Deserts in the Deluge: IPUMS Terra and Spatial Demography of Big Data

S. M. Manson[1], T. A. Kugler[2], D. Haynes II[2]

[1] Department of Geography, Environment, and Society. University of Minnesota. 414 Social Sciences, 267 19th Avenue South, Minneapolis, MN 55455, USA; Email: manson@umn.edu

[2] Minnesota Population Center, 50 Willey Hall, 225 – 19th Avenue South, Minneapolis, MN 55455; Email: takugler; dhaynes@umn.edu

Abstract

IPUMS Terra is a cyberinfrastructure project that integrates, preserves, and disseminates massive data collections describing characteristics of the human population and environment over the last six decades. IPUMS Terra has made a number of advances in the spatial demography of big data by making information interoperable between formats and across scientific communities. In this paper, we describe some of the challenges in dealing with these data, or 'deserts in the deluge' of big data, that are common to data and information sciences more broadly, and explore computational solutions specific to microdata, raster, and area data models. We examine in particular data and issues that are germane to spatial demography.

Introduction

Over the past half century, interregional differences in population growth rates, unprecedented urbanization, and international migration have led to profound shifts in the spatial distribution of global population. Economic changes have been dramatic as well, with global per-capita gross domestic product roughly doubling, while economic disparities grew in many countries (Rosa et al. 2010; Bloom 2011). These extraordinary demographic and economic changes have had positive impacts for many but also many troubling consequences, including environmental degradation, resource depletion, and climate change (Ehrlich, Kareiva, and Daily 2012). Changes in population size, characteristics, and behavior lie at the heart of these environmental challenges. The key drivers of change—especially fossil fuel emissions and deforestation—stem from population growth and economic development. At the same time, environmental change has profound implications for human behavior in that changes such as flooding, erosion of coastal areas, and drought are already profoundly affecting human societies and economies, and those effects will grow sharply in coming decades.

Demographic and economic data closely integrated with data on the environment are essential to describe the unfolding transformation of human and ecological systems. Data on the human population are crucial for understanding changes in the Earth's biological and climate processes; equally important, data on climate and the physical landscape are essential for understanding the impact of environmental change on human behavior and well-being (Millett and Estrin 2012). There is particular interest in big data on human environment systems. Big data in general refers to research centered on datasets that are much larger than those typically used in most fields. The size of these data in turn entail new forms of processing and analysis. There is a particular need for spatial big data in support of demography, seeing as most pressing environmental challenges are at their core population-environment ones.

However, there are deserts in the deluge of data. A recent special issue of *Science* argues that we need new solutions for dealing with the deluge of huge and complex human and environmental datasets in the face of critical shortcomings in our abilities for data manipulation, archiving, and discovery (2011). Social science data such as censuses and surveys are complex and unwieldy in terms of attributes (e.g., different questions and measures between survey instruments), space (e.g., census geographies that vary in space and scale), and time (e.g., changing measurement periods and frequencies). Raster-format environmental data pose different data integration challenges, including variation in projections and resolutions in addition to their massive sizes.

IPUMS Terra directly addresses these big-data challenges through data integration, machine-processable metadata, and sophisticated data discovery and access tools. Along the way, it addresses gaps in our ability to store, manipulate, and analyze spatial big data (after Wang and Liu 2009). The infrastructure dramatically cuts the costs of research by reducing redundant effort, encourages interdisciplinary research spanning the social and natural science divide, and enables access for a broad public audience. At broader conceptual levels, IPUMS Terra is helping address unresolved challenges in representing social and biophysical entities and relationships that operate at multiple levels of organization, over space, and through time. IPUMS Terra addresses these deserts in deluge of big spatial data and puts important tools into the hands of spatial demographers.

Deserts in the Deluge

Our understanding of the interactions between population and environment has been hampered by the paucity of data in some cases and the existence of data silos in others. Big data as a research area for human-environment dynamics is of great interest to many scientific communities, but there are deserts in the data deluge. Gaps in human-environment big data include:

- *Data*. For all the excitement about big data, scholars wanting to untangle human-environment interactions face a dearth of spatially-detailed multidecadal data. While some relevant data are available, such as climate observations or online opinions about global warming, there is surprisingly little detailed information about many social and natural features for most of the globe before the year 2000 (Manson and Kernik 2017).

- *Methods*. There are shortfalls in our ability to store, manipulate, and analyze the big data of human-environment systems. Most big data research is done on fairly simple data, in the sense they involve straightforward measures or a single research domain. In contrast, we face many unresolved challenges in representing social and biophysical entities and relationships that operate at multiple levels of organization, over space, and through time (Shook 2015).

- *Theory*. A growing number of big data proponents argue that the data deluge augers the ‘end of theory’ because this approach offers a powerful black-box approach that creates knowledge without needing domain experts or engagement with existing research, method, or theory. This proposition is profoundly at odds with the core conceptual precepts of many disciplines that seek to advance understanding of human-environment systems of interest to many spatial demographers (O'Sullivan and Manson 2015).

In terms of dealing with data needs, IPUMS-Terra is on track to be the largest curated source for global human-environment data. In addition to drawing in a range of environmental data, it draws on the larger IPUMS project, which is the largest population database in the world, with records on over half a billion individuals described by 270 billion data points. We identify, acquire, and develop various data sources ranging from historic handwritten census forms to current satellite observations of the earth. We research new ways to standardize these data and make them comparable across space and through time. We then preserve these data and make them internet-accessible so they are readily used by thousands of scholars and many other students, policy makers, and members of the public. We work with over a hundred national statistical agencies and other organizations around the world on the science of large, complex spatiotemporal datasets.

In terms of methodological demands, a major challenges lies in the fact that data describing characteristics of people and places are ordinarily maintained and analyzed in one of three major classes of data: microdata, area data, or raster data (Table 1). Each of these data classes has its own specialized analytic techniques and data manipulation tools, and very few investigators are fluent with the software and methods needed to use all three. Use of each data class tends to vary by discipline. Economists and sociologists most frequently describe behavior at the individual level, using microdata. Geographers are more likely to describe human behavior using area-level data, often using geographic units defined by statistical agencies, such as census tracts. Environmental scientists frequently analyze data in raster format, especially data derived from remote sensing or environmental modeling, and use area-level data to describe environmental features such as watersheds or floodplains.

Table 1. Spatiotemporal data classes and attendant representations

Spatiotemporal data classes	Spatial representations
Microdata: Each record represents an individual person, household, or firm	Each record includes a code identifying a geographic polygon (e.g., state, zip code) or a point (street address, longitude-latitude)
Area-level data: Each record contains information about a place. Data may include census tabulations, environmental categories, or laws and policies	Each record includes a code identifying a geographic polygon or point (e.g., state, zip code, environmental zone)
Raster data: Grid of pixels organized in rows and columns, where each pixel provides a value for a categorized or continuous variable (e.g., land cover category or mean precipitation)	Each grid cell corresponds to a rectangular area on the ground

Data integration transforms heterogeneous data from multiple sources into a unified schema. Most data integration focuses on a particular *class* of data. For example, a microdata integration project can harmonize the classifications, layout, metadata, and format of multiple microdata sets; a raster data integration project can reconcile projections, resolutions, format, and classifications of multiple raster datasets. Spatial demographers and other scholars interested in population-environment dynamics face two challenges, making heterogeneous data interoperable *within* each of the three major classes of spatiotemporal data, as well as making data easily interoperable *across* the major classes of data. Because each class of data is mainly used by particular scientific domains, breaking down these data silos also breaks down barriers to interdisciplinary research.

A primary challenge in data integration is the fact that basic data transformations—such as transforming raster data to area (technically, usually topological vector) data—are highly computationally intensive. Parallel computation, the usual solution for such problems, is fundamentally difficult for big spatiotemporal data (Eldawy 2015). Many computational problems are “embarrassingly parallel” because they can be solved by partitioning and distributing data among nodes in a computing cluster, solving the problem for a subset of data on each node, and then collating the results. However, such approaches are rarely workable for spatiotemporal data, due to the difficulty of preserving spatial and temporal relationships across nodes (Ding and Densham 1996). *Microdata* require a distribution algorithm that preserves relationships between individuals and their households. *Raster data* and *area-level data* embody complex spatial and topological relationships that are essential for answering most spatial problems, relationships that must be preserved when partitioning these data across nodes.

Parallel computing for spatial big data is an area of active research, but most existing computing platforms cannot handle multiple spatial data models or perform spatial data handling and analytics commonly found in even the most basic geographic information systems (Ray et al. 2015). In particular, most existing approaches to parallelization are limited in the data they can handle or provide extensions to existing specialized frameworks that are not well-suited to spatial data (such as MapReduce and column-store databases). While this research on extending existing parallelization approaches to new domains, including the spatial, is very exciting, existing systems are neither robust nor wide-ranging enough for a production environment like IPUMS Terra (Haynes et al. 2015).

Microdata. Microdata are typically stored as hierarchical fixed-width text or binary files in which each line represents an individual person or set of household characteristics. Challenges to high-speed processing of these individual-level data stem from the size and complexity of the data and having to conduct complicated queries across multiple samples with thousands of attributes and multiple relationships. We implemented Apache Spark’s Parquet columnar storage database and found significant performance gains across these queries over standard Java-based approaches. Query execution speed has improved by a factor of 10 to 300 for most common operations (per Armbrust et al. 2015).

Area-level data. Large area-level datasets are difficult to move into a parallel computing framework because spatial relationships such as adjacency and connectivity must be preserved across the nodes among which problems are decomposed (Ray et al. 2013; Puri and Prasad 2013). We use an industry-standard open-source spatial computing framework, PostgreSQL/PostGIS, because it offers deep data handling and analytical capabilities. However, PostgreSQL does not natively support parallel queries, although multiple projects are trying to scale PostgreSQL onto machine clusters (e.g., CitusDB, GridSQL, Postgres-XC, Postgres-XL, and Stado). We extensively tested the offerings of these projects and determined that they do not yet adequately support parallel spatial processing well. In the meantime, we developed a prototype vector analytic engine that divides PostgreSQL databases across computing nodes. This approach derives from research that demonstrates that parallel relational databases like PostgreSQL can perform significantly better than MapReduce systems (Pavlo et al. 2009). Our efforts have significantly improved performance in analyzing vector datasets, offering near linear speedup (when adding computing nodes) for simple topographical operations such as determining whether a polygon intersects a line or other polygon (Haynes et al. 2015; Ray et al. 2014). This research, in addition to helping IPUMS Terra users, also advances fundamental research in spatial high performance computing (Vo, Aji, and Wang 2014).

Raster data. Big-data raster datasets are difficult to make parallel in a computing environment because of the need to preserve spatial adjacency among grid cells and the complex kinds of computational operations needed to manipulate data. Joining PostgreSQL and PostGIS offers IPUMS Terra access to a comprehensive set of raster analysis tools but this combination does not handle large rasters well because row limit sizes (where rows are the primary data format in relational databases) are often exceeded by raster datasets (Stonebraker et al. 2011; Shook 2015). We have experimented with the use of array data structures in PostgreSQL, finding a three-times speed up in performance for many operations. We are also developing web applications to offer easy and fast access to these data via textual and web mapping interfaces, reducing the need for data users to know much about the exact characteristics of the data they seek (Manson et al. 2012).

A key conceptual challenge of big data is the need to allow domain scientists to develop theory outside of purely-inductive black box frameworks. To address this challenge, IPUMS Terra infrastructure offers tools that perform transformations between microdata, area data, and raster data in a way that allows domain experts to focus on answering conceptual questions in their fields without having to spend valuable time and resources on managing and manipulating data. Many of the transformations are conducted within a PostGIS database that houses much of the data collection. Stored procedures within the database convert area data to raster data by distributing variable values associated with each geographic unit across the grid cells that fall within the unit. Raster to area data transformations are performed by summarizing the values of the raster grid cells that fall within each geographic unit. Area data and summarized raster data are attached to microdata records based on the geographic unit. We have also developed a Spark-based tabulation engine that will allow users to create customized area-level tables from population microdata, bringing the power and flexibility of microdata to a new audience of researchers.

IPUMS Terra

IPUMS Terra is designed to make microdata, area-level data, and raster data easily interoperable. Using the IPUMS Terra extract builder, researchers can combine data from the three major data classes and obtain the results in any of the three classes. For example, IPUMS Terra can tabulate raster data derived from satellite images to determine the percentage of each census tract covered by trees and then attach that contextual information to each record of census microdata. Conversely, the software can tabulate population microdata to determine the density of in-migrants to each geographic unit and then convert those statistics into a gridded raster format.

The IPUMS Terra data collections are very large. They include microdata describing 250 billion microdata characteristics, 300 billion vector data points, and over a trillion pixels of raster data. The currently accessible IPUMS Terra data collection includes population data for over 170 countries, global long-term average and monthly time series climate data, a variety of global land cover and land use datasets, and the geographic boundaries necessary to support integration across the collection. IPUMS Terra incorporates global census microdata from IPUMS-International, covering 85 countries, 301 censuses and 672 million person records, as well as area-level tabulations of these data. For an additional 88 countries not yet participating in IPUMS-International, IPUMS Terra has acquired, processed, and made available aggregate subnational data on population by sex. IPUMS Terra raster data currently include Global Land Cover 2000, NASA's MODIS Land Cover, WorldClim long-term climate averages, Climate Research Unit monthly time series data, and Global Landscape Initiative data on cropland, pasture areas, and harvested area and yield for 175 crops (Friedl 2010, GLC 2000, Harris et al. 2014, Hijmans 2005, GLI 2015). We have developed boundary data for first-level administrative units in 169 countries and second-level administrative units for 91 countries. For 72 countries with multiple years of population data, we have created temporally harmonized boundaries providing geographic units that are stable over time.

To integrate and disseminate this vast data collection, we developed workflows and supporting software tools for processing data and metadata. We developed a suite of Python-ArcGIS tools that enable efficient boundary data processing (Kugler 2015). These tools match digital boundaries to the geographic units represented in current and historic population datasets, automate temporal harmonization, and manage regionalization necessary to protect respondent confidentiality (Kugler 2017). We are developing additional Python tools transform aggregate population data published by national statistical offices in diverse formats into standard data structures that can be ingested into the IPUMS Terra database. IPUMS Terra tools and workflows are supported by a custom-built metadata management application that tracks data provenance from the original sources through all IPUMS Terra processing steps and produces complete descriptions of the final data.

IPUMS Terra's integrated data are disseminated to researchers, educators, policy analysts, and others through three public-facing applications that communicate with the IPUMS Terra database and extract engine through an Application Programming Interface (API). The primary application is the IPUMS Terra extract builder (Figure 1). The extract builder allows users to browse the entire IPUMS Terra data collection, access rich metadata describing the data, select variables and datasets from any of the three data classes, and combine diverse data sources into a single integrated dataset in the format appropriate for a particular investigation.

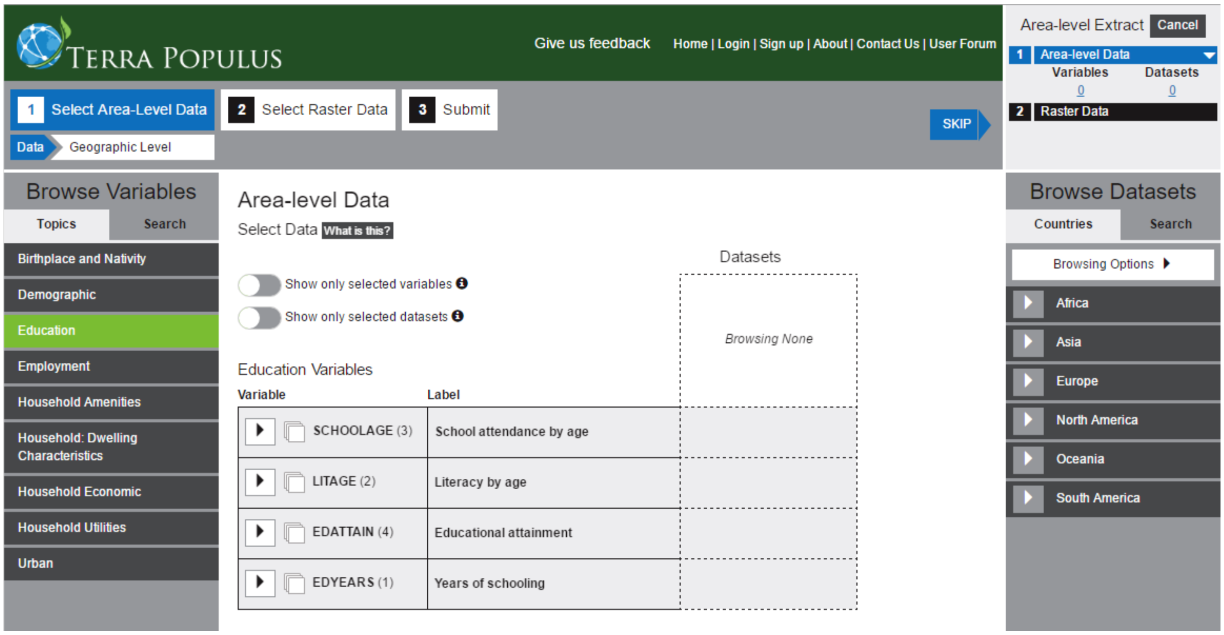

Figure 1. IPUMS Terra web interface for combining and abstracting data.

In addition to the main extract builder, two other applications provide alternate ways to access IPUMS Terra data. For users primarily interested in raster data, TerraClip provides an easy way to construct country-level subsets of global raster datasets, bypassing the need to identify and stitch together tiles or deal with otherwise cumbersome data access systems. TerraScope provides a map-based interface with which users can explore the area and raster portions of the IPUMS Terra collection, produce basic visualizations such as choropleth maps, and download the area-level data used in the visualizations.

Conclusion

IPUMS Terra offers one of the largest and most comprehensive publicly-available collections of data on human activities and behavior. These data are integrated with high-resolution geospatial information describing the environment, filling a major substantive gap. The IPUMS Terra data collection and integration infrastructure are a powerful resource for understanding the causes and consequences of the transformations in society and the environment that are reshaping the planet. By providing richly detailed, tightly integrated data spanning the world over many decades, IPUMS Terra is creating a unique test bed for social, economic, ecological, and climate models.

Because of the richness of the data and the ease with which it can be accessed and exploited, the audiences for IPUMS Terra are extraordinarily diverse. In addition to social and environmental scientists, IPUMS Terra is used by planners, journalists, students and their teachers, and the broader public. The data collection makes important contributions to education and public understanding.

IPUMS Terra also advances data science. The challenges of the project derive not only from the large scale of the data collections but also from their complexity. The population and environmental data are multiscale over time and space, have multiple levels of hierarchy, and cover a remarkable range of topics. To manage the scale, complexity, and heterogeneity of the data, the project engages the leading edge of data science and develops new technologies and processes. Innovative solutions are needed through the entire data life cycle, including collection, preservation, analysis, dissemination, and long-term access and management. IPUMS Terra provides open-source software, metadata, and workflows that can overcome these challenges and that can be readily adapted to spatiotemporal data in multiple scientific domains. In particular, our work on spatial high-performance computing will address critical bottlenecks in the integration and dissemination of massive spatiotemporal datasets. The infrastructure is directly applicable to urgent and compelling scientific and policy issues and meets the needs of large and diverse audience that extends beyond the academy.

Acknowledgements

This work is supported in part by the National Science Foundation OCI: Terra Populus: A Global Population/Environment Data Network (0940818), the National Institutes of Health supported Minnesota Population Center (R24 HD041023), and the Resident Fellowship program of the Institute on the Environment.

References

Armbrust, Michael, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K Bradley, Xiangrui Meng, Tomer Kaftan, Michael J Franklin, and Ali Ghodsi. 2015. “Spark Sql: Relational Data Processing in Spark.” In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 1383–94. Melbourne, Australia: ACM.

Bloom, David E. 2011. “7 Billion and Counting.” Science 333 (6042). American Association for the Advancement of Science: 562–69.

Ding, Yuemin, and Paul J Densham. 1996. “Spatial Strategies for Parallel Spatial Modelling.” International Journal of Geographical Information Systems 10 (6). Taylor & Francis: 669–98.

Ehrlich, Paul R, Peter M Kareiva, and Gretchen C Daily. 2012. “Securing Natural Capital and Expanding Equity to Rescale Civilization.” Nature 486 (7401): 68–73.

Eldawy, Ahmed, and Mohamed F Mokbel. 2015. “The Era of Big Spatial Data: A Survey.” Information and Media Technologies 10 (2). Information and Media Technologies Editorial Board: 305–16.

Friedl, M.A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley and X. Huang 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, 2001-2012, Collection 5.1 IGBP Land Cover, Boston University, Boston, MA, USA.

Global Land Cover 2000 database. 2003. European Commission, Joint Research Centre, http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php

Global Landscapes Initiative. 2015. Accessed at http://gli.environment.umn.edu/, 3/26/2105.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. 2014. Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *International Journal of Climatology* 34, 623-642. doi:10.1002/joc.3711

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Haynes, David, Suprio Ray, Steven M Manson, and Ankit Soni. 2015. “High Performance Analysis of Big Spatial Data.” In Big Data 2015: IEEE International Conference on Big Data, 1953–57. Santa Clara, California: Institute of Electrical and Electronics Engineers.

Kugler, Tracy A, David C Van Riper, Steven M Manson, David A Haynes II, Joshua Donato, and Katie Stinebaugh. 2015. “Terra Populus: Workflows for Integrating and Harmonizing Geospatial Population and Environmental Data.” Journal of Map & Geography Libraries 11 (2). Taylor & Francis: 180–206.

Kugler, Tracy A., Steven M. Manson, and Joshua R. Donato. "Spatiotemporal aggregation for temporally extensive international microdata." Computers, Environment and Urban Systems 63 (2017): 26-37.

Manson, S. M., L. Kne, K. Dyke, J. Shannon, and S. Eria (2012). Using eye tracking and mouse metrics to test usability of web mapping navigation. Cartography and Geographic Information Science 39 (1): 48-60

Manson S.M., and M. Kernik (2017). Human–Environment Interactions and Scalable Remote Sensing, Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2017. DOI: 0.1016/B978-0-12-409548-9.10413-

Millett, Lynette I, and Deborah L Estrin. 2012. Computing Research for Sustainability. National Academies Press.

Nelson, E., H. Sander, P. Hawthorne, M. Conte, S M Manson, and S. Polasky. 2010. “Projecting Global Land Use Change and Its Effect on Ecosystem Service Provision and Biodiversity with Simple Techniques.” PLoS ONE 5 (12): e14327.

O'Sullivan, D. and S. M. Manson (2015). Do Physicists Have ‘Geography Envy’? And What Can Geographers Learn From It? Annals of the Association of American Geographers 105 (4): 704-722.

Pavlo, Andrew, Erik Paulson, Alexander Rasin, Daniel J Abadi, David J DeWitt, Samuel Madden, and Michael Stonebraker. 2009. “A Comparison of Approaches to Large-Scale Data Analysis.” In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, 165–78. Providence, Rhode Island: ACM.

Puri, Shruti, and Sushil K Prasad. 2013. “Efficient Parallel and Distributed Algorithms for GIS Polygonal Overlay Processing.” In Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 27th International, 2238–41. Boston, Massachusetts: IEEE.

Ray, Suprio, Angela Demke Brown, Nick Koudas, Rolando Blanco, and Anil K Goel. 2015. “Parallel in-Memory Trajectory-Based Spatiotemporal Topological Join.” In Big Data (Big Data), 2015 IEEE International Conference on, 361–70. Santa Clara, California: Institute of Electrical and Electronics Engineers.

Ray, Suprio, Bogdan Simion, Angela Demke Brown, and Ryan Johnson. 2013. “A Parallel Spatial Data Analysis Infrastructure for the Cloud.” In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 284–93. ACM.

———. 2014. “Skew-Resistant Parallel in-Memory Spatial Join.” In Proceedings of the 26th International Conference on Scientific and Statistical Database Management, 6–12. Aalborg, Denmark: Association for Computing Machinery.

Rosa, Eugene A, Andreas Diekmann, Thomas Dietz, and Carlo Jaeger. 2010. Human Footprints on the Global Environment: Threats to Sustainability. Cambridge, MA: MIT Press.

Science Magazine. 2011. Dealing with Data. Special Issue. Science 331: 649–789.

Shook, E. 2015. Parallel Cartographic Modeling. In Onsrud, H. & Kuhn, W. (eds) Advancing Geographic Information Science: The Past and Next Twenty Years, 319–24. Washington DC: GSDI Association Press.

Stonebraker, Michael, Paul Brown, Alex Poliakov, and Suchi Raman. 2011. “The Architecture of SciDB.” In Scientific and Statistical Database Management, 1–16. Berlin: Springer.

Vo, Hoang, Ablimit Aji, and Fusheng Wang. 2014. “SATO: A Spatial Data Partitioning Framework for Scalable Query Processing.” In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 545–48. Dallas, Texas: Association for Computing Machinery.

Wang, Shaowen, and Yan Liu. 2009. “TeraGrid GIScience Gateway: Bridging Cyberinfrastructure and GIScience.” International Journal of Geographical Information Science 23 (5). Taylor & Francis: 631–56.