ELSEVIER

Comprehensive Geographic Information Systems

Editor in Chief Bo Huang

COMPREHENSIVE GEOGRAPHIC INFORMATION SYSTEMS

This page intentionally left blank

COMPREHENSIVE GEOGRAPHIC INFORMATION SYSTEMS

EDITOR IN CHIEF

Bo Huang

The Chinese University of Hong Kong, Hong Kong

VOLUME 1

GIS METHODS AND TECHNIQUES

VOLUME EDITORS

Thomas J. Cova

The University of Utah, Salt Lake City, UT, United States

Ming-Hsiang Tsou

San Diego State University, San Diego, CA, United States

ELSEVIER

AMSTERDAM BOSTON HEIDELBERG LONDON NEW YORK OXFORD
PARIS SAN DIEGO SAN FRANCISCO SINGAPORE SYDNEY TOKYO

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Notice
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-804660-9

Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Laura Escalante Santos
Associate Content Project Manager: Paula Davies and Katie Finn
Cover Designer: Mark Rogers

Printed and bound in the United States

EDITOR IN CHIEF

Bo Huang

Dr. Bo Huang is a professor in the Department of Geography and Resource Management, The Chinese University of Hong Kong, where he is also the Associate Director of Institute of Space and Earth Information Science (ISEIS). Prior to this, he held faculty positions at the University of Calgary, Canada, and the National University of Singapore. He has a background and experience in diverse disciplines, including urban planning, computer science, Geographic Information Systems (GIS), and remote sensing. His research interests cover most aspects of *GIScience*, specifically the design and development of models and algorithms in spatial/spatiotemporal statistics, remote sensing image fusion and multiobjective spatial optimization, and their applications in environmental monitoring and sustainable land use and transportation planning. The Geographically and Temporally Weighted Regression (GTWR) model (available in his ResearchGate) that was developed by him in 2010 has now been widely used in a wide range of areas, including economics, environment, geography, and urban planning. Dr. Huang serves as the Asia-Pacific Editor of *International Journal of Geographical Information Science* (Taylor & Francis), the Executive Editor of *Annals of GIS* (Taylor & Francis), and the Chief Scientist of the Joint Laboratory of Smart Cities (Beijing). He was awarded Chang Jiang Chair Professorship in 2016 by the Ministry of Education of PR China.

This page intentionally left blank

VOLUME EDITORS

Georg Bareth

Georg Bareth studied Physical Geography at the University of Stuttgart and graduated in 1995. From 1996 to 1999 he received a PhD scholarship from the German Research Foundation (DFG) and worked on his thesis "Emissions of Greenhouse Gases from Agriculture – Regional Presentation and Estimation for a dairy farm region by using GIS" at the University of Hohenheim. In 2004, he habilitated in Agroinformatics, and since 2004, he holds a professorship for Geoinformatics at the University of Cologne.

Kai Cao

Kai Cao is a lecturer in the Department of Geography at National University of Singapore (NUS), an affiliated researcher in the Institute of Real Estate Studies, a research associate in the Center for Family and Population Research, and a member of the steering committee of the Next Age Institute at NUS. He is serving on the Board Committee and as the chair of Newsletter Committee in the International Association of Chinese Professionals in Geographic Information Science. He had also been a member of the National Geographic's Committee for Science and Exploration for one year. He obtained his BSc degree in Geography (Cartography and Geographic Information Science) and MPhil degree in Geography (Remote Sensing and Geographic Information Science) from Nanjing University in China, and his PhD degree in Geography from The Chinese University of Hong Kong. Prior to joining the Department of Geography at NUS, he had worked in the Center for Geographic Analysis at Harvard University, in the Department of Geography at the University of Illinois at Urbana–Champaign, and in the World History Center at the University of Pittsburgh, respectively. He was also a visiting research scholar in the Department of Human Geography and Spatial Planning at Utrecht University in 2009, and a visiting scholar in the Center for Spatial Studies and Department of Geography at University of California, Santa Barbara (UCSB) in 2012.

Dr. Kai Cao specializes in GIScience, spatial simulation and optimization, urban analytics, and spatially integrated social science. He has published numerous internationally referred journal articles, book chapters, and conference papers in his field and had also been a guest editor of a special issue in the *International Journal of Geographical Information Science* on the topic of "Cyberinfrastructure, GIS and Spatial Optimization", together with Dr. Wenwen Li from Arizona State University and Prof. Richard Church from UCSB.

Tom Cova

Tom Cova is a professor of Geography at the University of Utah and director of the Center for Natural and Technological Hazards. He received a BS in Computer Science from the University of Oregon and an MA and PhD in Geography from the University of California, Santa Barbara where he was an Eisenhower Fellow. Professor Cova's research and teaching interests are environmental hazards, emergency management, transportation, and geographic information science (GIScience). His initial focus was regional evacuation modeling and analysis, but this has since been expanded to include emergency preparedness, public warnings, and protective actions. He has published in many leading GIS, hazards, and transportation journals including the *International Journal of Geographical Information Science* (IJGIS), *Transactions in GIS*, *Computers, Environment and Urban Systems*, *Transportation Research A and C*, *Natural Hazards*, *Geographical Analysis*, *Natural Hazards Review*, and *Environment and Planning A*. His 2005 paper in *Natural Hazards Review* resulted in new standards in the United States for transportation egress in fire-prone regions (National Fire Protection Association 1141). Concepts drawn from his 2003 paper on lane-based evacuation routing in *Transportation Research A: Policy and Practice* have been used in evacuation planning and management worldwide, most notably in the 2012 Waldo Canyon Fire evacuation in Colorado Springs.

Professor Cova was a coinvestigator on the National Center for Remote Sensing in Transportation (NCRST) Hazards Consortium in 2001–04. Since then most of the support for his research has been provided by the National Science Foundation on projects ranging from evacuation versus shelter-in-place in wildfires to the analytical derivation of warning trigger points. He chaired the GIS Specialty Group for the Association of American Geographers in 2007–08 and the Hazards, Risks and Disasters Specialty Group in 2011–12. In 2008 he served as program chair for the International Conference on Geographical Information Science (GIScience, 2008) in Park City, Utah. He was a mentor and advisor for the National Science Foundation project "Enabling the Next Generation of Hazards Researchers" and is a recipient of the Excellence in Mentoring Award from the College of Social & Behavioral Science at the University of Utah.

Elisabete A. Silva

Elisabete Silva, BA, MA (Lisbon), PhD (Massachusetts), MRTPI, is a University Senior Lecturer (Associate Professor) in Spatial Planning and a Fellow and DoS of Robinson College, University of Cambridge, UK. Dr. Silva has a research track record of 25 years, both at the public and private sector. Her research interests are centered on the application of new technologies to spatial planning, in particular city and metropolitan dynamic modeling through time. The main subject areas include land use change, transportation and spatial plans and policy, the use of Geographic Information Systems (GIS), spatial analysis, and new technologies/models in planning (i.e., CA and ABM). She is the coauthor of the Ashgate book *A Planners' Encounter With Complexity* (2010) and *The Routledge Handbook of Planning Research Methods* (2014).

Chunqiao Song

Chunqiao Song received his BS degree from Wuhan University in 2008 and his MS degree from the Chinese Academy of Sciences in 2011, respectively. Both major in geographic information science. He received his PhD degree in geography from the Chinese University of Hong Kong in 2014. He is currently working as a researcher in the University of California, Los Angeles.

He focuses his research on developing the applications of remote sensing and geographic information techniques in large-scale environment monitoring and process modeling. It aims to contribute to the development of novel scientific, theoretical, and methodological aspects of geoinformatics techniques to understand how the key environment elements (e.g., water, ice, and ecosystem) respond to a changing climate and human intervention in High Mountain Asia and worldwide. His current research work includes (1) developing high-resolution satellite-based lake hydrographical datasets, which are available at global scale, and (2) understanding lake water storage dynamic and its hydrological processes and cryosphere on the Tibetan Plateau (Earth's "Third Pole") and high mountainous regions. He is the author of more than 50 primary research articles, reviews, and book chapters in hydrological, remote sensing, ecological, or environmental fields.

Yan Song

Yan Song is a full professor at the Department of City and Regional Planning and director of the Program on Chinese Cities at the University of North Carolina at Chapel Hill. Dr. Song's research interests includes low-carbon and green cities, plan evaluation, land use development and regulations, spatial analysis of urban spatial structure and urban form, land use and transportation integration, and how to accommodate research in the above fields by using planning supporting systems such as GIS, big data, and other computer-aided planning methods and tools.

Ming-Hsiang Tsou

Ming-Hsiang (Ming) Tsou is a professor in the Department of Geography, San Diego State University (SDSU), and the founding director of the Center for Human Dynamics in the Mobile Age (HDMA) (http://humandynamics.sdsu.edu/). He received a BS (1991) from the National Taiwan University, an M.A. (1996) from the State University of New York at Buffalo, and a PhD (2001) from the University of Colorado at Boulder, all in Geography. His research interests are in Human Dynamics, Social Media, Big Data, Visualization, and Cartography, Web GIS, High Performance Computing (HPC), Mobile GIS, and K-12 GIS education. He is a coauthor of *Internet GIS*, a scholarly book published in 2003 by Wiley, and served on the editorial boards of the *Annals of GIS* (2008–), *Cartography and GIScience* (2013–), and the *Professional Geographers* (2011–). Tsou was the chair of the Cartographic Specialty Group (2007–08), the chair of the Cyberinfrastructure Specialty Group (2012–13) in the Association of American Geographers (AAG), and the cochair of the NASA Earth Science Enterprise Data System Working Group (ESEDWG) Standard Process Group (SPG) from 2004 to 2007. He has served on two US National Academy of Science Committees: "Research Priorities for the USGS Center of Excellence for Geospatial Information Science" (2006–07) and "Geotargeted Alerts and Warnings: A Workshop on Current Knowledge and Research Gaps" (2012–13). In 2010, Tsou was awarded a $1.3 million research grant funded by National Science Foundation and served as the principal investigator (PI) of the "Mapping ideas from Cyberspace to Realspace" (http://mappingideas.sdsu.edu/) research project (2010–14). This NSF-CDI project integrates GIS, computational linguistics, web search engines, and social media APIs to track and analyze public-accessible websites and social media (tweets) for visualizing and analyzing the diffusion of information and ideas in cyberspace. In Spring 2014, Tsou established a new research center, Human Dynamics in the Mobile Age (HDMA), a transdisciplinary research area of excellence at San Diego State University to integrate research works from GIScience, Public Health, Social Science, Sociology, and Communication. Tsou is the founding director of the HDMA Center. In Fall 2014, Tsou received an NSF Interdisciplinary Behavioral and Social Science Research (IBSS) award for "Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks" (Award#1416509, $999,887, 2014–18, http://socialmedia.sdsu.edu/). This large interdisciplinary research project studies human dynamics across social media and social networks, focusing on information diffusion modeling over time and space, and the connection between online activities and real-world human behaviors (including disaster evacuation, vaccine exemption, etc). Tsou is also involved with several GIS education projects for K-12 and higher education. He has served on the AP GIS&T course advisory board at AAG and as a senior researcher in the National GeoTech Center, and the Geospatial Technology Coordinator in California Geographic Alliance to promote GIS education in universities, community colleges, and high schools. Tsou has conducted professional GIS training workshops for GIS teachers annually at the San Diego State University during the last 10 years (http://geoinfo.sdsu.edu/hightech/).

This page intentionally left blank

CONTRIBUTORS TO VOLUME 1

Jochen Albrecht
Hunter College, City University of New York, New York, NY, United States

Reem Y Ali
University of Minnesota Twin Cities, Mississippi, MN, United States

Li An
San Diego State University, San Diego, CA, United States

Marc P Armstrong
The University of Iowa, Iowa City, IA, United States

Hyowon Ban
California State University, Long Beach, CA, United States

Saad Saleem Bhatti
University of Cambridge, Cambridge, United Kingdom

Kai Cao
National University of Singapore, Singapore

Jeremy W Crampton
University of Kentucky, Lexington, KY, United States

Andrew Crooks
George Mason University, Fairfax, VA, United States

Kevin M Curtin
George Mason University, Fairfax, VA, United States

Jie Dai
San Diego State University, San Diego, CA, United States; and University of California, Santa Barbara, CA, United States

Emre Eftelioglu
University of Minnesota Twin Cities, Mississippi, MN, United States

Rob Feick
University of Waterloo, Waterloo, ON, Canada

Colin J Ferster
University of Victoria, Victoria, BC, Canada

Shaun Fontanella
Ohio State University, Columbus, OH, United States

Sven Fuhrmann
George Mason University, Fairfax, VA, United States

Song Gao
University of California, Santa Barbara, CA, United States

Rina Ghose
University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Michael F Goodchild
University of California, Santa Barbara, CA, United States

Jacob Hartz
Rochester Institute of Technology, Rochester, NY, United States

Nick Hedley
Simon Fraser University, Burnaby, BC, Canada

Alison Heppenstall
University of Leeds, Leeds, United Kingdom

Paul Holloway
University of York, York, United Kingdom

Yingjie Hu
University of Tennessee, Knoxville, TN, United States

Miaoqing Huang
University of Arkansas, Fayetteville, AR, United States

Eric M Huntley
University of Kentucky, Lexington, KY, United States

Zhe Jiang
University of Alabama, Tuscaloosa, AL, United States

Emily C Kaufman
University of Kentucky, Lexington, KY, United States

Angelina Konovitz-Davern
Rochester Institute of Technology, Rochester, NY, United States

Dapeng Li
Michigan State University, East Lansing, MI, United States

Linna Li
California State University, Long Beach, CA, United States

Yan Li
University of Minnesota Twin Cities, Mississippi, MN, United States

Gengchen Mai
University of California, Santa Barbara, CA, United States

Jacek Malczewski
Western University, London, ON, Canada

Nick Malleson
University of Leeds, Leeds, United Kingdom

Ashely Miller
Rochester Institute of Technology, Rochester, NY, United States

Jennifer A Miller
University of Texas at Austin, Austin, TX, United States

Atsushi Nara
San Diego State University, San Diego, CA, United States

Trisalyn Nelson
Arizona State University, Tempe, AZ, United States

José Pedro Reis
University of Cambridge, Cambridge, United Kingdom

Colin Robertson
Wilfrid Laurier University, Waterloo, ON, Canada

David Schwartz
Rochester Institute of Technology, Rochester, NY, United States

Dara E Seidl
San Diego State University, San Diego, CA, United States

Shashi Shekhar
University of Minnesota Twin Cities, Mississippi, MN, United States

Xuan Shi
University of Arkansas, Fayetteville, AR, United States

Lena Siedentopp
United Nations University Institute for Environment and Human Security, Bonn, Germany

Elisabete A Silva
University of Cambridge, Cambridge, United Kingdom

Scott Simmons
Open Geospatial Consortium, Fort Collins, CO, United States

Joerg Szarzynski
United Nations University Institute for Environment and Human Security, Bonn, Germany

Zhenyu Tan
Wuhan University, Wuhan, China

Xun Tang
University of Minnesota Twin Cities, Mississippi, MN, United States

Brian Tomaszewski
Rochester Institute of Technology, Rochester, NY, United States

Ming-Hsiang Tsou
San Diego State University, San Diego, CA, United States

Fahui Wang
Louisiana State University, Baton Rouge, LA, United States

Suzanne P Wechsler
California State University, Long Beach, CA, United States

David W S Wong
George Mason University, Fairfax, VA, United States

Ningchuan Xiao
Ohio State University, Columbus, OH, United States

Bo Xu
California State University, San Bernardino, CA, United States

Chen Xu
University of Wyoming, Laramie, WY, United States

Xinyue Ye
Kent State University, Kent, OH, United States

Eun-Hye Yoo
University at Buffalo, Buffalo, NY, United States

Peng Yue
Wuhan University, Wuhan, China

CONTENTS OF VOLUME 1

Spatial Analysis and Modeling

Space-Time GIS

Spatial Data Quality

Cyberinfrastructure and GIS

Virtual GIS

Mobile GIS

Public GIS

GIS Design and Project Management

This page intentionally left blank

CONTENTS OF ALL VOLUMES

VOLUME 1: GIS METHODS AND TECHNIQUES

New Perspectives on GIS (Multidisciplinary)

Data Management

Spatial Analysis and Modeling

Public GIS

GIS Design and Project Management

VOLUME 2: GIS APPLICATIONS FOR ENVIRONMENT AND RESOURCES

GIS for Biophysical Environment

GIS for Resources

GIS for Energy

PREFACE

Since its inception in the 1960s, Geographic Information System (GIS) has been undergoing tremendous development, rendering it a technology widely used for geospatial data management and analysis. The past several decades have also witnessed increasing applications of GIS in a plethora of areas, including environment, energy, resources, economics, planning, transportation, logistics, business, and humanity. The rapid development of GIS is partly due to the advances in computational technologies and the increasing availability of various geospatial data such as satellite imagery and GPS traces.

Along with the technological development of GIS, its underlying theory has significantly progressed, especially on data representation, data analysis, uncertainty, and so on. As a result, the theory, technology, and application of GIS have made great strides, leading to a right time to summarize comprehensively such developments. Comprehensive Geographical Information System (CGIS) thus comes.

CGIS provides an in-depth, state-of-the-art review of GIS with an emphasis on basic theories, systematic methods, state-of-the-art technologies, and its applications in many different areas, not only physical environment but also socioeconomics. Organized into three volumes, GIS theories and techniques, GIS applications for environment and resources, and GIS applications for socioeconomics and humanity, the book comprises 79 chapters, providing a comprehensive coverage of various aspects of GIS. In particular, a rich set of applications in socioeconomics and humanity are presented in the book. Authored and peer-reviewed by recognized scholars in the area of GIS, each chapter provides an overview of the topic, methods used, and case studies.

The first volume of the book covers a wide spectrum of topics related to GIS methods and techniques, ranging from data management and analysis to various new types of GIS, e.g., virtual GIS and mobile GIS. While the fundamental topics in GIS such as data management, data analysis, and data quality are included, the latest developments in space–time GIS, cyber GIS, virtual GIS, mobile GIS, and public GIS are also covered. Remarkably, new perspectives on GIS and geocomputation are also provided. The further development of GIS is driven by the demand on applications, and various new data may be required. Big data has emerged to provide an opportunity to fuel the GIS development. Mike Goodchild provides an overview of such data, which is followed by voluntary geographic information, an important part of big geodata. Closely related to big data, open data is, however, accessible public data; they are not the same. Spatial analysis is indispensable for a GIS. After an overview of spatial analysis methods, big data analytics, spatial metrics, spatial optimization, and other relevant topics are included. Space and time are interrelated information, and their integration has long been an active research area in GIS. This section covers space–time data mining, space–time GIS, and time geography. Drawing on the developments in computer science and engineering, GIS has evolved to become more powerful through the integration with virtual reality and wireless technologies. Clearly, this volume provides new insights into different designs of GIS catering to the widespread needs of applications. This volume of the book will be of great interest not just to GIS researchers, but also to computer scientists and engineers.

Environment and resources are fundamental to human society. The second volume of the book focuses on GIS applications in these areas. GIS has been widely used in the areas related to natural environments; hence various such applications using GIS, such as vegetation, soil, hydrology, geomorphology, wetland, glaciers and glacial landforms, and paleolimnology, are covered. Resources and energy are closely related to the environment and so applications in these aspects are also covered. Climate change represents a challenge to human sustainable development. One reason for this is that climate change is increasing the odds of more extreme weather events taking place. It is apparent that GIS has been capitalized on to address the related issues, starting from climatology and meteorology to disaster management and vulnerability analysis. Parallel to applications

for natural environment, resources, energy, and climate, GIS has also applied to human production activities, such as agriculture and aquaculture, which has also been covered in this volume. In addition to natural environment, built environment and its associated topics such as place-making, public transit, and land use modeling and planning are also included.

Parallel to the second volume, the third volume of the book covers the applications of GIS in socioeconomics and humanities. Comparatively such applications are not as many as those in environment and resources. However, due to the increasing availability of data that can capture human activities, more applications have emerged in the areas, including economics, business management, history, linguistics, politics, law, human behavior, and policy making. Starting from Dan Griffith's overview of GIS and spatial statistics/econometrics, GIS applications in real estate, local economic development, and carbon mitigation are then covered. Innovation drives economic growth in today's knowledge-based economy; their relationship is covered in both the economics section and business management section. In addition to economics, GIS has also been widely applied to humanities. Such GIS applications as in history, linguistics, politics, and law are included. Human behavior has been given renewed emphasis due to the advent of social media and other types of big data. The first chapter in this section provides an overview of urban dynamics and geographic information science; several chapters are devoted to this topic. Finding evidence to support socioeconomic policy making is a highly important contribution that GIS can make. This volume also covers several chapters to find evidences for policy making.

This book could have not been completed without the help and advice of many people. In this regard we would like to thank a number of people who were instrumental in bringing this project to fruition. First, I would like to acknowledge the enthusiastic support of an outstanding editorial team including Thomas Cova and Ming-Hsiang Tsou (Volume 1), Yan Song, Georg Bareth and Chunqiao Song (Volume 2), and Kai Cao and Elisabete Silva (Volume 3). From the initial discussions of the structure of the book, the selection of authors for chapters in different volumes, to the encouragement of authors and review of chapters, they have made significant contributions at each stage of the book. I am very grateful for their invaluable input and hardwork.

I would also like to express my sincere gratitude to the production team at Elsevier, Priscilla, Paula, Katie, and in particular Laura, for their many efforts, perseverance, and skillful management of every aspect of this project. Last and certainly not least, I am hugely indebted to all of our authors. We have been extraordinarily fortunate in attracting individuals from all over the world to take time from their busy schedules to prepare this set of contributions.

Finally, my special thanks go to my wife Rongrong and our daughter Kate for their love, help, and understanding. Without their endless support, this book would have never come to the end.

Bo Huang, Editor in Chief

PERMISSION ACKNOWLEDGMENTS

The following material is reproduced with kind permission of Taylor & Francis

Figure 6 Spatial Analysis Methods
Figure 7 Spatial Analysis Methods
Figure 12 Spatial Analysis Methods
Figure 2 GIS for Linguistic Research
Figure 3 GIS for Linguistic Research
Figure 4 GIS for Linguistic Research
Figure 5 GIS for Linguistic Research
Figure 6 GIS for Linguistic Research
Table 2 GIS for Linguistic Research
Table 3 GIS for Linguistic Research
Figure 1 GIS and Scenario Analysis: Tools for Better Urban Planning
Figure 4 GIS Applications in Geomorphology
Figure 8 GIS for Glaciers and Glacial Landforms
Figure 18 GIS for Glaciers and Glacial Landforms
Table 1 Spatial Metrics - The Static and Dynamic Perspectives
Table 2 Spatial Metrics - The Static and Dynamic Perspectives
Figure 6 Urban Dynamics and GIScience
Figure 7 Urban Dynamics and GIScience
Figure 8 Urban Dynamics and GIScience
Table 2 Using GIS to Understand Schools and Neighborhoods
www.taylorandfrancisgroup.com

1.01 The Future Development of GISystems, GIScience, and GIServices

Ming-Hsiang Tsou, San Diego State University, San Diego, CA, United States

1.01.1 Introduction

In the last decade, innovative computing technologies and new software applications have transformed GIS from a centralized, function-oriented Geographic Information Systems (*GISystems*) into distributed, user-centered Geospatial Information Services (*GIServices*). Many new web services, open data, big data, geospatial cyberinfrastructure, mobile apps, and web map application programming interfaces (APIs) have become essential components within the GIS ecosystem. The fundamental knowledge of Geographic Information Science (*GIScience*) is also changing dramatically. GIS databases have shifted from relational databases to NoSQL databases. Data collection methods have been changed from paper-based digitization procedures to GPS tracking, to volunteered geographic information (VGI) and crowdsourcing. GIS software is transforming from desktop standalone programs to mobile app design, to Cloud-based web services. This article introduces some prominent future development directions from three aspects of GIS: GISystems, GIServices, and GIScience. Before we can describe these future technological advances of GIS in detail, it is important to provide a clear definition of the three aspects of GIS and their associated contents as follows:

- *GIS* is the abbreviation for geographic information systems or geospatial information services or geographic information science. It is a multifaceted research and technology domain and a generalized concept for describing geospatial technologies, applications, and knowledge.
- *Geographic Information Systems (GISystems)* focus on the development of computing software/hardware for conducting mapping and spatial analysis functions. Run-time performance, system architecture, information process flow, geocoding, user interface design, and database management are several key issues for the development of GISystems.
- *Geospatial Information Services (GIServices)* represent the service perspective of GIS, that is, delivering geospatial information, mapping services, and spatial analysis tools to end users over the Internet or mobile devices. Usability and User Experience (UX) are essential components for evaluating the effectiveness of GIServices.
- *Geographic Information Science (GIScience)* is "*the development and use of theories, methods, technology, and data for understanding geographic processes, relationships, and patterns*" (Mark, 2003; UCGIS, 2016 [2002], p. 1). GIScience is question-driven and follows scientific methods (questions, hypothesis, testing, analysis, and falsification).
- *Geospatial cyberinfrastructure* is the combination of distributed high-performance geospatial computing resources, comprehensive geospatial data coverages, wireless mobile networks, real-time geotagged information, geoprocessing web services, and geographic knowledge. The goal of geospatial cyberinfrastructure is to facilitate the advancement of GIScience research, geospatial information services, and GIS applications (modified from Zhang and Tsou, 2009).

The main driven force of future GIS development will be the advancement of geospatial cyberinfrastructure, which can enable fast and robust GISystems, provide smart and intelligent GIServices, and transform GIScience from a specialized scientific discipline into an important research domain bridging data science, computer science, and geography together. The following section provides some prominent predictions about the future development of GISystems, GIServices, and GIScience.

1.01.2 The Future Development of GISystems

There are four unstoppable trends in the future development of GISystems: (1) Web-based and Cloud-based GIS; (2) personalized data collection methods via mobile apps, drones, digital cameras, and portable LIDAR devices; (3) high-performance computing (HPC) and dynamic data storage services; and (4) lightweight and responsive mapping APIs with lightweight geodata exchange formats.

In the future, traditional desktop GIS software (such as ArcGIS, ArcGIS Pro, QGIS, gvSIG, uDIG, and MapInfo) probably will be used only by a small group of GIS professionals (20%) who need to handle sensitive or protected geospatial data within local and secured workstations. Most GIS users (80%) will utilize Web GIS and Cloud computing frameworks, such as ArcGIS online, Google Maps, Google Earth, MapBox, and CartoDB toolboxes, to conduct GIS tasks and spatial analysis functions.

Mobile GIS apps will be the main personalized data collection tool for future GIS applications. Several popular tools, such as ESRI Survey 123, ESRI Collector, and GIS Cloud, can enable GIS data collection via mobile phones and combine photos, videos, surveys, and GPS coordinates into online databases directly. Collected GIS data via mobile devices will be uploaded or synced via wireless communication to Cloud-based databases or storage services. Other personalized geospatial data collection devices, such as Unmanned Aircraft Systems (UAS) or Drones, digital cameras, portable 3D LIDAR scanning systems, and mapping vehicles (such as Google Street View Cars), will be integrated into Web GIS or Cloud GIS platforms seamlessly to provide high-resolution aerial photos, street views, or digital elevation models for various GIS applications.

Many GIS operations are computational intensive and require huge sizes of memories or data storage spaces. The recent development of big data and HPC framework, such as Hadoop, Apache Spark, and MapReduce, can be applied in future GIS data models and databases. These big data computing framework will enhance the performance of GIS operations significantly. However, the main challenge will be how to convert GIS data and spatial analysis operations suitable for parallel operations and how to set up the cluster-computing frameworks for various GIS applications. Another promising direction is to utilize graphics processing units (GPU) for the intensive 3D or animation display of GIS applications.

The future development of GIS software programs will also become more light-weighted and customizable for different applications. Some new web mapping service APIs and libraries, such as Leaflet, MapBox, CartoDB, and ArcGIS online, can provide dynamic mapping or spatial query functions for lightweight web apps or mobile apps (Tsou et al., 2015). GIS is no longer a large standalone system equipped with hundreds of functions inside a box but rather a customizable service framework, which can provide fast and simple GIS functions and services to end users (Tsou, 2011). Along with the development of lightweight mapping functions (such as Leaflet), lightweight data exchange formats (such as GeoJSON) will become very popular in Web GIS applications. GeoJSON is a JSON (JavaScript Object Notation)-based geospatial data-interchange format for web apps or mobile apps. It is a text-based data format utilizing JSON, decimal coordinate systems, and a predefined projection framework. Software developers can easily develop dynamic and responsive Web GIS by using lightweight mapping APIs and GeoJSON.

In summary, GISystems have evolved from the mainframe computers (in 1970s and 1980s) to desktop GIS (in 1990s), to Web GIS (in 2000s), and to mobile apps (in 2010s). The performance and functionality of GISystems have been improved significantly to meet the needs from various GIS users and applications. In the near future, every single GISystem can be linked and integrated together into a global geospatial cyberinfrastructure (with hundreds of thousands of GIS nodes across the whole world) (Tsou and Buttenfield, 2002). These dynamic GIS nodes can provide personalized and customizable GIServices for various users. The next section provides some good examples of future GIServices.

1.01.3 The Future Development of GIServices

GIServices are essential in our daily life. This section focuses on four types of important GIServices and discusses their future development: navigation services, web mapping services, spatial query and analysis services, and location-based services (LBS).

Navigation services are probably the most popular and heavily used GIServices in both mobile apps and web apps today. Popular navigation service platforms include Google Maps, Apple maps, HERE, Navigator for ArcGIS, MapQuest, and Bing maps. Uber app is another good example of navigation services for both drivers and passengers. Navigation services required comprehensive base road maps with detailed points of interests (POIs), real-time road condition, and traffic updates. One major application of navigation services in the future will be the development of self-driving cars (autonomous car). Self-driving cars will require a seamless integration between the navigation services and the sensor data (cameras, LIDAR, etc.) collected in real time on each vehicle. The future development of navigation services will need to integrate with all traffic cameras, weather stations, and the sensors collected from nearby vehicles. Hundreds of nearby autonomous cars will create a "mesh network" dynamically, and each nearby self-driving car can provide and relay traffic data via wireless communication to each other. The mesh network can provide real-time traffic and road condition updates automatically. All nearby autonomous cars can cooperate in the distribution of traffic data and navigation services together.

Recently, web mapping services have been applied in various mobile apps and GIS applications. For example, Pokémon GO utilized popular Google Mapping services to create a virtual world for users to catch monsters, eggs, and treasures. Zilliow and Foursquare used Google Mapping services to provide locational information and maps for their customers. Several prominent web mapping service developers, such as MapBox and CartoDB, have developed interactive, responsive, and fast mapping services to different GIS applications. One challenge of web mapping services is to provide effective map display on multiple devices using the same map contents. For example, users will need to display a campus map on his/her smart watches (320×320), mobile phones (750×1334), high-resolution computer screen (3840×1600), and smart 8K UHD TV (7680×4320) simultaneously. Advanced map generalization and intelligent cartographic mapping principles will be developed to transform web maps into responsive display for fitting different devices and screen resolutions. Web mapping services will provide both 2D and 3D display functions for next generation of web map applications for virtual reality (VR) and augmented reality (AR) applications.

In terms of spatial query and analysis services, one future application for utilizing these services will be the development of Smart Cities and Smart Transportation Systems (Smart Traffic Controls). For example, a visitor will be able to use his/her smart phone to query the best walking/running route nearby the hotel and to avoid unsafe areas and heavy traffic zones in real time. Car drivers will get advice and warning about possible traffic jams nearby and provide alternative routing options (which is already available in Google Maps now). One major future application of spatial analysis services could come from a virtual personal assistant in

a mobile phone, who can provide recommended shopping, eating, driving and parking, movie watching, dating, and exercising choices nearby the locations of users. Some advanced spatial analysis functions, such as clustered dots and hot spot analysis, can be applied to the crowd management for music concerts, conference meetings, and popular events.

LBS focus on the collection of consumer information based on the location of users and the nearby environment. LBS can include or combine with navigation services, web mapping services, and spatial analysis services. However, LBS will only focus on the nearby information or POI, rather than providing information far away from the users. Currently, the outdoor locations of users can be defined by GPS signals, Wi-Fi signatures, and cellular tower signal triangulation. One current technological challenge of LBS is how to provide a better and accurate indoor positioning system (IPS). Several possible technological frameworks of IPS include Wi-Fi access point signal triangulation, magnetic positioning, iBeacon, RFID tags, etc. However, most of IPSs require the setup of indoor environment labels in advance or the 3D scanning of each room before the positioning process. Some potential LBS applications for IPS are hospital patient room arrangement, conference exhibit halls, and popular event promotions.

1.01.4 The Future Development of GIScience

The knowledge domain of GIScience will change dramatically in the next decade driven by new types of GIServices and new design of GISystems. Four research topics in GIScience are highlighted in this section as representative trends: machine learning methods, crowdsourcing data, new data models for big data, and human dynamics.

Machine learning methods are derived from the development of artificial intelligence (AI) and statistic models. The GIScience community has developed a few applications utilizing AI and expert systems before (Openshaw and Openshaw, 1997). However, due to the lack of programming skills and suitable HPC frameworks, very few GIS researchers have developed fully functional AI or expert systems for GIS applications. Some cartographers have developed very limited expert systems for providing intelligent mapping, text labeling, and symbolization functions before. The recent development of geospatial cyberinfrastructure and easy-to-learn programming languages, such as Python and R, has enabled GIScientists to utilize powerful *machine learning methods* to develop intelligent web mapping and spatial analysis functions. Several machine learning methods (such as K-means, logistic regression, decision tree, deep learning, principal component analysis (PCA), support vector machine (SVM), and Naïve Bayes) can be applied in GIS data classification, map symbolization, spatial analysis, spatial pattern detection, and geovisualization. For example, dasymetric mapping methods can be improved by using SVM or Naïve Bayes to estimate the population density based on different types of land use and land cover. Geographic weighted regression (GWR) models can adopt PCA to provide a better explanation of multi-variables' contribution to the targeted data layer.

Crowdsourcing and citizen science have become major data input methods in GIScience. VGI is one popular type of crowdsourced data. Some VGI applications include OpenStreetMap, Waze, and iNaturalist. Other crowdsourced data input methods include geotagged social media data, wearable sensor data for mHealth, or GPS tracking data from bikes or taxi, which are not VGI. One major challenge of crowdsourced data is how to assess the credibility and accuracy of collected data. Since there are many errors in crowdsourced data, it is extremely important to develop effective data filtering, data cleaning, and data validation procedures for crowdsourced data. Sampling problems and user biases are other major concerns in crowdsourced data. For example, social media users (such as Twitter and Instagram) are mostly under age 35 and live in urban areas. Most volunteers working in OpenStreetMap are white male persons with full-time jobs.

Traditional GIS data models include vector-based object data model and raster-based field data model. However, very few geodatabases can provide effective space–time relationship for advanced spatiotemporal geography analysis. Along with new types of big data collections (such as social media and crowdsourced data), many traditional GIS models are no longer suitable for big geodata. NoSQL databases (as MongoDB) and new space-time data models will become more popular in the future, and researchers can utilize new data models to build more effective and customizable geospatial data analytics.

Another emerging research topic in GIScience is human dynamics, which can be defined as a transdisciplinary research field focusing on the understanding of dynamic patterns, relationships, narratives, changes, and transitions of human activities, behaviors, and communications. Many scientific research projects (in the fields of public health, GIScience, civil engineering, and computer science) are trying to study human dynamics and human behaviors. One main goal of these projects is to develop effective intervention methods to modify or change human behaviors and to resolve public health problems (such as obesity, disease outbreaks, and smoking behaviors) or transportation problem (traffic jams and vehicle incidents). Several innovative data collection methods can be applied to study human dynamics. For example, researchers can use computer vision algorithms to analyze Google Street Views and to estimate the built environment index and neighborhood social status. Combined CCTVs in urban areas and street traffic cameras can be used to analyze the usage of bike lanes and biking behaviors in different communities/neighborhoods. The frequency of geotagged social media check-ins can be used to estimate dynamic changes of population density for supporting disaster evacuation decision support systems.

1.01.5 The Future Societal Impacts of GIS Development

This article highlighted several prominent applications and topics in the future development of GISystems, GIServices, and GIScience. Many GIS researchers may think that the advancement of future GIS applications can provide better information services

for the general public and improve quality of life for everyone. However, the spatial disparity of geospatial technology and the potential digital discrimination between rural and urban areas could trigger serious social problems and social unrest in the future. Since the development of geospatial cyberinfrastructure is expensive and unequal, there are huge gaps of cyberinfrastructure between rural and urban areas, between developed and developing countries, and between the rich and the poor. For example, major cities in the United States have the most updated high-resolution aerial photos compared to some African regions, which only have low-resolution satellite images 10 years ago. Google Street Views are updated frequently in New York and San Francisco, but many small US cities have no Google Street Views at all. The spatial disparity of geospatial infrastructure can trigger "digital discrimination" to the people who live in low-income and rural areas. Along with the development of future GIServices, such as self-driving cars and smart transportation systems, people who live in rural and low-income areas will not be able to access these advanced GIServices. The advancement of geospatial technology will exaggerate the digital discrimination and the digital divide between urban and rural areas. The rich get richer and the poor get poorer.

To solve these potential social unrest and social problems, local and federal governments need to make significant investment of geospatial cyberinfrastructure in rural and low-income areas to reduce the disparities of GIServices across different regions. Hopefully, everyone can enjoy the progress of GISystems and GIServices without worrying potential social unrest in the future.

References

Mark, D.M., 2003. Geographic information science: Defining the field. Foundations of Geographic Information Science 1, 3–18.

Openshaw, S., Openshaw, C., 1997. Artificial intelligence in geography. Wiley, Chichester.

Tsou, M.H., 2011. Revisiting web cartography in the United States: The rise of user-centered design. Cartography and Geographic Information Science 38 (3), 249–256.

Tsou, M.H., Buttenfield, B.P., 2002. A dynamic architecture for distributing geographic information services. Transactions in GIS 6 (4), 355–381.

Tsou, M.H., Jung, C.T., Allen, C., Yang, J.A., Gawron, J.M., Spitzberg, B.H., Han, S., 2015. Social media analytics and research test-bed (SMART dashboard). In: Proceedings of the 2015 International Conference on Social Media & SocietyACM, New York, p. 2.

University Consortium for Geographic Information Science (UCGIS) (2016 [2002]). UCGIS bylaws (revised in 2016). http://www.ucgis.org/assets/docs/ucgis_bylaws_march2016.pdf (accessed 6 March 2017).

Zhang, T., Tsou, M.H., 2009. Developing a grid-enabled spatial Web portal for Internet GIServices and geospatial cyberinfrastructure. International Journal of Geographical Information Science 23 (5), 605–630.

1.02 Geocomputation: Data, Methods, and Applications in a New Era

Shaun Fontanella and Ningchuan Xiao, Ohio State University, Columbus, OH, United States

1.02.1 Introduction

Geographers embraced computing technology in their research and applications in the very early years of computers (Chrisman, 2006). Thoughts of using computational methods to solve geographic problems started to emerge in the 1960s (Chorley and Haggett, 1967). Such a trend gave rise to geographic information systems (GIS), which quickly became the dominating terminology in almost every field that involves spatial data. However, many quantitative geographers appeared to refuse to equate GIS with the computational needs and applications in geography. As a consequence, 1996 saw the first GeoComputation conference held by the Department of Geography at the University of Leeds. The organizers of this conference described geocomputation as a new paradigm and declared the "dawning of a new era of geocomputation" (Openshaw and Abrahart, 1996). Themes of the first GeoComputation conference include high-performance computing, artificial intelligence, and GIS. But the extent and scope of the conference sequence, along with the research field, have quickly changed to embrace broader technological advances and application domains (Gahegan, 1999).

Recent years saw a clear trend in computing technology that has started to shape the field of geocomputation. Many of the traditional geocomputational tasks such as mapping and spatial analysis have moved off the desktop and onto Web-based GIS platforms. Entire computational workflows, such as data collection, analysis, and mapping, can be done on hardware agnostic webpages. In addition to these new geocomputational capabilities, data-driven geography (Miller and Goodchild, 2015) has replaced the data paucity of early geocomputation and GIS. To be sure, traditional desktop platforms will still have their place, but for those with limited resources and technical skills, the Web offers a powerful geocomputational platform at costs significantly less than in the past.

In this new era of Web GIS, location-aware mobile devices, inexpensive cloud computing, and widespread broadband connectivity tend to dominate the conversation. However, in the background, the Web connects all of these technologies together. Without the Web, many of these technologies would be stranded on islands of spatial data computing as they were in the previous era; unable to communicate because of incompatible programs and protocols. The modern Web has connected the many islands of computing together and become the home of much of the data and processing power.

This article will discuss significant steps in the progress of computing technology and then present a case study that embodies many of the modern aspects of geocomputation. In the next section, we discuss computation in GIS by providing a narrative of data gathering and management techniques. In section "**Computing on the World Wide Web**", cloud infrastructure will be explained. We reinforce the idea that the cloud is what makes many of the new computing capabilities possible as it provides the facilities to build and host the Web. Cloud infrastructure has removed much of the friction from the process of deploying GIS applications. It

has lessened the barriers of entry to the Web and allowed individuals and small companies access to what was once prohibitively expensive technology. In section "**The Move to the Cloud**", we continue the discussion of cloud computing with a focus placed on the enabling techniques. In section "**Computational Methods**", we overview some of the computational methods that can be used to provide behind-the-scene analysis of spatial data. In the final part of this article we will present a case study that employs many of the themes described above. The case study application lives in the cloud where it actively collects data for updated analysis and visualization in the cloud.

It must be pointed out that in this article we often use the terms geocomputation and GIS in an almost interchangeable fashion when discussing computational issues in handling spatial data. We recognize that GIS is a much broader term that involves issues beyond computation. But we also recognize that it is difficult to clearly delineate the strictly computational aspects from what we would normally call GIS, even though the term geocomputation did not become part of the literature until the late 1990s. In the text that follows, when the term GIS is used, we intend to focus on the computation aspect of GIS.

1.02.2 Early Stage of Geocomputation and GIS

For those new to geocomputation and GIS, it is easy to think that the current environment, where computing resources and spatial data are abundant, has always been the norm. This is not so. The explosion of available data and geocomputation resources is a phenomenon as recent as the last decade. There were two previous eras of geocomputation and GIS that were much different. It is important to have some historical perspective to give context to the current environment. As this collection has an interdisciplinary scope, a review of the geocomputation landscape will be useful to the reader coming to this work from other disciplines.

We roughly identify three eras of developments in geocomputation and GIS. In the first era, computing resources and spatial data were scarce. In the very beginning of the discipline, computation on spatial data was done on large and highly expensive mainframe computers. It was mostly governments and universities who could afford these mainframe computers and so they were some of the few who could produce and analyze digital spatial data. Indeed, the term GIS was coined in the 1960s by Roger Tomlinson's group who were working on the Canada Geographic Information System (Tomlinson, 1998). The Canada GIS was a large government program that used expensive computer equipment to solve a spatial problem that could not be feasibly completed with manual analysis by humans using paper maps. It was only at that scale that computing on spatial data made economic sense. In this early era, computers were as large as whole rooms. Programs and data were stored and entered into a computer through punch cards and magnetic tape. Most output was to a printer, not to a screen, which was an expensive and optional add-on. The fundamental ideas of geocomputation also saw their roots in this early era but resources were so scarce that little progress could be made.

Much of the early data were digitized from existing paper maps. These maps could be stitched together to form larger data sets. Areal imagery taken with film cameras was also a source of spatial data (Campbell and Wynne, 2011). Photos could be digitized into files for analysis using scanning equipment to create digital copies (Faust, 1998). Areal imagery produced from photographs changed to remote sensing when electronic sensors began to take images without film and outside the range of the visible spectrum. Just like spatial data analysis, data collection required expensive equipment and labor. Availability of spatial data and computing resources was very limited for most researchers in this first era.

In the second era of geocomputation and GIS, Moore's Law (Moore, 1965) eventually made computing cheap enough to enable GIS on desktop computers. New commercial as well as free software became available to do mapping and analysis (Hart and Dolbear, 2013; Goran, 1998). This move from the mainframe to the desktop was a significant shift in GIS. The drop in price was significant enough that both public and private sectors started building capacities to collect and process geographic data sets. However, during this second period, many of these computers were still islands of computing. If they were attached to networks, those networks often used LAN (local area network) protocols like IPX/SPX to share resources like file servers and printers on small isolated networks. These protocols could not communicate with the larger Internet. If data were to be shared, they often had to be put on physical media and transferred.

The data island effect could (and still does) happen for reasons other than technical. GIS interoperability can also be restricted for of a number of institutional reasons. For instance, in the United States there has been little coordination or standardization among mapping agencies. Maps created at the county level often do not align their data or share standards with adjacent counties. This happens again at the state level. Even if the same software and file formats are used, researchers trying to create regional maps have been faced with collecting data from multiple entities. Despite powerful computers and fast networks, computing spatial data can still be impeded by institutional barriers.

GPS (global positioning system) started to be used in this second period of geocomputation and GIS development. GPS has become an important tool as it makes spatial data easier and cheaper to collect. GPS accelerated the collection of spatial data but was limited in use in early applications. GPS was originally restricted to use by the military. When GPS was made available to the public the signal was intentionally degraded to decrease accuracy to 50 m. It wasn't until the year 2000 that the signal degradation was turned off and GPS receivers could get good accuracy without assistance from ground stations. GPS greatly increased the speed of accurate data collection. However, in the early period of GPS use, it could still be an expensive, technical task to take a unit into the field, collect data, and upload those data to specialized and often expensive mapping software in order to do analysis and mapping. It wasn't until later when GPS points could be automatically uploaded and mapped through Web connectivity that GPS truly exploded as a collection device.

The Internet slowly started to extend into the islands of computing, first by phone lines and later using broadband. The Internet's communication protocol, TCP/IP, weeded out competing standards and eventually most computers were connected the Internet. At this point though, data were still mostly static files transferred through traditional Internet protocols like file transfer protocol (FTP) for larger files and email for smaller files. More complex data containers like spatial databases existed but were often secured behind firewalls, thus limiting their usefulness.

During these early periods, computation of spatial data was conducted using a variety of static data files. Some of the most common early versions of static data, such as ASCII text files, would be recognized today. These were efficient means of sharing tabular data that associated attributes with geographic entities. They also worked well for describing where points were located. However, as more complex features and geographies emerged with better collection methods, ASCII files started to show their limitations. For discrete vector data, combining binary and text formats was a more efficient way of storing data. Formats like the shapefile (ESRI, 1998) were able to store complex geometries and attributes in compressed binary formats while still linking to text data for attribute data.

Static data files have served the purpose of spatial data computation well but they have many limitations. These limitations have become more obvious as networks have tied together millions of computers. One of the problems with static files is that the entire file has to be downloaded and parsed even if just one row of information is needed. Static files may be very large. Some data from the US Census are multiple gigabytes in size. Getting these files may not be a problem on a fast network with a powerful computer. However, on a mobile network or with a resource-limited device, it may take a long time to download and parse the data.

The growing size of static files poses problems for many commercial off-the-shelf programs. Most productivity applications like Microsoft Excel are not designed for large data sets. They have built-in limits to the amount of data they can handle. These limitations are traded for performance and simplicity. Large static files may require special programs or programming scripts for computation or for extracting more manageable data sets that can be manipulated with productivity software.

Another limitation of static files is that they have problems holding multiple types of data efficiently. They are usually set up with a particular type of data intended for them. Researchers can now collect much more data from a diverse landscape of sources. One of the increasingly important trends is to gather data from individual silos and combine them for new analysis. Static files are poorly suited to complex queries across multiple tables of data. These types of research are better suited to databases with the ability to index data in multiple dimensions.

Static files still have a place for publishing many data types and they will still be around for a long time, but they are increasingly sidelined by databases and Web services. Particularly in Web mapping applications, sending all possible data to a client and having the client computer sift through them is a problematic model on bandwidth-limited mobile devices. Spatial databases are becoming more popular because they now have the utility and security of Web technologies which we will discuss below.

1.02.3 Computing on the World Wide Web

The Web has been one of the greatest influences on geocomputation and GIS, and its effects can be observed in multiple ways. To begin with, it has become a platform for collecting, analyzing, and visualizing GIS data. Online mapping platforms from Carto, Mapbox, Tableau, and Esri allow users to present spatial data on platform agnostic Web interfaces from any computer while the computational tasks are performed on the server side, behind the scene. In addition, the Web has standardized communication protocols and formats so that data can flow easily between systems. This important bridging function will be explained in greater detail below.

It is important to make the distinction between the Web and the Internet as to most people the Web and the Internet are the same. Speaking simplistically, the Web is an application that runs through the communication pipe provided by the Internet. It may be a surprise to some but the first four nodes of the ARPANET, the precursor to the Internet, were connected in the summer of 1969 (Lukasik, 2011; Fidler and Currie, 2015). The Web was not developed by Tim Berners-Lee until the early 1990s. By this time email, file transfer, newsgroups, and other networked programs were well established on the Internet. At that time too, there were other competing systems for sharing documents like Gopher and wide area information server (WAIS). Berners-Lee credits the flexibility, open standards, and decentralized nature of the Web for making it the eventual winner for document sharing (Berners-Lee et al., 2000). These attributes also made it well suited for spatial data.

Tim Berners-Lee's description of one of the first real successful uses of the Web is a great example of the transition in data access created by the Web. This transition is echoed with data in the GIS world. At CERN, there was frequent staff turnover because research projects were often temporary. Printed phone books could not keep up with the constant change and were inevitably inaccurate. An electronic up-to-date version of the phone book could be accessed through the mainframe computer. However, accessing the mainframe required credentials. Even when logged in, sessions would time out to free up the limited number of licenses. Researchers had to log in every time they wanted to check a phone number. A webpage that accessed the phone number database was created that would allow read-only access to the mainframe from any computer on the network running the Web software (Berners-Lee et al., 2000). Requests were instantaneous and stateless so they didn't tie up connections. They also didn't require authentication as the phone book was public data. Many more hosts could be served with the same resources. This was a great example of the Web providing access to public information that was limited through authentication and commercial licensing.

The three most crucial components that make the Web work are a set of communication protocols and file standards (Fu and Sun, 2010) that govern how data are transmitted. The first component is the hypertext transfer protocol (HTTP). HTTP transfers data

between client and server using an IP address and a port at that address. Ports differentiate applications on the same host. If a server is thought of as a warehouse, the IP address is the street address and the port is the dock door. Applications have officially designated ports maintained by the Internet Assigned Numbers Authority so that applications know the default port to communicate on. For instance, email servers use port 25 to send mail, and PostgreSQL servers use port 5432. The game DOOM is assigned port 666. These ports are open on local networks where access to a network can be physically controlled but it is unwise to open many application ports to the wider Internet for security reasons. These ports that serve out data are blocked by firewalls where the local network interfaces with the Internet. This is where the Web serves a valuable function.

One of the important capabilities of HTTP is the bridging function it performs for many applications. Web servers run on port 80 for HTTP and 443 for encrypted traffic on HTTPS. Since this port is already open for webpages, Web servers can act as proxies for resources secured behind a firewall. This is commonly the case for GIS servers. They take requests on port 80, contact a database such as a PostgreSQL server on port 5432 to request data, and return the results on port 80 to the client. Web servers allow protected resources on a network to be more securely shared with the Internet. This is one of the most important functions they serve.

The second component of the Web, hypertext markup language (HTML), is the base file format for sending webpages. HTML contains content such as the stock quotes and recipe ingredients that are on a webpage. The presentation of the content can be enhanced with cascading style sheets (CSS), Javascript, and other browser extensions like Flash or Silverlight, but the base container is HTML. HTML is an evolving standard. It is expanding to embrace capabilities that were once provided by add-ons like Flash or Silverlight. As HTML evolves, it continues to bring more sophisticated capabilities to the platform agnostic Web.

The final important part of the Web is the address system that allows resources to be accessed on the Web. Uniform resource locators (URLs) are the Web addresses that describe where a resource is located. The first part of a URL is the hostname of the computer that has the desired data. This hostname part of the URL builds upon the already established domain name system that allows Internet-connected computers find each other. The second part of the URL describes the resource requested from the server. This part can be a file name, a search query, or data being returned to the server. This last part of the URL is one of the things that makes the Web so powerful. As long as the server getting the request knows how to parse the URL, it can contain a wide array of text. This gives developers great latitude when programming applications for Web servers.

The Web has come a long way since its inception. Beautiful and well-designed webpages have replaced earlier, much cruder ones. But the facade of sophisticated webpages hides the even more important changes to the way webpages communicate data. Modern webpages rely on data being passed between client and server silently in the background. This communication often happens automatically without any interaction from the user. Predictive typing that guesses what a user is searching for on Google and prepopulates the search box is an example of data communication that is transparent to the webpage user. In the case of maps, the page needs data to update the map when a user interacts with it by panning, zooming, or switching layers on and off. The collection of protocols and technologies that communicate in the background to update the map seamlessly are part of what is known as Web 2.0 Goodchild (2007). Web 2.0 includes many themes of two-way communication between actors instead of one way consumption from few servers to many consumers (O'reilly, 2007), but this work will limit itself to a handful of protocols that make Web 2.0 work for Web GIS.

These "Slippy" maps are a great example of Web 2.0 in the realm of geocomputation and GIS. Maps on the Web existed before Web 2.0 but they were much more difficult to use. Panning or zooming required reloading the entire webpage with the new map extent and zoom, and of course new ads. Web 2.0 makes maps much more usable than early versions. Web 2.0 transmits data in the background and only updates the region of the webpage that has the map on it. This function is provided mostly through protocols like asynchronous Javascript and XML (AJAX). These behind-the-scenes protocols have made data communication agnostic to the various operating systems, databases, and applications that share data all over the world. This standardization is helping provide access to many more data sets that were previously only available through tedious manual collection methods. It is one of the primary reasons that the problem of Big Data (Miller and Goodchild, 2015) exists. The variety and volume of data that can now be collected are exceeding our ability to process and store them.

All of these available data have pushed researchers to use more sophisticated storage methods than traditional static files. Just as computer hardware has decreased in price, software, much of it through open-source development, has also become cheaper while at the same time becoming more sophisticated. In geocomputation and GIS, this has been the case with spatial databases. Spatial databases are the best method to maximize the research potential of the growing tide of collected data. Spatial databases will be discussed in the next section.

1.02.3.1 Spatial Databases

Spatial databases address many of the limitations of static data files. Spatial databases can contain large amounts of data in multiple tables with linking mechanisms that maintain data integrity. They can enforce restrictions on data entry to limit collection of inconsistent data. As they grow, they can span multiple physical machines as well as maintain copies of themselves for redundancy. Spatial databases can also maintain changes to a set of spatial data and track which users are making edits for approval and auditing.

It is important to make the distinction between typical databases and spatial databases. Spatial databases are standard databases that have been extended to accept spatial data types and queries. Spatial data types store feature geometry that describes shape and location. The geometry of spatial features is compressed and stored in a binary field along with the attribute data that describe the feature. In addition, the database application code is extended so that typical queries using alphanumeric characters and logical operators are extended to take advantage of location, proximity, and topology. For instance, in a typical customer database

a company might query for all customers whose last name begins with a certain letter. In a spatial database, they can query for all customers within proximity of a particular store or find clusters of customers. These types of spatial transactions are not available in typical databases.

Spatial databases existed long before the Web but they were often hidden behind firewalls or authentication and were unavailable to most users. Because spatial databases are usually hosted on powerful servers and are always connected to the Internet, they may become targets for hackers. Even with user authentication and read-only access, malicious users may attempt to gain unauthorized access to the database using bugs in the database or operating system programming code. That was the case when SQL Slammer used a buffer overrun exploitation to take control of over 75,000 SQL servers and brought the Internet to a crawl (Microsoft, 2003).

Traditional software vendors like Microsoft, Oracle, and IBM all have database offerings that can be used for geocomputation and GIS applications. Database software is often very expensive to implement in terms of both the cost of the software and the labor to implement them. GIS software has benefited greatly from open-source software, and spatial databases are no exception. For researchers prototyping applications on little or no budget, open-source software offers several free and robust spatial database alternatives. The most popular open-sourced spatial database is PostgreSQL, extended with PostGIS.

An emerging trend in databases is a more flexible "noSQL" database. Also called "document databases," these databases store data in a format much like a JSON file. These databases don't have the schemas that enforce data integrity in typical databases. Schemaless databases are not as structurally stringent and have flexibility with the data they can hold. This flexibility allows the storage of dissimilar data but requires that queries be designed to deal with dissimilar data. The most popular schemaless open-source spatial database software is called MongoDB and has a growing set of spatial capabilities.

Spatial databases have many advantages over static files but they too come with disadvantages. Database servers require significant hardware resources to host large data sets. The servers have to fit into existing networks with security policies and firewalls. In most cases, users have to be authenticated which requires maintaining user accounts and permissions.

1.02.3.2 Spatial Data as Web Services

The place where the Web and GIS really connected is the joining of spatial data and Web services. This function is provided by GIS servers that take spatial data from static files, file geodatabases, and enterprise databases and serve them out using HTTP. These endpoints are called Web services. These services can serve geographic data as text that describes features, as rendered map tiles, or as raster coverages. In addition to read-only publishing, Web services can also allow editing of data from Web-connected devices.

These Web services have greatly expanded the amount of data available for geocomputation. In addition to providing data for maps, they also allow users to access data programmatically through code. Software like Python or R can pull data from any Web service and perform analysis on them. The Web has essentially connected the entire Internet to the command line.

Data transmitted by Web services may take several forms as spatial data can represent the same set of features in multiple ways. For instance, a feature representing air quality can be represented by geoJSON text with latitude and longitude values, or a set of jpg images stitched together to make a map. A browser can consume the spatial data in any of these formats and make a map that looks the same. The method of transfer is transparent to the user.

An example that demonstrates the various ways that data may be served for maps can be seen in a simple map of a city's sidewalks. Individual sections of sidewalks may exist in a spatial database and have attributes associated with them. These details can track a section's condition, ADA compliance, or inspection date. In a large city, there will be hundreds of thousands of these segments. When exposing this data through a Web service, the segments can be transmitted as text describing the hundreds of thousands of sidewalk segments, or they can be transmitted as images rendered from spatial data in the database and broken into map tiles showing the thousands of sidewalk segments as images. From a performance perspective, it is much quicker to send the tiles. However, you may lose the ability to get feature attributes. If the segments are sent by text, care must be taken to filter the number of features transmitted, or the map will become useless as it hangs trying to download too many records. There are multiple factors that must be considered to determine how spatial data are transmitted. A GIS must balance detail, function, and performance. Too much data and a GIS will be too slow to be usable. Too little data and a map is incomplete or has no context.

The format of spatial data provided by a service depends on the type of Web service it is. Three common Web services are Web mapping services (WMS), Web feature services (WFS), and Web coverage services (WCS). Most people are familiar with WMS. WMS transmit tiles rendered from spatial data to a client. If these tiles are prerendered, they can be cached on the server to cut down the time needed to render the images. It may take significant resources to render and store a cached tile service but there is a significant performance increase in services with many features. Google Maps and Open Street map are examples of a WMS. The slippy map that looks like a solid canvas is actually many tiles that are stitched together. Web feature services (WFS) return text data that describe features. Geometry and attributes are transmitted in GML, JSON, or other format. The browser renders the text data into features and draws a map. The attribute data is often used to symbolize each feature or create hover effects **Fig. 1**.

Web services do not have to implement a full GIS server to serve geographic data. Web services can serve spatial data in text form that have simple geographic information encoded like latitude and longitude. For instance, Web services are often used as a front end for scientific instruments such as air quality meters or weather stations. These instruments provide real-time spatial data through a Web service. Many manufactures of equipment have switched from proprietary software and interfaces to standards-based interfaces using Web services and formats like XML and JSON. These interfaces can then be accessed from any computer connected to the Internet using a browser or programmatically through languages like Python or R.

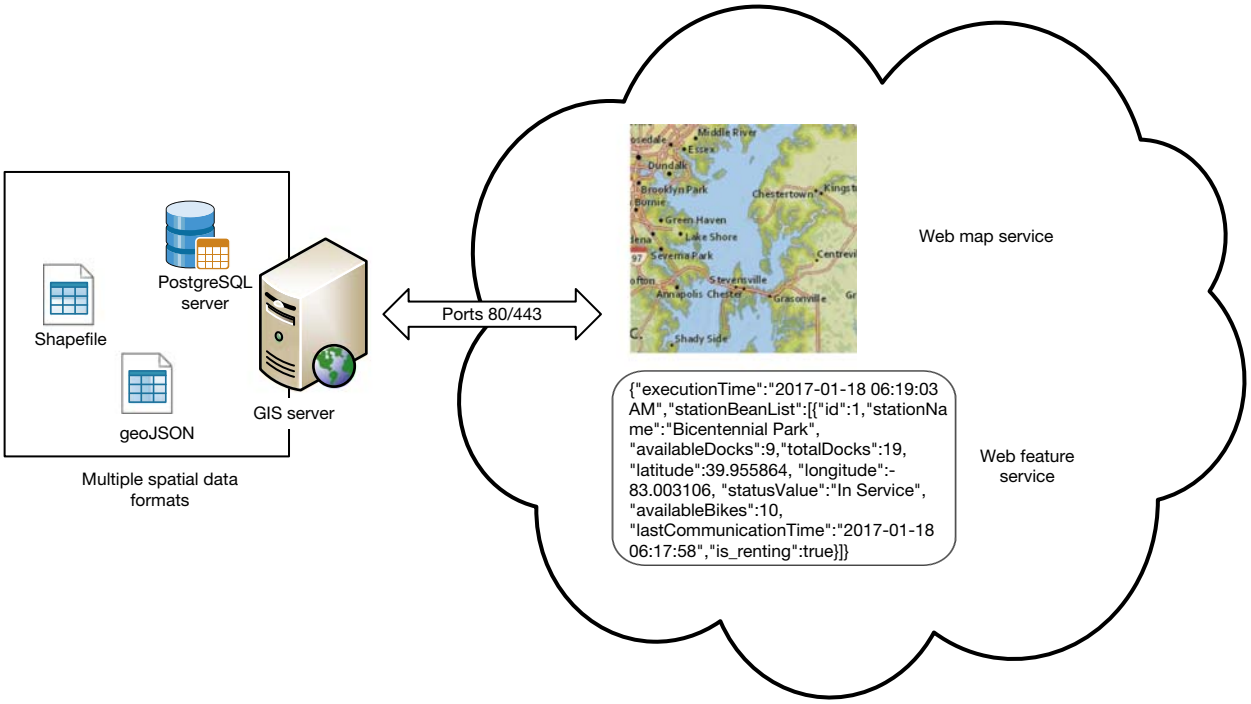

Fig. 1 Spatial data translated by a server into tiles and JSON.

In addition to simply serving data, some Web services can perform geoprocessing tasks (OGC, 2016). A geoprocessing Web service can take input from a webpage form or URL request, calculate results using spatial data, and return data to be visualized on a map. For instance, entering a value in a form or clicking on a map will calculate a drive time radius around a selected point. These geoprocessing tasks were once only available through desktop software. They can now be done over the Web in a browser.

Even if a website does not have a formal Web service, researchers can still collect data from regular webpages using a technique called scraping. Scraping uses software to go to a website and parse through the HTML code that makes that page look for particular data. For instance, data on gas prices can be scraped from crowd-sourced gas price sites like gasbuddy.com. These scraped data can be stored in a database for further analysis.

Data accessed through any of these sources may be immediately plotted on a map for display or captured and saved. Saved data can be analyzed for temporal patterns that may not be immediately apparent seen through simple display. It can also be correlated with other collected data. For example, air quality may be correlated with traffic data, precipitation, or electricity use by factories. Each of these data sources is published by different entities but combining them may reveal new patterns.

1.02.3.3 New Data Formats

The advances in data sharing made on the Web have led to new file formats for transferring data. Unlike many file types of the past, these formats are generally based on open standards. For instance, geography markup language (GML) is used to transfer data from Open Geospatial Consortium (OGC) Web feature services.

Another new file format is Keyhole Markup Language (KML). KML files are a form of extensible markup language (XML). KML can store spatial data as well as information used for visualization. KML files can store the location and orientation of a viewer of the spatial data. This viewer position is used in mapping software to define how the viewer sees the data. KML is often associated with Google as they bought the company that created the standard. KML is used in several Google products including their mapping API and Google Earth. However, KML has become an open standard and was accepted by the OGC for official standardization in 2008. (OGC, 2008)

XML-based files are useful for small collections of features. They can be opened and read with any text editor and file contents can be easily comprehended. A drawback of KML files is that they often have a significant amount of redundant data. In XML files, data are arranged in hierarchical trees that help describe the data. This is not a problem if the data set is small. However, if thousands of features are in a KML file, there is a large amount of redundant data that has to be transmitted and parsed. XML files are being used less frequently as other formats have become more popular.

The inefficiencies of XML led to a search for more efficient transfer formats. One format that became popular was Javascript object notation (JSON). The JSON format is more efficient to transfer than KML. It stores data as sets of key value pairs. One of the early sites to implement JSON and popularize its use was Twitter. The large number of sites integrating Twitter data exposed many programmers to JSON and spread its use.

JSON is useful for transmitting tweets and sports scores but it originally did not have a formal format for spatial data. In order to address the issues of spatial data, an extension of JSON called GeoJSON was created (Butler et al., 2016). GeoJSON was explicitly created for spatial data and has become one of the most popular formats for transmitting feature data. One limitation of geoJSON is that it does not store topology. To address this issue and add a more compact format, TOPOJSON was created. TOPOJSON is an extension of GeoJSON (Bostock and Metcalf, 2016).

1.02.3.4 New Challenges of the Web

The standardization of data transfer that the Web enabled made much more data available to researchers. In addition to traditional providers of geographic data, many instruments and devices are also sharing data. This is not a new problem to GIS, or going farther back, to general cartography. Arthur Robinson wrote about cartography in 1952 at the very beginning of digital computers, "The ability to gather and reproduce data has far outstripped our ability to present it" (Robinson, 1952). That was before GPS was available on billions of smart devices. The explosion of data is not a new problem but it is getting closer to the end user who in the past was not involved in the data process. It used to be that the cartographer filtered and generalized data to create a paper map. When desktop GIS came along, the cartographic process of filtering and generalization of data became a dynamic process for the GIS operator. In order for a GIS to be usable, data are often filtered and symbolized at different scales but this process was still hidden from the end user. The varying symbologies that make GIS analysis work well on the desktop are often invisible in the maps that are the output of a GIS. With the Web platform, the end user is ever closer to the data, switching layers and basemaps, searching, filtering, and performing analysis.

Free data published on the Web is a boon for researchers, but it also comes with risks. There is no guarantee that the data will be available online at all the times. Sites may go down for maintenance or upgrades. Site upgrades may change Web service URL syntax or paths to data. These changes may break a data collection process or a map based on published data. Researchers must think about the long-term goals of their data collection and research. If they feel the data sources may not be stable, it may be necessary to extract and self host the data.

With all the available data just a few lines of code away, there is a temptation for the Web GIS builder to want to put all the available data into a GIS application and let the user find what they want. This type of application has come to be known as "kitchen sink" GIS. These applications are often confusing and laden with domain-specific terms and acronyms. Strategies are emerging to produce more curated GIS applications.

As described above, data may be published in multiple formats. In some cases the researcher can collect whichever format is desired. In many cases though, data are published in a format that is inconvenient or unusable to the researcher. In these cases, data must be downloaded and transformed to a format usable by the researcher. This can take significant technical skills.

1.02.4 The Move to the Cloud

Research using Web services described above has been aided by the move of computing infrastructure into the cloud. The term cloud infrastructure has become a buzzword as of late. In general it means purchasing computing resources as services instead of physical entities from centralized providers connected to the Internet. These services may be hosted by an off-premise commercial provider, or they can be localized data centers within a large organization. They can even be a combination of the two with some data and services on premises and older data in "cold storage" offsite. With cloud infrastructure, customers do not have to consider any of the complex details of building advanced networks. They just rent capacity.

Amazon was the first major provider of cloud infrastructure. Amazon began providing cloud services in earnest when they started renting out excess capacity they had installed for their online store (Wingfield, 2016). Other large providers soon followed and cloud infrastructure is now available from many providers including Microsoft, Google, Digital Ocean, and Rackspace. Infrastructure has now been commoditized to such a degree that entire clusters of servers can be purchased and deployed in hours with nothing more needed than a credit card. As Moore's Law and competition take hold of the market, infrastructure services are getting even cheaper. These new resources available outside the restrictions and limitation of institutional network policies offer many opportunities for researchers to experiment, make small prototypes to demonstrate feasibility, and to conduct resource-intensive research on temporary basis without having to buy permanent hardware that will soon be obsolete.

Infrastructure services exist at different levels. Perhaps the most common example is putting files into Dropbox, Google Drive, or Microsoft Onedrive, and having them available through a browser on any computer or mobile device. In the past, it was common for these files to be stored on a file server on a school or work network and not be available off the network. This is another example of an old technology being made more useful and flexible through Web services. File storage is just a small part of the stack of available services. At the base of the stack is "bare metal" hosting where hardware or virtual hardware is provided for a customer. Moving up the stack, applications like databases and Web servers can be rented that share space on common hosts. At the top are services like computation and file storage. We will discuss some of these in more depth.

1.02.4.1 Host Virtualization

Host virtualization refers to running operating systems on a virtual set of hardware instead of physical hardware (Fig. 2). Each virtual machine (VM) thinks that it is running on physical hardware. Multiple VMs, each with their own configuration,

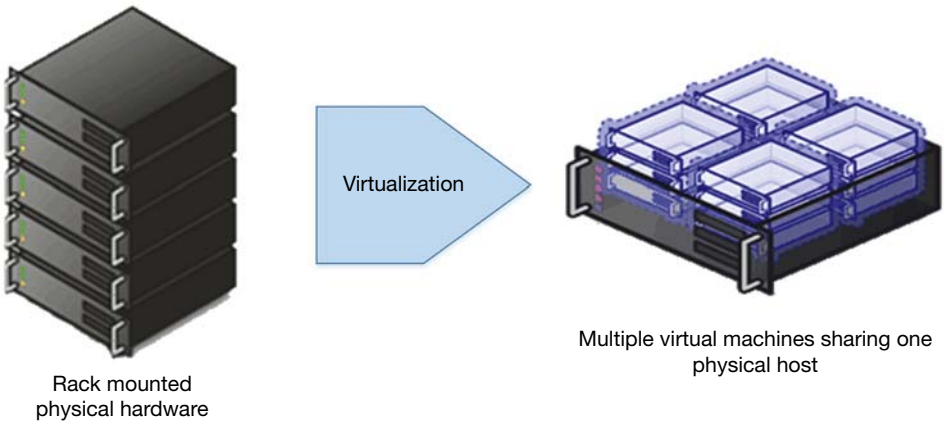

Fig. 2 Several virtual machines on a physical host.

applications, and security, can share a single set of hardware. Each VM has a complete installation of files just like a physical machine and can be addressed from the network as an individual server. However, these VMs all share the same physical host machine.

1.02.4.1.1 Why virtualization

In the past, it was often the case that data centers were full of servers that only ran a single application that had to have its own hardware because it conflicted with other software. These servers used very little of a computer's resources most of the day when they were being used and idled the rest of the day. The hardware for these servers was expensive and when many of them were put in one room, they took a large amount of energy to run and cool. Today, data centers full of physical hardware are being condensed onto powerful VM host servers that can contain hundreds of separate VMs. Increasing capacity provided through Moore's Law has meant that hardware has vastly outpaced the needs of most users and applications. Server hardware is now powerful enough to host multiple operating systems without performance degradation.

There are many advantages of host virtualization. A VM exists as one large file. All of its operating system files, applications, and data are contained in that one file. Because they are one file, they can be backed up or transitioned to different hardware. In addition, a differencing file can be used to snapshot the state of a VM before any significant change like software installation or configuration modification is made to the VM. If the change encounters problems, it is easy to roll back the system state by simply deleting the checkpoint snapshot.

Another advantage to using virtualization is the ability to create VMs from images that already have particular configurations of software installed. Base VM configurations can be stored in libraries so that new machines can be provisioned by simply copying a VM image from a library and configuring it. The time and expertise needed to complete a fresh install of all of the software are avoided. These libraries can be hosted or public. Both Amazon and Microsoft keep libraries of VM images.

VMs have virtual hardware that can be adjusted, prioritized, and shared. Small test VM servers can share limited resources while important production VMs can be provisioned more processors and memory. Provisioning may be done on a temporary basis. If a research project is going to be doing resource-intensive activities, it can be scheduled for off-peak hours when the other VMs on a physical host will have little activity. In the commercial world, seasonal companies like tax preparers or holiday merchants may pay for more virtual capacity during their peak season and throttle back during the off season.

Prototyping is another useful application of virtualization. VMs can be prototyped on a client and then moved to a server environment with more resources when the VM is ready. This allows flexibility of configuration and testing while not on a production network.

VMs may be used on fast networks disconnected from the Internet or low bandwidth connections and then migrated to high-speed networks. For instance, one workflow might have a researcher build a virtual machine while disconnected from the Internet on a long airplane flight. After arrival on a remote disconnected site, the researcher may use local wireless to share and collect data with other researchers. During remote data collection, the VM can be temporarily moved to a high-speed internet connection to sync data with a home institution's computers. When the researcher returns to their institution, they may choose to migrate the VM to robust hardware to conduct resource-intensive data processing. The entire VM and all of its data and configuration can be moved to a server infrastructure by moving one file.

Virtualization has environmental advantages as well. Since network connectivity is the only requirement for management, data centers full of VM hosts can be located with their environmental footprint in mind. Locating where renewable energy is abundant decreases a data center's carbon footprint. A huge part of a data center's energy consumption is often in cooling the many servers it contains. If a data center can be located in cool climates or where ample chill water is available, the footprint can be further reduced. Repurposing power-intensive industrial sites located next to hydro power like aluminum smelting plants is an ideal case (Wilhem, 2015).

For those with a historical perspective of computer operating systems, the interoperability of OSs with virtualization may raise some eyebrows. There has been a long-standing animosity between vendors of computer operating systems, particularly between Microsoft and the open-source community. This has recently changed as Microsoft has joined the Linux Foundation (Bright, 2016), supports open-source operating systems on their cloud platform Azure, and has recently released a version of Microsoft SQL server for Linux. While Linux has yet to take over the desktop market, its server versions are increasingly popular as free, stable, and high-performing operating systems. One further note about operating systems and virtualization. While other OSs can be virtualized on Apple hardware, it is a violation of the license agreement to virtualize Apple operating systems on non-Apple hardware, so it is not discussed below.

1.02.4.1.2 Implementing host virtualization

For the researcher who wants to use virtualization for their research, the three most common choices are Microsoft Hyper-V, VMware, and Oracle VirtualBox. Microsoft Hyper-V is built into most modern versions of Windows Server and has recently been added to the Windows 10 desktop operating system making it ideal for prototyping VMs on clients and migrating to servers. Hyper-V supports virtualization of both Linux and Windows operating systems. The host software only runs on Microsoft Windows computers, so while VMs can be moved and shared among Windows hosts they cannot be moved to other operating systems. Hyper-V is often used because it is the default on Windows and is a key technology that is strongly supported and actively developed by Microsoft.

VMWare is the most featured virtualization software and the most expensive. It is targeted toward large enterprise installations. VMware does have a limited version that can be used for free. Large institutions that host many servers or have their own on premises cloud often use VMWare to manage their virtualization infrastructure. VMware can host VMs on Windows, Linux, and Apple operating systems. Because VMware is a licensed product, one drawback of VMware is that it may restrict the sharing of VMs between researchers at different institutions.

VirtualBox is a very popular choice because it is free and it supports guests on Windows, Linux, and Apple operating systems. This gives it the greatest flexibility for sharing VMs between researchers. VirtualBox can host Windows and Linux operating systems. It is the base of the Docker containerization technology that will be discussed below.

1.02.4.2 Containerization

Containerization is a form of virtualization that is becoming increasingly popular. Containerization is an even denser version of virtualization. As described above, virtualization takes an entire operating system and condenses it down to one file. With full virtualization, all the files for the entire operating system exist within each VM. If a host server has 10 VMs, it has 11 copies (including its own) of all of the operating system files. A container is a VM that exists as a separate OS from the host OS but it only has the files that are unique to it. Containerization removes the redundant files and only maintains configuration files necessary to maintain the system state. This sharing of base files means that the VMs take even fewer resources to host. Also, library images of containers with stock configurations are much smaller and can be shared more efficiently. Containerization allows researchers to spin up many VMs to keep applications logically separate while not having to worry about idling hardware.

1.02.4.3 Application Hosting

Cloud infrastructure can be moved further up the stack from whole machines to applications and services. For instance, database servers require a significant amount of technical expertise to install and maintain. While researchers may know how to programmatically interact with a database server, installation and configuration is often a difficult task. Databases on large networks have to conform to institutional policies. They have to sit behind institutional firewalls. This is not the case with cloud services. Researchers can provision a database server with a credit card. The database server is built and maintained by a cloud host. Communications with the database can be encrypted through use of a virtual private network to maintain security during transmission of data.

1.02.4.4 Cloud Computation

Cloud computing is another example of services in the cloud. Significant but temporary computation services can be rented to process large data sets. Like the database example above, the infrastructure is hidden and the researchers use only what is needed. Prices may vary based on the time of day so further savings are possible with the correct timing.

It is important to note that most of these services have free tiers. For researchers this is important as it can allow them to complete a limited proof of concept in order to apply for funding to build a full-scale application. If successful, an application or project can be scaled up when funding is available. This fast prototyping in the cloud allows researchers to iterate through multiple configurations that may be slowed by institutional friction on a home network.

1.02.4.5 Software as a Service

Software as a service (SaaS) is another category of cloud infrastructure. One of the software as a service categories that has significantly advanced geocomputation is software version control. Software version control allows multiple programmers to work on

the same software by managing the changes that are made in the programming code. Multiple programmers can collaborate on the same software without conflicting. Version control may also implement project management tasks like task lists, documentation, and time lines. Many version control systems allow developers to share their software with the rest of the world. Collaborative version control has greatly advanced the progress of research by allowing researchers to build on each other's work.

The most popular software version control website is Github. Github is a software version control website that allows users to publish unlimited projects if they are open to the public. Private repositories are also available for a fee. From a GIS perspective, cloud-based version control plays an important role in developing new capabilities. Many of the libraries used for Web GIS are open sourced and are hosted on Github. Leaflet.js, one of the most popular open-source mapping software on the Web, is hosted on Github as is Esri Leaflet, the library that allows Leaflet maps to talk to the most used proprietary servers from Esri.

One final example of SaaS that is particularly useful in a geocomputation and GIS context is Python Anywhere (Anywhere, 2016). Python is an open-source programming language that is particularly popular with spatial research. Python can be used to collect data automatically on a set schedule from Web services and deposit them in a database. Data collection using Python requires few resources but must be done from a stable, always online host. That is where Python Anywhere can be useful to researchers. Python Anywhere offers a programming and application hosting environment in the cloud. It can be reached from any Web browser on any platform. The service has a free tier and very inexpensive paid tier plans. Python Anywhere can be used to build simple GIS data collection applications or to prototype applications to use as a proof of concept. In a classroom context, instructors can get students programming on the Web without setting up complicated infrastructure onsite. Students can retain their work after classes end or use them across multiple classes.

1.02.5 Computational Methods

The previous sections described the evolution of computation and its move onto the Web. These resources have been linked up with a tremendous amount of data. New questions are emerging as researchers take advantage and cope with these new opportunities.

With the advances in spatial–temporal data and computing technology, an important direction of handling such data is to discover useful and nontrivial patterns from the data using various computational methods. This is a vast area that encompasses a diverse set of methodological spectrum, and it is a daunting task to even try to categorize the methods. Here, we summarize these computational methods into roughly four groups: visualization, machine learning, spatial optimization, and simulation. We note that such a categorization is far from being ideal. However, this allows us to effectively identify salient tasks in geocomputation that have emerged in recent years.

1.02.5.1 Visualization

An important task in geocomputation, or any area that involves the use of data, is to know your data. Such a need has led to the emergence of the research field of exploratory data analysis, or EDA, in statistics. EDA stresses the importance of seeing the data in order to understand and detect patterns in the data, instead of just analyzing the data (Tukey, 1977). Shneiderman (1996) summarized a three-step approach to EDA: "Overview first, zoom and filter, and then details-on-demand." EDA not only requires computational and visualization tools (as in software packages) to process the data, but also relies on principles and methods that guide the development of those tools. The extension of EDA that deals with spatial and temporal data has a root in cartography and has fully fledged into an interdisciplinary field (Andrienko and Andrienko, 2006).

In any spatial data set, we call the indications of space and time the references and the measures at these references a set of characteristics. The goal of spatial and temporal exploratory data analysis is to use graphics to address various tasks (Bertin, 1967). These tasks ask questions that can be as simple as "what is the air quality of Columbus, Ohio today?" or as complicated as "which areas in Columbus, Ohio have the worst air quality and also have the highest density of minority ethnicity groups?" While answering these questions requires intimate knowledge about the data, it is also important to develop tools that provide interactivity for users to explore the data by means that can be as simple as looking up and comparing the data values of different references, or as complex as finding patterns and associations between data values and their referenced spatial and/or temporal entities. Much of this can only be achieved in an interactive environment where the user can choose the scope and extent of the data to be displayed and further examined.

While tremendous progress has been made in spatial exploratory data analysis in the last two decades (Andrienko and Andrienko, 1999), recent years have seen a rapid change in the computational community where data can be closely bound with graphic elements on a visualization, especially for data visualization on a Web-based platform. Techniques enabling such a trend include D3 (Bostock, 2016) and LeafLet (Agafonkin, 2016), both JavaScript libraries. D3 utilizes the vector-based image data format called scalable vector graphics (SVG) that supports detailed description of graphic elements (such as shapes, text, color, and other graphic effects) for different visualizations. A key feature in D3 is to bind the data into various graphic elements. For example, we can bind a one-dimensional array with circle elements so that the size of the array determines the number of circles, and the value of an array element determines the size of its corresponding circle. In addition to binding the data array with circles, D3 can also bind the range of the data in the array with the Y-axis and the number of elements in the array with the Y-axis, which will effectively construct a bar chart. This kind of data binding can be used between multidimensional data and other graphic elements. Beyond data binding, D3 can also be used for subsetting (filtering) data that allows the user to change the focus of data exploration.

Maps can be made in various ways today. Packages such as D3 can support mapping as a general graphic representation method. However, here we specifically note that LeafLet provides a wide range of tools for mapping data from different formats and can be used to integrate a custom map with data from many sources. LeafLet enables basic interactive functions such as zooming and panning. It can also allow the user to filter, query, and highlight features on the map. More importantly, these interactive functions can be linked to D3 visualizations by establishing the match of the unique identifications of the features on the map and the data items bound with D3 graphic elements.

1.02.5.2 Machine Learning

The root of machine learning dates back to as early as the 1950s when Alan Turing asked the fundamental question of "can machines think?" Turing (1950) stipulated a computer program that simulates a child who learns while developing the brain. The start of machine learning follows the footsteps of artificial intelligence in logical inference. In recent years, however, machine learning has shifted more toward the computational aspect of artificial intelligence, inspired by advances of algorithms and software in statistics, ecology, and other disciplines (Mitchell, 1997). The essential task of machine learning is the ability to classify input data into categories that can be learned from.

In general, there are two main camps of machine learning: supervised and unsupervised. In supervised learning, the user must prepare a data set that includes the desired output to train the algorithm. For example, to develop a supervised machine learning algorithm to assign remote sensing image pixels into different land use types, the training data must include the land use types of each pixel, which will allow the algorithm to learn about the rules that can be used to classify new pixels. Supervised machine learning methods include various parametric and nonparametric statistical models, many neural networks, decision trees, and support vector machines. Unsupervised learning methods do not rely on the use of training data sets. Instead, the learning algorithm is designed to reveal the hidden structure in the data. Some neural networks and most of the clustering methods belong to this category.

A widely used machine learning method in spatial data is the k-means clustering method (Lloyd, 1982). This method can be used to search for the cluster of points in an area. This is an iterative method that starts from a random set of k locations and each of these locations is used to serve as the center of a cluster. Points are assigned to their closest center to form the k clusters. In the next step, the points assigned to each cluster are used to compute a new center. If the new centers are significantly different from the previous k centers, the new centers will be used and we repeat the process until no significant changes can be made. At the end, the method will yield a partition of the points into k clusters. While there are obvious applications of the k-means method in two-dimensional data, a k-means method applied on a one-dimensional data is similar to the very widely used Jenks classification method in choropleth mapping.

A supervised machine learning works in a different way. A neural network, for example, utilizes a series of weights that convert a set of inputs to outputs. By comparing with the known result, an algorithm is used to adjust the weights in order to minimize the error between the model output and known output. A support vector machine, on the other hand, utilizes an optimization method to try to determine if a line (or a hyperplane for multidimensional data) can be considered to be best separating input data into two classes. Some of these machine learning methods have been used to process data from nontraditional sources. For example, support vector machines were used to geocode tweets by identifying location expression in the text by finding the best match between locations retrieved from the text and place names in a gazetteer (Zhang and Gelernter, 2014).

1.02.5.3 Spatial Optimization

Spatial optimization refers to the need of finding an optimal set of spatial configurations such that a goal or objective can be reached while some constraints are satisfied. There are various applications of spatial optimization. For example, one may wish to find a subset of land parcels in an area under a budget constraint such that the total benefit of the selected land parcels is maximized. In general, spatial optimization problems can be categorized into two major groups (Xiao, 2008). A selection problem aims to find a subset of spatial units. We may impose spatial constraints on these selected spatial units. For example, some problems may only consider contiguous spatial units to be selected. A second type of spatial optimization problem is partitioning problems. These problems aim to separate the spatial units into a set of regions. For example, the political redistricting problems require an area to be partitioned into a number of contiguous districts.

Solving spatial optimization problems generally requires tremendous amount of computing resources as many of these problems are NP-hard, meaning there may not exist a solution method that can be used to find the optimization solution in a reasonable amount of time. For this reason, researchers have developed a special type of solution approach called heuristics; methods that can be used to find quickly good, but not necessarily optimal, solutions to the problem. The computational efficiency of heuristic methods is the key (Xiao, 2016).

Traditional heuristic methods are typically designed to solve one type of optimization problem. For example, the p-median problem is a representative optimization problem in location-allocation analysis, where the goal is to locate facilities or services on p nodes on a network so that the distance from each node to its nearest facility or service node is minimized (Hakimi, 1964). A commonly used heuristic method to solve the p-median problem is the vertex exchange algorithm (Teitz and Bart, 1968). This algorithm starts with randomly selected p nodes from the network and keeps switching these nodes with unselected ones until such exchange can no longer improve the solution.

In contrast to traditional methods, metaheuristic methods aim to provide a solution framework for a wide range of problems. The increasing list of metaheuristic methods includes genetic algorithms (Holland, 1975), tabu search (Glover et al., 1997), and simulated annealing (Kirkpatrick et al., 1983). It is noticeable that these methods are often inspired by some natural processes and they can be used to solve different kinds of problems. To use a genetic algorithm (GA) to solve the *p*-median problem, for example, one could use an array of integers to represent the indices of the nodes selected for locating the facilities. The GA will start with a set of arrays, each containing randomly selected indices. This is called the population of solutions. The GA will evaluate each individual in the population by assigning it a fitness value. A high fitness value means a good solution in the population with a small total distance. Individuals with high fitness values have a high chance to be selected by the GA to participate in an operation called crossover, where the two individuals selected mix their contents to create two new individuals. New individuals with fitness values better than the current individuals will be inserted into the population. A mutation operation may also be used to randomly change the content of an individual. By continuously doing these selection, crossover, and mutation operations, the population will evolve toward a new state where individuals with better fitness values (thus better solutions) will be obtained.

1.02.5.4 Spatial Simulation

Models used to solve optimization problems are often called normative models because they prescribe what should be done. But results coming from such models may not work as prescribed because situations in the real world may not necessarily match exactly as formulated in the models (Hopkins et al., 1981). To have a better understanding of the behavior of the system, it is therefore necessary to identify the processes or mechanisms in the system, which may require various types of simulation models.

Two types of simulation approach have been widely adopted in geocomputation research. The first approach is derived from John Conway's Game of Life (Gardner, 1970), where a set of simple transition rules govern the change of state for each location (a cell) based on the states of the neighbors of that location. This type of simulation model is called cellular automata where each cell will automatically change its state if the transition rules are met. Researchers extended such model to simulate the dynamics of spatial systems such as urban sprawl, and land use and land cover change (Clarke and Gaydos, 1998).

A second type of spatial simulation approach explores a more explicit representation of the processes linked to spatial systems, where the important players in the system must be identified and, more importantly, the interactions between these players must be understood and simulated. These are called agent-based models (Epstein and Axtell, 1996), where each agent is a player in the system and agents proactively seek to maximize their own benefits. A particular application of the agent-based modeling approach is found in the land use and land cover change literature (Parker et al. 2003) where different players such as land owners and other stakeholders can be identified to act through interactions such as land use planning, land market, and development.

The relatively straightforward modeling concept has made agent-based modeling highly popular across a wide range of disciplinary boundaries. More importantly, implementing such a model has become increasingly intuitive. Programming platforms such as NetLogo (Wilensky, 1999) and MASON (Luke et al., 2005) not only support the necessary coding environment but also a comprehensive visualization tool to help present the simulation results. These tools have played an important role in making agent-based modeling accessible to researchers and professionals without much training in programming.

1.02.6 Visualizing Twitter Data

Twitter data represents a significant challenge to geocomputation. By sending a message of less than 140 characters, millions of Twitter users around the world are constantly connected with each other by sending new tweets or retweeting tweets from other users. The volume of Twitter data is growing rapidly. Among all the tweets, a subset of them also contains locational information that can be used to help identify where they were sent from. Though these geocoded tweets only account for a small portion of all the tweets, they can be used to help understand the dynamics of a region or the world. Many researchers have also started to analyze tweets in order to categorize these tweets in terms of the emotion of the sender and the kinds of topics conveyed in the tweets (Mitchell et al., 2013; Steiger et al., 2015). This is a bona fide big spatial temporal data set.

Here we concentrate on the tweets around the Central Ohio area where the city of Columbus is located. Geocoded tweets from this region were collected for the time period of May to July 2015 using the public Twitter API, which provided a small sample of about 1% of the tweets. In this period of time, more than 200,000 tweets are collected and stored in a PostgreSQL database. A spatial temporal index is developed for efficient data retrieval from the database. Each tweet stored in the database is assigned a set of tags, indicating the theme of the tweet. Seven themes are identified for this region: food, business, move (mobility), sports/fitness, current events, LGBT, and local interests. A set of keywords is used to identify a theme. For example, a tweet is tagged as "Food" if it contains words such as pizza, barbecue, and brewing. A tweet is also tagged as happy, unhappy, or neutral by calculating the average happiness index of the words in the tweet (Dodds and Danforth, 2009).

Fig. 3 is a Web-based graphical user interface of a prototype geocomputational approach to visualizing the tweets collected for the central Ohio region. The left side of the screen visualizes the temporal aspects of the tweets. A user can define a slice of time for visualization using the sliding bar underneath the plot. Each curve in the plot shows the number of tweets of each category. Each of the curves can be turned on and off. On the very right side of the screen is a small window that can be used to show random individual tweets in the selected category within the timeframe defined in the left window. These are renderings of the Twitter data that

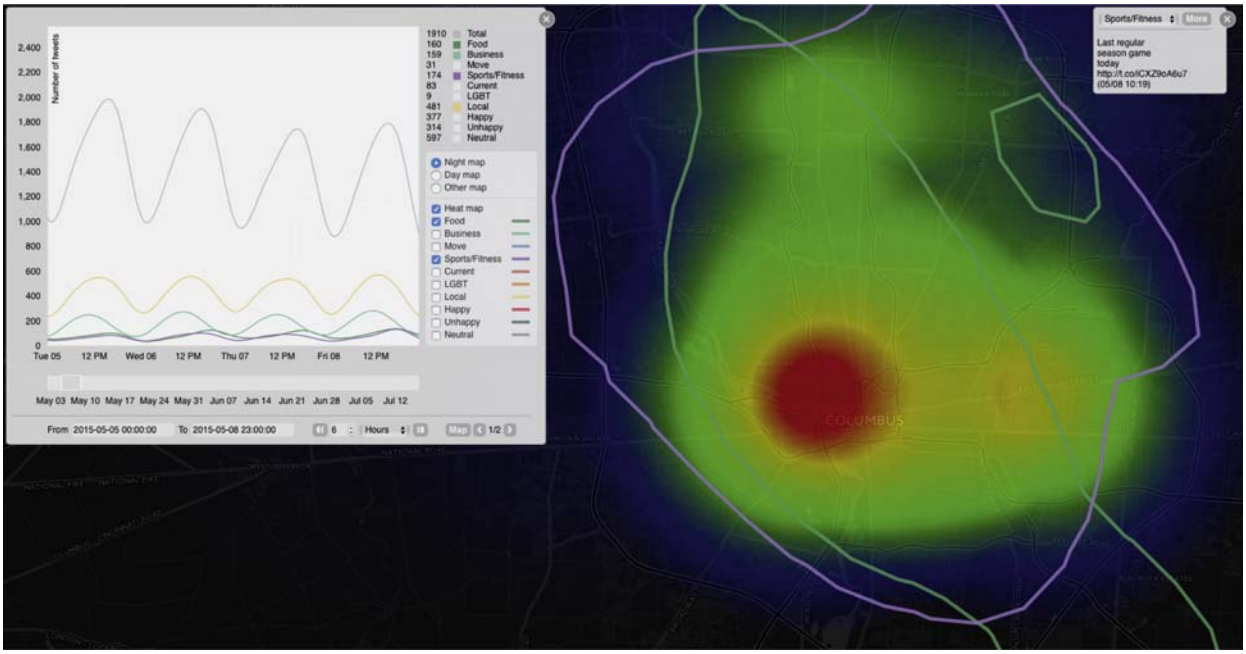

Fig. 3 A geocomputational approach to visualizing Twitter data.

are retrieved from the database through an AJAX framework. The request made in the Web browser is routed through a Web server that is connected to the database.

When the user clicks on the Map button in the left window, the Web browser gathers the information from the screen, including the start and end time period specified by the user. A request is then sent to the server which will fire up server side programs to (1) compute the heat map of all the tweets in the region in the specified timeframe and (2) calculate a spatial clustering method called kernel density estimation (Wand and Jones, 1995; Winter and Yin, 2010) for each of the tweet categories. The results of these computational tasks are then sent back to the browser as JSON data that are subsequently used to create Leaflet layers for the heat map and the kernel curves. The computation of the kernel density estimation is a nontrivial task because a total of 10 kernels must be calculated for each request. To make the interface effective, only the 50% kernel is shown for each category, meaning that 50% of the tweets in each category were sent within the enclosed curves on the map. The user can create up to 20 sets of these heat maps and kernels; the forward and backward buttons allow the user to loop through these sets.

1.02.7 Conclusion

For much of the history of geocomputation and GIS, we witnessed incremental improvements in geoprocessing capacity. The same workflows maintained as geographic data were created and analyzed in isolated computing units and GIS software installations. Data were only shared through extracts and static data files. Recently, though, the modern Web has standardized data transfer between hosts and facilitated much easier data sharing. This ability to share data came along just as GPS, broadband, and wireless have connected many islands of computing and mobile devices. The physical implementation of the Web has been hastened by cloud infrastructure. The cloud has decreased the costs to develop and host applications on the Web. It has granted resource-constrained parties access to powerful Web geocomputation tools. These new data and capabilities offer researchers many new opportunities for investigation but come with challenges all their own.

References

Agafonkin V (2016) Leaflet.js javascript library. https://www.leafletjs.com. (Date accessed 2/6/2017).
Andrienko, G.L., Andrienko, N.V., 1999. Interactive maps for visual data exploration. International Journal of Geographical Information Science 13 (4), 355–374.
Andrienko, N., Andrienko, G., 2006. Exploratory analysis of spatial and temporal data: A systematic approach. Springer Science & Business Media, Heidelberg.
Anywhere P (2016) Python anywhere website. https://www.pythonanywhere.com. (Date accessed 2/6/2017).
Berners-Lee, T., Fischetti, M., Foreword By-Dertouzos, M.L., 2000. Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor. HarperInformation, New York.
Bertin, J., 1967. Semiology of graphics: Diagrams, networks, maps. University of Wisconsin Press, Madison.
Bostock M (2016) Data-driven documents. https://www.d3js.org. (Date Access 2/6/2017).
Bostock M and Metcalf C (2016) The topojson format specification. https://github.com/topojson/topojson-specification/blob/master/README.md. (Date Accessed 2/6/2017).

Bright P (2016) Microsoft, yes, microsoft, joins the Linux foundation. *Ars Technica*. https://arstechnica.com/information-technology/2016/11/microsoft-yes-microsoft-joins-the-linux-foundation/. Date Published:11/16/2016. (Date accessed 2/6/2017).
Butler H, Daly M, Doyle A, Gillies S, Hagen S and Schaub T (2016) *The geojson format*. Technical report. Internet Engineering Taskforce. https://tools.ietf.org/html/rfc7946. (Date accessed 2/6/2017).
Campbell, J.B., Wynne, R.H., 2011. Introduction to remote sensing. Guilford Press, New York.
Chorley, R.J., Haggett, P., 1967. Models in geography. Methuen, London.
Chrisman, N., 2006. Charting the unknown: How computer mapping at Harvard become GIS. ESRI Press, Redlands.
Clarke, K.C., Gaydos, L.J., 1998. Loose-coupling a cellular automaton model and GIS: Long-term urban growth prediction for San Francisco and Washington/Baltimore. International Journal of Geographical Information Science 12 (7), 699–714.
Dawn C Parker, Steven M Manson, Marco A Janssen, Matthew J Hoffmann, Peter Deadman (2003) Multi-agent systems for the simulation of land-use and land-cover change: a review *Annals of the association of American Geographers*. Taylor & Francis 93(2): 314–337.
Dodds, P., Danforth, C., 2009. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness Studies 11, 441–456.
Epstein, J.M., Axtell, R., 1996. Growing artificial societies: Social science from the bottom up. Brookings Institution Press, Washington, DC.
ESRI, 1998. ESRI shapefile technical description. Environmental Systems Research Institute, Redlands.
Faust, N., 1998. Chapter 5: Raster based GIS. In: Foresman, T.W. (Ed.), The history of geographic information systems: Perspectives from the pioneers. Prentice Hall, Oxford.
Fidler, B., Currie, M., 2015. The production and interpretation of Arpanet maps. IEEE Annals of the History of Computing 37 (1), 44–55.
Fu, P., Sun, J., 2010. Web GIS: Principles and applications. ESRI Press, Redlands.
Gahegan, M., 1999. What is geocomputation? Transactions in GIS 3, 203–206.
Gardner, M., 1970. Mathematical games: The fantastic combinations of John Conway's new solitaire game life. Scientific American 223 (4), 120–123.
Glover F, Laguna M, et al. (1997) Tabu search—part I. ORSA Journal on computing 1.3 (1989): 190–206.
Goodchild MF (2007) In the world of web 2.0. *International Journal* **2**(2): 27–29.
Goran, W., 1998. Chapter 12: GIS technology takes root in the department of defense. In: Foresman, T.W. (Ed.), The history of geographic information systems: Perspectives from the pioneers. Prentice Hall, Oxford.
Hakimi, S.L., 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research 12 (3), 450–459.
Hart, G., Dolbear, C., 2013. Linked data: A geographic perspective. CRC Press, Boca Raton.
Holland, J.H., 1975. Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. University of Michigan Press, Ann Arbor.
Hopkins, L.D., Brill, E.D., Wong, B.D., et al., 1981. Generating alternative solutions for dynamic programming models of water resources problems. University of Illinois, Water Resources Center, Urbana.
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., et al., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.
Lloyd S (1982) Least squares quantization in PCM. *IEEE transactions on information theory*. 28.2: 129–137.
Lukasik, S.J., 2011. Why the Arpanet was built. IEEE Annals of the History of Computing 33 (3), 4–20.
Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., Balan, G., 2005. MASON: A multi-agent simulation environment. Simulation: Transactions of the Society for Modeling and Simulation International 82, 517–527.
Microsoft (2003) Microsoft security bulletin ms02-039—critical. https://technet.microsoft.com/library/security/ms02-039. (Date accessed 2/7/2017).
Miller, H.J., Goodchild, M.F., 2015. Data-driven geography. GeoJournal 80 (4), 449–461.
Mitchell, T.M., 1997. Artificial neural networks. Machine Learning 45, 81–127.
Mitchell, L., Frank, M.R., Harris, K.D., Dodds, P.S., Danforth, C.M., 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One 8, e64417.
Moore G (1965) Cramming more components onto integrated circuits. *Electronics* **38**(8): 83–84.
OGC (2008) OGC keyhole markup language. http://www.opengeospatial.org/standards/kml. (accessed 2/7/2017).
OGC (2016) Web service common standard. http://www.opengeospatial.org/standards/common. (accessed 2/7/2017).
Openshaw, S., Abrahart, R.J., 1996. GeoComputation. In: Abrahart, R.J. (Ed.), Proceedings of the First International Conference on GeoComputation. University of Leeds, Leeds, pp. 665–666.
O'reilly T (2007) What is web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies* 65(1): 17–38. ISSN: 1157–8637.
Robinson, A., 1952. The look of maps. University of Wisconsin Press, Madison.
Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. In: *Proceedings of the IEEE Symposium on Visual Languages*. Boulder: Colorado. pp. 336–343. IEEE.
Steiger, E., de Albuquerque, J.P., Zipf, A., 2015. An advanced systematic literature review on spatiotemporal analyses of Twitter data. Transactions in GIS 19, 809–834.
Teitz, M.B., Bart, P., 1968. Heuristic methods for estimating the generalized vertex median of a weighted graph. Operations Research 16 (5), 955–961.
Tomlinson, R., 1998. Chapter 2: The Canada geographic information system. In: Foresman, T.W. (Ed.), The history of geographic information systems: Perspectives from the pioneers. Prentice Hall, Oxford, p. 2.
Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley, Reading.
Turing, A.M., 1950. Computing machinery and intelligence. Mind 59 (236), 433–460.
Wand, M.P., Jones, M.C., 1995. Kernel smoothing. Chapman & Hall, London.
Wilensky U (1999) Netlogo. http://ccl.northwestern.edu/netlogo/ (accessed 2/7/2017).
Wilhem S (2015) Power shift: Data centers to replace aluminum industry as largest energy consumers in Washington state. http://www.bizjournals.com/seattle/blog/techflash/2015/11/power-shift-data-centers-to-replace-aluminum.html. (accessed 2/7/2017).
Wingfield N (2016) Amazon's cloud business lifts its profit to a record. *New York Times*, April 28.
Winter, S., Yin, Z., 2010. Directed movements in probabilistic time geography. International Journal of Geographical Information Science 24 (9), 1349–1365.
Xiao, N., 2008. A unified conceptual framework for geographical optimization using evolutionary algorithms. Annals of the Association of American Geographers 98 (4), 795–817.
Xiao, N., 2016. GIS algorithms. Sage, London and Thousand Oaks.
Zhang, W., Gelernter, J., 2014. Geocoding location expressions in twitter messages: A preference learning method. Journal of Spatial Information Science 2014 (9), 37–70.

1.03 Big Geodata

Michael F Goodchild, University of California, Santa Barbara, CA, United States

1.03.1 Definitions

1.03.1.1 Geodata

Geodata are normally defined as data about the surface and near-surface of the Earth. More precisely, geodata are observations about what is present at some location. Since the number of possible locations is infinite, geodata are often observed or captured in the form of aggregated or summary observations about areas (e.g., states, forest stands), lines (e.g., rivers, highways), or volumes (e.g., oil reservoirs, buildings); or geodata may be sampled at selected locations. A host of types of geodata exist, ranging from data about such physical variables as ground elevation or surface temperature, to data about the numbers of inhabitants in an area, or their average income. Geodata may be structured in a range of common formats, and are conveniently handled in readily available software. Synonyms for geodata include geospatial data, geospatial information, and geographic information. Spatial data is normally assumed to be a superset that includes data about phenomena embedded in other spaces besides geographic space.

1.03.1.2 Big Data

Big Data is a term of comparatively recent coinage, and has been the focus of a remarkable outpouring of energy and innovation in the past decade. The most obvious meaning of the term relates to data volume, and to the very rapid expansion of data storage and processing capacity in recent years. Whereas a gigabyte (roughly 10^9 bytes or 8×10^9 bits) might well have stretched the capacity of most computers in the 1970s, today the average laptop has approaching a gigabyte of random-access memory and a terabyte (roughly 10^{12} bytes) of hard-drive storage. It has always been possible to imagine a quantity of data larger than a given device can handle, and thus one convenient definition of Big Data is a volume of data larger than can readily be handled by a specified device or class of devices. For example, the volume of geodata collected by the Landsat series of satellites at their inception in the early 1970s was well beyond the processing capacity of the computers at the time. Already there are research projects that must deal with petabytes (roughly 10^{15} bytes) of data; we are living, we are told, in the midst of an "exaflood" of data (one exabyte is roughly 10 to the power 18 bytes); and 10^{24} has already been given an internationally agreed prefix ("yotta").

But although it is important, volume is not the only distinguishing characteristic of Big Data, which is why the term is capitalized here, to distinguish it from run-of-the-mill voluminous data. Networks of sensors, the Internet of Things, and increasingly sophisticated data collection, transmission, and aggregation systems have created a new and abundant supply of dynamic data in close-to-real time. The average citizen now expects near-instantaneous delivery of information on such topics as traffic congestion, international news, and sports results. Thus "velocity" is often cited as a second defining characteristic of Big Data.

Finally the last three decades have seen a steady transformation from a world dominated by single, authoritative sources of information to a proliferation of multiple and often contentious or conflicting sources. To cite a typical geodata example, a search for information about the elevation of a given point used to produce a single result, with the authority behind it of a national mapping

agency. Now, however, multiple technologies for measuring elevation that include surveys using GPS (the Global Positioning System), measurements by hikers, traditional maps, LiDAR (Light Distancing and Ranging), and SRTM (the Shuttle Radar Topography Mission) create a plethora (a "variety") of answers with a range of values. How to choose among them, and whether an improved estimate can be obtained by combining them, for example, by averaging, is one of the new concerns raised by the advent of Big Data. Thus Big Data is often defined by the "Three Vs": volume, velocity, and variety.

A fourth "V" is often suggested, to capture the various forms of uncertainty associated with Big Data, especially Big Data that come from nonauthoritative sources that have not been subjected to quality control. This fourth V might stand for validity or veracity, but unfortunately validity or veracity is what Big Data often *lack*, rather than a distinguishing property. Uncertainty is an especially important issue for geodata, and the subject of a large and growing literature (e.g., Zhang and Goodchild, 2002). Position is an essential element of geodata, and is measured using one of a range of techniques, each of which introduces its own level of uncertainty. For example, the GPS receiver in an average smart phone produces latitude and longitude with an error that is commonly in the 10 m range, but may be as high as 100 m if measurement is impacted by tall buildings, tree canopies, and many other factors. The local properties recorded in geodata (commonly termed the *attributes*) are also often subject to errors of numerous kinds, and when what is recorded is a class (of vegetation cover, e.g., or land use) the definition of the class will include uncertainty, such that two observers cannot be guaranteed to record the same class at a given point. In summary, geodata can never be perfect, never the truth.

1.03.1.3 Big Geodata

The concept of Big Geodata is comparatively recent, and deals with the well-defined and important subset of Big Data that are geodata. As the Landsat example cited above illustrates, volume is no stranger to geodata, and today our ability to collect and acquire geodata vastly exceeds our ability to store or process them. But the traditional process of acquiring geodata, through surveying, photogrammetry, or satellite-based remote sensing, has been slow and painstaking. Thus velocity in acquisition is a much more recent concern, with impacts that are disruptive. Similarly variety is novel, given the past reliance on single, authoritative sources, and thus also disruptive. The nature of these disruptive impacts is discussed at length later.

If volume is no stranger to geodata, how have the problems of excessive volume been addressed in the past? Several long-accepted techniques have been used, allowing researchers and others to deal with what would otherwise have been impossible volumes of data. Geographers have long practiced the use of *regions*, by dividing the world into a number of areas that can reasonably be assumed to be uniform with respect to one or more properties. For example, the Midwest is part of the central United States, with large areas devoted to raising corn and soybeans. Of course it is not uniform, and significant variation exists within it, but it is nevertheless useful as a way of simplifying what otherwise might be overwhelming detail. Similarly geographic data is often abstracted or generalized, omitting detail in the interests of reducing volume, such as detail below a specified spatial resolution. Geographic data may also be sampled, on the principle that the phenomena in the gaps between the samples are likely to be similar to those at the sampled points. *Spatial interpolation* makes use of a range of techniques to estimate values of properties such as elevation or atmospheric temperature between sampled points.

In addition, researchers and others have frequently addressed the volume problem using techniques that are generally termed *divide-and-conquer*. Landsat data, for example, is acquired and stored as a series of approximately 50,000 *scenes*, each covering an area of about 100 km by 100 km, and when combined providing complete coverage of the Earth's 500,000,000 km^2. For the Thematic Mapper sensor each scene contains about 3000 by 3000 cells, each roughly 30 m by 30 m. In order not to overload storage capacity, much research using Landsat proceeds one scene at a time. The weakness of this approach stems from the inability to identify and examine patterns that extend from one scene to its neighbors—but the benefit lies in the ability to process Landsat with modest computing facilities. In summary, standard techniques widely practiced across the sciences have long made it possible to process Big Geodata using conventional means. It follows that the kinds of novel, unconventional computing described later open numerous opportunities for new discoveries.

This entry is organized as follows. The next section discusses concepts that are related to Big Geodata, and broader but related trends that are impacting science and society. This is followed by a section on the disruptive impacts of Big Geodata, and by another on the technical advances associated with Big Geodata. The final section addresses the research issues that Big Geodata raise, and the prospects for their resolution in the near future.

1.03.2 Related Concepts

Big Geodata, with their characteristics of volume, velocity, and variety, have appeared at a time of major disruption in both science and society. Some of the more important and relevant of these are discussed in the following subsections.

1.03.2.1 Data-Driven Science

The volumes of data now being captured in digital form from sensors, satellites, social media, and many other sources have led to the suggestion that we are on the verge of a new era of *data-driven science*. Instead of the often ponderous process of theory-building and theory-testing, we should rely on analytic tools to discover patterns and correlations, and thus to make discoveries. This concept

has often been termed the *Fourth Paradigm* (Hey et al., 2009), emphasizing a progression in the history of science from empirical observation to theory to simulation, and now elevating data to the primary role. In the world of data-driven science there would be no need for theory, as methods of machine learning and artificial intelligence would be sufficient to search for and discover all important patterns. Miller and Goodchild (2014) discuss the prospects for a data-driven geography based on geodata.

The notion of automated pattern detection predates Big Data by several decades. Dobson (1983) was arguing for an automated geography in the 1980s, and techniques from artificial intelligence such as artificial neural nets (e.g., Schalkoff, 1997) and self-organizing maps (Agarwal and Skupin, 2008) are widely used. But while such techniques are elegant ways of finding patterns, in the absence of theory they provide no basis on which to interpret those patterns, and no basis for assuming that the patterns discovered from one sample of data or one geographic area can be generalized to other samples or areas. Similarly they may be successful at predicting certain events, but they provide no reason to expect that they will always be equally successful.

Moreover, the underlying mantra of data-driven science, that we should "let the data speak for themselves," assumes that the data are perfectly representative of reality, and thus that what is discovered about the data is also being discovered about reality. But measurements are always subject to measurement error. The positions that are an essential element of geodata are measurements, and moreover geodata are universally subject to uncertainties of many additional kinds, as noted earlier. Thus if we "let the geodata speak for themselves" we are never at the same time letting geography speak for itself. The patterns and correlations discovered in this way may be true of the real world, but they may also be spurious and uninteresting artifacts of the data. Moreover the data may miss aspects of the real world that are essential to understanding. For example, if the data were acquired by satellite imaging with a spatial resolution of 100 m, any important features or patterns with spatial resolution much finer than 100 m will be effectively invisible, and undiscoverable from the data.

1.03.2.2 Real-Time Analytics

With increasing volumes of data available in near-real time, Big Geodata appear to offer the possibility of a new kind of activity that is very different from the somewhat protracted and even leisurely traditions of geodata analysis. Rather than spend as much as 2 years gathering data, conducting analyses, writing and finally publishing the results, it is now possible to imagine geodata being continuously monitored, providing early warning of such events as disease outbreaks or earthquakes, and making discoveries in minutes that might previously have taken months. Moreover the Internet provides the means for almost instant dissemination of results.

Yet while early warning can clearly be extremely valuable, the broader implications of velocity for science are more challenging. Science has always given highest value to knowledge that is true everywhere and all times (*nomothetic* science). Thus there would be little value given to the discovery of some principle that was true only at certain times, in certain places (*idiographic* science). Such discoveries might be described by many scientists using such pejorative terms as "journalism" or "mere description." But while this attitude might be prevalent in the physical sciences and perhaps even in the environmental sciences, the situation in the social sciences is more nuanced. In recent decades the concept of "place-based analysis" has received significant attention, in the form of techniques such as local indicators of spatial association (LISA; Anselin, 1995) and geographically weighted regression (GWR; Fotheringham et al., 2002). Such techniques are driven by the notion that while a single mathematical principle may be (more or less) true everywhere, the parameters of the principle (e.g., the constants in a linear regression) may vary from place to place.

1.03.2.3 The Changing Nature of Science

Early science was dominated by the lone investigator, the researcher who "stood on the shoulders of giants" in the words of Isaac Newton, to create new knowledge through empirical investigation or theoretical reasoning. Science was organized into disciplines, with the underlying assumption that the giants in any area were members of a researcher's own discipline, or one closely related to it.

That model worked well for the likes of Newton, Darwin, or Einstein, as long as there were major advances to be made by solving comparatively simple problems. Today, however, there is a growing sense that the simple problems have been solved, and that future progress in science must involve researchers from many disciplines, working in teams, and untangling the many aspects of complex systems. Science is becoming multidisciplinary, with teams that integrate the contributions of many individual investigators. One consequence of this emerging pattern of science is that no one individual member of a team is able to know and understand all aspects of the study. Yet scientific methodology, which emerged in the days of Newton, Darwin, and Einstein, made every investigator responsible for all aspect of his or her work.

For Darwin, for example, it was essential that virtually all of the observations that led him to develop the theory of natural selection were made personally by him. To be sure the prior work of others, including von Humboldt and Wallace, was influential, but in no sense did Darwin have to put his trust in data collected by others. This new world of collaborative research and the sharing of data is challenging, and threatens to disrupt the very foundations of the scientific method. Metadata, and documentation generally, may be held up to be the answer, but however complete they may be, metadata are never a perfect substitute for personal engagement in data acquisition.

Science is also becoming more computational, and the computer has become an indispensable part of every project. Much of the software used in a project was likely written by someone who was not one of the team of investigators, and the data may have been acquired from a source outside the team, without complete documentation of the data's provenance. The effect is that science

conducted in this environment of multiple investigators, acquired data, and acquired software may not be fully replicable. Thus the changing nature of science, and especially science based on Big Geodata, may no longer be adhering to the long-established principles of the scientific method.

1.03.2.4 Open Data and Open Software

In a small group, collaborating scientists will be able to build a level of trust in each other's work and expertise. Ideally each member of the group should be able to question openly the work of the others, so that in effect the principle of individual responsibility for science is transferred into a principle of group responsibility. But open sharing of data and software, often between individuals who will never come into direct or even electronic contact, makes this principle much less tenable. The open-data movement, which advocates publication and widespread dissemination of data, cannot possibly establish the same level of trust that is achievable in a small group of colocated collaborators. In principle, open data should be accompanied by complete documentation of provenance and quality, but this is rarely the case. Moreover the level of detail required for adequate documentation of provenance expands along with the bounds of dissemination: for example, the data-collection practices of one discipline may need much more detailed explanation if the data are to be shared with members of another discipline. If an *information community* is defined as a community that shares terminology and practices, then sharing of data and software across information communities clearly requires greater documentation and care than sharing within an information community. Especially problematic is the case where data or software originates in the commercial sector, which may well be concerned about its proprietary interests and less willing to share details of the data's provenance or the software's detailed algorithms.

1.03.3 Disruptions

It is clear from the discussion thus far that Big Geodata is capable of being disruptive or transformative, that is, of changing traditional practices. This section examines some of those disruptions in greater detail.

1.03.3.1 Publication

In the traditions of science, the results of a study are distilled into one or more papers or books, which are then reviewed by peers, edited, published, distributed, and made available in libraries. There has always been pressure to speed the process, but because so many stages involve humans it has been difficult to reduce the time of publication to much less than a year. Yet the "velocity" aspect of Big Geodata and the near-instantaneous communication offered by the Internet are having disruptive effects on the publication process. Moreover the "volume" aspect is removing many of the constraints on the amount of information that can be published.

Once papers, books, maps, and atlases had been published and printed, their contents necessarily remained constant (though they might be subjected to later publications of errata, and a single copy might be modified by personal annotation). In this sense there was a clear end to a specific process of scientific investigation or data compilation. On the Internet, however, velocity has come to mean the demise of that clear end, as results can often be disseminated in draft form prior to full review, and modified later to reflect improvements in the author's thinking or new results. Online maps can be modified instantaneously as new geodata become available. Today we are surprised when a restaurant found using an online service turns out to have been closed, whereas a generation ago few maps bothered to show information that was likely subject to change.

Variety is also having a disruptive impact on publication. The processes of editing and peer review were largely successful at ensuring that published information was correct. Today, however, there are few if any checks on the validity of information that is published through social networks, blogs, Wikis, and other Internet-era media.

Finally the advent of Big Data implies the removal of effective constraints on the volume of information that can be published. Traditional publication involved an intensive process of distilling, given the cost of the process and the limited capacity of books and papers. Books were effectively limited to a few hundred pages, and papers to a few thousand words, making it inconvenient to publish extensive raw data, complex software, or the detailed results of analysis. Today investigators are expected to make their data and software available through the Internet, so that replication or re-analysis by others becomes much more feasible, subject of course to issues of confidentiality and intellectual property. In short, the arrival of Big Data has dramatically disrupted the publication process.

1.03.3.2 Production of Geodata

The traditional processes of acquiring, compiling, publishing, and disseminating geodata were expensive, slow, and often manual. There was a high fixed cost of entry into the production process, with the result that only well-financed government agencies and large corporations could be producers. Maps and other products were designed to serve as many purposes as possible, for as long as possible, in order to offset the high costs. Thus the phenomena that were captured in geodata tended to be those that were comparatively static, and broadly useful.

Beginning in the early 1990s, with the advent of mapping tools on personal computers, the costs of entry into the process of geodata production began to fall, eventually to close to zero. Economies of scale were no longer as important, and it became

possible to make maps of almost anything, for purposes that were often very specific. Maps could be centered on the individual, rather than on an abstract system of tiles. Maps could be oriented to show the view from close to the ground, rather than from vertically above. Maps could be used to represent data that were valid only at the time of collection, rather than valid for an extended period into the future. Moreover maps could be made by anyone equipped with a personal computer, some software, and data gathered personally through the use of GPS and a variety of techniques typified by the camera flown on a personal drone, or downloaded from numerous Web portals. Turner (2006) has termed this *neogeography*, a new and disruptive world in which the individual citizen is empowered both to collect and to use geodata, and in which the old distinctions between the expert professional and the amateur are no longer as significant.

Neogeography is a highly disruptive impact of the advent of Big Geodata, with its volume, velocity, and variety. It calls into question the value of long-established practices in the production of geodata, and the authority of mapping agencies and corporations that has for generations been the source of trust in published data. Geodata can now be *crowdsourced* (Sui et al., 2012), providing a very competitive product that benefits from the efforts of individual citizens rather than those of a small number of professional experts.

1.03.3.3 New Questions

Big Geodata differs from its precursors in many significant ways. First, it offers the possibility of finer spatial resolution. Widely available GPS tools can determine location to 10 m with the simplest and cheapest devices, and to decimeters with more advanced versions. Satellite imagery now offers spatial resolutions of well under 1 m, giving far more detail than the comparatively coarse resolutions of the past. When rules protecting confidentiality allow, cities can be studied and simulated at the level of the individual rather than in the aggregate, and new questions can be asked about human movement patterns and interactions. For example, the tracking of taxis and cell phones is being used to reach new understandings of the social structure of cities, and to build new models of the transmission of disease or the problems of evacuation during disasters. Fine-resolution imagery can be used to monitor crops around the world, and to observe the destructive impacts of earthquakes and hurricanes.

Similar disruptions to past practice can be attributed to improved resolution in time, and to the benefits of near-real-time data. The average citizen now has access to current traffic congestion, and to tools that allow routes around congestion to be found. Weather maps of near-real-time data allow us to watch the development of storms and monitor the distribution of rainfall. All of these examples illustrate how improving temporal resolution is changing the way we do things, and vastly improving the information to which we have access.

In some instances improved spatial and temporal resolution allows us to make better decisions and to correct the mistakes of the past that may be attributable to the coarse nature of the data then available. But the truly interesting impacts are on the new questions that improved data, and the technology of Big Data, allow us to ask and investigate. Here, however, it is necessary to confront the essential conservatism of science, or what Kuhn (1970) has called "normal science." In Kuhn's analysis, science continues along a largely predictable path until it is transformed in some way. Such transformations may be triggered by new data, new tools, new theories, or new concepts. Often such new ideas come from outside a given discipline, and must therefore fight an uphill battle against the established norms of the discipline.

1.03.3.4 Consumerization

As noted earlier, a central tenet of neogeography is that the relationship between amateur and professional expert is becoming blurred. The average citizen is now empowered both to consume and to produce geodata. Yet the tools and technology of geodata are not designed for this emerging world. Locations are defined using coordinate systems, such as latitude and longitude, that are not part of the everyday life of the citizen. Instead, people tend to learn about and describe the geographic world in terms of named places, which may range from entire continents ("Asia") to rooms in one's home ("the kitchen"). Places lack the precise boundaries of officially recognized places, such as cities or counties, and are often culturally, linguistically, or context-specific. Thus "Los Angeles" has different meaning to an Angeleno, a New Yorker, or a resident of China; and "The English Channel" and "La Manche" are used to refer to the same feature by the British and the French respectively.

Traditionally it was necessary to limit and standardize the names of places, through the work of national toponymic committees. In the world of Big Geodata, however, there is ample potential to capture and represent the vernacular names of places, and their associations, and to provide interfaces to geotechnology that accommodate the ways people work with geography. Thus the interface to Google Maps, for example, includes a wide range of names that lack official recognition and would not have appeared on traditional maps. Researchers have developed techniques for identifying references to places in text, and for linking such references to maps (see, e.g., Li et al., 2013; Jones et al., 2008).

1.03.3.5 Spatial Prediction

Much of the enthusiasm for Big Data emerged in the world of commerce, where vast new sources of data began to provide a basis for useful prediction. For example, in a celebrated paper O'Leary was able to predict the winner of the Eurovision Song Contest from Big Data. Other work has been more skeptical, especially about the generalizability of such predictions, and about the general lack of attention to uncertainty, but the potential benefits of such predictions in the commercial world are undoubtedly huge.

As noted earlier, prediction is not generally a highly valued activity in science. Moreover the very basis of the term "prediction" implies a concern with time, not space. Thus the advent of Big Data has the potential to change the balance between discovery and prediction in science, raising the latter from a somewhat peripheral to a central role.

Traditionally prediction has meant estimation of *what* will occur *when*. Spatial prediction might similarly be defined as estimation of *what* will occur *where*, and perhaps also *when*. In line with scientific norms, however, both forms of prediction have been given little attention. Some exceptions can be found in the literature of geographic information systems (GIS). For example, the geologist Bonham-Carter was concerned with prediction of where gold deposits were likely to be found in Canada (Bonham-Carter, 1991) based on layers of geologic data. Lloyd and Greatbatch (2009) were able to unravel the clues spread throughout the novels of P.G. Wodehouse to predict the geographic location of the imaginary Blandings Castle. Common practical applications of spatial prediction include estimates of real-estate value based on the characteristics of the house and its neighborhood.

The second V, velocity, is likely to make spatial prediction much more valuable by opening the possibility of early warning in near-real time. That, together with the volume of data now available and the variety of its sources, suggests that it would be worth developing more extensive and sophisticated tools for spatial prediction in this era of Big Geodata.

1.03.4 The Technology of Big Geodata

As noted earlier, part of the drive toward Big Data is technological. Big volume requires high performance, and the use of the world's most powerful computers. But velocity and variety also create technological challenges, as discussed later.

1.03.4.1 High-Performance Computing

We now have the ability to collect, process, store, and disseminate unprecedented quantities of data. We now have storage for petabytes and access to supercomputers that operate at petaflop rates (1 petaflop is roughly 10^{15} floating-point operations per second). Even the cheapest personal computers now employ multiple processors, while the number of CPUs (central processing units) and GPUs (graphics processing units) in the largest supercomputers is increasingly in the thousands or even millions.

Several papers have discussed the potential of such massive computation for the geographical sciences. Operations that were impossible or would have taken prohibitively long on earlier devices are now within the capabilities of today's supercomputers. These tend to be problems at fine spatial resolution and covering large areas, with algorithms that are not readily amenable to divide-and-conquer.

For example, the US Geological Survey faced a significant computational problem in its National Map project when dealing with the rasterized versions of its 1:24,000-scale topographic maps. These maps use a projection with several distinct zones, such that pairs of maps that cover adjacent areas that lie in different zones will not fit cleanly along their common border. The problem does not arise in the case of vector data, so vector features such as roads and coastlines continue as expected across zone boundaries—but for raster data, early versions of the National Map showed unacceptable "rips" in the raster data. Reprojection of the raster data on the fly was computationally intensive, but a solution was found ("pcRasterBlaster"; Finn et al., 2012) in high-performance computing through parallelization of the reprojection algorithm.

The CyberGIS project headed by the University of Illinois at Urbana-Champaign, now in its fifth and final year of funding from the National Science Foundation, has systematically explored the application of high-performance computing and research collaboration for problems in the geographical sciences. As such, it represents a major investment in the technology that is making it possible to handle Big Geodata, and to explore the new questions that Big Geodata allows researchers to investigate. The possibilities of applying high-performance computing to geospatial problems have long intrigued researchers (Healey, 1998), but have remained tantalizingly out of reach for a variety of reasons. Today, however, it seems that the age of high-performance GIS has finally arrived.

1.03.4.2 Synthesis

The transition from a world of single, authoritative sources to multiple sources of often unknown provenance has drawn attention to the general lack of tools for the synthesis of Big Geodata. In the past, the multiple sources of data that might contribute to a geographic fact, such as the elevation of Mount Everest, have been largely compiled and synthesized by experts. Thus the estimate of 8848 m is the result of a long series of measurements of increasing accuracy, extending over more than a century, from early triangulation from the plains of India to the latest techniques of photogrammetry, radar interferometry, and satellite-based positioning. The trust we place in that estimate derives from the authority of the experts and their mapping agencies.

Things could not be more different in the current neogeographic world. A search of the Web would produce perhaps thousands of pieces of information that might bear on a given fact; and the highly paid experts that in the past would have synthesized those pieces no longer exist in sufficient numbers to conduct the painstaking synthesis of every needed fact. Instead we are forced to rely on automation, in the form of techniques of fusion that create estimates from raw inputs. Although numerous techniques have been described, they are not yet at the level of availability to rival our techniques of analysis. In short, the geospatial world remains locked in a paradigm of analysis of single authoritative sources, rather than one of the synthesis of an abundance of sources of highly variable quality. To cite one very simple example, everyone knows how to compute a mean and many could estimate the uncertainty of the mean under common assumptions, but far fewer are trained in how to produce an optimum estimate from a number of sources of varying reliability, or how to estimate the uncertainty in the result.

1.03.5 Conclusion

In recent years Big Data has captured the imagination of many researchers, and there has been very rapid growth in the demand for data scientists. Big Geodata is a well-defined subset of Big Data, sharing many of its concerns and priorities. It is also extreme in some respects: in the importance of uncertainty and the impossibility that any geodata can represent the truth; and in the efforts that have been expended in the geospatial community over the past quarter century on topics such as metadata, provenance, data sharing, interoperability, archiving, and other concerns of data science.

There can be no doubt that Big Geodata will continue to grow in significance. Volume, velocity, and variety will continue to increase, taking advantage of broader developments in information technology and in the fundamental technologies of geodata. Spatial prediction offers some very significant practical applications. Progress will continue to be made in data integration and synthesis. Some of those problems will be solved. On the other hand if one defines Big Geodata as having volume, velocity, or variety that is beyond our current ability to handle, then Big Geodata will continue to remain just beyond our reach, and a major invitation to cutting-edge research.

References

Agarwal, P., Skupin, A. (Eds.), 2008. Self-organizing maps: Applications in geographic information science. Wiley, Chichester, UK.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical Analysis 27 (2), 93–115.

Bonham-Carter, G.F., 1991. Integration of geoscientific data using GIS. In: Maguire, D.J., Goodchild, M.F., Rhind, D.W. (Eds.), Geographical information systems: Principles and applications, vol. 2. Longman, Harlow, UK, pp. 171–184.

Dobson, J.E., 1983. Automated geography. The Professional Geographer 35 (2), 135–143.

Finn, M.P., Liu, Y., Mattli, D.P., Guan, Q., Yamamoto, K.H., Shook, E., Behzad, B., 2012. pRasterBlaster: High-performance small-scale raster map projection transformation using the Extreme Science and Engineering Discovery Environment. In: The XXII Congress of the International Society for Photogrammetry and Remote Sensing. Melbourne, Australia.

Fotheringham, A.S., Charlton, M., Brunsdon, C., 2002. Geographically weighted regression: The analysis of spatially varying relationships. Wiley, Hoboken, NJ.

Healey, R.G., 1998. Parallel processing algorithms for GIS. Taylor and Francis, Bristol, PA.

Hey, A.J.G., Tansley, S., Tolle, K.M., 2009. The fourth paradigm: Data-intensive scientific discovery. Microsoft Research, Redmond, WA.

Jones, C.B., Purves, R.S., Clough, P.D., Joho, H., 2008. Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science 22 (10), 1045–1065.

Kuhn, T.S., 1970. The structure of scientific revolutions. University of Chicago Press, Chicago.

Li, L., Goodchild, M.F., Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and Geographic Information Science 40 (2), 61–77.

Lloyd, D.A., Greatbatch, I.D., 2009. The search for blandings. Journal of Maps 5 (1), 126–133.

Miller, H.J., Goodchild, M.F., 2014. Data-driven geography. GeoJournal 80 (4), 449–461.

Schalkoff, R.J., 1997. Artificial neural networks. McGraw-Hill, New York.

Sui, D.Z., Elwood, S., Goodchild, M.F. (Eds.), 2012. Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice. Springer, New York.

Turner, A., 2006. Introduction to neogeography. O'Reilly, Sebastopol, CA.

Zhang, J.X., Goodchild, M.F., 2002. Uncertainty in geographical information. Taylor and Francis, New York.

1.04 Current Themes in Volunteered Geographic Information

Colin J Ferster, University of Victoria, Victoria, BC, Canada
Trisalyn Nelson, Arizona State University, Tempe, AZ, United States
Colin Robertson, Wilfrid Laurier University, Waterloo, ON, Canada
Rob Feick, University of Waterloo, Waterloo, ON, Canada

1.04.1 Introduction

Data collection can be expensive, and sometimes practical constraints limit the ability to map in detail or keep maps up to date. Take the example of collecting data about species distributions of birds. Traditional surveys can provide detailed data on bird locations, but given the limited number of trained people on research teams, the large number of species, and the wide migratory distributions, it is impossible to get a detailed and comprehensive record of bird locations over a large area and a long period of time using standard approaches. Roads are another great example of phenomena that are difficult to map because they are always changing. Imagine trying to keep road maps up to date in a city that is developing quickly. Maintaining road datasets requires constant updating by city technical staff. By the time the map is completed it is out of date. Furthermore, the data may not be available for the public to access, use in the way it wants, or make edits and additions based on first-hand knowledge. With the growing popularity of mobile devices equipped with location sensors, there has been increasing demand for geographic data in applications and the possibility for individuals to use these tools to gather data. Issues of data timeliness, limitations in spatial and temporal resolutions, restrictions on use of official data, and difficulty capturing large spatial extents have led to growing interest in having individuals be part of networks for collecting data. When linked to a map, citizen science or crowdsourced data is called volunteered geographic information (VGI). VGI can be defined as individuals using the Web to create, assemble, and disseminate geographic information (Goodchild, 2007).

VGI's popularity is also supported by new feasibility for map data collection made possible by major advances in the way people create, view, and use maps. Improvements in mobile computing, in particular, personal communication devices (e.g., smartphones) have made digital maps accessible to an unprecedented number of people. Smartphones are increasingly common, and many are equipped with global positioning systems (GPS) to measure location, in addition to tools for capturing images (camera), recording text, sound, and other input (touch screen and microphone), and sharing over networks (data connectivity). Social networks are also important in that they connect many people to facilitate information sharing. It is now common to have measurements of location attached to messages and images shared over social networks (e.g., geotagged Tweets or images on Flickr) (Robertson and Feick, 2015) or for scientists to engage untrained volunteers in collecting and analyzing data (citizen science) (Haklay, 2013). Many aspects of scientific inquiry are geographic in nature, and therefore relate to VGI. Never before have as many advanced mapping tools been available to so many people, even while often this mapping task is hidden, creating a growing cadre of "accidental geographers" (Unwin, 2005).

Geospatial tools have embedded geography into people's lives. While generating geographic information may be a passive activity (e.g., a city routing app. tracks your progress while you drive through traffic to calculate the fastest route), there are also intentional efforts to collect data (Harvey, 2013). Scientists increasingly turn to citizen engagement to enhance data collection and outreach efforts in their communities (Dickinson and Crain, 2014). Also, the collection of technologies and methods known

as the "geoweb" create opportunities for people to rapidly collaborate on geographic issues of interest. Because tools are available to so many people, there are opportunities for many people to contribute, and VGI has several possible advantages. First, because there are opportunities for so many people to contribute, data collection can be more spatially and temporally extensive than traditional approaches coordinated through a central organization (Sui et al., 2013). Social media tools can engage an audience who is interested, experienced, skilled, and invested in a topic to contribute observations, interpret results, and share findings (Cooper and Lewenstein, 2016). Therefore, VGI can represent unique perspectives, experiences, and knowledge. Funding limitations for traditional government mapping initiatives may lead organizations to look outward for help keeping data up to date or motivate people to collect their own data to meet a need. VGI may cover topics that are difficult, or impossible, to address using traditional approaches (Nelson et al., 2015). VGI can be more rapidly responsive and adaptive to local needs than centralized efforts (Goodchild and Glennon, 2010).

Despite the strengths of VGI, research has focused on data collection at the expense of information use and decision making, and many concerns remain in these areas. An obvious first concern is related to the quality of data contributed by volunteers, as often no credentials or training are required to participate in a project (Burgess et al., 2016). Another concern is that VGI may represent an incomplete segment of society (Romanillos et al., 2015). Barriers may be in place to having everyone represented; for example, many people do not have access to smartphones, computers, and networks due to available infrastructure and personal cost, the ability to effectively use the tools, free time to participate, and motivation to take part (Sanchez and Brenman, 2013). Addressing these concerns will increase the ability to use VGI to inform science and management, lead to new discoveries, and represent a wider range of experiences than traditional approaches.

The purpose of this article is to introduce and discuss major themes in VGI, present three recent case studies where geographers used the Internet and smartphone tools to engage wide audiences using VGI. Our discussion of the themes of VGI is divided into two sections: the process of generating VGI and the resultant data products. Related to the process of generating VGI, first we discuss definitions, history, types, motivations of both participants and project organizers, and potential barriers to participating in VGI projects. Related to VGI products, we discuss data quality, data ownership, and the use in governance, in particular for cities. In the next section, we present three recent VGI case studies that demonstrate considerations for applying theory to real-life VGI projects. In the final section, we relate the projects back to the themes discussed in the article and summarize the main points.

1.04.2 Themes

1.04.2.1 Process of Generating VGI

1.04.2.1.1 History

It has long been a goal of some geographers to include a wide variety of people and interests in maps (Miller, 2006). Local people have local knowledge, are geographically close to a phenomenon, and may dedicate effort to topics and outcomes they are directly invested in, perspectives that may be missed by more centralized approaches (Feick and Roche, 2013). Devices like smartphones and personal computers are equipped with input tools for text, sound, images, location, and movement. Networks seamlessly connect individual devices, online tools, and extensive "cloud-based" storage. In the process of using these tools, masses of data are generated, much of which is geographic (Sui et al., 2013). These tools have presented new opportunities for collecting and sharing geographic information.

Our understanding of the term VGI has evolved to embrace both a spectrum of types of geographic data that citizens create and share, as well as a growing range of technologies and social practices that enable these data and information resources to be created (Elwood et al., 2012). When viewed as geographic data, VGI can be seen to include both citizens' objective recording of their environments (e.g., local stream temperature readings, counts of amphibians) and more subjective information that relates to their perceptions of opinions of places and features (e.g., georeferenced narratives). VGI can reference locations in either an explicit (e.g., geographic coordinates) or an implicit (e.g., vernacular regions such as "downtown") manner. The data that individuals collect can vary substantially in format and include data that correspond with traditional GIS data, such as points, lines, and polygon features in OpenStreetMap that are characterized with descriptive text tags (Haklay, 2010). Other types of VGI, such as geotagged photographs (e.g., Flickr), text messages (e.g., Twitter), and videos (e.g., YouTube) are products of communication and present new opportunities to document and analyze human and natural phenomena (Shelton et al., 2015; Quesnot and Roche, 2014).

1.04.2.1.2 Types of VGI

Diversity can also be seen in the processes that underlie how VGI is created and used. Stefanidis et al. (2013), for example, make a useful distinction between VGI that individuals create deliberately and data that are generated passively as a byproduct of other activities. Active VGI is characteristic of citizen science activities as well as more routine municipal issue reporting (e.g., graffiti, potholes) where citizens are engaged in deliberately collecting information for a set problem or interest (e.g., bird sightings) and usually across a predefined set of variables (e.g., species, sex, time observed). In this way, active VGI projects seek to enlist citizens in helping experts to collect, curate, and share information that can be used to monitor environmental conditions and/or address applied research questions. In contrast, ambient and passively generated VGI are typically created without user intervention as an outcome of another process or activity. Considerable attention has been directed at examining how communication data, such as Twitter, microblog texts, and geotagged photographs, and videos, can be used to infer new insights about human behavior, movement, and perceptions (Li and Goodchild 2014; Shelton et al., 2015; Robertson and Feick, 2015).

Even within these bounds, the diversity of VGI data and authoring processes introduces interesting challenges and opportunities to further research. In terms of challenges, VGI producers often have different reasons for creating data and engaging in a VGI project. Some individuals, for example, may be interested in enhancing their personal reputation and status within a community, while others may have more altruistic motivations (Coleman et al., 2009). Since individuals' motivations and expertise differ, the quality of data contributions can vary substantially from person to person (Foody et al., 2013; Li and Goodchild 2014; Devillers et al., 2010). Data coverage can be uneven as more people are willing to collect data about features and places that are popular and accessible than their counterparts that are seen to be less interesting or more difficult to capture. Notwithstanding challenges of this nature, the use of VGI offers several key advantages for advancing citizen science. The three VGI examples in section "**Examples**" illustrate several of these challenges and opportunities in more detail.

1.04.2.1.3 Motivation

VGI created through an active and purposeful engagement is dependent on the participatory process. Through the evolutions of public participation geographic information systems (PPGIS) and VGI with roots in participatory planning, much has been learned about processes of participation. Arnstein's (1969) ladder of participation has been used to link degrees of participation of citizens to issues of power and control. While PPGIS projects were firmly rooted in urban and regional planning, VGI has a wider scope in terms of the types of projects and forms of participation enabled by more recent advances in geotechnologies (Goodchild, 2007).

The reasons why people participate in VGI projects are intimately tied to the project's objectives, and often many motivations exist for participants within single projects and even within single individuals (Coleman, 2009). Understanding participant motivations is a critical need for project designers that want to foster and build tools that cater to specific participant motivations. Coleman (2009) characterized user motivations in VGI, drawing from experiences in the open source software community, listing categories of motivations including altruism, professional/personal interest, intellectual stimulation, personal investment, social reward, enhancing personal reputation, an outlet for creativity and self-expression, and pride of place, as well as negative motivations of mischief, having a hidden agenda, or even criminal intent. These user motivations are tied to the nature of the information contributed, and may be a predictor of data quality. Note also that motivations are not static and can and will change throughout the life of a project (Rotman et al., 2012). For example, eBird, a popular citizen science project in ornithology, changed slogans from "Birding for a Cause," which aimed to engage the altruistic motivations of volunteers, to "Birding in the 21st Century" with a focus on providing digital tools for birders to become better at their hobby. This change was associated with large increases in the number of contributions and improvements in the quality of contributed data (Cooper and Lewenstein, 2016). In projects that solicit citizen reporting of plants and animals, many people may be motivated to contribute data simply because they understand how better data may lead toward improved research, decision making, and/or conservation efforts. Researchers can therefore encourage submissions by using the data for research, publishing papers, presenting at scientific conferences and developing knowledge mobilization activities that translate the findings back to the participant community. User motivations also relate to mechanisms that can be built into a VGI project, such as data standards and sampling designs (Goodchild, 2009).

A second and much less theorized aspect of VGI relates to the motivations of project designers or researchers. For citizen science, the most widely expressed motivation is expanding sampling effort through the use of citizen data collectors (Dickinson et al., 2010). However, research objectives may extend beyond data collection, for example, testing web-based participation tools (Nuojua, 2010) or evaluating spatial cognition tasks, in which case researcher and participant motivations may not align explicitly, and strategies such as gamification might be employed to target a general class of participants. Deeper participation in citizen science takes an approach "higher in the participatory ladder," whereby participants are engaged in defining project objectives and how and what data gets collected (Haklay, 2013). Participatory action research may have much to inform citizen science and VGI more broadly in this regard, which has a long history of linking researcher and "practitioner" interests in research projects (Argyris et al., 2002). While participatory action research is rooted in social research, citizen science and VGI encompass both social and natural science research questions, often at scales not possible within the participatory action research model (Cooper et al., 2007). The continuum of control for VGI defines power structures that influence how actors in the project relate to each other.

The interlinkages between user motivations, researcher and/or designer objectives, data quality characteristics, and project design choices can ultimately determine the characteristics of a VGI project and the information it produces. In the case studies investigated here, we see several motivations at play for participants. In the case of RinkWatch (see section "**RinkWatch**"), research objectives were both educational/outreach in nature (linking climate change to meaningful cultural ecosystem services) and to provide data for research relating temperature variability to outdoor skating. Through qualitative interviews of participants from RinkWatch, many participants revealed their motivation was driven by interest in outdoor skating. Outdoor skating, and in particular rink-making, is an activity individuals engaged in individually and RinkWatch served as an online community. In response to this realization, researchers implemented several bulletin boards to cater to these users, giving them a forum to exchange ideas and tips related to rink-making and stories about outdoor skating.

1.04.2.1.4 Equity in VGI

We have addressed the benefits of VGI as a novel data source that can satisfy unmet social or scientific needs and also as a vehicle for citizens to create and share information that is interesting or important to them. However, these benefits (and any costs) are not distributed equally. Individuals, social groups, and geographic areas differ in terms of access to technical resources that enable VGI production and use (e.g., Internet connectivity, open spatial data), as well as social, economic, and societal factors (e.g., financial resources, digital literacy, sociopolitical environments, legal structures) that condition how and whether digital data and tools are used (Sui et al., 2013). These inequities in access to digital tools and the ability to use them effectively have been described as a digital divide.

Access to the Internet is an easily understood prerequisite for generating VGI. A large proportion of the world's population have limited or no Internet access due to a lack of infrastructure and the costs exceeding the income of many people (International Telecommunications Union (ITU), 2016). In North America, computer and smartphone ownership have increased dramatically, even for many in lower-income groups (Sanchez and Brenman, 2013). However, there are other nontechnical factors that impact how and if individuals can engage in VGI and citizen science projects. For example, many people do not have the free time needed to create VGI, particularly if they need to work long hours at multiple jobs or care for young or elderly family members (Wiggins, 2013). Similarly, disadvantaged groups may encounter social and educational barriers that limit their capacity to organize data collection projects and how effectively they can interact with governments through online tools (Sanchez and Brenman, 2013). While participation in VGI projects may not be possible for everyone, there is some hope that the outcomes (e.g., better data for planning) may benefit broader groups of people beyond those who directly participated.

It is important to note that even within advantaged groups in society that have capacity to contribute to crowdsourcing and VGI projects, rates of participation differ leading to participation inequality. For example, in Wikipedia, a small subset of participants is responsible for generating the vast majority of the content, while many people make few contributions or only consume content (Bruns et al., 2013; Quattrone et al., 2015). The most active contributors for OpenStreetMap have been predominantly young, male, educated, and focused their efforts to mapping urban centers (Stephens, 2013; Quattrone et al., 2015). Different types of projects can attract contributions by different groups; for example, females made most of the contributions to the citizen science project EyeWire, a puzzle-like game to map neurons (Kim et al., 2014). Questions have arisen about the consequences of groups with similar demographic profiles, and likely common experiences and world views, generating crowdsourced products that are increasingly ubiquitous in use (Lam et al., 2011). Haklay (2016) emphasized that "[w]hen using and analysing crowdsourced information, consider the implications of participation inequality on the data and take them into account in the analysis."

1.04.2.2 Products

1.04.2.2.1 Data quality

Generally, definitions of data quality relate to fitness of the data for a given use (Chrisman, 1984), and VGI have opened discussion about more complex dimensions of data quality. The standards developed through the International Standards Organization (ISO) and specifically Technical Committee 211 provide a good starting point for examining spatial data quality (ISO, 19157, 2013). These cooperatively developed standards cover data quality elements such as lineage (history of the data generation and processing), positional accuracy, attribute accuracy, temporal consistency, and completeness, among others.

Applying spatial data quality standards to VGI can be challenging. Unlike spatial data that are created by experts in government, private, or nongovernment organizations, VGI are often authored by many dispersed contributors who differ in expertise, interests, and methods for creating and documenting data (Poore and Wolf, 2013). As a result, data quality can vary from contributor to contributor within a single data set. Critiques of volunteered data quality have focused on concerns over spatial and attribute accuracy. In particular, inexperience in scientific protocols, the use of low-cost consumer devices (e.g., smartphone GPS sensors) rather than dedicated instruments, and differing motivations of volunteers as a source of bias in volunteered science (Show, 2015). These views are challenged on the basis that professional scientists also demonstrate nonobjectivity in research design and implementation, narrow views of data quality (as simple adherence to scientific protocols) are limiting, and differing approaches are enriching for the greater goal of discovery (Newman et al., 2015). Certainly, demonstrating adherence to measurement protocols can add authority to arguments backed by volunteered data (Ottinger, 2009). Technology can also be used to support curation of data to ensure unusual measurements are flagged for review and areas needing more detailed observations identified (Ferster and Coops, 2014). Similar to the measures of lineage, the dispersed contributors of VGI require strategies such as training for new participants and filtering and reviewing unusual values to ensure logical consistency of submitted data (Sullivan et al., 2014).

Another challenge for VGI is data completeness. VGI is heterogeneous by nature and the density of data contributions varies by where volunteers choose to concentrate their efforts. However, VGI also presents opportunities to collect data that are spatially and temporally extensive. In some cases, having more data that are extensive and uncertain may be more valuable than having few very accurate data, or possibly no data at all, from official sources (Hochachka et al., 2012; Goodchild and Glennon, 2010). In general, VGI projects and citizen science initiatives will likely continue to operate with fewer protocols to control data quality than traditional scientific measures. However, the volume of data opens up potential and future research should consider how to implement confirmatory approaches to allow consistent data to be highlighted when repeated reports indicated similar patterns.

Emergent considerations for dimensions of data quality are related to the extent and opportunity for contributions by a diversity of people, as the value of data can be enriched if a diversity of people have a chance to make unique contributions, chance discoveries, and opportunities to follow up on insights and chance discoveries (Lukyanenko et al., 2016). The key example in citizen science highlighted by Lukyanenko et al. (2016) was the discovery of a rare type of astronomical object by a school teacher in Holland, Hanney Van Arkel, in the Galaxy Zoo project ("Hanney's Voorwerp") (Raddick et al., 2010). The task being performed was a rather mechanical manual classification of shape within telescope images. Opportunities were provided to ask follow-up questions in an Internet forum attended to by experts, leading to a major outcome for science and the individual involved.

The connection between participating citizens and the representativeness of data is a growing area of interest for VGI. For example, in cycling research there is a concern about how technology-based VGI may exclude participants and generate biased data (Romanillos et al., 2015). Equity has been discussed in terms of access to forms of active transportation and associated health benefits for different social groups (Lee et al., 2016). There is a strong appetite for cycling route data for city planners, and novel and

crowdsourced origins are being considered. For example, Strava is a cycling smartphone game where participants can track, compare, and compete their rides (Strava, 2017). The data on Strava represent actual cycle trips, but there may be a bias toward different types of routes compared to recreational or utility cyclists (e.g., competitive riders may seek out hilly routes in rural areas for training) (Griffin and Jiao, 2015). In other areas, such as urban centers, there may be less difference between Strava users and other types of cyclists (Jestico et al., 2016). Given appropriate modeling constraints, these novel data sources can be complementary to traditional data sources by offering covariates to make estimates over larger areas (Jestico et al., 2016). Interpretation within context is important to ensure the equitable allocation of public resources for developing active transportation facilities (Le Dantec et al., 2015).

1.04.2.2.2 Data ownership and open data

Data and algorithms are increasingly important in society and the two are intimately linked; the performance of algorithms is often tied to the size and quality of the training data used to develop them. The dominant mode of big data ownership has normalized the practice of individuals trading personal information services for access and ownership to personal data. In social media, for example, each exchange implies a trade in service (e.g., posting a message to friends) for a piece of data (e.g., access and ownership of the digital representation of that message). This trade positions corporations in opposition to individual users, in what many consider to be an imbalanced relationship (Smith et al., 2012; Kitchin, 2014). In citizen science, data ownership can take a variety of forms. One model is the previous model, where participants have no or little ownership or control over the data they create. Often, in academic projects, this model is not tenable because research ethics boards require participants to be able to withdraw from projects and have their data removed from the larger database. In some cases, full data access is granted on an individual and aggregate basis.

For many users, access to raw data is less relevant than access to information products that relate to participant interests and motivations. In the case of citizen science with higher levels of citizen participation, participants can have direct roles in the management and access of data. Only in this case, where researchers and participants comanage data access policies, can the project be considered an open source citizen science project. Yet there are several important barriers to "opening" citizen science data. Firstly, many projects are designed to target a specific research or societal issue, and public access could have negative consequences. For example, reporting observations of an endangered species could incite others to seek it out, causing damage to habitat and possibly conflicting with conservation aims of the wider project. Similarly, health-related projects are particularly susceptible to ethical issues associated with open data (Goranson et al., 2013). Privacy concerns are also an issue, whereby home-based observations and usernames can be linked to other information (e.g., social media profiles) and risk harm to participants. Participation in determining data access policies is one way around this, where potential risks are discussed early and on an ongoing basis among researchers and participants (Haklay, 2013). A joint researcher–participant oversight committee is one tool projects can use to realize this level of participatory design in citizen science.

1.04.2.2.3 VGI in cities and governance

OpenStreetMap (OSM) is perhaps the most established VGI project and the focus on roads results in a strong emphasis in cities (Haklay and Weber, 2008). OSM is a community of people that map and update worldwide data on roads, which are notoriously difficult to keep current through traditional mapping workflows. Through time, OSM has grown to include other infrastructure and services. The strength of OSM is the huge number of contributors. Neis et al. (2011) estimated that in 2011 there were over 0.5 million contributors and that the number of contributors grew by 150 people each day. There is huge power in collectively harnessing local knowledge, especially involving something as dynamic as roads. OSM is a great example of a key strength of VGI: by compiling bits of data from a massive number of individuals, a new type of information is generated.

As is typical of VGI projects, a key discussion of OSM has been around the quality of data (Ward et al., 2005; Jackson et al., 2013). The quality of road data is difficult to assess given lack of data to "truth" maps. Comparing OSM data to official data is helpful for assessing congruence and variability but it is not possible to know which is right. Additional quality discussions have emphasized assessment of the amount of OSM data, which varies spatially and is influenced by access to technology and skills of the population (Neis et al., 2011). The longevity and number of applications that use OSM is an indication that even with concerns about quality, VGI datasets can be the best available and are growing from fringe data into mainstream sources.

The ubiquity of smartphones has created new opportunities for urban and city embedded VGI projects. Mobility and transportation focused VGI initiatives are becoming particularly prevalent, as personal GPS devices track where people move with unprecedented resolution (Misra et al., 2014; Griffin and Jiao, 2015). Cycling research is at the forefront of discussions about the validity of VGI as a source of data, given the plethora of fitness apps for tracking where people ride (Krykewycz et al., 2011). Personal fitness apps (e.g., Strava) are used by cyclists to track where they ride, cycling distance, and speed. With a gaming element, Strava encourages use by allowing people to compete with themselves and friends on distance and speed or undertake challenges, such as riding the elevation gain of Everest in a given time period. Perhaps unintentionally, Strava users are contributing to a massive global dataset on where people ride (**Fig. 1**). Strava Metro is now pursuing a new business model where they curate and sell the data to cities and researchers interested in ridership data. Derivative products that map ridership and visualize cyclist flow through a city are now possible. The primary criticism of using Strava is that data are primarily collected by recreational bike riders and biased toward young men (Jestico et al., 2016). However, recent research also indicates that in midsized North American cities the ridership patterns of Strava riders are similar to overall ridership patterns (Jestico et al., 2016). While there is much to be done to understand

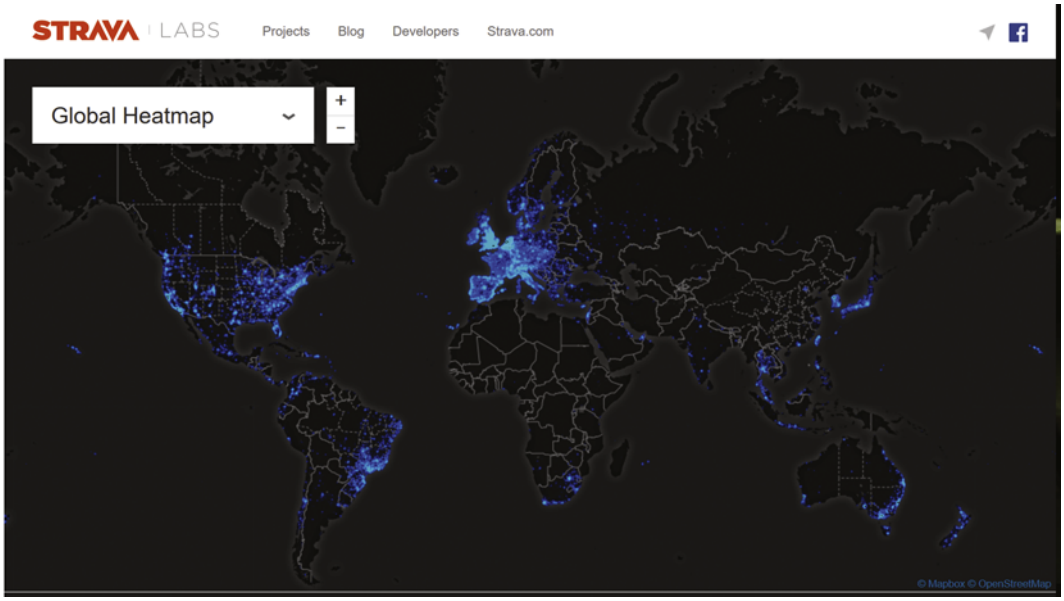

Fig. 1 Strava global heatmap. Brighter blues indicate higher rates of Strava application use to record cycle trips. *Source:* Strava.com – Global Heatmap. http://labs.strava.com/heatmap/.

the appropriate use of fitness app data as a source of VGI for urban planning and population health research, interest in data will continue given the proliferation of personal apps that leverage GPS capabilities of phones (Romanillos et al., 2016).

Emergency response and disaster management have been another focus of VGI research. There are very compelling examples of the benefit of using VGI when responding to floods (Bruns et al., 2011) and wildfires (Goodchild and Glennon, 2010). Particularly, in developing countries where authoritative data may be limited or void, such as the 2010 Haiti earthquake, volunteer data can be the only available source of information for evacuation, rescue, and recover (Meier, 2012). Beyond overcoming a paucity of data, VGI can have the advantage of real-time updating. Due to the potential for a large number of data contributors, as well as the lack of requirement to verify data before publishing, VGI can be updated much more quickly than official sources. During a disaster, this leads to both benefits and challenges (Roche et al., 2013). The positive benefit of knowing quickly what is happening and how change is occurring can be diminished if rumors or false data are provided. Ultimately, unverified data will have inaccuracies. To optimize utility of VGI, data need to be compiled and accessible, which does happen when a skilled developer of VGI quickly sets up a website to respond to an issue. One of the main strengths of VGI, the responsiveness and flexibility to provide information at the times and places where it is needed most, makes it a key tool for disaster response and management (Meier, 2012, Zook et al., 2010).

Another aspect of emergency response where VGI is having an impact is preparedness. As an example, the BikeMaps.org (see section "**BikeMaps.org**") initially launched in a city prone to earthquakes. A group of emergency responders quickly reached out to see if the technology could be broadened to support bike-based evacuation during an earthquake. Though an interesting application of VGI, cell phone service could not be guaranteed during an earthquake, making it a poor choice for this particular application. One challenge that may be met by VGI in the future is motivating community action for disaster mitigation and preparedness. Among volunteers in the ForestFuelsApp (see section "**The ForestFuelsApp**"), there were very low levels of knowledge and action for the existing wildfire mitigation programs (non-VGI), even among the highly engaged and knowledgeable audience. Salience and motivation for tasks related to wildfire preparedness are often highest immediately following a large fire event, while engaging communities at other times can be challenging (Monroe et al., 2006). Many recent wildfire apps are designed to disseminate information about active wildfires, while only a few provide information about mitigation or preparedness (Kulemeka, 2015). More broadly, there is a gap between the many natural disaster-related VGI efforts that are directed at response and the few that are directed at mitigation or preparedness (Horita et al., 2013; Klonner et al., 2016). Among traditional public outreach efforts for wildfire mitigation, interactive and participatory approaches have been the most effective, but have limited audiences (Toman, 2006). If VGI can be used to reach larger audiences, there is potential to use it as a tool to increase disaster mitigation and preparedness.

1.04.3 Examples

1.04.3.1 RinkWatch

RinkWatch is a citizen science project that engages citizens in climate change research through the reporting of ice skating conditions on outdoor community and backyard rinks. The website RinkWatch.org launched in 2012 with a simple web map interface and user

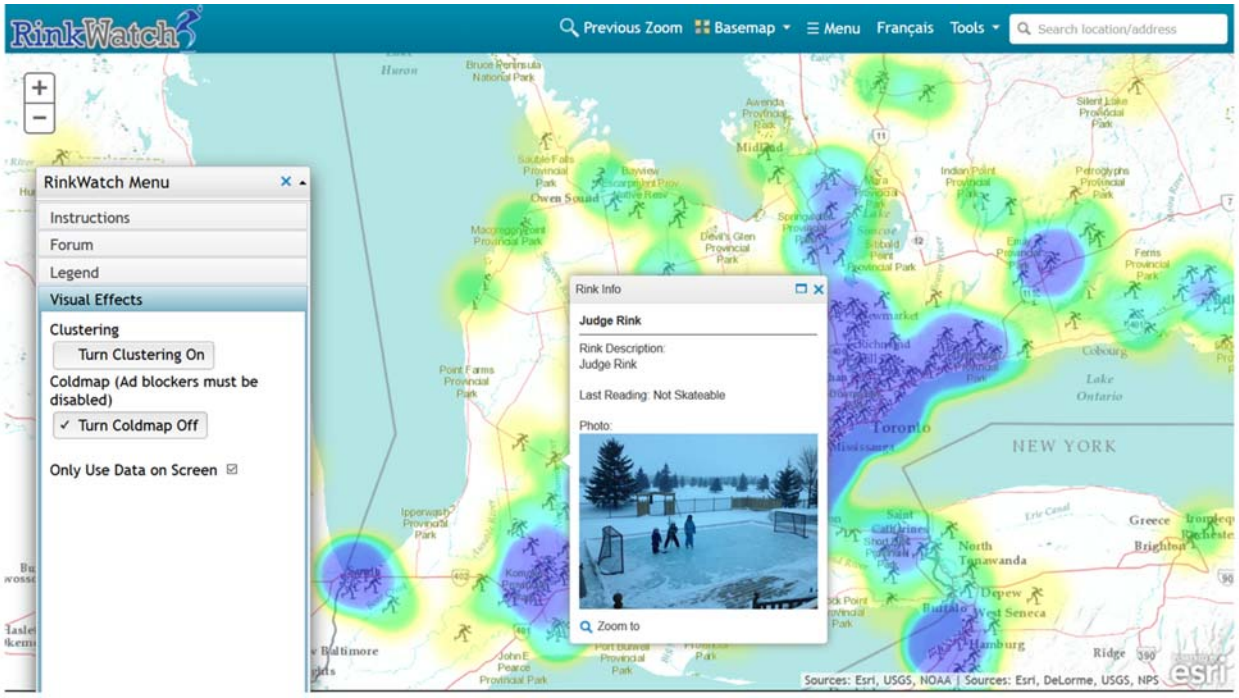

Fig. 2 RinkWatch "coldmap" showing percentage of skateable reports submitted by participants and rink information for a selected rink. *Source:* RinkWatch – http://www.rinkwatch.org/.

management system that allowed people to register with an email address, identify the location of their rink, and then to update continuously throughout the winter which days they could skate (or not) as temperatures changed (**Fig. 2**).

RinkWatch helps people make the link between climate change and impacts to their daily life. Despite increasing availability of information and growing public awareness of climate change, it remains difficult to get the general public to take actions that enhance their ability to adapt to its potential impacts (Burch, 2010). The link between climate change and the feasibility of outdoor skating was made by Damyanov et al. (2012) through a modeling study that analyzed changes in weather station data and forecasted future change to the outdoor skating season in Canada. Given the cultural importance of outdoor skating to many northern communities, RinkWatch was formed to examine climate change impacts on daily life and the culture connection to climate change through citizen science. The RinkWatch project has three interrelated objectives: to better understand how temperature changes are impacting people's ability to engage in outdoor skating, to engage and inspire interest in climate change and climate change research in a meaningful way, and to provide a testing bed for developing the "science of citizen science" (Robertson et al., 2015).

The response to the RinkWatch project was immediate and widespread, with hundreds of participants signing up in the first weeks, and over 500 in its first season. Currently, now approaching the fifth year of operation, there are over 1900 registered rinks with the project and over 30,000 skating reports. Partly responsible for this successful recruitment was widespread media interest in Canada and the northeast United States, providing over 100 media opportunities to help publicize the project. We have since been able to analyze the data in relation to local temperature records and couple relationships between skatability and temperature to climate model scenarios (Robertson et al., 2015). While these projections are derived from relationships learned from only two seasons of data, they enable the translation of climate model projections from units of temperature (e.g., change centigrade) to units of days suitable for outdoor skating (see Brammer et al., 2014 for additional empirical work on this)—a potentially more personally relevant metric to inspire changes that will reduce personal carbon footprints (Whitmarsh et al., 2011).

Since the original launch of the website, the original web map interface has been replaced with a more comprehensive web mapping framework including spatial data visualizations (e.g., "cold" maps, point clustering—see **Fig. 2**), full open access data export for individuals and for all data, time series graphs, and some expert analyses of the data in relation to local temperature data. We have also added user-engagement tools through user forums, and ability to post and share photos of rinks. We have collaborated with the sustainability arm of the National Hockey League (NHL Green) on communicating and promoting the project, and sponsored student research projects into outdoor skating and climate change.

RinkWatch has been successful in generating attention and interest in the topic of outdoor skating and climate change. As well, we have been able to leverage the data for research, showing regional variation in the skating-temperature relationship and how that may change with projected changes in climate (Robertson et al., 2015). Data quality has been assessed qualitatively through comparing plots of "skateability" to temperature recorded at local weather stations, finding generally consistent patterns and evidence to support $-5°C$ as a skating threshold (rinkwatch.org). Since we expect some variability in within-city temperatures due to microclimates (e.g., shading), the spatial variability of observations within cities is an area of current investigation. Handling variable submission rates among users has led to the development of new methods for dealing with messy, heterogeneous

observations (Lawrence et al., 2015). In future seasons we plan to investigate the notion of climate fatigue, and how winter temperature variability impacts outdoor activities (e.g., rebuilding rinks). Challenges have also been managerial in nature: keeping the website current and updated with new content and features; finding resources to support software development (rinkwatch has never been supported by research funding), and maintaining engagement with users through social media and online forums.

1.04.3.2 The ForestFuelsApp

The ForestFuelsApp. was a regional citizen science project to collect data about the fuel available to burn wildfires in the Wildland-Urban Interface (WUI), where human development meets natural areas. Populations are expanding in the WUI (Radeloff et al., 2005), and when wildfire occurs, there can be devastating human impacts (e.g., stress, injury, loss of life, and loss of homes and other infrastructure) (Gray et al., 2015). The app was tested in Kelowna, BC, where rural and scenic ideals often lead to people living in places where potential harm due to wildfire may occur. For example, the 2013 Kelowna Mountain Park Fire destroyed 239 homes and forced 27,000 people to evacuate (City of Kelowna, 2016). Kelowna, BC is located in the Very Dry variant of the Ponderosa Pine Biogeoclimatic Zone, where "[frequent low intensity] fires have played an important role in the ecology" (Hope et al., 1991). With increased development, suppression of low-intensity fires has resulted in open stands characterized by Ponderosa pine (*Pinus ponderosa* Dougl. ex. Laws.) being succeeded by more closed stands of Lodgepole pine (*Pinus contorta* var. *latifolia* Engelm. ex S. Wats.) abundant ground and ladder fuels and closed canopies. In these stands, wildfires can burn at high intensity and spread rapidly (Hope et al., 1991).

The aim of the ForestFuelsApp was to make tools to assess forest fuels loading accessible to a broader population, collect consistent data, and increase awareness of WUI wildfire issues. The approach was inspired by ocular assessment methods, which provide a rapid and accessible way to make a general assessment of forest conditions (e.g., distinguishing open and closed canopies) by comparing field conditions with reference photographs (Keane, 2013). Forms from provincial protocols (Morrow et al., 2008) were coded with reference photographs and illustrations developed by the research team (**Fig. 3**). Location was measured using the device GPS. Photographs were acquired using the device camera and accelerometer to ensure consistent framing and leveling. The device compass and accelerometer were used to measure slope and aspect. When initially opened, a brief tutorial was presented with illustrations and text describing wildfire fuel assessments and use of the app.

Eighteen volunteers were recruited from the community through media coverage, classified advertisements, posters, and contact with local hiking clubs and neighborhood associations. Questionnaires were administered before and after using the application. Smartphones (Apple iPhone 4) were provided to volunteers for testing with the ForestFuelsApp loaded. Volunteers were accompanied by a member of the research team and instructed to collect measurements at locations of their choice in the general vicinity of

Fig. 3 An example of estimating the crown closure of conifer trees using (A) reference photographs of different stand conditions and coding from official protocols, (B and C) instructions and illustrations, and (D) capturing reference imagery using the accelerometer to ensure consistent acquisition angles.

the University of British Columbia Okanagan campus, to simulate a volunteered and opportunistic dataset collected over the Internet. Observational notes were collected. Finally, the locations chosen by volunteers were revisited by the research team to collect reference measurements for comparison.

Many of the volunteers were recruited through hiking or neighborhood groups or had professional experience and interest in WUI fuels conditions (50%). While the interest from professional foresters was higher than anticipated, this provided an opportunity to both solicit professional feedback on the application, and compare differences between people with and without professional experience in wildfire and forestry topics. Both groups expressed similar motivations for volunteering in the project (the most common reason was related to values—wanting to help solve a community problem), while people with professional forestry experience more frequently expressed career motivations (e.g., learning new tools that may be useful for their job). One tension that existed was that some professional foresters expressed concerns about nonprofessionals coming to incorrect conclusions or setting unrealistic expectations for treatments that exceeded available resources. This concern was not reflected by the participants without professional experience, who generally indicated that they were more interested in helping out with the tedious task of collecting data to help their community, than setting priorities for stand treatments. Some of the volunteers indicated that they would have found more physical and demanding tasks, such as covering greater distances and submitting more data, more rewarding. The most frequently cited reward for participation was related to understanding, both related to wildfires and technology. People over the median age more frequently reported that they learned a new skill related to either the use of smartphones or wildfire management (Ferster et al., 2013).

For many of the fuel components, the volunteered measurements were consistent with the reference measurements. For measurements of slope and aspect, measurements by people without professional experience were less accurate than people with professional forestry experience, likely due to less practice with compass and inclinometer. This would be expected to improve over time. Observations of height to live crown were more consistent when made by people without professional forestry experience. People with professional forestry experience were observed to have differing working definitions of this attribute based on a range of experiences, while those without working experience more closely followed the instructions. People with no previous professional experience with wildfires collected data that covered a greater spatial extent and a wider range of conditions, while people with professional fire experience identified high priority locations near buildings with higher fire loads. For model building, the two sets of measurements were complementary (Ferster and Coops, 2014).

The ForestFuelsApp followed a very traditional volunteering approach, requiring high levels of engagement and effort from volunteers. One participant stated "tools are needed for people living in the [WUI], including communication, steps, and actions. I could see this being useful for work parties in the community." As a result, a number of volunteers were limited and volunteers were highly dedicated. While a broader audience may have been reached using less intensive activities, at the same time, highly engaged participants could have been given more demanding tasks. The implementation did not fully utilize the potential for social connectivity; for example, volunteers could not see the data collected by other volunteers to find out where more measurements were needed, and there were not opportunities to interact with other volunteers using social models (e.g., using social media to promote, connect participants, analyze data, and discuss and share results). Concerns about liability and community conflict restricted further growth. For example, there were concerns about people using the application to document fuel threats on private land where landowners lacked resources to perform treatments and may suffer financial liabilities, leading to community conflict. However, initial outcomes were promising, with useful data collected, positive responses from participants, and expression of collective goals between people with different experiences in the community.

1.04.3.3 BikeMaps.org

BikeMaps.org is a global VGI tool that is filling the gap in available data on cycling safety. It is estimated that only 30%–40% of cycling related incidents are recorded in official databases. Official databases are typically generated by police departments and insurance reports, and primarily represent bike incidents that involved vehicles. However, in a study of injured adult cyclists, treated in emergency departments, only 34% of incidents were collisions with motor vehicles and another 14% were a result of avoidance of a motor vehicle (Teschke et al., 2014). When a bike collision occurs with infrastructure, another bike, or a pedestrian there is typically no mechanism for reporting.

Another gap in cycling data is the lack of near miss reporting. Near miss events are critical for safety management in general (Gnoni et al., 2013) and have the potential to provide early warning of high-risk areas. When compared to the number of human errors or near miss incidents, a crash is a relatively rare event. Thus collecting sufficiently large near-miss databases can enable earlier detection of problematic areas (Reason, 1991) and support robust statistical analysis. Near-miss information also provides critical link to overcoming deterrents to ridership. Concerns about safety are a primary barrier to new ridership (Winters et al., 2012) and with many cities setting goals to increase ridership understanding real and perceived safety concerns are critical. In cycling, near misses can have significant physiological impacts that deter ridership (Aldred, 2016).

Through BikeMaps.org citizens can report cycling crashes, near misses, hazards, and thefts. BikeMaps.org includes a webmap, smartphone apps, and visualization tools (**Fig. 4**) (Nelson et al., 2015). Citizens identify an incident location by clicking a "submit new point" button and adding the location on the map where the incident occurred. Details of collisions and near misses are reported via a digital form through pull-down options. All reports are anonymous. The attributes captured through the pull-down menus are designed to enable research on determinants of cycling injury (Teschke et al., 2012). There are three categories of attributes: incident details, conditions, and personal details, with a balance of required and optional questions to manage citizen mapper burden. BikeMaps.org is also supported by Apps for both Android and iPhone devices. In addition to allowing mobile

Fig. 4 BikeMaps.org visualization tools. The visualizations are dynamic, adjusting with display extents as well as selection of incident types and time periods. *Source:* BikeMaps.org – https://bikemaps.org/vis/.

mapping, the Apps provide feedback to users through push notifications that alert cyclists to new mapping in their area. The website also includes a visualization page where a spatial extent can be selected and temporal trends in crashes and near misses summarized by day of the week and hour. The visualizations are dynamic enabling queries of data trends with the click of the mouse.

Launched in the fall of 2014, BikeMaps.org has over 3200 locations mapped in 35 countries (Fig. 5). The global response to the website is an indication that BikeMaps.org is filling an important gap in data available to study cycling safety. Sixty percent of

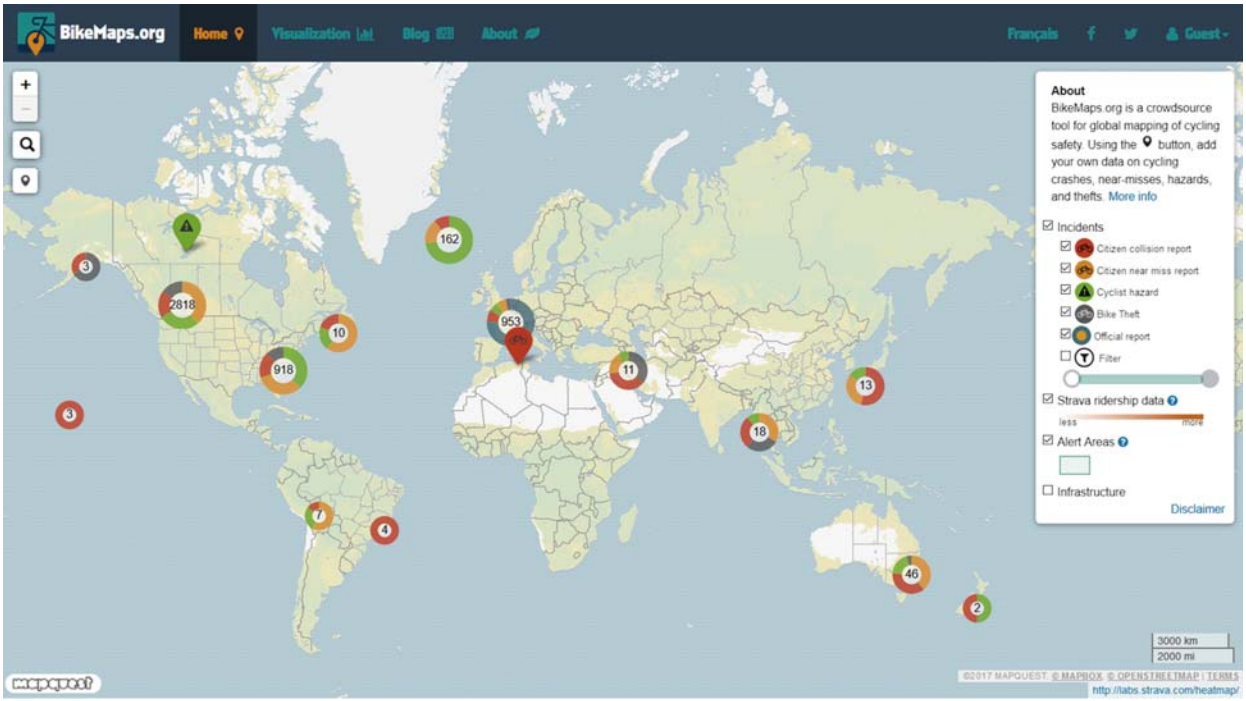

Fig. 5 Global reports of BikeMaps.org incidents. *Source:* BikeMaps.org – https://bikemaps.org/.


NEWS

BikeMaps charts course across the country

BikeMaps.org team member Colin Ferster, executive director Karen Laberee and intern Conner Leverett show off the new bike seat covers used to market the cycling app. — *Image Credit: Travis Paterson/News Staff*

Fig. 6 Local print media covering BikeMaps.org is often associated with increased use. *Source:* Paterson, T. (2016) 'BikeMaps charts course across the country', Saanich News, 19 May, p.1.

locations are mapped in Victoria and Vancouver, Canada, where outreach efforts were initially focused. In most other locations, uptake has been more organic, resulting from social and earned media. In Victoria and Vancouver, we have undertaken a range of outreach activities. The biggest gains in data are typically associated with print media (Fig. 6). While social media, Twitter (Fig. 7) especially, have been important for broader communication of our message, visits to the website and data submissions are highest when local newspapers feature a story on BikeMaps.org. Guerrilla marketing strategies have also proven effective (Fig. 8). In one campaign, we delivered 500 branded water bottles to parked bikes around the city. When cyclists returned to their bikes, they found a note to encourage them to contribute to safer cycling by mapping their experience.

A strength of the BikeMaps.org VGI project is the use of data for community engagement, research, and policy decisions. With expertise in spatial analysis, the BikeMaps.org team is able to create map products from data, such as maps of cycling safety hot spots, and these have been invaluable for ongoing engagement of users. Maps are also a great way to generate earned media as they tell a story of broad interest. As data sets increase in size, we are also using BikeMaps.org data for peer-reviewed publications on cycling safety, bringing credibility to the project. In areas where a substantial number of incidents have been reported, city

Fig. 7 BikeMaps.org Twitter feed. *Source:* BikeMaps.org Twitter – https://twitter.com/bikemapsteam?lang=en.

Fig. 8 Examples of BikeMaps.org guerrilla marketing include distributing branded saddle covers, water bottles, and other goods.

planners have requested data and used them for planning. For example, in 2015 BikeMaps.org data were used in the City of Victoria's bicycle network planning (Fig. 9).

1.04.4 Summary and Conclusions

VGI is a new source of data that is changing what we can study and how we explore our world. Growth in VGI is fueled by technology such as GPS, GIS, and the ability to quickly build and share maps over the Internet. Digital maps are everywhere and mobile GPS technology has been made mainstream through smartphones. As such, a huge proportion of individuals are carrying out day-to-day tasks with a device that is perfect for VGI collection, a smartphone. The power of VGI is that it leverages the fact that each of us has knowledge or can make observations. When we combine an individual's knowledge or experience within a coherent data structure, the whole becomes more than the sum of the parts. The types of VGI that are generated are diverse, from simple actions such as georeferencing other shared media (e.g., a Tweet or a photography) to intentional efforts to engage a wide audience in generating information. The motivations for both project organizers and participants of VGI projects are wide and ranging. It is informative to consider motivations when evaluating project popularity and outcomes for individuals, management, and science.

Issues of data quality and representation bias seem to be the primary criticism of VGI. However, even with limitations, VGI often represents the best available data. While we can and should design technology and methods that optimize the collection of high-quality and consistent data, VGI projects will benefit from research that develops approaches to working with uncertain data. Solutions to uncertain data may take several approaches. For example, an important solution could be to develop tools that enable tracking of confirmatory VGI that emphasizes patterns that are consistent. A second solution is to develop statistical approaches to integrating or conflating VGI with traditional data sets. For example, Jestico et al. (2016) leverage the spatial and temporal extents of Strava by integrating it with official counts that have complete attribution. Finally, awareness of barriers to VGI use by groups of people can lead to greater inclusion or alternate strategies to solicit input.

RinkWatch links climate and the culturally important activity of outdoor ice skating to engage interest and awareness of climate change. This approach attracted extensive media attention and garnered many contributions through submitted data, discussion forums, and data visualizations. The submitted reports are useful for exploring regional variation in climate and linking climate change models to the "skateability" of outdoor rinks, a metric that is relatable for many people.

The ForestFuelsApp showed that people who do not traditionally take part in forestry data collection can assist with data collection tasks, there are people in the community who are motivated to assist, and these people reported enjoyment and learning from some of the tasks. However, compared to the other projects (RinkWatch and BikeMaps.org), which had lower requirements in terms of entry levels of effort, a relatively small audience was engaged. More dynamic and less tedious forms of engagement may reach larger audiences, while deeper forms of engagement may still have a role for certain types of tasks.

A flagship success of BikeMaps.org is that within 1 year of project launch the data were used to support planning of cycling infrastructure in Victoria, Canada. The project goal was to overcome the lack of available cycling safety data and when data began being requested for decision making the BikeMaps.org project had begun to achieve its mission. Essential to success were the quality of VGI technology and the careful development of the attributes associated with data. As well, promotion efforts were substantive during the first year which generated a quantity of data that was sufficient to demonstrate utility of the site. The overarching reason for success is that the VGI generated by citizen cyclists fills a specific data niche, making it a valuable resource for planning and research.

As an author team we have run a variety of VGI projects and a key lesson learned is that VGI projects require maintenance. Both the technology and promotion of VGI tools require ongoing support. It can be costly to keep apps and websites maintained and the knowledge of the technology may need to transfer from one technician to another over time. As well, it is rare that a VGI project will become self-promoting. Rather, it is typical that continued use of a VGI tool requires ongoing outreach to the user community.

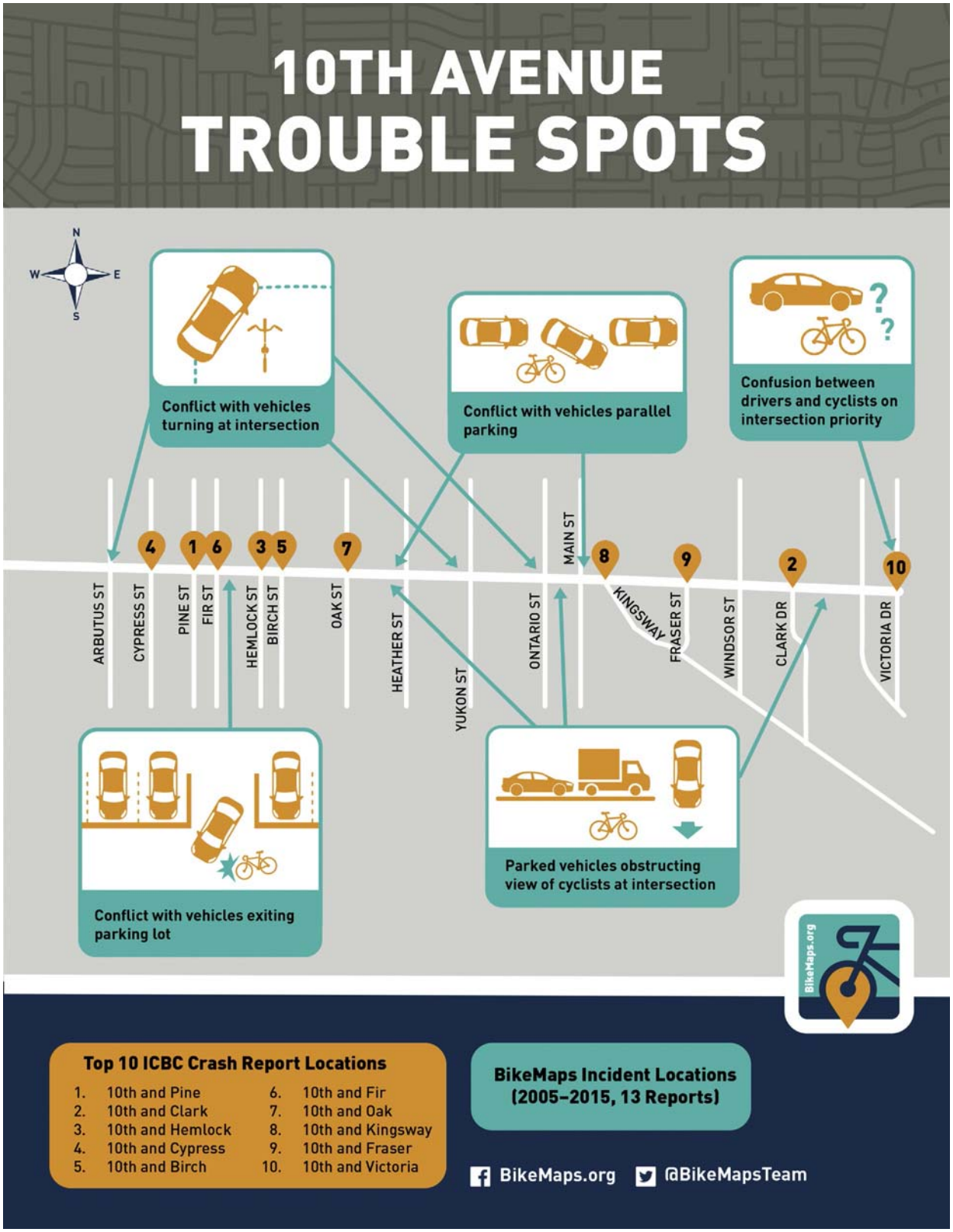

Fig. 9 An example of BikeMaps.org data presented for use in bicycle network planning in Vancouver, British Columbia, Canada. *Source:* BikeMaps.org – https://bikemaps.org/blog/post/10th-avenue-corridor-vancouver-bc-cycling-safety-trouble-spots.

Gamification of tools and generating products from data will help. However, a plan is required to ensure the investment in VGI technology will have long-term use and benefits. In the case of BikeMaps.org, the project initially began as a research project but is now morphing into a team that has both a research arm and a nonprofit outreach arm.

References

Aldred, R., 2016. Cycling near misses: Their frequency, impact, and prevention. Transportation Research Part A: Policy and Practice 90, 69–83. http://dx.doi.org/10.1016/j.tra.2016.04.016.

Argyris, C., Schön, D.A., 2002. Participatory Action Research and Action Science Compared. American Behavioral Scientist 32 (5), 612–623.

Arnstein, S.R., 1969. A Ladder Of Citizen Participation. Journal of the American Institute of Planners 35 (4), 216–224. http://dx.doi.org/10.1080/01944366908977225.

Brammer, J.R., Samson, J., Humphries, M.M., 2014. Declining availability of outdoor skating in Canada. Nature Climate Change 5 (1), 2–4. http://dx.doi.org/10.1038/nclimate2465.

Bruns, A., Burgess, J., Crawford, K., Shaw, F., 2011. qldfloods and @ QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods, ARC Centre of Excelence for Creative Industries and Innovation. Brisbane. http://dx.doi.org/10.1007/978-3-642-39527-7_16.

Bruns, A., Highfield, T., Burgess, J., 2013. The Arab Spring and social media audiences: English and Arabic Twitter users and their networks. In: McCaughey, M., Ebooks Corporation (Eds.), Cyberactivism on the participatory web. Routledge, New York, pp. 86–116.

Burch, S., 2010. Transforming barriers into enablers of action on climate change: Insights from three municipal case studies in British Columbia, Canada. Global Environmental Change 20 (2), 287–297. http://dx.doi.org/10.1016/j.gloenvcha.2009.11.009.

Burgess, H.K., DeBey, L.B., Froehlich, H.E., Schmidt, N., Theobald, E.J., Ettinger, A.K., HilleRisLambers, J., Tewksbury, J., Parrish, J.K., 2016. The science of citizen science: Exploring barriers to use as a primary research tool. Biological Conservation 208, 113–120. http://dx.doi.org/10.1016/j.biocon.2016.05.014.

Chrisman, N.R., 1984. Part 2: issues and problems relating to cartographic data use, exchange and transfer: the role of quality information in the long-term functioning of a geographic information system. Cartographica: The International Journal for Geographic Information and Geovisualization 21 (2-3), 79–88.

Coleman, D.J., Georgiadou, Y., Labonte, J., 2009. Volunteered Geographic Information: The Nature and Motivation of Produsers. International Journal of Spatial Data Infrastructures Research 4 (4), 332–358. http://dx.doi.org/10.2902/1725-0463.2009.04.art16.

Cooper, C.B., Lewenstein, B.V., 2016. Two meanings of citizen science. In: Cavalier, D., Kennedy, E.B. (Eds.), The rightful place of science: citizen science. Consortium for Science, Policy, & Outcomes, Tempe, Arizona, pp. 51–62.

Cooper, C.B., Dickinson, J., Phillips, T., Bonney, R., 2007. Citizen science as a tool for conservation in residential ecosystems. Ecology and Society 12 (2), 11.

Damyanov, N.N., Damon Matthews, H., Mysak, L.a., 2012. Observed decreases in the Canadian outdoor skating season due to recent winter warming. Environmental Research Letters 7, 14028. http://dx.doi.org/10.1088/1748-9326/7/1/014028.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., Shi, W., 2010. Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. Transactions in GIS 14 (4), 387–400. http://dx.doi.org/10.1111/j.1467-9671.2010.01212.x.

Dickinson, J., Crain, R., 2014. Socially Networked Citizen Science and the Crowd-Sourcing of Pro-Environmental Collective Actions. In: Agarwal, N., Lim, M., Wigand, R.T. (Eds.), Online Collective Action, 1st edn. Springer-Verlag Wien, Vienna, pp. 133–152. http://dx.doi.org/10.1007/978-3-7091-1340-0 (Lecture Notes in Social Networks).

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annual Review of Ecology, Evolution, and Systematics 41 (1), 149–172. http://dx.doi.org/10.1146/annurev-ecolsys-102209-144636.

Elwood, S., Goodchild, M.F., Sui, D., 2012. Researching volunteered geographic information: spatial data, geographic research, and new social practice. Annals of the Association of American Geographers 102 (3), 571–590.

Feick, R., Roche, S., 2013. Understanding the value of VGI. In: Sui, D.Z., Elwood, S., Goodchild, M.F. (Eds.), Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice, 2012th edn. Springer, Dordrecht, Netherlands, pp. 15–29.

Ferster, C.J., Coops, N.C., 2014. Assessing the quality of forest fuel loading data collected using public participation methods and smartphones. International Journal of Wildland Fire 23 (4), 585–590.

Ferster, C.J., Coops, N.C., Harshaw, H.W., Kozak, R.A., Meitner, M.J., 2013. An exploratory assessment of a smartphone application for public participation in forest fuels measurement in the wildland-urban interface. Forests 4 (4), 1199–1219.

Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D.S., 2013. Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. Transactions in GIS 17 (6), 847–860. http://dx.doi.org/10.1111/tgis.12033.

Gnoni, M.G., Andriulo, S., Maggio, G., Nardone, P., 2013. "Lean occupational" safety: an application for a near-miss management system design. Safety Science 53, 96–104.

Goodchild, M., 2009. NeoGeography and the nature of geographic expertise. Journal of Location Based Services 3 (2), 82–96. http://dx.doi.org/10.1080/17489720902950374.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221. http://dx.doi.org/10.1007/s10708-007-9111-y.

Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3 (3), 231–241.

Goranson, C., Thihalolipavan, S., Di Tada, N., 2013. VGI and Public Health: Possibilities and Pitfalls. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge, 1st edn. Springer Netherlands, Dordrecht, pp. 329–340. http://dx.doi.org/10.1007/978-94-007-4587-2_18.

Gray, R.W., Oswald, B., Kobziar, L., Stewart, P., Seijo, F., 2015. Reduce wildfire risks or we'll continue to pay more for fire disasters, Position statement of the Association for Fire Ecology, International Association of Wildland Fire, and The Nature Conservancy. Eugene, OR. Available at: http://fireecology.org/Reduce-Wildfire-Risks-or-Well-Pay-More-for-Fire-Disasters (Accessed April 3, 2017).

Griffin, G.P., Jiao, J., 2015. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. Journal of Transport and Health 2 (2), 238–247. http://dx.doi.org/10.1016/j.jth.2014.12.001.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. Environment and Planning B: Planning and Design 37 (4), 682–703. http://dx.doi.org/10.1068/b35097.

Haklay, M., 2013. Citizen science and volunteered geographic information: overview and typology of participation. In: Sui, D.Z., Elwood, S., Goodchild, M.F. (Eds.), Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice, 2012th edn. Springer, Dordrecht, Netherlands, pp. 105–122.

Haklay, M., 2016. Why is participation inequality important? In: Capineri, C., et al. (Eds.), European handbook of crowdsourced geographic information. Ubiquity Press, London, pp. 35–44.

Haklay, M., Weber, P., 2008. OpenStreetMap: user-generated street maps. IEEE Pervasive Computing 7 (4), 12–18.

Harvey, F., 2013. To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice. Springer, Dordrecht, Netherlands, pp. 31–42.

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W., Kelling, S., 2012. Data-intensive science applied to broad-scale citizen science. Trends in Ecology & Evolution 27 (2), 130–137.

Hope, G.D., Lloyd, D.A., Mitchell, W.R., Erickson, W.R., Harper, W.L., Wikeem, B.M., 1991. Ponderosa Pine Zone. In: Meidinger, D., Pojar, J. (Eds.), Ecosystems of British Columbia, 1st edn. Research Branch, BC Ministry of Forests, Victoria, British Columbia, pp. 139–151.

Horita, F., Degrossi, L., Assis, L., Zipf, A., Porto de Albuquerque, J., 2013. The use of volunteered geographic information and crowdsourcing in disaster management: a systematic literature review. In: Proceedings of the 19th Americas Conference on Information SystemsAIS, Chicago, pp. 1–10. Available at: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1591&context=amcis2013.

International Telecommunications Union (ITU), 2016. ITU ICT Facts and Figures 2016. Switzerland, Geneva. Available at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf (Accessed: 3 April 2017).

ISO 19157 (2013) Retrieved October 28, 2016, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575

Jackson, S.P., Mullen, W., Agouris, P., et al., 2013. Assessing completeness and spatial error of features in volunteered geographic information. ISPRS International Journal of Geo-Information 2 (2), 507–530.

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. Journal of Transport Geography 52, 90–97.

Keane, R.E., 2013. Describing wildland surface fuel loading for fire management: a review of approaches, methods and systems. International Journal of Wildland Fire 22 (1), 51.

Kim, J.S., et al., 2014. Space–time wiring specificity supports direction selectivity in the retina. Nature 509 (7500), 331–336.

Kitchin, R., 2014. The data revolution: Big data, open data, data infrastructures and their consequences. In: Rojek, R., Dickens, K., Haw, K. (Eds.), 1st edn. SAGE Publications, Croyden, England.

Klonner, C., Marx, S., Tomás, U., Porto de Albuquerque, J., Höfle, B., 2016. Volunteered geographic information in natural hazard analysis: a systematic literature review of current approaches with a focus on preparedness and mitigation. ISPRS International Journal of Geo-Information 5 (7), 103.

Krykewycz, G., Pollard, C., Canzoneri, N., He, E., 2011. Web-based "crowdsourcing" approach to improve areawide "bikeability" scoring. Transportation Research Record: Journal of the Transportation Research Board 2245, 1–7.

Kulemeka, O., 2015. A review of wildland fire smartphone applications. International Journal of Emergency Services 4 (2), 258–270.

Lam, S., Uduwage, A., Dong, Z., et al., 2011. WP: clubhouse?: An exploration of Wikipedia's gender imbalance. In: Proceedings of the 7th International Symposium on Wikis and Open CollaborationACM Press, New York, pp. 1–10. Available at: http://dl.acm.org/citation.cfm?id=2038560.

Lawrence, H., Robertson, C., Feick, R., Nelson, T., 2015. Identifying Optimal Study Areas and Spatial Aggregation Units for Point-Based VGI from Multiple Sources. In: Harvey, F., Leung, Y. (Eds.), Advances in Geographic Information Science, 1st edn. Springer International Publishing, Cham, Switzerland, pp. 65–84. http://dx.doi.org/10.1007/978-3-319-19950-4_5.

Le Dantec, C.A., Asad, M., Misra, A., Watkins, K., 2015. Planning with crowdsourced data: rhetoric and representation in transportation planning. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social ComputingACM Press, New York, pp. 1717–1727. Available at: http://dl.acm.org/citation.cfm?doid=2675133.2675212.

Lee, R.J., Sener, I.N., Jones-Meyer, S.N., 2016. A review of equity in active transportation. In: Transportation Research Board 95th Annual Meeting Compendium of Papers, no. 16-1835. Available at: http://amonline.trb.org/16-1835-1.2982186?qr=1.

Li, L., Goodchild, M.F., 2014. Spatiotemporal Footprints in Social Networks. In: Alhajj, R., Rokne, J. (Eds.), Encyclopedia of Social Network Analysis and Mining, 1st edn. Springer New York, New York, NY, pp. 1990–1996. http://dx.doi.org/10.1007/978-1-4614-6170-8_322.

Lukyanenko, R., Parsons, J., Wiersma, Y.F., 2016. Emerging problems of data quality in citizen science. Conservation Biology 30 (3), 447–449.

Meier, P., 2012. Crisis mapping in action: how open source software and global volunteer networks are changing the world, one map at a time. Journal of Map & Geography Libraries 8 (2), 89–100.

Miller, C.C., 2006. A beast in the field: the Google Maps mashup as GIS/2. Cartographica: The International Journal for Geographic Information and Geovisualization 41 (3), 187–199.

Misra, A., Gooze, A., Watkins, K., Asad, M., Le Dantec, C., 2014. Crowdsourcing and its application to transportation data collection and management. Transportation Research Record: Journal of the Transportation Research Board 2414, 1–8.

Monroe, M.C., Pennisi, L., McCaffrey, S., Mileti, D., 2006. Social science to improve fuels management: a synthesis of research relevant to communicating with homeowners about fuels management. In: General Technical Report NC-267USDA FS, St. Paul Minnesota. Available at: http://www.nrs.fs.fed.us/pubs/gtr/gtr_nc267.pdf.

Morrow, B., Johnston, K., Davies, J., 2008. Rating Interface Wildfire Threats in British Columbia, Report to BC Ministry of Forests and Range Protection Branch. Victoria, British Columbia.

Neis, P., Zielstra, D., Zipf, A., 2011. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4 (1), 1–21.

Nelson, T.A., Denouden, T., Jestico, B., Laberee, K., Winters, M., 2015. BikeMaps.org: a global tool for collision and near miss mapping. Frontiers in Public Health 3, 53.

Newman, G., Roetman, P., Vogel, J., Brocklehurst, M., Cappadonna, J., Cooper, C., Goebel, C., Haklay, M., Kyba, C., Piera, J., Ponti, M., Sforzi, A., Shirk, J., 2015. Letter in response to 'Rise of the Citizen Scientist'. European Citizen Science Association, Berlin, Germany. Available at: https://ecsa.citizen-science.net/sites/default/files/cs_associations_response_to_nature_editorial.pdf (Accessed: 3 April 2017).

Nuojua, J., 2010. WebMapMedia: A map-based Web application for facilitating participation in spatial planning. Multimedia Systems 16 (1), 3–21. http://dx.doi.org/10.1007/s00530-009-0175-z.

Ottinger, G., 2009. Buckets of resistance: standards and the effectiveness of citizen science. Science, Technology & Human Values 35 (2), 244–270.

Poore, B.S., Wolf, E.B., 2013. Metadata squared: enhancing its usability for volunteered geographic information and the GeoWeb. In: Sui, D.Z., Elwood, S., Goodchild, M.F. (Eds.), Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice, 2012th edn. Springer, Dordrecht, Netherlands, pp. 43–64.

Quattrone, G., Capra, L., De Meo, P., 2015. There's no such thing as the perfect map. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social ComputingACM Press, New York, pp. 1021–1032. Available at: http://dl.acm.org/citation.cfm?doid=2675133.2675235.

Quesnot, T., Roche, S., 2014. Measure of Landmark Semantic Salience through Geosocial Data Streams. ISPRS International Journal of Geo-Information 4 (1), 1–31. http://dx.doi.org/10.3390/ijgi4010001.

Raddick, M.J., Bracey, G., Gay, P.L., et al., 2010. Galaxy zoo: exploring the motivations of citizen science volunteers. Astronomy Education Review 9 (1), 10103.

Radeloff, V.C., Hammer, R.B., Stewart, S.I., Fried, J.S., Holcomb, S.S., McKeefry, J.F., 2005. The Wildland–Urban Interface in The United States. Ecological Applications 15 (3), 799–805. http://dx.doi.org/10.1890/04-1413.

Reason, J., 1991. Too little and too late: a commentary on accident and incident reporting systems. In: van der Schaaf, T., Lucas, D., Hale, A. (Eds.), Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford, pp. 9–26.

Robertson, C., Feick, R., 2015. Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas. Cartography and Geographic Information Science 43 (4), 283–300. http://dx.doi.org/10.1080/15230406.2015.1088801.

Robertson, C., McLeman, R., Lawrence, H., 2015. Winters too warm to skate? Citizen-science reported variability in availability of outdoor skating in Canada. Canadian Geographer 59 (4), 383–390. http://dx.doi.org/10.1111/cag.12225.

Roche, S., Propeck-Zimmermann, E., Mericskay, B., 2013. GeoWeb and crisis management: issues and perspectives of volunteered geographic information. GeoJournal 78 (1), 21–40.

Romanillos, G., Zaltz Austwick, M., Ettema, D., De Kruijf, J., 2016. Big data and cycling. Transport Reviews 36 (1), 114.

Romanillos, G., Zaltz Austwick, M., Ettema, D., De Kruijf, J., 2015. Big Data and Cycling. Transport Reviews 1647, 1–20. http://dx.doi.org/10.1080/01441647.2015.1084067.

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., Jacobs, D., 2012. Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12, pp. 217–226. http://dx.doi.org/10.1145/2145204.2145238. Seattle, WA.

Sanchez, T.W., Brenman, M., 2013. Public participation, social equity, and technology in urban governance. In: Silva, C.N. (Ed.), Citizen e-participation in urban governance. IGI Global, Hershey, PA, pp. 35–48.

Shelton, T., Poorthuis, A., Zook, M., 2015. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landscape and Urban Planning 142, 198–211. http://dx.doi.org/10.1016/j.landurbplan.2015.02.020.

Show, H., 2015. Rise of the citizen scientist. [online]. Nature 524 (7565), 265.

Smith, M., Szongott, C., Henne, B., von Voigt, G., 2012. Big data privacy issues in public social media. In: 2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST), pp. 1–6. http://dx.doi.org/10.1109/DEST.2012.6227909.

Stefanidis, A., Crooks, A., Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal 78 (2), 319–338. http://dx.doi.org/10.1007/s10708-011-9438-2.

Strava, 2017. How it works: Frequently asked questions. Available at: https://www.strava.com/how-it-works (Accessed: 3 April 2017).

Stephens, M., 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. GeoJournal 78 (6), 981–996.

Sui, D., Goodchild, M., Elwood, S., 2013. Volunteered geographic information, the exaflood, and the growing digital divide. In: Sui, D.Z., Elwood, S., Goodchild, M.F. (Eds.), Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice, 2012th edn. Springer, Dordrecht, Netherlands, pp. 1–12.

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.a., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.a., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K.V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.-K., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation 169, 31–40. http://dx.doi.org/10.1016/j.biocon.2013.11.003.

Teschke, K., Harris, M., Reynolds, C., et al., 2012. Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. American Journal of Public Health 102 (12), 2336–2343.

Teschke, K., Frendo, T., Shen, H., et al., 2014. Bicycling crash circumstances vary by route type: a cross-sectional analysis. BMC Public Health 14 (1), 1205.

Toman, E., Shindler, B., Brunson, M., 2006. Fire and Fuel Management Communication Strategies: Citizen Evaluations of Agency Outreach Activities. Society & Natural Resources 19 (4), 321–336. http://dx.doi.org/10.1080/08941920500519206.

Unwin, D.J., 2005. Fiddling on a different planet? Geoforum 36 (6), 681–684.

Ward, M., Nuckols, J., Giglierano, J., et al., 2005. Positional accuracy of two methods of geocoding. Epidemiology 16 (4), 542–547.

Whitmarsh, L., 2011. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. Global Environmental Change 21 (2), 690–700. http://dx.doi.org/10.1016/j.gloenvcha.2011.01.016.

Wiggins, A., 2013. Free as in puppies: compensating for ICT constraints in citizen science. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work and Social ComputingACM Press, San Antonio, TX, pp. 1469–1480. Available at: http://dl.acm.org/citation.cfm?id=2441942.

Winters, M., Babul, S., Becker, H.J., et al., 2012. Safe cycling: how do risk perceptions compare with observed risk? Canadian Journal of Public Health 103 (9), S42–S47.

Zook, M., Graham, M., Shelton, T., Gorman, S., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. World Medical & Health Policy 2 (2), 6–32. http://dx.doi.org/10.2202/1948-4682.1069.

1.05 Open Data and Open Source GIS

Xinyue Ye, Kent State University, Kent, OH, United States

1.05.1 Introduction

With the growing capability of recording individual's digital footprints and the emerging open culture, the open big data are flooding everywhere (Batty, 2012). Geospatial data are an important component of open data unfolding right in front of our eyes (Warf and Arias, 2008). Geographic information system (GIS) research is shifting toward analyzing ever-increasing amounts of large-scale, diverse data in an interdisciplinary, collaborative, and timely manner. Goodchild (2013) proposed the crowd, social, and geographic approaches to assess big data quality. Sui (2014) argued that open GIS should involve eight dimensions related to data, software, hardware, standards, research, publication, funding, and education facilitated by web-based tools and the growing influence of the open culture. The key pillars of open GIS have always been and will continue to be open source, open data, open modeling, open collaboration, and open publishing for future GIS research and applications (Rey, 2014). Sui (2014) noted that "the big data torrent will eventually be more powerful if they can be made to conform to open standards such as those developed by OGC over the years". This article places greater emphasis on open data and open source GIS. According to Sui (2014), open GIS offers four exciting opportunities for participation and collaboration among both GIS experts and volunteers: (1) technology-driven opportunities for addressing big data challenges; (2) application-led opportunities for improving decisions across all levels; (3) curiosity-inspired, crowd-powered opportunities for developing citizen science; and (4) education-focused opportunities for realizing a spatial university.

To realize science's powerful capacity for self-correction, it is critical to reproduce the outcomes of scientific research (Ye et al., 2015). However, if data and codes are not transparent, the reproductivity will not work due to bottlenecks or restrictions from copyright, patents, or other mechanisms of control. Open data and open source GIS aim to make GIS research open to everyone. In other words, data and codes should be made legally open and accessible to both professional and nonprofessional communities (Stodden, 2009).

The growing and evident interdisciplinary efforts dedicated to "open data and open source GIS" represent a transformative trend shaped by increased scholarly collaboration and research methods sharing (Sui, 2014). Instead of following the traditional proprietary approach, open data and open source GIS can lead to a large number of benefits to both citizens and businesses across the globe, with the success of GIS research, education, and applications.

The rest of this article is organized as follows. Elements of the emerging open data are described in section "**Open Data**". Section "**Open source GIS**" discusses how open source GIS plays a pivotal role in the research and education. Section "**Practicing Open Source GIS**" demonstrates the use of open source GIS for regional economic analysis. The article ends with a summary and conclusion for open GIS paradigm in section "**Summary**" toward the goal of reproducibility and the desired reuse.

1.05.2 Open Data

Knowledge is eventually derived from data. The term "open data" was coined in 1995 to deal with the disclosure of geophysical and environmental data realizing the idea of common good applied to knowledge (Chignard, 2013). Open data is gaining popularity with the launch of open government data initiatives. The volume of open space–time data in various disciplines and domains has dramatically increased due to the growing sophistication and ubiquity of information and communication technology (Jiang, 2011). Open data can be used and reused at no cost or restriction without mechanisms of control such as copyright and patents (Auer et al., 2007). The intention of open data movement is to make publicly acquired data available for direct manipulation such as cross tabulation, visualization, and mapping (Gurstein, 2011). It is clear that a space–time perspective in using such data has become increasingly relevant to our understanding of socioeconomic and environmental dynamics in the collaborative and transdisciplinary manner. As noted by (Rey, 2014), "Open data constitutes available, intelligible, accessible, and usable data. For science's error–correction mechanisms to kick in, data underlying research projects must be made accessible to the wider research community". Featured by the ever-growing volume, variety, and velocity of ubiquitous geospatial information, the big spatial data in the changing environmental, urban, and regional contexts demand innovative thinking that can capture the rich

information of patterns and processes and provide spatial strategies for sustainable development. Meanwhile, the volume of data created by an ever-increasing number of geospatial sensor platforms such as remote sensing and social sensing (including citizen sensors) to collect data at ever-increasing spatial, spectral, temporal, and radiometric resolutions currently exceeds petabytes of data per year and is only expected to increase. Data come from various sources, types, organizations (governments, military, NGOs, etc.), and purposes. Recent developments in information technology commonly referred to as big data along with the related fields of data science and analytics are needed to process, examine, and realize the value of the overwhelming amount of open geospatial data. The research agenda has been substantially redefined in light of open data, which have transformed the focus of GIS research toward dynamic, spatial, and temporal interdependence of human–environment issues across multiple scales. Metadata is information about data. The use of metadata enhances the opportunities for semantical interoperability involving open data, lowering the cost of access, and manipulating and sharing across data boundaries (Nogueras-Iso et al., 2004). As declared by (FGDC 2017), "Geospatial metadata describes maps, Geographic Information Systems (GIS) files, imagery, and other location-based data resources. The FGDC is tasked by Executive Order 12906 to enable access (see GeoPlatform.gov) to National Spatial Data Infrastructure (NSDI) resources and by OMB Circular A-16 and the A-16 Supplemental Guidance to support the creation, management, and maintenance of the metadata required to fuel data discovery and access".

Twelve critical factors were identified to city-level, regional, and transnational cases regarding the publication and use of open data (Susha et al., 2015): (1) a national guide on legal intellectual property right issues; (2) a clear data publishing process; (3) addressing of societal issues and publishing of related data; (4) interest for users; (5) where to publish datasets; (6) a virtual competence center for technical help; (7) a strategy for maintaining published datasets; (8) allowing citizens to post, rate, and work with datasets and web services; (9) a clear user interface; (10) standards for data, metadata, licenses, URIs, and exchange protocols; (11) integrate metadata schemas and federated controlled vocabularies; and (12) application programming interfaces for open data provision.

As (Maguire and Longley, 2005, p. 3) noted, "geoportals are World Wide Web gateways that organize content and services such as directories, search tools, community information, support resources, data and applications". The Geospatial One-Stop emerged as an easier, faster, and less expensive gateway for searching relevant geographic information sponsored by the US Federal Government (Yang et al., 2007). As (Goodchild et al., 2007, p. 250) pointed out, "humans have always exchanged geographic information, but the practice has grown exponentially in recent years with the popularization of the Internet and the Web, and with the growth of geographic information technologies. The arguments for sharing include scale economies in production and the desire to avoid duplication. The history of sharing can be viewed in a three-phase conceptual framework, from an early disorganized phase, through one centered on national governments as the primary suppliers of geographic information, to the contemporary somewhat chaotic network of producers and consumers". Many governments have been developing various programs to open data available via websites for public consumption such as the US, the UK, and Canada. These datasets typically contain records with spatial properties in democratizing public sector data and driving innovation (Arribas-Bel, 2014). Data.gov as a website launched in 2009 aims to enhance accessing the repository for federal government information and to ensure better accountability and transparency over 50 US government agencies with over 194,708 datasets (Hendler et al., 2012; Data.Gov, 2017).

In line with the spirit of crowdsourcing and citizen science, open data movement is that data should be legally and technically open to the scientific community, industry, and the public to use and republish. In other words, data should be provided in open machine-readable formats and readily located, along with the relevant metadata evaluating the reliability and quality of the data to promote increased data use and facilitate credibility determination. To advocate both transparency and innovation, open government data initiatives have been implemented in many countries from local to global level regarding accessibility, persistent identification, and long-term availability. The open data initiatives encourage peer production, interactivity, and user-generated innovation, which have stimulated the sharing and distribution of information across communities and disciplines. The underlying philosophy of open government or the theory of open source governance is that the interested citizens can access the documents of the government to facilitate effective public oversight and enable the direct involvement in the legislative process to promote openness, participation, and efficiency in government (Janssen et al., 2012). Civic hacking is utilizing government data to make governments more accountable through solving civic problems by those who care about their communities. For instance, Code for America is a nonprofit organization founded in 2009 to deal with the growing cross-sector gap in their effective use of technology and design for good (Wadhwa, 2011). Some civic hackers are employed by Code for America. National Day of Civic Hacking is a nationwide day of action where various developers come together to coordinate civic tech events dedicated to civic hacking (Johnson and Robinson, 2014). Transparency and participation through data integration and dissemination across domains and boundaries will facilitate collaboration among researchers, private sectors, and civilians leveraging their skills to help the society. (Gurstein, 2011) also examined the impact of such open data initiatives on the poor and marginalized and call for ensuring a wide opportunity for effective data use in the context of digital divide.

Papers with publicly available datasets usually have a higher citation rate and visibility than similar studies without available data, either by a direct data link or indirectly from cross-promotion (Piwowar and Vision, 2013). However, open data movement often faces the economic, legal, organizational, political, social, and technical challenges at either individual or institutional levels. Many researchers are still reluctant to share the original data used in their research to support the open access initiative, fearing the loss of credits, future publications, and competitive advantage, as well as the time to document and deal with questions for users (Fernandez, 2010; Sui, 2014). Their motivation for publication of datasets and intentions in doing so remain uncertain. Arguably, different person-based policies among stakeholders with various backgrounds and interests need to be developed to encourage sharing behavior in collaboration. Organizational support has been playing a substantial role in promoting researchers' intentions of sharing datasets through dealing with the heterogeneity of collaborators and the complexity of the data sharing process.

1.05.3 Open Source GIS

The Free and Open Source Software for Geospatial Conference has been playing a pivotal role in promoting the open science in software development. Open source GIS is gaining growing market shares in academia, business, and public administration. This recognition came at a time when many open source programming and scripting languages such as Python and R are starting to make major inroads in geospatial data production, analysis, and mapping (Ye and Rey, 2013). As a consequence, open source software development has been a crucial element in the GIS community's engagement with open GIS and the most well-developed aspect of open GIS (Rey, 2009). The availability and widespread use of codes and tools to support more robust data analysis will play a critical role in the adoption of new perspectives and ideas across the spatial sciences. The openness to scrutiny and challenge underlies the open source GIS movement through the release of the source code, which has subsequently influenced software functionality and support (Neteler and Mitasova, 2008). Users have the freedom to access, modify, and distribute the source code based on licensing agreements such as MPL, MIT, Apache, GPL, and BSD. Making source code both legally and technically open is the very first step of being promoted as a public good (Rey and Ye, 2010). In particular, scientists could benefit from the open source code, which would reduce code duplication and free up additional developer time to enhance the respective applications (Rey, 2009). Bonaccorsi and Rossi (2003) argued that "when programmers are allowed to work freely on the source code of a program, this will inevitably be improved because collaboration helps to correct errors and enables adaptation to different needs and hardware platforms". The credibility of research findings tends to be higher for papers with the available code. Third-party researchers might be more likely to adopt such papers as the foundation of additional research. In addition, coding repository platforms such as GitHub and BitBucket are making this open source tide stronger. Links from code to a paper might enhance the search frequencies of the paper because of accelerated awareness of the methods and findings.

The dramatic improvement in computer technology and the availability of large-volume geographically referenced data have enabled the spatial analytical tools to move from the fringes to central positions of methodological domains. By and large, however, many existing advanced spatial analysis methods are not in the open source context. The open source and free approach offer unprecedented opportunities and the most effective solution for developing software packages through attracting both users and developers. Instead of reinventing the wheel, we can study how the program works, to adapt it, and to redistribute copies including modifications from a number of popular alternatives. Anselin (2010) emphasized the role of the open source software movement in stimulating new development, transcending disciplinary boundaries, and broadening the community of developers and adopters. With accelerated development cycle, open source tools can give GIS users more flexibility to meet the user community needs that are only bound by our imaginations, which are aligned with more efficient and effective scientific progress. New theories and novel practices can thus be developed beyond narrowly defined disciplinary boundaries (Sui, 2014). Regarding open source efforts on spatial analysis, Arribas-Bel (2014) argued, "the traditional creativity that applied researchers (geographers, economists, etc.) have developed to measure and quantify urban phenomena in contexts where data were scarce is being given a whole new field of action". Sui (2014) also noted that a hybrid model integrating both open/free paradigm and proprietary practices (copyright and patent, IP stuff) would be the most realistic option and promising route to move GIS forward. Open source GIS can facilitate the interdisciplinary research due to "the collaborative norms involving positive spillover effects in building a community of scholars" (Rey, 2009; Ye et al., 2014).

During the past several decades, burgeoning efforts have been witnessed on the development and implementation of spatial statistical analysis packages, which continue to be an active area of research (Rey and Anselin, 2006; Anselin, 2010). The history of open source movement is much younger, but its impact on GIS world is impressive (Rey, 2009). As Rey (2009) commented, "a tenet of the free software (open source) movement is that because source code is fundamental to the development of the field of computer science, having freely available source code is a necessity for the innovation and progress of the field". The development of open source packages has been boosted. However, many duplicates and gaps in the methodological development have also been witnessed. The open source toolkit development is community-based with developers as well as casual and expert users located everywhere. Through the use of an online source code repository and mailing lists, users and developers can virtually communicate to review the existing code and develop new methods. However, Tsou and Smith (2011, p. 2) argued that "open source software is not well adopted in GIS education due to the lack of user-friendly guidance and the full integration of GIS learning resources". Some representative open source desktop GIS software packages include KOSMO, gvSIG, uDig, Quantum GIS (QGIS), Geographic Resource Analysis Support System (GRASS), and so on. KOSMO was implemented using the Java programming language based on the OpenJUMP platform and free code libraries. Developed by the European GIS community offering multiple language user interfaces, gvSIG is known for having a user-friendly interface, being able to access a wide range of vector and raster formats. Built upon IBM's Eclipse platform, uDig (user-friendly desktop Internet GIS) is an open source (EPL and BSD) desktop application framework. QGIS integrates with other open source GIS packages such as PostGIS, GRASS GIS, and MapServer, along with plugins being written in Python or C++. As a founding member of the Open Source Geospatial Foundation (OSGeo), GRASS offers comprehensive GIS functions for data management, image processing, cartography, spatial modeling, and visualization.

1.05.4 Practicing Open Source GIS

The study of economic inequality and convergence continues to attract enormous attention thus becoming a dynamic academic landscape where the interdisciplinary literature has evolved (Ye and Rey, 2013). This interest has been reflected in the analysis

of spatial patterns of economic convergence and the temporal dynamics of geographical inequality. However, the literature studies of process analysis and form analysis are mainly separated because most methods are standalone without the sharing of the code. At the same time, the increasing availability of open space–time data has outpaced the development of space–time analytical techniques across social sciences. The methodological integration of space and time call for open data and open source methods, which will help narrow the gap between growth theories and their empirical testing. While the substantive focus of this case study is on open source computing of regional income dynamics, the issues examined are relevant to the development of a wide class of methods based on open science. This section suggests some novel exploratory approaches to compare spatial pattern and temporal trend of regional development. The cross-fertilization between domain science and open source computing is identified and illustrated.

There is a growing list of papers using local indicator of spatial autocorrelation (LISA) to measure the spatial structure of socio-economic patters due to the following two reasons: the availability of open socioeconomic data across administrative units and the implementation of LISA indicators in software packages (Anselin, 1995). However, the release of spatial panel data calls for the extension of static spatial methods into the dynamic context. A LISA time path can be used to measure how economies covary over space and time in a regional context, giving insights to the debate about cooperative versus competitive regional growth. The challenging issue is that most domain users cannot handle the LISA time path due to the lack of programming skills.

Once the new space–time indicators are developed, an extensive set of inferential approaches is needed to evaluate their sampling distributions for comparative analysis between two regional systems. A tortuous LISA time path indicates that the focal economy and its average neighbor have instable convergence/divergence rates, while a frequent crossing suggests mixed convergence and divergence trends over time. With the growing awareness of the potential importance of the spatial dimension of economic structure, these space–time constructs can be implemented into empirical specifications to test the existence of poverty traps, convergence clubs, and spatial regimes. Two interesting questions arise from this analysis (the value hereafter refers to tortuous indicator or crossing ratio):

1. Is a LISA time path statistically more or less tortuous than what is expected if the path is randomly organized?
2. Is a LISA time path's crossing ratio statistically larger/smaller than what is expected if the path is randomly organized?

Both LISA time path and various simulation procedures can be modified based on codes from open source packages STARTS and PySal (Rey, 2009, 2014). It is expected that LISA coordinates are independent from each other; however, an individual region's economic growth at one time point relates to its history and its neighbors' temporal economic dynamics. An alternative to the aforementioned methods is to employ a Monte Carlo simulation approach and thus circumvent the assumption of independence that causes inferential problems. The presence of space–time effects needs to be considered when examining the distributional properties of a LISA time path indicator. Three sets of permutation approaches are suggested to test the independence of space, time, and trend through python code implementation.

The spatial independence test answers the following questions: Can the observed value (or the differences between two observed values) be used to reject the null hypothesis of spatial randomness? The temporal independence test answers the following question: Can the observed value (or the differences between two observed values) be used to reject the null hypothesis of temporal randomness? The trend independence test answers the following question: Can the observed value (or the difference between two observed values) be used to reject the null hypothesis of trend randomness?

Figs. 1–3 show the effects of these three independence tests, the top left view is a LISA time path of region i from Time 1 to Time 10. In Fig. 1, through random permutations on the spatial coordinates of all regions, the other three graphs show three different LISA time paths with different groups of points (LISA coordinates). In Fig. 2, through randomly relabeling time stamps, the other three graphs show different LISA time paths based on the same group of points. In Fig. 3, through randomly normalization the path segments to follow a normal distribution, the other three graphs show different LISA time paths that retain the trends of the original path.

We test LISA path significance using tortuosity and crossing ratio using provincial GDP data in mainland China from 1998 to 2008, consisting of 31 provinces in total. This work uses the k-nearest neighbors method to construct the spatial matrix, and default k is 4. The related code is adopted from PySal (http://pysal.readthedocs.io/en/latest/users/tutorials/weights.html). We compute the observations $\boldsymbol{o}_t$ and their lag values $\boldsymbol{l}_t$ for each time points. Both values are standardized as z-scores. Construct a dictionary $\boldsymbol{D}$ that uses the time point t as key, and a matrix comprised two rows transposed from $\boldsymbol{o}_t$ *and* $\boldsymbol{l}_t$. We then construct the observed LISA path with the values extracted from the dictionary $\boldsymbol{D}$. A LISA time path P_i for a given province i consists of several spatial coordinates with their temporal labeling, represented as:

$$P_i = \{(x_k,\ y_k,\ t_k)\},\ 0 < k < n+1$$

The first two elements x_k and y_k in the tuple are the attribute value and its spatial lag value, respectively, of the province i, while t_k is the k_{th} year when the value is measured. Finally, we obtain a set of paths Sp for all provinces. For each path, Pi in the path set Sp, two indexes are integrated in the following simulation procedures: tortuosity and crossing ratio. Tortuosity is represented as:

$$T_p = distance((x_1, y_1), (x_n, y_n))/len(P)$$

where (x_1,y_1) and (x_n,y_n) are the head and tail spatial coordinates of the LISA path P, and $len(P)$ measures the total length of all the segments of the path. The value of tortuosity is in the range [0, 1], with 0 representing higher degree of path tortuosity, and 1 for the path to be completely straight spatially, stretching in a stable direction.

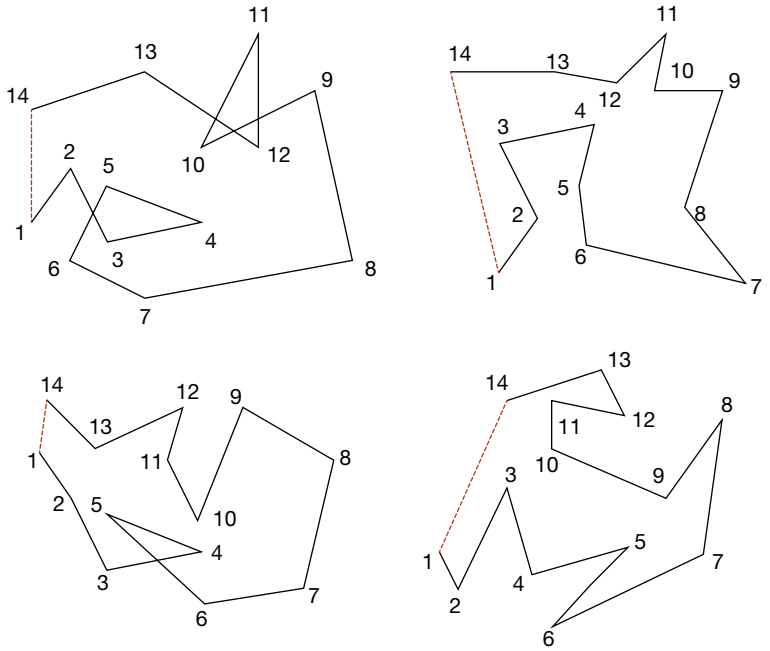

Fig. 1 Spatial independence test.

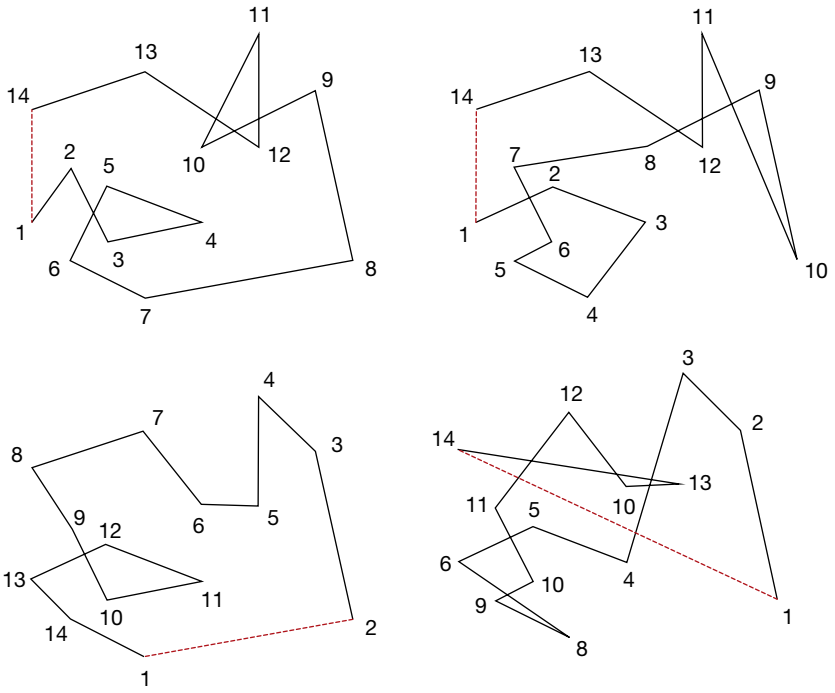

Fig. 2 Temporal independence test.

The crossing ratio is represented as:

$$c_p = 2 * ic_p \Big/ \left(n_p{}^2 - n_p\right)$$

where ic_p represents the self-intersection count of the LISA path p, and n_p represents the number of points in p. The self-intersection count is calculated by checking whether segments of this path P intersect with each other. The value of crossing ratio also lies in the range [0, 1], with 0 representing no crossing, which indicates that the path is highly stable, and 1 for another extreme case where all the segments in the path happen to intersect with each other, indicating highly unstable evolvements of the path over time.

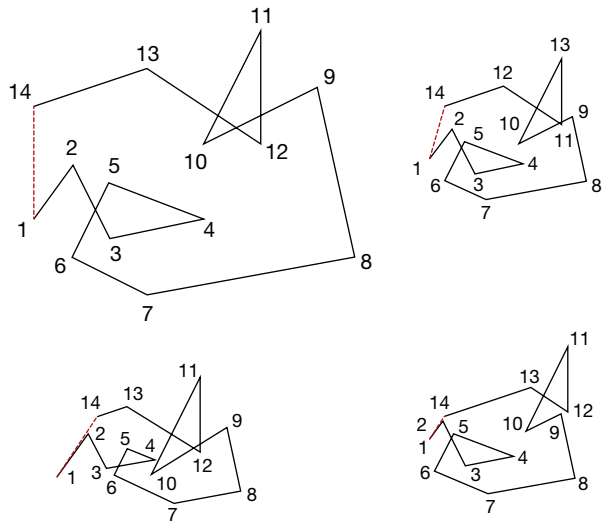

Fig. 3 Trend independence test.

We permute the original data 999 times using one of the following three modes: spatial independence, temporal independence, and trend independence.

(i) Spatial independence test

The spatial independence acts on all the provinces. For each time point *t*, randomly permutate the spatial coordinates for all the regions. This process rearranges the LISA and the lag values among the provinces, and a new LISA time path is constructed for the given region.

(ii) Temporal independence test

The temporal independence test uses the data solely from the observed LISA path. The test randomly permutates the temporal order of the spatial coordinates of the given region and form a new LISA time path.

(iii) Trend independence test

Similar to the temporal independence test, the trend independence test takes only the data from the observed LISA path. The test starts with breaking the observed LISA path into a set of vectors. These vectors are then normalized to follow a normal distribution centered at zero and with unit length as standard deviation while preserving their directions. A trend list is thus formed. The new LISA time path is generated by first randomly picking a starting coordinate and then uses the trend list to construct the whole path.

The values calculated in the simulation process are then ordered, and the pseudo significance level is then computed. This simply sorts the empirically generated 999 values and then develops a pseudo significance level by calculating the share of the empirical values that are higher than the actual value.

The results show that spatial independence and temporal independence tests show stronger significance level than trend independence test. The ranks of the tortuosity and crossing ratio in each test are not much correlated but are relatively stable respectively across three tests. Under the hypothesis of spatial or temporal independence, all provinces are significantly tortuous than expected, except three provinces (Jiangxi, Jilin, and Shanghai). However, none of the provinces shows significant crossing frequencies. In other words, there are no obvious mixed convergence and divergence trends over time.

1.05.5 Summary

Spatial turn in many socioeconomic theories has been noted in a vast field, encompassing both social and physical phenomena (Krugman, 1999; Goodchild et al., 2000; Batty, 2012). The fast growth in socioeconomic dynamics analysis is increasingly seen as attributable to the availability of space–time datasets (Rey and Ye, 2010; Ye et al., 2016). Rigorous space–time analysis and modeling open up a rich empirical context for scientific research and policy interventions. To help scholars and stakeholders deal with the challenges and issues, methodologies and best practice guidelines are needed at both international and national or local level. Making data and source code available for both replication and continuing research will have far-reaching and broader impacts in both GIS and domain communities, highlighting the growing recognition of the role of geography in interdisciplinary research (Karnatak et al., 2012). Such improved discoverability is beneficial for both investigators and the science community as a whole. Open source GIS research has received increased attention and will lead to the revolutionary advances in individual and collective decision-making processes (Sui et al., 2012). The goal of this article is to make a modest effort to synthesize an agenda surrounding and hopefully to stimulate further discussions that promote open GIS as the driving force to guide the development of

GIS at a finer scale (Sui, 2014). To gain momentum under the general umbrella of big data and new data, GIS should fully embrace the vision of an open data and open source GIS paradigm to enhance government efficiency and to improve citizen services (Wright, 2012). Although there are academic, legal, social/political, and environmental impediments for the practice, open GIS will provide numerous technology-driven, application-led, science-inspired, and education-focused opportunities (Sui, 2014).

Methods developed in the mainstream spatial science disciplines need to be progressed with more attention paid to the potential reuse of data and codes. Though a growing list of research papers have highlighted the increasing awareness of spatio-temporal thinking and action, the gap has been widening between a small group of method developers and users. Hence, a crucial step is to develop the dialog between computational scientists and domain users, seeking the cross-fertilization between the two fast-growing communities. As Rey (2009) suggested, "increased adoption of open source practices in spatial analysis can enhance the development of the next generation of tools and the wider practice of scientific research and education". The methods are built in open source environments and thus easily extensible and customizable. Hence, open source project can promote collaboration among researchers who want to improve current functions or add extensions to address specific research questions.

Acknowledgments

This research was funded by the National Science Foundation (1416509, 1535031, 1637242).

References

Anselin, L., 2010. Thirty years of spatial econometrics. Papers in Regional Science 89 (1), 3–25.
Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical Analysis 27 (2), 93–115.
Arribas-Bel, D., 2014. Accidental, open and everywhere: Emerging data sources for the understanding of cities. Applied Geography 49, 45–53.
Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In: *The semantic web* (Lecture notes in computer science, vol. 4825), p. 722. 10.1007/978-3-540-76298-0_52.
Batty, M., 2012. Smart cities, big data. Environment and Planning-Part B 39 (2), 191.
Bonaccorsi, A., Rossi, C., 2003. Why open source software can succeed. Research Policy 32 (7), 1243–1258.
Chignard, S. (2013). A brief history of open data. *ParisTech Review*, 29 March.
Data.Gov (2017) http://www.data.gov/ (accessed 20 February 2017).
Fernandez, R. (2010). Barriers to open science: From big business to Watson and Crick. http://opensource.com/business/10/8/barriers-open-science-big-business-watson-and-crick (accessed 16 April 2017).
FGDC.Gov (2017). https://www.fgdc.gov/metadata (accessed 20 February 2017).
Goodchild, M.F., 2013. The quality of big (geo) data. Dialogues in Human Geography 3 (3), 280–284.
Goodchild, M.F., Anselin, L., Appelbaum, R.P., Harthorn, B.H., 2000. Toward spatially integrated social science. International Regional Science Review 23 (2), 139–159.
Goodchild, M.F., Fu, P., Rich, P., 2007. Sharing geographic information: An assessment of the Geospatial One-Stop. Annals of the Association of American Geographers 97 (2), 250–266.
Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday* **16**(2). (accessed 16 April 2017).
Hendler, J., Holm, J., Musialek, C., Thomas, G., 2012. US government linked open data: Semantic. Data.Gov. IEEE Intelligent Systems 27 (3), 25–31.
Janssen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. Information Systems Management 29 (4), 258–268.
Jiang, B., 2011. Making GIScience research more open access. International Journal of Geographical Information Science 25 (8), 1217–1220.
Johnson, P., Robinson, P., 2014. Civic hackathons: Innovation, procurement, or civic engagement? Review of Policy Research 31 (4), 349–357.
Karnatak, H., Shukla, R., Sharma, V., Murthy, Y., Bhanumurthy, V., 2012. Spatial mashup technology and real time data integration in geo-web application using open source GIS – A case study for disaster management. Geocarto International 27 (6), 499–514.
Krugman, P., 1999. The role of geography in development. International Regional Science Review 22 (2), 142–161.
Maguire, D.J., Longley, P.A., 2005. The emergence of geoportals and their role in spatial data infrastructures. Computers, Environment and Urban Systems 29 (1), 3–14.
Neteler, M., Mitasova, H., 2008. Open source GIS: A GRASS GIS approach, 3rd ed. Springer, Berlin.
Nogueras-Iso, J., Zarazaga-Soria, F.J., Lacasta, J., Béjar, R., Muro-Medrano, P.R., 2004. Metadata standard interoperability: Application in the geographic information domain. Computers, Environment and Urban Systems 28 (6), 611–634.
Piwowar, H.A., Vision, T.J., 2013. Data reuse and the open data citation advantage. Peerj 1 (175), 1–25. http://dx.doi.org/10.7717/peerj.175.
Rey, S., 2009. Show me the code: Spatial analysis and open source. Journal of Geographical Systems 11 (2), 191–207.
Rey, S.J., 2014. Open regional science. The Annals of Regional Science 52 (3), 825–837.
Rey, S.J., Anselin, L., 2006. Recent advances in software for spatial analysis in the social sciences. Geographical Analysis 38 (1), 1–4.
Rey, S., and Ye, X. (2010). Comparative spatial dynamics of regional systems. In: Páez, A., Le Gallo, J., Buliung, R. & Dall'Erba, S. (eds.) *Progress in spatial analysis: Theory, computation, and thematic applications*, pp. 441–464. Springer: New York City.
Stodden, V., 2009. The legal framework for reproducible research in the sciences: Licensing and copyright. IEEE Computing in Science and Engineering 11 (1), 35–40.
Sui, D., 2014. Opportunities and impediments for open GIS. Transactions in GIS 18 (1), 1–24.
Sui, D., Elwood, S., Goodchild, M., 2012. Crowdsourcing geographic knowledge: Volunteered geographic information in theory and practice. Springer, Berlin.
Susha, I., Zuiderwijk, A., Charalabidis, Y., Parycek, P., Janssen, M., 2015. Critical factors for open data publication and use: A comparison of city-level, regional, and transnational cases. JeDEM-eJournal of eDemocracy and Open Government 7 (2), 94–115.
Tsou, M. H., & Smith, J. (2011). *Free and open source software for GIS education*. White paper written with the support from the National Geospatial Technology Center of Excellence (GeoTech Center, http://www.geotechcenter.org/).
Wadhwa, V. (2011). Code for America: An elegant solution for government IT problems. *The Washington Post*.
Warf, B., Arias, S. (Eds.), 2008. The spatial turn: Interdisciplinary perspectives. Routledge, London, UK.
Wright, D. (2012). Big data, GIS, and the academic community. http://blogs.esri.com/esri/esri-insider/2012/10/03/big-data-gis-and-the-academic-community/#more-1311.(accessed 15 April 2017).

Yang, P., Evans, J., Cole, M., Marley, S., Alameh, N., Bambacus, M., 2007. The emerging concepts and applications of the spatial web portal. Photogrammetric Engineering & Remote Sensing 73 (6), 691–698.

Ye, X., Rey, S.J., 2013. A framework for exploratory space-time analysis of economic data. Annals of Regional Science 50 (1), 315–339.

Ye, X., She, B., Wu, L., Zhu, X., Cheng, Y., 2014. An open source toolkit for identifying comparative space-time research questions. Chinese Geographical Science 24 (3), 348–361.

Ye, X., She, B., Zhao, H., Zhou, X., 2016. A Taxonomic Analysis of Perspectives in Generating Space-Time Research Questions in Environmental Sciences. International Journal of Applied Geospatial Research doi. http://dx.doi.org/10.4018/IJAGR.2016040104.

Ye, X., Yu, J., Wu, L., Li, S., Li, J., 2015. Open source point process modeling of earthquake. In: Geo-informatics in resource management and sustainable ecosystem. Springer, Berlin and Heidelberg, pp. 548–557.