Deserts in the Deluge: IPUMS-Terra the 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 (also called TerraPop) 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. TerraPop 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 challenges of these data, or 'deserts in the deluge' of data, that are common to spatial big data more broadly, and explore computational solutions specific to microdata, raster, and vector data models.

Introduction

Over the past six decades, the world's population more than doubled. Sharp interregional differences in growth rates—together with unprecedented urbanization and international migration—have led to dramatic spatial redistribution of population. Economic changes were equally remarkable, as world per-capita gross domestic product roughly doubled (Rosa et al. 2010; Bloom 2011). This extraordinary global demographic and economic growth has ushered in alarming environmental degradation, resource depletion, and climate change (Ehrlich, Kareiva, and Daily 2012).

Scientific and policy bodies have called for more richly-detailed data to support the research and informed decisions necessary to meet the challenges of rapid social and environmental change (Millett and Estrin 2012). There is particular interest in the 'data deluge' or 'big data', or research based on datasets that are vastly larger than those traditionally used in most fields, and which in turn entail new forms of processing and analysis. In particular, there is a deep need for a spatial demography of big data, seeing as most pressing environmental challenges are at their core population-environment ones.

However, there are deserts in the deluge of data. At the level of data as such, scholars untangling human-environment interactions face a dearth of spatially-detailed multidecadal data. While some relevant data are available, such as climate observations, there is surprisingly little detailed information about many social and natural features for most of the globe before the year 2000 (Nelson et al. 2010). At the level of methods, there are similar shortfalls in our ability to store, manipulate, and analyze spatial big data (Wang and Liu 2009). And at the level of theory, 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 (O'Sullivan and Manson 2015). TerraPop address these deserts in deluge of big spatial data and puts important tools into the hands of spatial demographers.

TerraPop

TerraPop addresses challenges in data, methods, and theory in the spatial demography of big data by using location-based data integration to make heterogeneous data interoperable and thereby break down barriers to interdisciplinary research. Researchers can combine data across three major data classes – microdata, raster, and area-level. For example, TerraPop can summarize raster data derived from satellite images to determine the percentage of each municipio in Brazil covered by trees, and then attach that contextual information to each record of census microdata (or data that represents an individual person or set of household). TerraPop has population data for over 170 countries, global long-term climate data, a variety of global land cover and land use datasets, and the geographic boundaries necessary to support integration across the collection. Many of these data sets are unique (especially those on demographics and socioeconomic characteristics) and help address one of the primary deserts in the deluge, the dearth of data on human populations for much of the globe prior to 2000. TerraPop makes these global datasets interoperable across time and space, disseminates them to the public and to multiple research communities, and preserves these resources for future generations.

Spatial high-performance computing

We addressed a number of fundamental challenges in the spatial demography of big data in order 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 of current and historic population datasets, automate temporal harmonization, and manage regionalization to protect respondent confidentiality (Kugler et al. 2015). We also developed a metadata management application that tracks data provenance from the original sources through all TerraPop processing steps and produces complete descriptions of the final data.

At the core of the TerraPop infrastructure is a set of spatial high-performance computing solutions that transform microdata, vector data, and raster data. We have microdata describing 250 billion microdata characteristics, 300 billion vector data points, and over a trillion pixels of raster data. These data are available via a web interface or application program interface (Figure 1). These large datasets create research opportunities and challenges. Parallel computation, the usual solution for such problems, is fundamentally difficult for big spatiotemporal data (Eldawy and Mokbel 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.

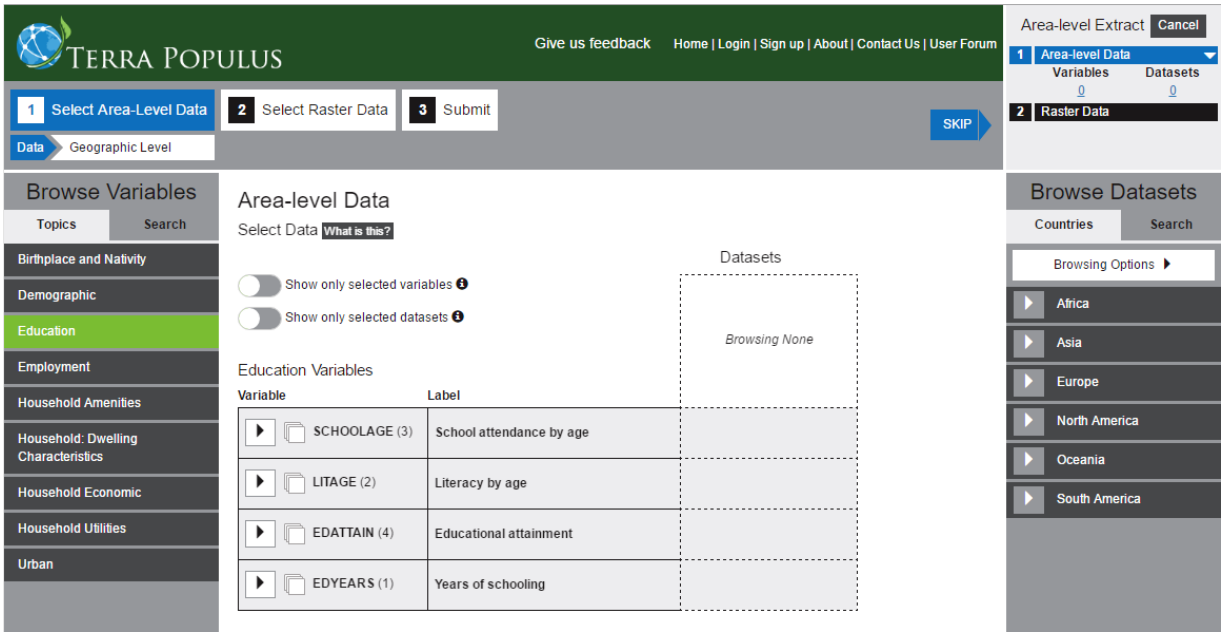

Figure 1. TerraPopulus web interface for combining and abstracting data.

Standard high-performance computing approaches are often inefficient or unworkable 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 vector data embody complex spatial and topological relationships that are essential for answering most spatial problems, relationships that must be preserved when partitioning 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 approaches to parallelization are limited in scope or provide extensions to existing frameworks such as MapReduce and column store databases. While this work is very promising, no existing systems are robust or wide-ranging enough for GIScience production environments like TerraPop (Haynes et al. 2015).

Microdata. Microdata are most often stored as hierarchical fixed-width text or binary files, where each line represents an individual person or set of household characteristics. Challenges to high-speed processing of individual-level data derive from the size and complexity of the data and the need to conduct complicated queries across multiple samples with thousands of attributes and multiple embedded 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 increased by a factor of 10 to 300 for a variety of common operations and using Parquet promises further gains because Parquet offers record shredding and assembly (Armbrust et al. 2015).

Vector data. Large vector datasets are difficult to parallelize because spatial relationships such as adjacency and connectivity must be preserved across nodes (Ray et al. 2013; Puri and Prasad 2013). We use the leading open-source spatial computing framework, PostgreSQL/PostGIS,

because it offers deep data handling and analytical capabilities. However, PostgreSQL does not natively support parallel queries, though multiple projects are trying to scale PostgreSQL onto machine clusters (e.g., GridSQL, Stado, Postgres-XC, CitusDB and Postgres-XL). We extensively tested these projects and determined that they do not support parallel spatial processing well (although several projects are working on the problem) so we have been developing a prototype vector analytic engine that partitions a PostgreSQL database across computing nodes. We chose this approach based on evidence that parallel relational databases like PostgreSQL can perform significantly better than MapReduce systems (Pavlo et al. 2009). Our work to date has significantly improved performance in analyzing vector datasets, offering near linear speedup when adding nodes by sharding spatial queries across a cluster of machines where a PostgreSQL database instance is run on each node 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 work thereby addresses fundamental research needs in spatial high performance computing (Vo, Aji, and Wang 2014).

Raster data. Large raster datasets are difficult to parallelize because of the sheer volume of data involved, the need to preserve spatial relationships among grid cells, and the large number of varying raster operations that are needed to manipulate data. While the combination of PostgreSQL/PostGIS offers a comprehensive set of raster analytics, that approach does not handle large rasters well because row limit sizes are often exceeded by raster datasets (Stonebraker et al. 2011). By experimenting with array data structures, we have doubled or tripled performance for most operations while ensuring that larger raster layers do not fail outright. We are also experimenting with web applications to offer easy and fast access to these data via textual and web mapping interfaces (Manson et al. 2012).

Conclusion

TerraPop incorporates the largest and most comprehensive available collections of data on human activities and behavior, along with important global environmental datasets. 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, we will engage the leading edge of data science and develop 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. TerraPop will provide open-source software, metadata, and workflows that can overcome these challenges and that can readily be 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.

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.

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.

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

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.

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.