	Examples of sentiment (positive/negative)
Products and Services Subcategories: Accommodation Food & Beverage Site attractions Events Entertainment Transportation Infrastructure Other	[d1]: Destination [D] offers a satisfying tourism product or service	Accommodation: hotel room, concierge. Restaurant: menu, valet. Sports: baseball game Package service: guided tour through city	"The waiter gave us excellent wine recommendations with our dinner"
	[d2]: [D] offers a pleasant atmosphere	Weather: comfort and seasonal aesthetics. Attractions: design, cleanliness Architecture: museums, concert halls	"Autumn in New York is a beautiful time to visit and take lots of photos"
	[d3]: [D] offers products and services that are good value	Accommodation: affordability and overall value for price of hotel rooms. Transportation: reasonability of fares and charges for time spent	"My taxi fare cost 30 USD. . . very expensive!"
	[d4]: [D] presents accurate information of their products and services	Attractions: insider guides to lesser-known points of interests, insight into daily life	"Don't listen to the guidebooks- I'll share my favorite galleries off the beaten path"
Society	[d5]: [D] offers interesting local culture and traditions	Attractions: festivals, holidays Sports: national teams and competitions People: diversity of food, drink, language, architecture, religion	"The pumpkin festival is an annual favorite amongst locals and tourists alike"
	[d6]: [D] has hospitable residents	Restaurants: welcome of tourists Accommodation: hospitality and value added recommendations; and delivery of standard room quality Transportation: standard rate cards for fares by zone Shopping: negotiations at public markets	"When the locals saw we were lost, they helped us with our directions on the map" "The blankets cost twice as much for tourists as for locals"
Governance	[d7]: tourism industry and organizations cooperate and interact	Public figures/government: regulation of industries related to tourism; Accommodation + Transportation: interaction between segments; Local population + tourists: welcome	"You could be fined for feeding wild animals, which disrupts their migration habits, regardless of whether you are a tourist or a local"

(continued)

Table 2 (continued)

Core dimensions	Drivers	Examples of topic expressed (can be communicated by text, image, video, or other symbol)	Examples of sentiment (positive/negative)
	[d8]: [D] presents innovative and/or improved products and services	Technology: improved websites and interactive experiences Accessibility: products for handicapped	"The new IMAX theater at the National Space Museum shows a 3-D scuba diving movie!"
Environment	[d9]: [D] has a high eco-awareness	Accommodations: green building, certifications Public figures/government: endorse new	"The heat in the building is provided by rooftop solar panels"
	[d10]: [D] has a favorable weather	Favorable weather conditions	"Summer is the best season to visit the destination: no rain and cold"
	[d11]: [D] offers a safe environment	Weather: shelter from inclement conditions; Accommodations: security Events: security News: reports of crime	"Women should not walk alone at night in this city"
Performance	[d12]: [D] presents an accurate image	News: dispelling or confirming rumors Accommodation: text, images or videos that maintain or prove inconsistent the official site's portrayal	"The destination website's photos may look nice, but see how dirty we found our room"
	[d13]: [D] meets my expectations	Accommodations: surprise or disappointment about quality before and after trip Events: surprise or disappointment about quality before and after trip	"I was disappointed at how crowded the park was after seeing such lovely photographs in books"
	[d14]: [D] offers a satisfying tourism experience	Accommodation + Restaurant + Touring: Destination as a holistic experience. (TBD) use of star ratings for packaged deals.	"The trip was amazing in every way. I'm so glad we chose New York for our vacation"

6.2 Applying DORM to Analyze Online Reputation of Tourism Destinations

DORM requires content analysis of the online contents presented on tourism-related online conversations. DORM is used as a pre-established model to classify the topics expressed online as it allows for a systematic analysis. The content analysis performed by a human coder, as proposed in the DORM model, follows four main steps (see Fig. 3) simulating the process a prospect does in order to collect information about a destination. A similar four-step procedure can be performed on social media platforms.

Step 1. Query selection and links (URLs) collection: search activities with relevant keywords should be performed using a search engine (e.g. Google) considering the first 3 pages of results in order to gather the tourism destination's online representation. An example of the definition of tourism-related keywords to associate to the name of a tourism destination is: "visit + Lugano". Other popular keywords are: accommodation, holidays, travel, tourism, events, things to do, things to see.

Step 2. URL coding: per each keyword-combination it is required to analyse the URLs presented on the first 3 pages of search results. In this step it is required to identify the links containing user generated contents/online conversations (UGC). The coder(s) has to identify if in the landing page under analysis a UGC is published.

Step 3: Media classification: if the page contains user generated contents, it needs to be classified into specific social media types in order to describe the information market around the online tourism domain. Social media types can be: Virtual Community (e.g. Lonely Planet); Consumer Review (e.g. Tripadvisor.com); Blogs and blog aggregators (e.g. personal blog, blogspot); Social Networks (e.g. Facebook); Media Sharing (Photo/Video sharing—e.g. Flickr, YouTube);

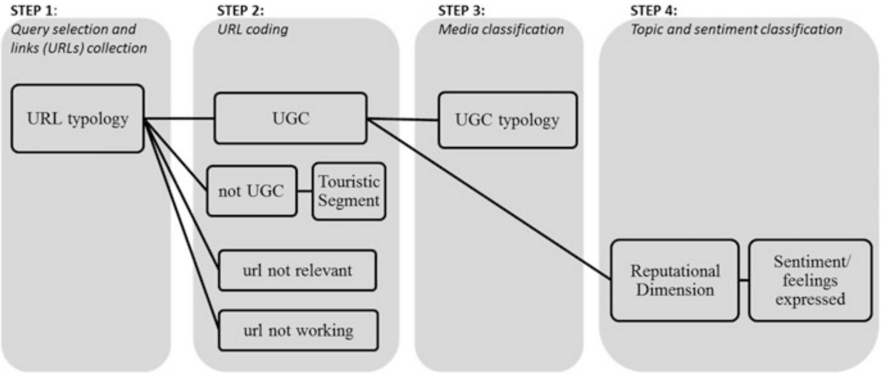


Fig. 3 The four-step procedure to perform an online content analysis using DORM

Microblogging (e.g. Twitter); Wikipedia, Wikitravel, and Other, such an online magazine hosting users' comments.

Step 4: Topic and sentiment classification: the main topic expressed within each page has to be associated to a reputation dimension of the DORM model. Once the main topic is identified, the sentiment polarity, that is the main value of expressed judgments should be identified. A 5-point Likert scale ranging from 1 (= contents in the page express mainly negative value judgments) to 5 (= contents in the page express mainly positive value judgments), and N.A. (Not Applicable = The item does not express any value judgment) can be used. Moreover, a qualitative data analysis software such as NVivo 10 can be used to generate semantic networks from content analyzed.

Finally, it is suggested to perform an inter-coder reliability test on the content analyzed, in order to avoid biased results from a one-coder analysis. Thus, at least two independent coders should code randomly the same contents in order to verify if they agree on the coding, and apply the same coding scheme.

The following sections present a case study of an online content analysis performed using the DORM model regarding Ticino as a tourism destination, which is the Italian-speaking region in southern Switzerland. First, the following keywords have been used to investigate the online reputation of Ticino on Google search engine: Ticino tourism, Ticino restaurants, Ticino activities, Ticino events, visit Ticino, Ticino attractions, Ticino holiday, Ticino shopping, and Ticino accommodation. The data collection has considered the results of the first three pages of Google for each keyword, as those pages are considered the most important in online searches. This procedure allowed to collect 247 unique URL (cleaned of double/repeated URLs), and used for content analysis. Results showed that the URLs containing user-generated contents accounted for 30.8 %. 13.8 % was represented by sites that were not related to the tourism domain, and therefore treated as not relevant (see Fig. 4).

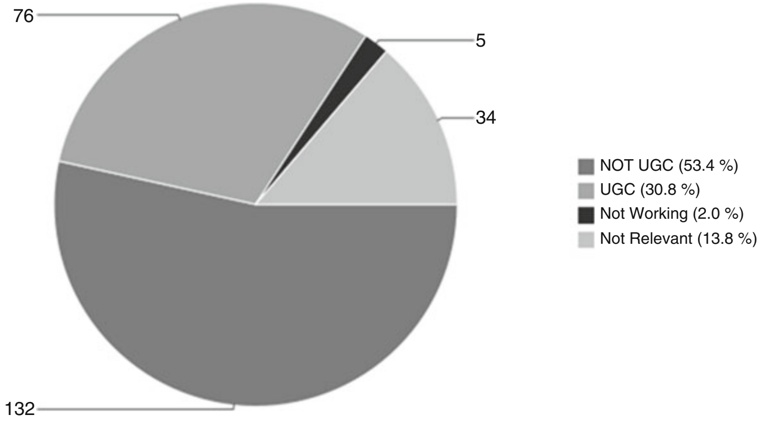


Fig. 4 URLs results obtained by Google search engine

Among the user generated contents, the main media types identified were consumer reviews (with Tripadvisor.com as the main consumer reviews platform present among the results), and media sharing, in particular videos (with YouTube.com as main video sharing platform). That is particularly relevant when it comes to the preparation of the kind of response. Indeed, knowing the kind of contents published online might help on preparing *ad-hoc* responses, such a response on reviews platforms, a video, etc. (see Fig. 5).

Regarding the main topic dimensions (see Fig. 6) present on the online pages analyzed, results showed that products and services-related contents were the majority (80 %). About 14.5 % were not-applicable contents, meaning that the pages resulted as empty. Interestingly, results indicated a lack of contents related to the society (e.g. local traditions, hospitality, and residents), and the governance-related dimension, while the environment dimension counted only for 3.9 %.

Among the products and services, it has been noticed that Ticino was mentioned mainly for site attractions, and just few times for events and entertainment. This is particularly relevant for the destination as it might decide to re-balance this information and to provide contents related to this topic that appeared to be missing within UGCs (see Fig. 7).

As can be seen in Fig. 8, negative comments were mainly related to complaints about the services offered at the destination. In particular, it was found that 15 negative contributions referred to the hotel industry communicating a poor cost/benefit ratio. Another negative aspect emerging from the analysis was related to a video presented on YouTube. The video was about a collection of images regarding Ticino and was accompanied by a voice over that devaluated the tourist

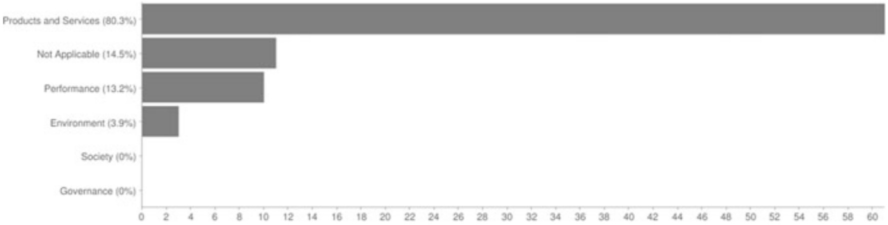


Fig. 5 Media types distribution among the UGC pages

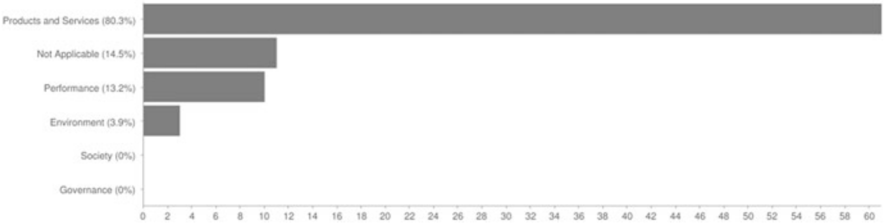


Fig. 6 Reputational topic dimensions

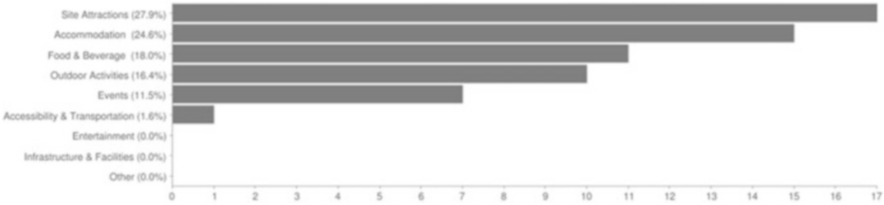


Fig. 7 Topics frequency for the reputational dimension "Products and Services"

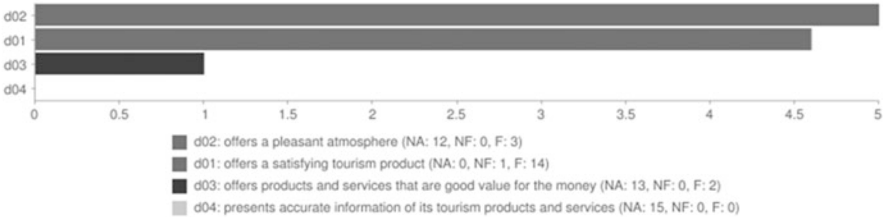


Fig. 8 An example of reputational drivers for the "Products and Services" dimension

offer of Ticino. Studies have shown that online videos and photos in general have a greater persuasive impact on tourists' decision-making compared to text, therefore this aspect should be considered by the managers of destinations.

Results thus showed that Ticino is represented online mainly by its products and services; in particular its attractions. Overall, results showed that Ticino is a more than satisfactory and an extremely enjoyable tourism destination; few comments (although very good) were registered for the weather, no comment related to the safety at the destination, or to entertainment. The lack of mentions regarding the entertainment aspect depicts the destination as mainly related to outdoor-nature driven attractions. Site managers can therefore consider this result on their future marketing campaigns, reflecting if it is good for the destination to establish this kind of online reputation and/or if it is important to balance missing/biased information.

7 Conclusions

UsERA and DORM models represent two possible measurement techniques to assess quality issues related to a DMO's online communication, and to its overall online reputation. In particular, the UsERA model gives relevant information on the studied website/app, highlighting the possibility of a poor user experience; while DORM provides indications on which topics are more sensitive to a positive or a negative sentiment expressed in online conversations. Insights provided by those approaches can give useful indications to destination managers to prioritize online communication interventions. Indeed, results from both approaches might be discussed with website's managers and should be aligned with the overall goals of the online communication as indicated in the OCM framework. In particular, destinations' managers could find in the UsERA model a way to associate different quality assessment tools for their online communication in order to tackle quality issues and take more informed decisions about their online communication strategy. With DORM, destinations' managers could benefit from an easy to use step-by-step procedure to analyze online contents about their destination. Moreover, as DORM allows for a systematic analysis, the same analysis procedure is suggested to be performed comparing similar destinations in order to check pros and cons against competition.

UsERA has demonstrated to provide a powerful framework for the identification of risks of negative user experiences within a destination website, determined by three factors: threats, vulnerability and resilience. The model emphasizes the relationship among usability and usages and aims to provide indications to website's managers on how to re-design their digital presence. The DORM model proposes a human-coding procedure and might be more time consuming compared to an automatic tool as it requires to read and interpret reviews, comments, blogs, watch videos, etc., and it is limited to a restricted number of data. Nonetheless, performing content analysis with a human coder provides a more accurate picture of a destination's online presence, helps to better learn the needs of guests, and allows managers to stay updated and get more familiar with current trends. Moreover, it has to be considered that automatic tools are not 'mature' enough to understand certain contents such as humor, irony and sarcasm, therefore a human-coding approach is crucial in the online reputation analysis in order to avoid biases in data interpretation. Specific guidelines on sentiment expressed evaluation on social media pages are suggested for future research. Future research in the online reputation analysis should consider who was writing the comment, and where this person is based in order to prepare an accurate response and be part of the conversation, adding the destination voice and experience, moreover a combination of qualitative and quantitative methods should be adopted in order to ensure larger samples, and cross-validate content analysis performed through human and automatic tools. Furthermore, a longitudinal analysis of the topics and related feelings expressed over time is suggested in order to track the evolution of online conversations over longer periods of time, and to understand their role in informing specific tourism behaviors, such as the perception of a dominant opinion and its effects over time.

Finally, it is possible to identify synergies between the two models: while both of them are able to identify risks related to a user experience, UsERA differs from DORM as it only identifies the in-house communication (e.g. a DMO website), and DORM only identifies the external communication (e.g. online conversations). However, destination managers can use the results from both approaches in order to build a solid online presence, and correct missing/biased information that can be present online. For example: the online reputation analysis revealed that a destination is depicted online as mainly related to entertainment offers, and few comments refer to its cultural heritage attractions. A DMO can therefore balance the lack of such information participating in the online conversations and strengthen this aspect in its website.

References

Adukaite, A., Inversini, A., & Cantoni, L. (2013, July 21–26). Examining user experience of cruise online search funnel. In A. Marcus (Ed.), *Design, user experience, and usability. Web, mobile, and product design*. Proceedings of the second international conference, DUXU 2013 (held as Part of HCI International 2013. Las Vegas, NV, USA, pp. 163–172). Berlin: Springer.

Atterer, R., Wnuk, M., & Schmidt, A. (2006). *Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction*. Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland.

Au Yeung, T., & Law, R. (2003). Usability evaluation of Hong Kong Hotel websites. In A. J. Frew, M. Hitz, & P. O'Connor (Eds.), *Information and communication technology in tourism*. New York: Springer.

Berens, G., & van Riel, C. B. M. (2004). Corporate associations in the academic literature: Three main streams of thought in the reputation measurement literature. *Corporate Reputation Review, 7*(2), 161–178.

Brinck, T., Gergle, D., & Wood, S. D. (2002). *Usability for the web*. San Francisco: Morgan Kaufmann.

Buhalis, D. (2003). *eTourism: Information technology for strategic tourism management*. Harlow: Prentice Hall.

Cantoni, L., & Ceriani, L. (2007). *Fare comunicazione online, analisi dell'attività di un sito internet attraverso i file di log*. Roma: Comunicazione Italiana.

Cantoni, L., & Tardini, S. (2006). *Internet*. London: Routledge.

Douglas, A., & Mills, J. E. (2004). Staying afloat in the tropics: Allying a structural equation model approach to evaluating national tourism organization websites in the Caribbean. In R. Law & J. E. Mills (Eds.), *Handbook of consumer behavior, tourism and the Internet* (pp. 269–293). Binghamton, NY: Haworth Hospitality Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fombrun, C. J., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of Management Journal, 33*, 233–258.

Govers, R., & Go, F. (2009). *Glocal, virtual and physical identities, constructed, imagined and experienced*. Basingstoke: Palgrave Macmillan.

Gretzel, U. (2006). Consumer generated content—trends and implications for branding. *e-Review of Tourism Research, 4*(3), 9–11.

Inversini, A., & Cantoni, L. (2009). Cultural destination usability: The case of visit bath. In *Information and communication technologies in tourism 2009*. Proceedings of the international conference in Amsterdam, The Netherlands (pp. 319–331). New York: Springer.

Inversini, A., Cantoni, L., & Buhalis, D. (2009). Destinations' information competition and web reputation. *Itt—Journal of Information Technology and Tourism, 11*, 221–234.

Inversini, A., Cantoni, L., & Bolchini, D. (2010, February 10–12). Presenting UsERA: User experience risk assessment model. In *Information and communication technologies in tourism 2010*. Proceedings of the international conference in Lugano, Switzerland (pp. 99–110). New York: Springer.

Inversini, A., Cantoni, L., & Bolchini, D. (2011). *Connecting usages with usability analysis through the user experience risk assessment model: A case study in the tourism domain*. In

A. Marcus (Ed.), Design, user experience, and usability [Pt II, HCII 2011, LNCS 6770] (pp. 283–293). Berlin: Springer.

Inversini, A., Marchiori, E., Dedekind, C., & Cantoni, L. (2010, February 10–12). Applying a conceptual framework to analyze online reputation of tourism destinations. In *Information and communication technologies in tourism 2010*. Proceedings of the international conference in Lugano, Switzerland (pp. 321–332). New York: Springer.

ISO (1998) ISO 9241. Ergonomic requirements for office work with Visual Display Terminals (VDTs)—Part 11: 'Guidance on usability'.

Marchiori, E., & Cantoni, L. (2012). The online reputation construct: Does it matter for the tourism domain? A literature review on destinations' online reputation. *Journal of Information Technology and Tourism, 13*(3), 139–159.

Marchiori E., & Cantoni L. (2013, April 17–19). *Cues affecting the recognition of the dominant topic and sentiment expressed on social media pages*. TTRA 2013 Europe. Proceedings of the International Conference in Dublin, Ireland.

Marchiori E., & Cantoni L. (2015). The role of prior experience in the perception of a tourism destination in user-generated content. *Journal of Destination Marketing & Management*, 4/3, 194–201.

Marchiori, E., Inversini, A., Cantoni, L., & Dedekind, C. (2010, April 18–20). *Towards a tourism destination reputation model. A first step*. Proceedings of the 6th International Conference Thought leaders in brand management. Lugano, Switzerland. ISBN 978-88-6101-006-2.

Marchiori, E., Inversini, A., & Cantoni, L. (2011, February 18–20). *Credibility in the online tourism: An analysis of the aspects of reception and consumption of imaginaries produced in Web 2.0 Tourism Services*. Proceedings of the international conference tourism imaginaries. Berkeley, California.

Marchiori, E., Pavese, G., & Cantoni, L. (2012). eTcoMM—eTourism communication maturity model. A framework to evaluate the maturity of a DMO when it comes to the online communication management. The case of Canton Ticino and Lombardy. In *Information and communication technologies in tourism 2012*. Proceedings of the international conference in Helsingborg, Sweden (pp. 215–226). New York: Springer.

Marcus, A., Schieder, T. K., & Cantoni L. (2013, July 21–26). The travel machine: mobile UX design that combines information design with persuasion design. In A. Marcus (Ed.), *Design, user experience, and usability. web, mobile, and product design*. Proceedings of the Second International Conference, DUXU 2013 (held as Part of HCI International 2013. Las Vegas, NV, USA), Berlin: Springer.

Money, K., & Hillenbrand, C. (2006). Using reputation measurement to create value: An analysis and integration of existing measures. *Journal of General Management, 32*(1), 1–12.

Nielsen, J. (2006) *Severity ratings for usability problems*. Retrieved November 2015, from http://www.useit.com/papers/heuristic/severityrating.html

Nielsen, J., & Mack, R. (1994). *Usability inspection methods*. New York: Wiley.

Passow, T., Fehlmann, R., & Grahlow, H. (2005). Country reputation from measurement to management: The case of Liechtenstein. *Corporate Reputation Review, 7*, 309–326.

Pike, S. (2005). *Destination marketing organizations*. Oxford: Elsevier.

Pitkow, J. (1997). In search of reliable usage data on the WWW. In *Sixth International World Wide Web Conference*, Santa Clara, CA, pp. 451–463.

Qi, S., Buhalis, D., & Law, R. (2007). Evaluation of the usability on Chinese destination management organisation websites. In M. Sigala, L. Mich, & J. Murphy (Eds.), *Information and communication technologies in tourism 2007* (pp. 267–278). Wien: Springer.

Tardini, S., Adukaite, A., & Cantoni L. (2014). How to do things with websites. Reconsidering Austin's perlocutionary act in online communication. *Semiotica*, *202*, 425–437.

Tardini, S., & Cantoni, L. (2015). Hypermedia, internet and the web. In L. Cantoni & J. A. Danowski (Eds.), *Communication and technology* (pp. 119–140). Berlin: De Gruyter Mouton.

Triacca, L., Inversini, A., & Bolchini, D. (2005). *Evaluating web usability with MiLE+*. Web site evolution IEEE Symposium, Hungary, Budapest.

Tussyadiah, I., & Fesenmaier, D. (2008). Mediating tourist experiences. Access to places via shared videos. *Annals of Tourism Research, 36*(1), 24–40.

Vidaver-Cohen, D. (2007). Reputation beyond the rankings: A conceptual framework for Business School Research. *Corporate Business Review, 10*(4), 278–304.

Walsh, F., & Wiedmann, K. P. (2004). A conceptualization of corporate reputation in Germany: An evaluation and extension of the RQ. *Corporate Reputation Review, 6*(4), 304–312.

Yang, S. U., Shin, H., Lee, J. H., & Wrigley, B. (2008). Country reputation in multidimensions: Predictors, effects, and communication channels. *Journal of Public Relations Research, 20*(4), 421–440.

Market Intelligence: Social Media Analytics and Hotel Online Reviews

Zheng Xiang, Zvi Schwartz, and Muzaffer Uysal

1 Introduction

The competitiveness of the hotel industry has been well documented in the literature (e.g., Olsen & Roper, 1998). It is also well recognized that in this increasingly challenging environment, thorough understanding of the market conditions is required for effective decisions and to sustain short-term and long-term competitive advantages (Enz, 2009; Pizam, Lewis, & Manning, 1982). In this context, market (AKA business) intelligence is defined as information and knowledge obtained from external sources that can be used for identifying problems, changes and opportunities in the marketing environment (Wood, 2001). A variety of market information systems have been proposed and developed to assist the industry to identify opportunities and threats resulting from market dynamics (e.g., Minghetti, 2003; Wöber, 2003). The rapidly evolving technological landscape facilitates a growing interest in employing new data sources and developing new measurement tools to help managers engage with the new reality in hospitality and tourism (Xiang & Law, 2013; Xiang, Pan, Law, & Fesenmaier, 2010). In particular, the tremendous growth of social media and consumer-generated content on the Internet has further transformed the information landscape for businesses, and is providing rich sources of data to understand market conditions and develop market intelligence tools in the new contexts. A business intelligence and analytics approach

Z. Xiang (✉)
Department of Hospitality and Tourism Management, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
e-mail: philxz@vt.edu

Z. Schwartz
University of Delaware, Newark, DE, USA

M. Uysal
University of Massachusetts, Amherst, MA, USA

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_16

stresses and leverages the capacity to collect, analyze, and interpret data with an unprecedented breadth, depth, scale and speed, to solve real-life problems (Chen, Chiang, & Storey, 2012; Mayer-Schönberger & Cukier, 2013). Especially, due to its growing significance in the information ecosystem, social media has recently attracted much attention with huge potential for harvesting the "wisdom of crowds", i.e., social media analytics, in a variety of fields (Fan & Gordon, 2014; Wood, Guerry, Silver, & Lacayo, 2013; Zeng, Chen, Lusch, & Li, 2010). In the hospitality industry, commercial tools provided by companies such as Revinate (see http://www.revinate.com) and IDeaS (see http://www.ideas.com) are now incorporating social media contents by linking consumer sentiment to hotel revenue management decisions.

This study demonstrates the potential and usefulness of applying the social media analytics principles to develop market intelligence for the hotel industry based upon online hotel reviews. It is argued that online reviews reflect customers' actual hotel experiences, can be used to recognize guest perceptions, and in turn provide valuable new insights about the industry's market structure. That is, we show how one can move from specific words customers use to describe their experience with a hotel, combined with their expressed satisfaction levels, to classify hotels into distinct clusters. This realization is unique and important because it paves the way for an alternative market structure analysis, one that is closer to a *market commonality* approach as opposed to the more traditional, and perhaps less relevant *resource similarity* one. From a practical perspective the implications are clear. Current practices of understanding the competitive environment's structure, both on the strategic and tactical levels, are slow, cumbersome and less responsive. Analysis using customer reviews not only provides insightful new layer of information to better describe the competitive structure of the industry, but it also represents an unprecedented opportunity for ongoing, real time dynamic analysis. In other words, we argue that this highly efficient approach has the capacity to continually monitor the environment and respond to changes, all with high level of automation, using huge amount of data and with minimal human intervention in the process. The question of who the hotel is competing with, is relevant to multiple layers of practicality: from tactical daily and sometime hourly decisions on room rates, and allocation of room inventories to distribution channels, all the way to strategic positioning decisions. Therefore, social media analytics can provide not only a highly relevant perspectives, but it can also do it as frequently as needed and at a considerably low cost.

2 Research Background and Framework

In the highly competitive hotel industry, firms offer essentially homogeneous products and services and, thus, they must find ways to distinguish themselves among their competitors. To examine one's position in the consumer market it is essential to understand how the product is perceived in the consumer's mind in relation to others (Kotler, Bowen, & Makens, 2006). Research has traditionally considered the hotel as a bundle of service offerings. In the widely used conceptual framework, the hotel product can be deconstructed into several levels, including the core product, the facilitating product, the supporting product, as well as the augmented product (Kotler et al., 2006). Alternatively, the hotel product is composed of a set of attributes as suggested by Saleh and Ryan (1992) and others (e.g., Qu, Ryan, & Chu, 2000). These attributes include services, location, room, price/value, food & beverage, image, security, and marketing. The most frequently-cited single items in the literature are friendliness of staff, price, professionalism/quality, and cleanliness of room, location, room service, comfort of bed, reputation, restaurant facilities and service speed. These hotel attributes have been shown to induce various levels of hotel guests' satisfaction (dissatisfaction). The question is then if these attributes would also operate in the same manner across different hotel segments. The frequently-cited Two Factor Theory (Herzberg, 1966) postulates that hygiene factors like cleanliness and maintenance may not positively contribute to satisfaction, although dissatisfaction may result from their absence, while motivator factors such as the expressive aspects of staying at a hotel result in positive satisfaction (Noe & Uysal, 1997).

More recently, scholars have adopted the service-dominant logic to argue that what a guest gains from staying at a hotel is not limited to what the hotel offers, but instead it is co-created by both the service provider and the guest (Chathoth, Altinay, Harrington, Okumus, & Chan, 2013; Shaw, Bailey, & Williams, 2011). For example, Prahalad and Ramaswamy (2004) criticized the strategy of "staged experience" developed by tourist attractions and luxury hotels as reflected in the well-known concept of the experience economy (Pine & Gilmore, 1999). They argued that these firms treated their consumers as passive receivers of their service during all aspects of engagement, from self-checkout to participation in a staged experience. As such, these firms are still primarily product-centric, service-centric, and, therefore, company-centric. Shaw et al. (2011) cited examples in the hotel industry showing that highly personalized services and involvements of the guests can lead to guest satisfaction. Grissemann and Stokburger-Sauer (2012) demonstrated that a co-created experience can positively affect customer satisfaction and customer loyalty with a travel service company. These recent findings suggest that the guest's multifaceted experience, including the co-created aspects, may serve as a better conceptual basis for understanding the true values and benefits perceived by hotel guests than simply using elements and attributes of hotel services (Walls, Okumus, Wang, & Kwun, 2011; Wu & Liang, 2009). As such, it is argued that individual consumers' heterogeneous experiences, which have impact on their perceptions of the hotel product, can be used to understand a hotel's position among its competitors in the market.

Within the realm of social media, consumer ratings and online reviews of travel and hospitality products have been found highly influential on online consumer behavior. Particularly, online reviews of travel experiences posted on reliable websites are perceived as unbiased and trustworthy because they reduce the likelihood of later regretting a decision as well as allow readers to easily imagine what products look like (Gretzel & Yoo, 2008; Park & Nicolau, 2015; Sparks &

Browning, 2011). Studies have found that various aspects of online reviews (i.e., star ratings, review richness, and valence of reviews) and characteristics of review providers (i.e., personal identity disclosure and level of expertise) assist purchase decisions and could have positive impact on a company's revenue (Park, Xiang, Josiam, & Kim, 2014; Sparks & Browning, 2011; Vermeulen & Seegers, 2009). Therefore, understanding what constitutes the guest experience in online reviews allows us to understand what leads to guest (dis)satisfaction and thus the value perception of the hotel product, which subsequently allows us to better understand how hotel properties can be distinguished based upon these value perceptions. Given the amount of information generated on a daily basis, online hotel reviews, therefore, should be considered ideal sources of data for understanding the market structure of the hotel industry based upon consumer perceptions of the hotel product in a potentially real time fashion. Ultimately, this leads to a better understanding of a firm's competitive position within the market resulting in the (re)formulation of competitive strategies and practices for the firm.

With this in mind, we describe a general framework specifically focused on using online reviews to generate insights into the market structure of the hotel industry. This research framework is based upon the general principles of social media analytics, which is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application (Fan & Gordon, 2014; Zeng et al., 2010). Social media analytics follows a normative process that involves three steps: (1) capture, which includes gathering data from social media sources and preprocessing (e.g., reducing noise) and extracting pertinent information from the data; (2) understand, which uses usually quantitative analytical methods such as sentiment analysis, topic modeling and social network analysis, etc., to identify patterns in the data; and, (3) present, which focuses on summarizing, interpreting, and presenting (oftentimes through visualization) the findings (Fan & Gordon, 2014). As can be seen in Fig. 1, this framework represents a process wherein textual online reviews are translated into various constructs in hotel marketing and management in order to describe the nature and characteristics of the hotel product and, ultimately, help hotel managers to formulate strategies in response to their current positions among others. This process consists of two

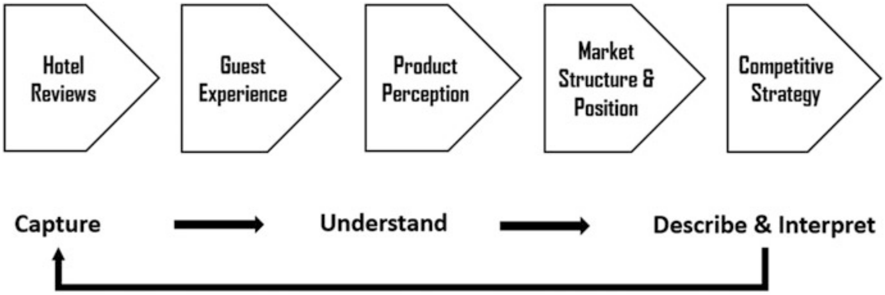


Fig. 1 Research framework for developing market intelligence from online hotel reviews

“layers”: the first (bottom) layer represents the analytical/text mining mechanisms that include three phases of operations, i.e., “capture”, “understand”, and “describe/interpret”. The second (top) layer represents the process that translates raw data (online review) into business intelligence (i.e., market structure and position) and potentially competitive strategy for the hotel companies through the understanding of guest experience and product perception. From the analytical point of view, online reviews can be deconstructed into words and semantic structures to describe and represent the guest experiences (Gretzel et al., 2008; Woodside, Cruickshank, & Dehuang, 2007). These individual stories, when aggregated, can be used to derive hotel guests’ perceptions of the hotel product, which then allow us to understand the hotel product at the market level. This framework incorporates a feedback loop indicating that it can be an ongoing process in order to capture the market dynamics following possible strategic and tactical responses based upon the market intelligence.

Within this framework, Xiang, Schwartz, Gerdes, and Uysal (2015) recently conducted a study applying text analytics to a large quantity of customer reviews extracted from Expedia.com to deconstruct hotel guest experience, and examine its association with satisfaction ratings. Exploring the words hotel guests used to evaluate their hotel stay, the study identified several dimensions of guest experience with novel, meaningful semantic compositions. This paper aims to further explore the usefulness of this social media analytics approach by applying the guest experience dimensions, and level of satisfaction rating, to delineate the characteristics of the hotel properties and gain insights into the hotel market structure from the consumers’ standpoint.

3 Methodology

A large-scale text analytics study was conducted based upon publicly available data in Expedia.com. Details of the research design including rationale for using Expedia.com to collect online reviews, preprocessing of the textual contents, identification of guest experience-related words and underlying dimensions as well as the examination of the association between guest experience and satisfaction rating can be found in Stringam and Gerdes (2010) and Xiang et al. (2015). The data were collected during a period of 12 consecutive days using a crawler to extract customer reviews for all hotels listed by Expedia for the 100 largest U.S. cities, as defined by then the most recent U.S. Census Bureau population estimate (Census Bureau & Population Division, 2007). For each city the crawler gathered all available textual content of customer reviews, overall star rating for the hotel, average guest overall satisfaction, and all available data for each customer review. Data were collected for a total of 10,537 hotels, which represented more than one-fifth of the entire hotel population nationwide, resulting in 60,648 customer reviews.

Data analysis followed the process outlined in Fig. 1 with two phases. The first phase replicated Xiang et al. (2015) deconstruction of the online reviews, and the generation of the hotel guest experience dimensions. Our discussion of this first phase is brief since both the data and the methods are described in great details in the two papers listed above. Textual data pre-processing involved a series of operations such as stemming, misspelling identification, and identification and removal of stop words such as certain pronouns, adverbs, and conjunctions. For domain identification a coding schema was used to guide the process in order to extract words related to hotel guest experience. This coding schema took into account the existing literature on each stage of the guest's experience with hotels services, i.e., the pre-trip stage, arrival and on-site experience, and departure, resulting in a dictionary of 416 primary words used by consumers to describe their experiences at a specific hotel. To identify a robust data structure that yields a strong association between guest experience and satisfaction, a linear regression model was tested by adjusting the hotel and word frequency thresholds to maximize the explanatory power on satisfaction rating. In an iterative way, the dataset was reduced to 529 hotel cases and 80 guest experience-related words. Table 1 lists the 80 guest experience-related words extracted from customer reviews in Expedia.com, along with their average frequency per hotel. Among these hotel properties the vast majority (>96 %) were mid- and up-scale hotels ranging between two and four stars. These hotels were located in over 30 states, with half of which from California, Florida, New York and Washington, DC.

Finally, cluster analysis and correspondence analysis were conducted to understand the market structure of the hotel industry using the previously identified guest experience dimensions. Considering that these dimensions reflect what hotel guests talk about when they describe their experience, along with the associated satisfaction rating, they can be seen as reflecting the value perception of the hotel product (Nasution & Mavondo, 2008; Oh, 1999). These factor scores, along with the satisfaction ratings, were entered as input variables in cluster analysis to reveal hotel segments that can be distinguished by guest's value perceptions of the hotel product. The general goal of cluster analysis is to identify homogeneous groups (clusters) that are different from all other groups. To develop distinct segments of the hotel industry, non-hierarchical clustering approach, specifically k-mean was chosen in order to place each property into only one specific cluster. While there is a lack of standard criteria to determine the optimum number of clusters, it has been suggested that the best way to validate a clustering solution is to search for cues on the validity of the clusters, that is, to demonstrate their usefulness or value for hospitality and tourism managerial practices (Frochot & Morrison, 2000). In our case, we examined several clustering solutions in an iterative fashion to make certain that (1) the clusters were meaningful and practically useful and (2) there were not any extremely large or tiny clusters. Although our approach introduced a level of subjectivity into the analysis, we believe this was a sensible approach to finding meaningful segments within the hotel market, especially since this is a first attempt to apply this method of creating hotel clusters based on words customers use in their social media discussions. A correspondence map was constructed using

Table 1 Top 80 primary words in hotel customer reviews

Word	Avg. Freq. per Hotel	Word	Avg. Freq. per Hotel	Word	Avg. Freq. per Hotel
Room	10.7	Food	0.9	Kids	0.5
Clean	5.9	Distance	0.9	Tv	0.5
Staff	5.5	Shuttle	0.8	Attractions	0.5
Location	5.4	Street	0.8	Water	0.5
Comfortable	4.1	Shopping	0.8	Coffee	0.5
Service	3.2	Maintained	0.8	Amenities	0.5
Friendly	3.1	Beach	0.8	Experience	0.5
Close	3.0	Access	0.8	Suite	0.4
Breakfast	2.9	Park	0.7	Money	0.4
Helpful	2.6	Floor	0.7	Carpet	0.4
Bed	2.5	Check in	0.7	Courteous	0.4
Price	2.5	Spacious	0.7	City	0.4
Restaurants	2.2	Bar	0.7	Expensive	0.4
Walking	1.9	Lobby	0.7	Dirty	0.4
Area	1.6	Internet	0.7	Renovated	0.4
Parking	1.5	Trip	0.6	Tub	0.4
Bathroom	1.4	Pay	0.6	Safe	0.4
Pool	1.4	Door	0.6	Far	0.4
Free	1.3	Shops	0.6	Air	0.4
Convenient	1.3	Sleep	0.6	Refrigerator	0.4
Downtown	1.3	Business	0.6	Quality	0.4
Airport	1.2	Complaint	0.6	Decor	0.4
Desk	1.2	Shower	0.6	Wait	0.4
View	1.1	Family	0.6	Freeway	0.4
Recommend	1.0	Value	0.5	Elevator	0.4
Noise	0.9	Cheap	0.5	Accommodation	0.2
Quiet	0.9	Smelled	0.5		

the frequencies of the top 80 words in relation to each of these hotel clusters to visualize how these hotel segments are related to each other within these online reviews.

4 Findings

A six-cluster solution to the K-mean cluster analysis emerged as most adequate based on distances between clusters and resulting cluster member numbers. These six clusters, listed in Table 2, were distinctively associated with the five guest experience factors. It seems that Clusters 1, 2, 3, 4, and 6 were deemed satisfactory, with the average rating around or above the grand mean, while Cluster 5 was the only group with a distinctly less satisfactory score of 3.2. It is however quite

Table 2 Hotel Clusters identified using satisfaction rating and guest experience factors

	Means of cluster centre							
	C1 (N = 85)	C2 (N = 101)	C3 (N = 95)	C4 (N = 87)	C5 (N = 76)	C6 (N = 85)	F-ratio	Sig.
Satisfaction rating	3.996	4.209	4.077	4.216	3.207	4.304	93.100	0.000
Hybrid	0.575	−0.848	−0.170	0.548	1.054	−0.881	118.116	0.000
Deals	0.969	−0.375	−0.261	0.739	−1.281	0.158	113.540	0.000
Family friendliness	−0.976	0.311	−0.602	1.248	−0.081	0.075	102.835	0.000
Core product	−0.863	−0.643	1.177	0.436	−0.291	0.126	102.267	0.000
Staff	0.004	0.823	0.437	0.155	−0.372	−1.297	88.570	0.000
Total N = 529								

revealing to examine the salient factors in association with the average satisfaction ratings. While some of the hotel clusters display similar levels of satisfaction, it is driven by different aspects of the guest experience. Specifically, Cluster 1 hotels seem to be positively associated with "deals", even though these hotels were not necessarily family friendly or have good core products. Both Clusters 2 and 6 are similar in one aspect as both are high on the Hybrid factor (the negative sign suggests these hotels are positive on the experiential aspects) but almost differ on the Staff factor. This reveals that these two types of hotels offered quite similar experiential aspects for the guests but their staff was perceived in a very different way and consequently the impact of their staff on their guest satisfaction. Cluster 3 seemed to have good Core Product with less helpful and friendly staff (the positive sign implies negative experience with staff). Cluster 4 seemed to be predominantly distinguished by Family Friendliness. Among all these hotel segments Cluster 5 was rated unsatisfactory largely because of negative maintenance and hygiene factors (note the positive sign of the Hybrid factor) as well as the lack of deals. It is particularly interesting that these groups of hotels were rated either satisfactory or unsatisfactory due to one or two dominant guest experience factors.

A correspondence map to visualize their position in the semantic space consisting of the 80 words that define the underlying dimensions of guest experience wherein the map illustrates the co-occurrence matrix of the hotel clusters and frequencies of specific words. As can be seen in Fig. 2, the first two extracted

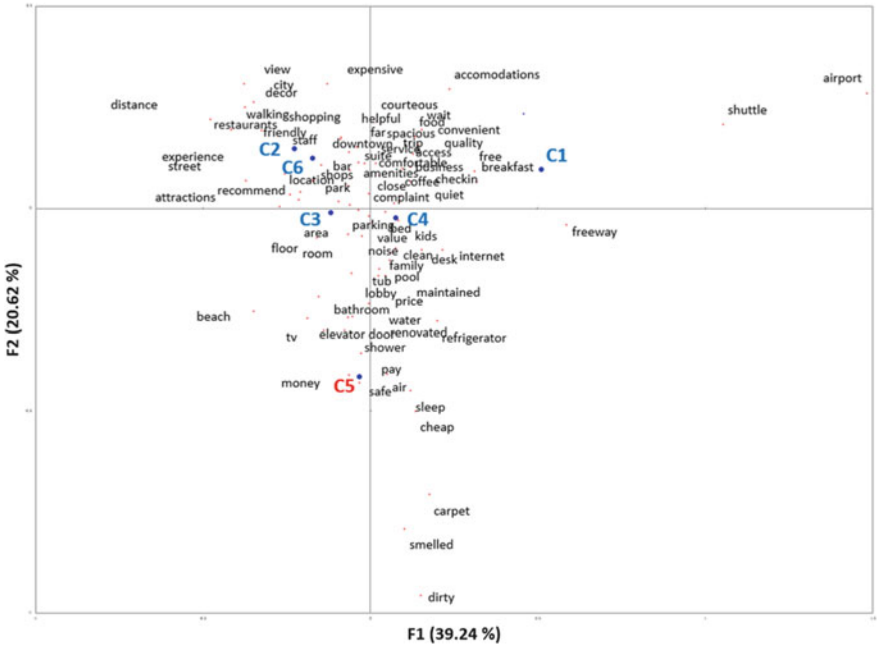


Fig. 2 Hotel Cluster profiles in the semantic space (correspondence analysis using symmetric plot (axes F1 and F2: 59.86 %))

factors explain approximately 60 % of inertia in the data. This semantic space has a large set of words representing the core attributes of products and services of a hotel, densely distributed in the center of the map and shared between the hotel clusters. More interestingly, most of the individual hotel clusters (except for Clusters 2 and 6) are associated with some distinct words that "stretch" the semantic space of guest experience. In the case of Cluster 1, words such as "airport", "shuttle", "free", and "breakfast" are prominent, representing the underlying factor of "deals". Cluster 3 is closely associated with words such as "room" which represents the core product of the hotel. Cluster 4 is surrounded by words such as "kids", "family", and "value" which signify the family friendliness of this hotel cluster. Apparently, Clusters 2 and 6 are associated with words such as "restaurants", "walking", "shopping", "experience", "bar", "décor", "view", and "location", which represent the experiential aspect of the stay. Note that while these two clusters overlap almost entirely in the semantic space, based upon the cluster scores in Table 2 the primary difference between the two lies in their customers' experiences with hotel staff, with one negative (Cluster 2) and the other positive (Cluster 6). Cluster 5 is the only hotel segment rated unfavorably by customers which is closely associated with words representing issues related to hotel maintenance or the hygiene factors. Also, Cluster 5 seems to be "isolated" in the lower half of the map, suggesting maintenance-related issues are important attributes that distinguish a hotel that makes its guests unhappy from those that make their guests happy (for different reasons).

Hotel properties within each cluster were checked against their levels of service (i.e., star ratings) to gain a better understanding of the compositions of these hotel clusters. As can be seen in Fig. 3, Cluster 1 seems to be dominated by low- and mid-scale hotels (between two- and three-star). Considering Cluster 2 is strongly associated with the "Deals" dimension, this suggests that a hotel, even with

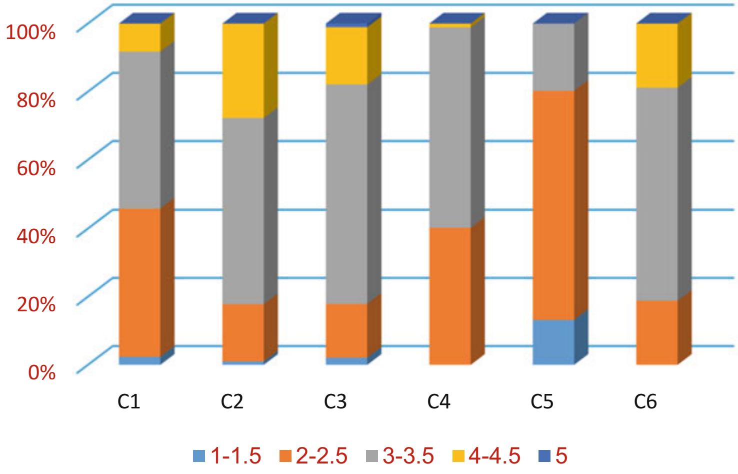


Fig. 3 Hotel Cluster profiles with star ratings

limited services, can still make their guests happy by offering good deals such as free breakfast and transportation. Clusters 2, 3, and 6 consist of predominantly mid- and up-scale ones (between three- and four-star). These hotels appear to be quite similar and almost identical (especially Clusters 3 and 6) in terms of their star ratings. The vast majority of Cluster 4 hotels consists of mainly low-and mid-scale hotels (family-friendly). This suggests that, while star rating is, to a certain degree, indicative of the level of satisfaction, hotel customers may be happy for a variety of reasons regardless of star rating as in the cases of Clusters 2, 3, and 6. Finally, Cluster 5, which was rated unsatisfactory by their customers, appears to consist of lower-end hotels (i.e., majority of them are two-star or below).

5 Conclusions

In order to understand market conditions of the hotel industry, we applied previously identified guest experience dimensions and satisfaction ratings based upon a large quantity of authentic online customer reviews to explore whether hotels can be distinguished by these dimensions. The findings suggest that there were different types of hotels with unique salient traits such as good deals, family friendly amenities, as well as opportunities for experiential encounters that satisfy their customers, while those who failed to do so mostly had issues related to cleanliness and maintenance-related factors. The correspondence map visually confirms how hotels are associated with words describing guest experiences in the semantic space. This study shows that the hotel product can be distinguished by the combination of satisfaction rating and guest experience as reflected in online customer reviews. As demonstrated by the cluster analysis, the combination of level of guest satisfaction and the determinants of satisfaction, i.e., salient experience dimensions, is similar within, but dissimilar across hotel clusters. This indicates that, the hotel sector can be "segmented" based upon what drives the customers' post-purchase evaluation, as reflected in online reviews even without knowing much about who the reviewers are (e.g., their demographics). This study makes several genuine contributions to the literature both theoretical and practical.

First, a growing amount of hospitality and tourism research examines users' responses to social media content, and identifies correlations between online reviews and hotel performance (e.g., Crotts, Mason, & Davis, 2009; Li, Law, Vu, Rong, & Zhao, 2015; Li, Ye, & Law, 2013; Liu, Law, Rong, Li, & Hall, 2012; Stringam, Gerdes, & Vanleeuwen, 2010). In this study we proposed and applied an analytics framework that incorporates guest experience (i.e., what consumers talk about) as the basis for text analysis, which, in combination with satisfaction ratings, yields guests' value perceptions of the hotel product. The proposed framework delineates a clear roadmap of knowledge creation, i.e., from unstructured text to guest experience, to product perception, to market structure, and ultimately to strategic decision, which enables hospitality firms to generate insights into the market dynamics in the hotel industry. We believe that, as shown in this study, social media analytics in hospitality should build upon the rich, profound domain knowledge in order to realize its potential to contribute to both theory and practice.

Second, this study has the potential to contribute theoretically and practically to the emerging debate on the proper way to form hotels competitive sets. While hotels are traditionally classified using hotel amenities, service attributes and location, there has been criticism that these classification systems may not truthfully reflect consumers' perceptions (e.g., Li & Netessine, 2012; López Fernández & Serrano Bedia, 2004) and consequently be somewhat misleading. We demonstrate that a text analytic consumer-centric approach to understanding the market structure of the hotel industry is plausible and, perhaps valid, especially when consumer-generated data becomes abundant. This type of consumer-based hotel clustering approach can assist in the more granular hotel operational level of forming more meaningful hotel competitive sets, sets that better reflect the consumer's perspective and consequently are more appropriate in evaluating the hotel performance, and in formulating its strategies in the competitive market place. The current standard practice in the industry in forming the hotel's competitive set (s) largely focuses on the hotel's characteristics: The average daily rate, location, size, scale, food and beverage outlets, meeting space, brand affiliation status, etc. (see, for example, STRanalytics, 2014). However, the increasing reliance on comparative (relative) performance measures such as the occupancy, ADR and RevPAR indices (the widely used STAR report) to shape tactical revenue management decisions give rise to the notion of making the competitive sets, and consequently the performance indices, better reflect the "true" competitors in the eyes (and actions) of the customers.

Lastly, this study offers several practical implications for hotel managers. Although our analysis focused on mapping the entire hotel market at the national level, our approach can certainly be applied to individual properties or brands at a more local level to develop a variety of business intelligence. For example, post-purchase behavioral studies examining customer satisfaction can help practitioners effectively realign their strategies in service delivery and product development (Kozak & Rimmington, 2000). With the knowledge about different determinants of guest satisfaction, hoteliers can have the leverage to make up for service attribute deficiency, which may extract from guest satisfaction, by focusing on providing unique features that would help tangibilize intangible attributes. Also, the importance of co-creation of experience in driving guest satisfaction suggests that hotels should not limit their strategy to providing desirable attributes and services; rather, they must also consider playing a facilitator's role in helping guests to identify and create what they see as meaningful experiences (Grissemann & Stokburger-Sauer, 2012; Shaw et al., 2011). Compared to conventional approaches such as surveys and focus group studies, which are oftentimes expensive, time consuming and backward looking (e.g., Dev, Morgan, & Shoemaker, 1995), social media analytics offers not only a cost effective but also a dynamic (real time) solution to develop market intelligence.

This study has several limitations. In addition to the limitations identified in Xiang et al. (2015) this dataset was collected several years ago and obviously does not reflect the current market conditions in the US hotel industry. More importantly, it was essentially a snapshot of one of the many online travel agency websites and, therefore, did not represent social media in a comprehensive, dynamic way. Nonetheless, this study points to several directions for future research. As an important theoretical construct the structure of guest experience need to be further explored and validated. Specifically, our analysis in the previous study indicates that, if a threshold level of hygiene variables is not met, it prevents customers from self-fulfillment through experiential/co-production elements of their stay. As shown in this study, once this threshold level is surpassed, other determinants of guest satisfaction become compensatory to each other. However, whether these determinants as a whole are compensatory or non-compensatory (hierarchical) in nature remains to be substantiated. Furthermore, given the limitations of the data we do not have much knowledge about certain hotel characteristics such as location, size, and amenities as well as characteristics of consumers. It would be interesting to find out whether these differences between hotel clusters are due to inherent product or customer characteristics in order to improve the validity of the social media analytics approach.

References

Chathoth, P., Altinay, L., Harrington, R. J., Okumus, F., & Chan, E. S. (2013). Co-production versus co-creation: A process based continuum in the hotel service context. *International Journal of Hospitality Management, 32*, 11–20.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Crotts, J. C., Mason, P. R., & Davis, B. (2009). Measuring guest satisfaction and competitive position in the hospitality and tourism industry an application of stance-shift analysis to travel blog narratives. *Journal of Travel Research, 48*(2), 139–151.

Dev, C. S., Morgan, M. S., & Shoemaker, S. (1995). A positioning analysis of hotel brands: Based on travel-manager perceptions. *The Cornell Hotel and Restaurant Administration Quarterly, 36*(6), 48–55.

Enz, C. A. (2009). *Hospitality strategic management: Concepts and cases*. Hoboken, NJ: Wiley.

Fan, W., & Gordon, M. D. (2014). Unveiling the power of social media analytics. *Communications of the ACM*, In Press (June 2014), 26.

Frochot, I., & Morrison, A. M. (2000). Benefit segmentation: A review of its applications to travel and tourism research. *Journal of Travel & Tourism Marketing, 9*(4), 21–45.

Gretzel, U., Xiang, Z., Wöber, K., Fesenmaier, D. R., Woodside, A. G., & Martin, D. (2008). Deconstructing destination perceptions, experiences, stories and internet search: text analysis in tourism research. In *Tourism management: Analysis, behaviour and strategy*, 339–357.

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In *Information and communication technologies in tourism 2008* (pp. 35–46). Vienna: Springer.

Grissemann, U. S., & Stokburger-Sauer, N. E. (2012). Customer co-creation of travel services: The role of company support and customer satisfaction with the co-creation performance. *Tourism Management, 33*(6), 1483–1492.

Herzberg, F. (1966). *Work and the nature of man*. Cleveland, OH: World Publishing.

Kotler, P., Bowen, J. T., & Makens, J. C. (2006). *Marketing for hospitality and tourism*. New Delhi: Pearson Education India.

Kozak, M., & Rimmington, M. (2000). Tourist satisfaction with Mallorca, Spain, as an off-season holiday destination. *Journal of Travel Research, 38*(3), 260–269.

Li, J., & Netessine, S. (2012*). Who are my competitors?-Let the customer decide*. Working paper. Retrieved on July 20, 2014, from http://www.insead.edu/facultyresearch/research/doc.cfm?did=50311

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management, 46*, 311–321.

Li, H., Ye, Q., & Law, R. (2013). Determinants of customer satisfaction in the hotel industry: An application of online review analysis. *Asia Pacific Journal of Tourism Research, 18*(7), 784–802.

Liu, S., Law, R., Rong, J., Li, G., & Hall, J. (2012). Analyzing changes in hotel customers' expectations by trip mode. *International Journal of Hospitality Management, 34*, 359–371.

López Fernández, M. C., & Serrano Bedia, A. M. (2004). Is the hotel classification system a good indicator of hotel quality? An application in Spain. *Tourism Management, 25*(6), 771–775.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Minghetti, V. (2003). Building customer value in the hospitality industry: Towards the definition of a customer-centric information system. *Information Technology & Tourism, 6*(2), 141–152.

Nasution, H. N., & Mavondo, F. T. (2008). Customer value in the hotel industry: What managers believe they deliver and what customer experience. *International Journal of Hospitality Management, 27*(2), 204–213.

Noe, F. P., & Uysal, M. (1997). Evaluation of outdoor recreational settings: A problem of measuring user satisfaction. *Journal of Retailing and Consumer Services, 4*(4), 223–230.

Oh, H. (1999). Service quality, customer satisfaction, and customer value: A holistic perspective. *International Journal of Hospitality Management, 18*(1), 67–82.

Olsen, M. D., & Roper, A. (1998). Research in strategic management in the hospitality industry. *International Journal of Hospitality Management, 17*(2), 111–124.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research, 50*, 67–83.

Park, H., Xiang, Z., Josiam, B., & Kim, H. (2014). Personal profile information as cues of credibility in online travel reviews. *Anatolia, 25*(1), 13–23.

Pine, B. J., & Gilmore, J. H. (1999). *The experience economy: Work is theatre & every business a stage*. Boston: Harvard Business Press.

Pizam, A., Lewis, R. C., & Manning, P. (1982). *The practice of hospitality management*. New York, NY: AVI Publishing.

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of Interactive Marketing, 18*(3), 5–14.

Qu, H., Ryan, B., & Chu, R. (2000). The importance of hotel attributes in contributing to travelers' satisfaction in the Hong Kong Hotel Industry. *Journal of Quality Assurance in Hospitality & Tourism, 1*(3), 65–83.

Saleh, F., & Ryan, C. (1992). Client perceptions of hotels: A multi-attribute approach. *Tourism Management, 13*(2), 163–168.

Shaw, G., Bailey, A., & Williams, A. (2011). Aspects of service-dominant logic and its implications for tourism management: Examples from the hotel industry. *Tourism Management, 32*(2), 207–214.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323.

STRanalytics. (2014). Who's in your competitive set? Retrieved November 9, 2014, from https://www.strglobal.com/Media/Default/Samples/NNA_samples/NNA_CompSetSuite_Details.pdf

Stringam, B. B., & Gerdes, J., Jr. (2010). An analysis of word-of-mouse ratings and guest comments of online hotel distribution sites. *Journal of Hospitality Marketing & Management, 19*(7), 773–796.

Stringam, B. B., Gerdes, J., Jr., & Vanleeuwen, D. M. (2010). Assessing the importance and relationships of ratings on user-generated traveler reviews. *Journal of Quality Assurance in Hospitality & Tourism, 11*(2), 73–92.

U.S. Census Bureau, Population Division. (2007, June 28). *Table 1: Annual estimates of the population for incorporated places over 100,000*, Ranked by July 1, 2006 Population: April 1, 2000 to July 1, 2006 (CSV). http://www.census.gov/popest/states/NST-ann-est2006.html

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management, 30*(1), 123–127.

Walls, A. R., Okumus, F., Wang, Y. R., & Kwun, D. J. W. (2011). An epistemological view of consumer experiences. *International Journal of Hospitality Management, 30*(1), 10–21.

Wood, E. (2001). Marketing information systems in tourism and hospitality small-and medium-sized enterprises: A study of Internet use for market intelligence. *International Journal of Tourism Research, 3*(4), 283–299.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, *3*.

Woodside, A., Cruickshank, B. F., & Dehuang, N. (2007). Stories visitors tell about Italian cities as destination icons. *Tourism Management, 28*(1), 162–174.

Wöber, K. W. (2003). Information supply in tourism management by marketing decision support systems. *Tourism Management, 24*(3), 241–255.

Wu, C. H. J., & Liang, R. D. (2009). Effect of experiential value on customer satisfaction with service encounters in luxury-hotel restaurants. *International Journal of Hospitality Management, 28*(4), 586–593.

Xiang, Z., & Law, R. (2013). Online competitive information space for hotels: An information search perspective. *Journal of Hospitality Marketing & Management, 22*(5), 530–546.

Xiang, Z., Pan, B., Law, R., & Fesenmaier, D. R. (2010). Assessing the visibility of destination marketing organizations in Google: A case study of convention and visitor bureau websites in the United States. *Journal of Travel & Tourism Marketing, 27*(7), 694–707.

Xiang, Z., Schwartz, Z., Gerdes, J., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16.

Part VI
Closing Remarks

Big Data Analytics, Tourism Design and Smart Tourism

Zheng Xiang and Daniel R. Fesenmaier

1 Introduction

In a recent article published in the Harvard Business Review, Porter und Heppelmann (2014) wrote:

> Information technology is revolutionizing products. Once composed solely of mechanical and electrical parts, products have become complex systems that combine hardware, sensors, data storage, microprocessors, software, and connectivity in myriad ways. These 'smart, connected produces'—made possible by vast improvements in processing power and device miniaturization and by the network benefits of ubiquitous wireless connectivity—have unleashed a new era of competition.

With this the authors move on to paint a picture of today's economy wherein information technology (IT) redefines the meaning of production and, consequently, the structure of competition as the new conditions for corporate strategy. While this view certainly reflects the free-market, capitalistic philosophy primarily focused upon the so-called competitive advantage as the end outcome of strategy, Porter and his colleague offer an intriguing vision of the transformative effect of IT's reaching into every facet of products and becoming the driver for the restructuration of an industry. And similar to the manufacturing industry, travel and tourism is likely to go through substantial transformation because of today's information technology. Indeed, imagine a world full of embedded sensors that are digitally-connected to form the Internet of Things (Atzori, Iera, & Morabito, 2010);

Z. Xiang (✉)
Department of Hospitality and Tourism Management, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
e-mail: philxz@vt.edu

D.R. Fesenmaier
National Laboratory for Tourism & eCommerce, Department of Tourism, Recreation and Sport Management, University of Florida, Gainesville, Florida, USA

Z. Xiang, D.R. Fesenmaier (eds.), *Analytics in Smart Tourism Design*, Tourism on the Verge, DOI 10.1007/978-3-319-44263-1_17

a world where every traveler using a variety of interfaces and devices (wearables, smartphone, tablets, and laptops and so forth) to actively engage in (and create) travel-related activities, to actively interact with both physical and virtual environments (Xiang, Wang, O'Leary, & Fesenmaier, 2015), and to connect with their everyday life and social circles before, during, and after travel (Wang, Xiang, & Fesenmaier, 2016). And even further, a world with computer programs (i.e., artificial intelligence) capable of understanding each traveler's needs and making real time personalized recommendations. No wonder there is a growing consensus that we are entering an era of the so-called smart tourism (Gretzel, Sigala, Xiang, & Koo, 2015).

As the use of IT evolves, so has our means of understanding and designing today's new reality. The emergence of big data analytics is not simply a buzzword; instead, it is a logical result of advancements in computer engineering (in both hardware and software), the wide adoption and use of IT by consumers, and the industry's search for efficiency and new ways to measure productivity and performance, especially in the last two decades. It is also the logical result of the desire by individuals to somehow measure themselves (using new tools to monitor the status of exercise, etc.) and to measure many artefacts within nature and society and are discussed within the notions of the 'quantified self' and 'people as sensors.' Within these wide range of contexts, big data analytics has been proposed as a new paradigm and a toolbox for tourism design, tourism marketing and destination management. And, these tools are radically different from the conventional methods of research and development in travel and tourism. The collection of chapters within this book reflects such thinking and fits into the overall vision of strategic use of IT for tourism development. We are hopeful that the ideas illustrated here will further motivate all of us to ask fundamental questions such as "how does the tourism adapt to this new business reality?", "how does the traveler adapt to this new reality?"; and, of course, "how should we design and manage tourism places?"

2 Information Technology and Tourism Development

Much has been said about the impact of IT on the economy as an essential driver of change. Early intellectual efforts since the 1990s provided a complex vision of how firms could realize the promises of the development of the Internet (e.g., Friedman, 2005; Negroponte, 1995; Tapscott, Ticoll, & Lowy, 2000). Parallel to these developments, a few books focusing on the role of the IT in travel and tourism were written; most notable were Poon's *Tourism, Technology and Competitive Strategies* (1993), Sheldon's *Tourism Information Technology* (1997), and Werthner and Klein's *Information Technology and Tourism—A Challenging Relationship* (1999), which reflected the new thinking regarding the nature and impact of IT. Propelled by information technology, tourism development has gone through three stages where the first stage of development roughly occurred between the years 1991 and 2000 when leaders in the tourism industry began to realize that they were largely information arbitrators and that the Internet enabled them to communicate easily and effectively with their existing and potential customers. During this time period the Internet was largely seen as a market communication tool. Many within the tourism industry envisioned new ways of meeting the information needs of this market where websites replaced travel brochures for essentially every destination and attraction, and for every travel-related service worldwide. In the United States, for example, essentially every tourism organization had developed a website by the early 2000s, and many had gone through the evolution from a simple 'electronic brochure' to highly interactive systems that supported reservations, search and even virtual tours; importantly, the website had become the primary (and in many circumstances, the only) source of contact with potential visitors (Zach, Gretzel, & Xiang, 2010). In retrospect, this transformation can be easily understood as the computer framework already existed through the various global distribution systems (GDSs) linking travel agencies to the airlines. Also during this time, many innovative destination marketing organizations (DMOs) began to realize their new role as partners within the tourism system wherein they became "information brokers" as they sought to develop and coordinate a range of new systems that would be used by their stakeholders (Gretzel & Fesenmaier, 2002; Wang & Xiang, 2007).

Following the decade of the 1990s came the second stage of development (roughly 2000–2010) wherein the leaders of the tourism industry began to understand and appreciate that travel experiences are products that can be bundled and sold with the aid of IT. Exemplified by the success of *The Experience Economy* by Pine and Gilmore (1999), the core business model of many tourism organizations changed and the impacts of IT took hold. With this new perspective on the core product, the tourism industry was challenged to recognize that the "new consumer" demands highly personalized experiences, that competition for visitors would now be waged in global markets, and that the traveler largely took 'control' of this new marketplace. Traditional travel agencies were decimated by newly formed online firms such as Expedia and Travelocity; the large travel suppliers such as airlines and hotels could connect directly with potential customers; search engines such as Google became dominant as they provided instant access to websites, and therefore could be indexed, advertised and managed; on top of all this, meta search engines like Kayak further made the distribution of travel products more accessible and more transparent. In response, destination marketing organizations were forced to recalibrate again their role to become a different kind of intermediary whereby they largely focused on building the capacity necessary to assist small and medium tourism firms in adapting to this new and very challenging environment. And, as a result, they became destination managers by changing their business model where it focused on creating new forms of value within the tourism chain.

The third stage of development started circa 2010 and onward wherein the advancements in areas such as search engines, social media, the Internet of Things (IoT) and mobile technologies simulated further transformation of the tourism industry (Xiang, Wang, et al., 2015). In particular, the introduction of Web 2.0 signaled a new round of adaptation which required another new and even more transformational framework for tourism management. The more important feature of this stage is the development and maturity of new social systems which began to emerge as an "Army of Davids" (Reynolds, 2006). Further, the advent of smartphones, mobile computing systems that incorporate a variety of technologies including communications, GPS, and photography, enriched the social environment further such that it empowers users to control their travel experience. The combination of instrumented IT infrastructure (i.e. sensors' ability to measure use and conditions of the environment and tourism assets) and interconnected systems (e.g., smartphones, cloud computing, Internet of Things, RFID networks) effectively enable tourism destinations to gather, integrate, analyze, and ultimately support optimized decisions based on collective knowledge, which in turn, improves the operational efficiency and quality of life of a city (destination residents). In particular, the Internet of Things is crucial for creating a pervasive, "smart" technological environment that encompasses connected physical and digital infrastructures (Atzori et al., 2010).

Given the information-intensive nature of tourism and the resulting high dependence on IT, the concept of smart tourism has been proposed to describe this current stage of tourism development (Gretzel et al., 2015). In many ways, smart tourism can be seen as a logical progression from traditional tourism and the more recent e-tourism in that the groundwork for the innovations and the technological orientation of the industry and the consumers were laid early with the extensive adoption of IT. This stage of development continues with the widespread adoption of social media (Sigala, Christou, & Gretzel, 2012), and a shift of focus towards enhancing the tourism experience with reliance on the interconnectivity of physical/digital objects, high fluidity of tourism information and high mobility of travelers (Buhalis & Law, 2008; Wang, Park, & Fesenmaier, 2012). Within this stage, smart systems can be used to support travelers by: (1) anticipating user needs based upon a variety of factors, and making recommendations with respect to the choice of context-specific consumption activities such as points of interest, dining and recreation; (2) enhancing travelers' on-site experiences by offering rich information, location-based and customized, interactive services; and, (3) enabling travelers to share their experiences so that they help others in their decision making process, revive and reinforce their experiences as well as construct their self-image and status on social networks. From the destination's perspective, the emphasis is on process automation, efficiency gains, new product development, demand forecasting, crisis management, and value co-creation (Gretzel, 2011) defines the future of smart tourism. But from the traveler perspective, the empowerment of the traveler though active involvement in the creative process and the freedom of choice represents SMART tourism.

3 Big Data Analytics, Smart Tourism and Tourism Design

The vision of smart tourism clearly rests on the abilities of tourism businesses and destinations to not only collect enormous amounts of data, but to intelligently store, process, combine, analyze and use big data to design tourism operations, services and business innovation (Fesenmaier & Xiang, 2016). The technological foundations of smart tourism is multidimensional, consisting of the ubiquitous infrastructure, mobile and context-aware information systems, and the increasingly complex and dynamic connectivity that supports interactions not only with one's physical environment but also the community and society at large directly or indirectly related to the traveler. As shown in Fig. 1, smart tourism development is built upon the collection, exchange, and processing of data generated in different components of the system involving the consumer, the business, and the destination as a whole. Particularly, the networks that surround travelers in trip planning and their mobility encompass systems that capture and generate enormous amount of consumer data. Thus, the new systems supporting a variety of travel-related metrics enable tourism managers to better understand where and how potential and existing visitors live, the nature of information used to plan a trip, as well with whom travelers share their experiences before, during and after the trip. These business analytical applications support the design of smart tourism by offering enhanced customer intelligence, improving business processes, and, ultimately, enabling the implementation of new strategies for navigating an increasingly competitive environment.

As a toolbox, big data analytics is obviously diverse in terms of the nature of data, analytical operation, and business application (Xiang et al. 2015). Compared to traditional methods of research and development, big data analytics improves our capabilities to understand the consumer market at unprecedented scale, scope, and depth (Boyd & Crawford, 2012). While there is a lack of clear-cut definition of its epistemological boundaries and structures, smart tourism development can be used as a general framework that informs us of different contexts and conditions for big data analytics in tourism.

At the consumer level, the focus of smart tourism development is on providing intelligent support based upon the timely, comprehensive understanding of the tourism experience. In this regard tourism big data are intended to be more context

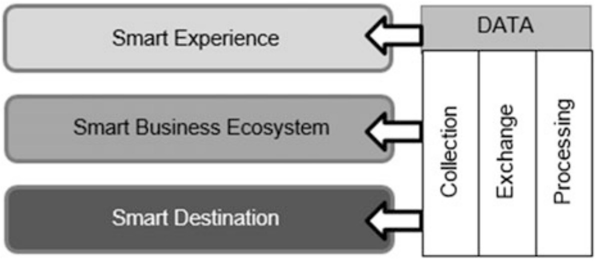


Fig. 1 Components and layers of smart tourism (Gretzel et al., 2015)

rich, more dynamic and potentially more reflective of the real time conditions, which potentially offers opportunities to understand travelers in more authentic ways. First, non-conventional data such as location-based transaction data can offer a moment-by-moment picture of interactions over extended periods of time, providing information about both the structure and content of economic relationships. In this regard, mobile, geo-based data offer opportunities to produce real time and context-rich insights in the consumer market, giving rise to the capabilities of "now-casting" (Scaglione, Favre, & Trabichet, 2016). Second, today's travelers are likely more socially-connected and therefore tourism big data, e.g., those collected from social media, can provide more information about travel as a social activity (Wood, Guerry, Silver, & Lacayo, 2013). New technologies, such as video surveillance, email, and smart name badges, offer a complete picture of social interactions over extended periods of time, which could provide information about both the structure and content of human relationships. The social dimension can also be recognized as smart objects embedded in the environment may automatically trigger the transmission of messages to family and friends to enable them to know what we are doing or what we have done in the past, such as moving from one site to another or meeting some common friends. Sensors embedded in the travel environment can help establish and assess group interactions over time with "sociometers", leading to a new understanding of travel groups and communities (Lazer et al., 2009; Olguın, Gloor, & Pentland, 2009). Third, wearable technologies such as smart watches play an important role in this as well as they not only collect data through their sensors and cameras but also communicate with the network and potentially the Internet of Things. This enables us to understand not only how people travel but also how their travel activities connect with their everyday lives and contribute to their personal and social well-being (Uysal, Sirgy, Woo, & Kim, 2016; Wang et al., 2016).

At the business level, smart destinations rely on an abundance of free information to be translated into business value propositions. Although tourism businesses (and their systems) can be characterized as heterogeneous, distributed, and even fragmented, the overarching goal of for system development should be open, scalable, and cooperative, enabling full autonomy of the respective participants of the industry as well as supporting the entire tourist experience and all business phases (Staab & Werthner, 2002). Traditionally, economic power in tourism development arises from the control over information sources and flows (e.g., in the case of online travel agencies). Within the context of big data, it is equally important to recognize that business value not only emerges from ownership but increasingly from access to shared data and other resources. Therefore, the practice of big data analytics can be seen as a catalyst which fosters partnership building and resource sharing among tourism businesses. For example, data from industry sectors that are conventionally considered not directly relevant to the tourism sector can now be used as indicators to measure a range of tourism activities including volumes and tourist flows through a destination.

At the destination level, the essence of smart tourism is the transformation of the tourist place (e.g., the smart city) wherein information technology serves as the bedrock for innovation in economic activities and societal wellbeing as the result of tourism. The ultimate goal of smart tourism is to support mobility, creativity, resource availability and allocation, sustainability and quality of life and visits through large-scale, coordinated efforts and strategic investments in technological infrastructure. To achieve this goal, smart destinations must build an "info-structure" which encourages both active and creative (e.g., creating and then sharing one's experiences) or implicit (through sensors or wearable devices) sharing of data by consumers. Open technological platforms can be established to harness social wisdom through crowdsourcing (Howe, 2006) and the so-called "citizen science" (Goodchild, 2007; Silvertown, 2009), whereby voluntary participation by individuals in the society contributes to system-wide knowledge and value co-creation. In this regard, big data analytics creates an environment of openness and serves as a critical foundation for innovation within the general framework of smart tourism (Egger, Gula, & Walcher, 2016).

4 Issues and Challenges

In this chapter and implicit throughout this book we argue that big data analytics in inherently connected with the recent emergence of tourism design and smart tourism development, which is a logical result of the advancements of IT and its wide adoption in both consumer market and the industry in the last 20 years. Data lies at the core of all smart tourism activities, and the utilization and exploitation of big data will likely result in new business models and industry-wide innovations in travel and tourism. However, there are many issues and challenges ahead in the use of big data. For example, privacy is an obvious concern in the context of smart tourism, especially location-based services, while extremely useful for tourists, also make consumers vulnerable (Anuar & Gretzel, 2011). Indeed, the European Union and other governing bodies have pressed many of the data related firms such as Google and essentially all telecoms to protect the privacy rights of users. The use of big data also raises significant new issues with respect to information governance and how we can correctly derive the value of information in tourism (Gretzel et al., 2015). The recent coverage of Target's use of data driven marketing provides a simple example of how such systems can easily create many unintended consequences including the loss of privacy (Duhigg, 2012). Further, there have been growing criticism about the data-driven approach (i.e., data mining) in terms of new epistemological dilemmas and inductive reasoning in the implementation of big data analytics (e.g., Frické, 2015; Tufekci, 2014) wherein researchers argue that big data analytics changes the fundamental nature of the research process to such a degree that 'science is gone.' While these very real and very important concerns are not addressed by the authors of the chapters in this book, they do make it clear that smart products will continue to challenge (i.e., cause huge economic, social and political problems) the basic building blocks of the industry and society as a whole. Further from a more optimistic perspective, big data and tourism analytics and smart tourism will support the tourism industry and travelers by improving their capabilities to capture, analyze and interpret data, and these new tools will drive the tourism industry's search for value creation, innovation and the ability to manage tourism destinations.

References

Anuar, F. I., & Gretzel, U. (2011, January 26–28). *Privacy concerns in the context of location based services for tourism*. ENTER 2011 Conference, Innsbruck, Austria. Retrieved March 1, 2015, from http://ertr.tamu.edu/enter-2011-short-papers/

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks, 54*(15), 2787–2805.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism Management, 29*(4), 609–623.

Duhigg, C. (2012). How companies learn your secrets. *New York Times*. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Egger, R., Gula, I., & Walcher, D. (Eds.). (2016). *Open tourism: Open innovation, crowdsourcing and co-creation challenging the tourism industry*. Vienna: Springer.

Fesenmaier, D. R., & Xiang, Z. (Eds). (2016). *Designing tourism places*. Vienna: Springer.

Friedman, T. (2005). *The world is flat: A brief history of the twenty-first century*. New York: Farrar, Straus and Giroux.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology, 66*(4), 651–661.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of Tourism Research, 38*(3), 757–779.

Gretzel, U., & Fesenmaier, D. R. (2002). Implementing knowledge-based interfirm networks in heterogeneous B2B environments: A case study of the Illinois Tourism Network. In K. Wöber, A. J. Frew, & M. Hitz (Eds.), *Information & communication technologies in tourism 2002* (pp. 39–48). Wien: Springer.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets, 25*(3), 179–188.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–4.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Life in the network: The coming age of computational social science. *Science, 323*(5915), 721.

Negroponte, N. (1995). *Being digital*. New York: Knopf.

Olguın, D. O., Gloor, P. A., & Pentland, A. S. (2009). Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS* (Vol. 9).

Pine, B. J., & Gilmore, J. H. (1999). *The experience economy: Work is theatre & every business a stage*. Boston: Harvard Business Press.

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard Business Review, 92*(11), 64–88.

Reynolds, G. (2006). *An army of Davids: How markets and technology empower ordinary people to beat big media, big government and other Goliaths*. Nashville: Thomas Nelson.

Scaglione, M., Favre, P., & Trabichet, J.-P. (2016, April 11–12). Using mobile data and strategic tourism flows: Pilot study MoniTour in Switzerland. In *Proceedings of the big data & business intelligence in the travel & tourism domain workshop*, Östersund, Sweden.

Sigala, M., Christou, E., & Gretzel, U. (Eds.). (2012). *Social media in travel, tourism and hospitality: Theory, practice and cases*. London: Ashgate.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution, 24*(9), 467–471.

Staab, S. & Werthner, H. (2002). Intelligent systems for tourism. *IEEE Intelligent Systems*, November/December, 2002, 53–55.

Tapscott, D., Ticoll, D., & Lowy, A. (2000). *Digital capital: Harnessing the power of business webs*. Boston: Harvard Business Press.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. arXiv preprint arXiv:1403.7400.

Uysal, M., Sirgy, M. J., Woo, E., & Kim, H. L. (2016). Quality of life (QOL) and well-being research in tourism. *Tourism Management, 53*, 244–261.

Wang, D., Park, S., & Fesenmaier, D. (2012). The role of Smartphones in mediating the tourism experience. *Journal of Travel Research, 51*(4), 371–387.

Wang, Y., & Xiang, Z. (2007). Toward a theoretical framework of collaborative destination marketing. *Journal of Travel Research, 46*(1), 75–85.

Wang, D., Xiang, Z., & Fesenmaier, D. R. (2016). Smartphone use in everyday life and travel. *Journal of Travel Research, 55*(1), 52–63.

Werthner, H., & Klein, S. (1999). *Information technology and tourism: A challenging relationship*. Vienna: Springer.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3.

Xiang, Z., Schwartz, Z., Gerdes, J., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*(1), 120–130.

Xiang, Z., Wang, D., O'Leary, J. T., & Fesenmaier, D. R. (2015). Adapting to the internet: Trends in travelers' use of the web for trip planning. *Journal of Travel Research, 54*(4), 511–527.

Zach, F., Gretzel, U., & Xiang, Z. (2010). Innovation in the web marketing programs of American convention and visitor bureaus'. *Information Technology and Tourism, 12*(1), 47–63.