Geocomputational infrastructure for population-environment data

S. M. Manson,*[1] T. Kugler, D. Van Riper, J. Schroeder, and S. Ruggles [2]

,[1] Department of Geography, Environment, and Society, University of Minnesota, Minneapolis, MN 55455
[2] Institute for Social Research and Data Innovation, University of Minnesota, Minneapolis, MN 55455
*Email: manson@umn.edu

Abstract

Geocomputation is increasingly integrated with spatial data infrastructure to develop and deliver massive datasets and attendant analysis and visualization capacity to a wide range of users. IPUMS Terra is spatial data infrastructure that develops and uses geocomputational approaches to provide one of the largest collections of integrated population and environment data in the world. In this paper, we describe new efforts to fundamentally change the landscape of population-environment data by integrating, preserving, and disseminating vast amounts of aggregate census and agricultural census data. We are developing data manipulation tools and workflow management approaches to transform and standardize data as well as capture metadata. These developments in turn facilitate the processing, documenting, and intake of tens of thousands of data tables into IPUMS Terra, which then are shared with the scientific community and the broader public to advance understanding of the population and agricultural systems that are central to many complex human-environment systems.

Keywords: Big data, population, environment, aggregate census, agricultural census

1. Geocomputation and population-environment data

Geocomputation is increasingly integrated with spatial data infrastructure to develop and deliver massive datasets and attendant analysis and visualization capacity to a wide range of users. As Batty (2017) notes, geocomputation has a long pedigree of merging the digital, computational, and geographical. Paralleling the evolution of geocomputation has been the development of spatial data infrastructure, sociotechnical frameworks that bring together locational data and metadata with users and tools over computer networks (Hendriks, Dessers and Van Hootegem, 2012). With the advent and rapid growth of the internet, geocomputation has encompassed spatial data infrastructure elements including cyberinfrastructure, web mapping, and high-performance computing (Longley *et al.*, 1998; Abrahart and See, 2014; Brunsdon and Singleton, 2015).

Geocomputation and spatial data infrastructure are increasingly integral to addressing the pressing challenges posed by rapidly changing coupled human-environment systems. Extraordinary levels of population and economic growth since the 1950s have come at the cost of environmental degradation and climate change. Changes in population size, characteristics, and behavior lie at the heart of these environmental challenges (Bloom, 2011). The key drivers of change—especially fossil fuel emissions and deforestation—are directly tied to population growth and economic development (Rosa *et al.*, 2010). Attendant environmental change has profound implications for human societies through impacts of flooding, erosion of coastal areas, and drought among many other existing or potential threats (Ehrlich, Kareiva and Daily, 2012).

Many scientific and policy bodies are calling for better data to support critical research and inform decisions needed to meet the challenges of rapid human-environmental change (Millett and Estrin, 2012). Social data closely integrated with earth systems data are essential to describe the unfolding transformation of human and ecological systems. Of particular interest is big data, datasets that are larger and more difficult to handle than those typically used, which in turn require new forms of processing and analysis for research and policy development (Holm *et al.*, 2013).

IPUMS Terra exemplifies how integrating geocomputation and cyberinfrastructure can provide the data necessary to address a range of human-environment challenges. IPUMS Terra (https://terra.ipums.org/) is spatial data infrastructure that offers one of the largest collections of integrated population and environment data in the world. This paper describes a new IPUMS Terra project to fundamentally change the landscape of population-environment data by capturing vast new datasets of population and agricultural census data. In doing so, we are advancing geocomputation and spatial data infrastructure. This paper offers an overview of IPUMS Terra and then looks at progress in an effort to integrate, preserve, and disseminate vast troves of aggregate census and agricultural data currently scattered around the globe. We are developing data manipulation tools to transform diversely formatted published tables into a standardized structure, including extending our workflow management tools to capture metadata. These advances in turn allow us to process, document, and intake tens of thousands of these newly standardized tables into IPUMS Terra, which may then be shared with the scientific community and the broader public to advance scientific and public understanding of the population and agricultural systems that are the linchpins of many complex human-environment systems.

2. IPUMS Terra: geocomputation for spatial infrastructure

IPUMS Terra is a cyberinfrastructure project that integrates, preserves, and disseminates massive data collections that measure and describe characteristics of the human population and environment over the last six decades. IPUMS Terra uses a range of computation approaches, including geocomputation, to integrate data via their location and make heterogeneous (in the spatial, temporal, and attribute domains) data interoperable. IPUMS Terra has population data for over 170 countries, long-term high-resolution global scale climate data, global land cover and land use datasets, and the geographic boundary information necessary to support integration across the collection. IPUMS Terra makes these global datasets interoperable in the temporal, spatial, and attribute domains and disseminates them to a wide array of research communities and beyond, and preserves these resources for future generations.

IPUMS Terra employs a range of geocomputational approaches to address a number of fundamental research challenges in order to integrate and disseminate this vast data collection. At the core of this infrastructure are is spatial high-performance computing that transform data including microdata on 250 billion personal characteristics, 300 billion vector data points, and over a trillion pixels of raster data. IPUMS Terra uses location-based data integration to make area-level data, microdata, and raster data easily interoperable (Haynes, Manson and Shook, 2017). Although transformations between area-level and raster data classes can be carried out using GIS software, these software systems are difficult to learn and use for most domain scientists, and processing large or complex datasets can be prohibitively difficult. The project has created workflows and software for processing data and metadata, along with 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).

IPUMS Terra simplifies and streamlines key operations, reduces redundant effort, solves challenges of manipulating large datasets, and places powerful yet easy-to-use cyberinfrastructure in the hands of domain scientists. IPUMS Terra is fast and robust enough to accommodate the vast volumes of data required for even the most complex analyses (Haynes *et al.*, 2015). The infrastructure dramatically reduces the costs of research by reducing redundant effort, encourages interdisciplinary research spanning the social and natural science divide, enables access for a broad public audience, and preserves these data for future generations. In the citation for an award from the *Journal of Map and Geography Libraries*, IPUMS Terra was described as a broadly applicable model for future data infrastructure projects that will lead to a new and deeper understanding of research findings (JMGL, 2016).

3. Freeing aggregate census and agricultural data

This paper focuses on how IPUMS Terra is integrating data from over 1,000 population and agricultural censuses describing the characteristics of subnational regions around the world. Detailed subnational data on human population and agricultural production are useful for understanding changes in coupled human-environment systems. Ongoing expansion and enhancement of IPUMS Terra will make a trove of census data accessible for global-scale analyses, stimulating a new class of studies on the spatial organization of human activity. These public domain data are currently unusable for systematic scientific analysis because they are locked in tens of thousands of PDF files, spreadsheets, and paper documents dispersed across hundreds of websites and libraries and in heterogeneous formats with no machine-processable metadata. When these data are available through IPUMS, researchers will be able to discover and obtain these valuable data, along with environmental data describing land use, land cover, and climate, in customized integrated datasets ready for analysis.

Adding aggregate census data provides global coverage and geographic detail describing population characteristics and will introduce new area-level agricultural data. More than ninety countries have partnered with IPUMS International to prepare and disseminate integrated microdata (Ruggles, 2014), or data on specific individuals and households. However, we lack microdata for over 100 countries, including large nations such as Russia, Australia, and the Democratic Republic of the Congo. To make studies of the interrelationships of human activity and the environment throughout the entire world feasible, we must turn to area-level population data for administrative units within countries. Fortunately, such area-level data are available for virtually every country in the world across decades (Figure 1).

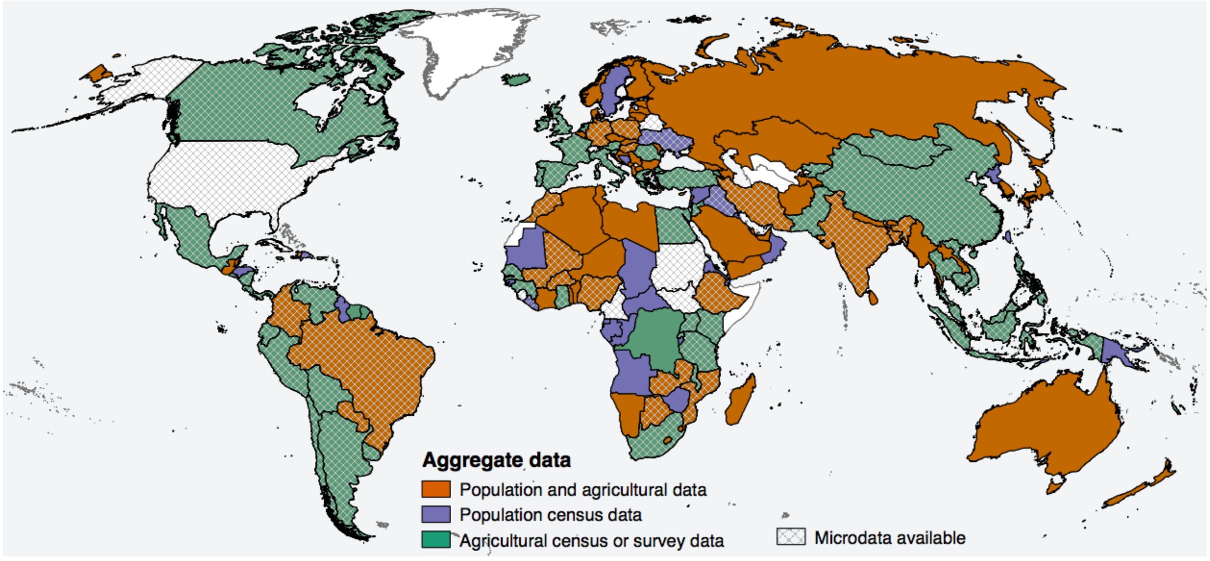
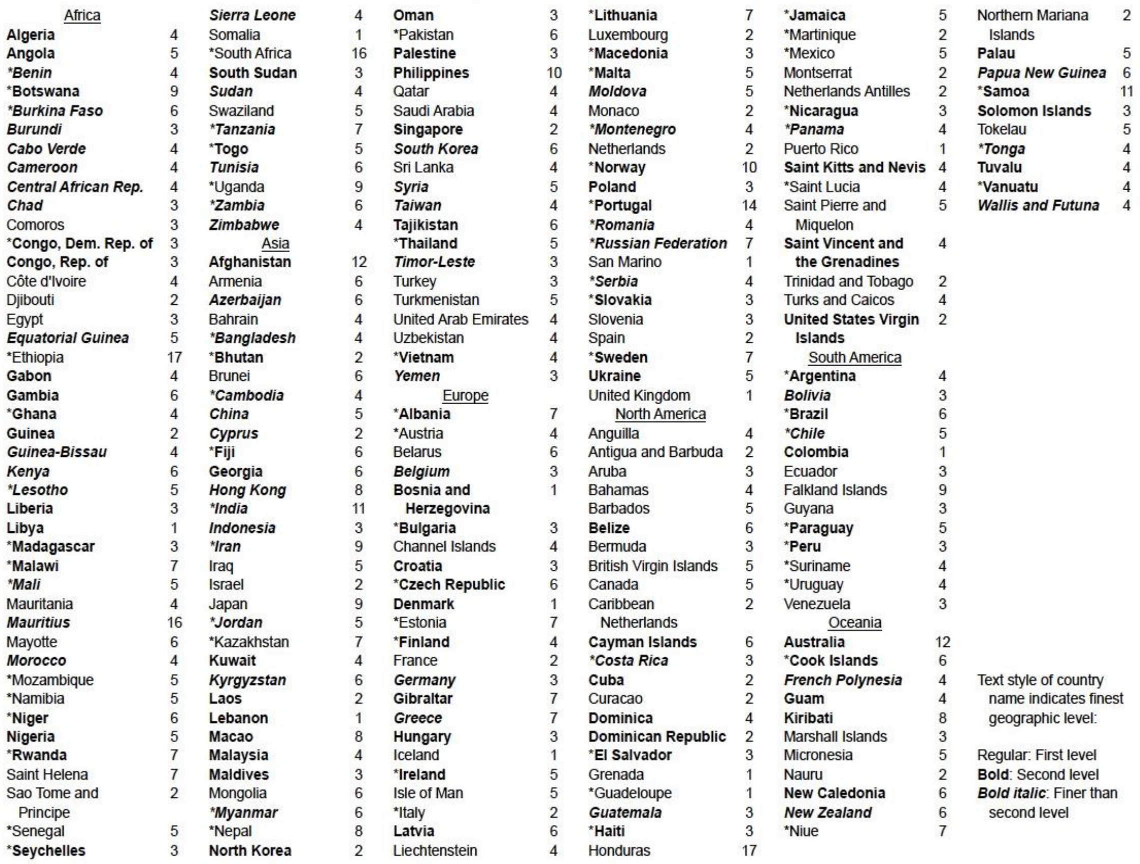

Aggregate data
Population and agricultural data
Population census data
Agricultural census or survey data
Microdata available

Figure 1: Aggregate data and microdata holdings

Even where microdata are available, area-level data provide important geographic detail. Due to the potentially sensitive nature of individual-level data, the geographic detail provided with microdata is restricted. In most countries, many second level administrative units (parallel to counties in the United States) must be combined to meet the population threshold. In contrast, aggregate census data are typically available for all second-level units and often for finer levels of geography. When these aggregate census data are linked to geographic boundaries delineating the places they describe, researchers can study spatial patterns for a broad array of human-environment dynamics.

IPUMS Terra draws on four key sources of international census data: 1) population census data published on national statistical office (NSO) websites, 2) population census reports assembled by the United Nations Population Division under their Data Archive project, 3) agricultural census reports made available by the UN Food and Agriculture Organization (FAO), and 4) agricultural census reports assembled by the International Food Policy Research Institute (IFPRI) and the University of Minnesota’s HarvestChoice project. These sources cover over 1,000 censuses, and include tens of thousands of tables (Table 1).



Africa	
__Algeria__	4
__Angola__	5
___*Benin___	4
*__Botswana__	9
___*Burkina Faso___	6
___Burundi___	3
___Cabo Verde___	4
___Cameroon___	4
___Central African Rep.___	4
___Chad___	3
Comoros	3
*__Congo, Dem. Rep. of__	3
__Congo, Rep. of__	3
Côte d'Ivoire	4
Djibouti	2
Egypt	3
___Equatorial Guinea___	5
*Ethiopia	17
__Gabon__	4
__Gambia__	6
*__Ghana__	4
__Guinea__	2
___Guinea-Bissau___	4
___Kenya___	6
___*Lesotho___	5
__Liberia__	3
__Libya__	1
*__Madagascar__	3
*__Malawi__	7
___*Mali___	5
Mauritania	4
___Mauritius___	16
Mayotte	6
___Morocco___	4
*Mozambique	5
*Namibia	5
*__Niger__	6
__Nigeria__	5
*__Rwanda__	7
Saint Helena	7
Sao Tome and Principe	2
*Senegal	5
*__Seychelles__	3
___Sierra Leone___	4
Somalia	1
*South Africa	16
__South Sudan__	3
___Sudan___	4
Swaziland	5
___*Tanzania___	7
*__Togo__	5
___Tunisia___	6
*Uganda	9
___*Zambia___	6
___Zimbabwe___	4
Asia	
__Afghanistan__	12
Armenia	6
___Azerbaijan___	6
Bahrain	4
___*Bangladesh___	4
*__Bhutan__	2
Brunei	6
___*Cambodia___	4
___China___	5
___Cyprus___	2
*__Fiji__	6
__Georgia__	6
___Hong Kong___	8
___*India___	11
___Indonesia___	3
___*Iran___	9
Iraq	5
Israel	2
Japan	9
___*Jordan___	5
*Kazakhstan	7
__Kuwait__	4
___Kyrgyzstan___	6
__Laos__	2
__Lebanon__	1
__Macao__	8
__Malaysia__	4
__Maldives__	3
Mongolia	6
___*Myanmar___	6
*Nepal	8
__North Korea__	2
__Oman__	3
*Pakistan	6
__Palestine__	3
__Philippines__	10
Qatar	4
Saudi Arabia	4
__Singapore__	2
___South Korea___	6
Sri Lanka	4
___Syria___	5
___Taiwan___	4
__Tajikistan__	6
*__Thailand__	5
___Timor-Leste___	3
Turkey	3
Turkmenistan	5
United Arab Emirates	4
Uzbekistan	4
*__Vietnam__	4
___Yemen___	3
Europe	
*__Albania__	7
*Austria	4
Belarus	6
___Belgium___	3
__Bosnia and Herzegovina__	1
*__Bulgaria__	3
Channel Islands	4
__Croatia__	3
*__Czech Republic__	6
__Denmark__	1
*Estonia	7
*__Finland__	4
France	2
___Germany___	3
__Gibraltar__	7
___Greece___	7
__Hungary__	3
Iceland	1
*__Ireland__	5
Isle of Man	5
*Italy	2
__Latvia__	6
Liechtenstein	4
*__Lithuania__	7
Luxembourg	2
*__Macedonia__	3
*__Malta__	5
___Moldova___	5
Monaco	2
___*Montenegro___	4
Netherlands	2
*__Norway__	10
__Poland__	3
*__Portugal__	14
___*Romania___	4
___*Russian Federation___	7
San Marino	1
___*Serbia___	4
*__Slovakia__	3
Slovenia	3
Spain	2
*__Sweden__	7
__Ukraine__	5
United Kingdom	1
North America	
Anguilla	4
Antigua and Barbuda	2
Aruba	3
Bahamas	4
Barbados	5
__Belize__	6
Bermuda	3
British Virgin Islands	5
Canada	5
Caribbean Netherlands	2
__Cayman Islands__	6
___*Costa Rica___	3
__Cuba__	2
Curacao	2
__Dominica__	4
__Dominican Republic__	2
*__El Salvador__	3
Grenada	1
*Guadeloupe	1
___Guatemala___	3
*__Haiti__	3
Honduras	17
*__Jamaica__	5
*Martinique	2
*Mexico	5
Montserrat	2
Netherlands Antilles	2
*__Nicaragua__	3
___*Panama___	4
Puerto Rico	1
__Saint Kitts and Nevis__	4
*Saint Lucia	4
Saint Pierre and Miquelon	5
__Saint Vincent and the Grenadines__	4
Trinidad and Tobago	2
Turks and Caicos	4
__United States Virgin Islands__	2
South America	
*__Argentina__	4
___Bolivia___	3
*__Brazil__	6
___*Chile___	5
__Colombia__	1
Ecuador	3
Falkland Islands	9
Guyana	3
*__Paraguay__	5
*__Peru__	3
*Suriname	4
*Uruguay	4
Venezuela	3
Oceania	
__Australia__	12
*__Cook Islands__	6
___French Polynesia___	4
__Guam__	4
__Kiribati__	8
Marshall Islands	3
Micronesia	5
Nauru	2
__New Caledonia__	6
___New Zealand___	6
*Niue	7
Northern Mariana Islands	2
__Palau__	5
___Papua New Guinea___	6
*__Samoa__	11
__Solomon Islands__	3
Tokelau	5
___*Tonga___	4
__Tuvalu__	4
*__Vanuatu__	4
___Wallis and Futuna___	4

Text style of country name indicates finest geographic level:

Regular: First level
__Bold__: Second level
___Bold italic___: Finer than second level

Table 1: Number of censuses by country/territory; asterisks indicate countries with agricultural censuses

The geographic detail and topical depth found in published population and agricultural censuses exceeds that of existing global scale data products. Sources such as the UN's FAOSTAT portal (www.fao.org/faostat) and the World Bank Open Data portal (data.worldbank.org) provide access to many socioeconomic and agricultural variables but their data are almost entirely at the national level. Several global gridded datasets provide spatial detail, but lack topical and temporal depth. The Gridded Population of the World (GPW) provides global extent grids of total population counts and densities from 2000-2020 at five year intervals (Doxsey-Whitfield *et al.*, 2015). WorldPop provides grids of total population counts, counts by age, counts of births and pregnancies, and internal migration flows for some regions from census and survey data from around 2010 (www.worldpop.org.uk.). EarthStat provides global extent grids on crop production, yields, and fertilizer consumption for 140 crops, fertilizer application for 17 crops, yield gaps for 16 crops, and water depletion and greenhouse gas emissions from croplands from agricultural census data from 2000 forward (www.earthstat.org). These are all high-quality data products and widely used, but their limitations restrict the types of research questions scientists may pursue.

Population censuses report on a variety of population characteristics, often at multiple scales of geography. At a minimum, census reports include counts of the population in each administrative unit of the country, typically broken down by age and sex. Most census reports also cover other demographic characteristics such as marital status and household structure; cultural characteristics such as birthplace, migration, language, ethnicity and religion; educational characteristics such as literacy, school attendance, and educational attainment; employment characteristics such as labor force participation, employment status, and industry sectors; and housing characteristics such as availability of utilities and amenities, age and size of housing units, and ownership status.

Agricultural censuses describe characteristics of agricultural production. These attributes include counts of farms and the demographics of those living or working on farms; land tenure; farm size and area devoted to different land use types (e.g., pasture, cropland, fallow); crop characteristics such as yields and areas under cultivation of various crop types; livestock characteristics such as the count, type, and age and sex distribution. Many countries also report on topics such as irrigation, pesticide use, fertilizer type and use, and access to financial institutions and farm equipment.

Incorporating this massive body of aggregate census data into IPUMS reduces redundant effort by researchers. Currently, if researchers wish to work with subnational census data for spatial analysis, they must 1) search for and locate the reports, often sifting through dozens of documents across multiple countries; 2) identify tables of interest from within the text document; 3) extract the data from the page (by copying and pasting or using conversion tools to move tables from PDF to Excel or another data structure); 4) carry out the manipulation needed to prepare the data for analysis, especially if using data from multiple tables or different censuses; 5) locate geographic boundary files for the units described in the tables; and 6) join the tables to the geographic data. By bringing tens of thousands of aggregate census tables together in a single searchable interface and supplying the data in a standardized structure with linked boundary files, we greatly streamline the process of acquiring and manipulating these rich data, allowing researchers to focus on their substantive analyses.

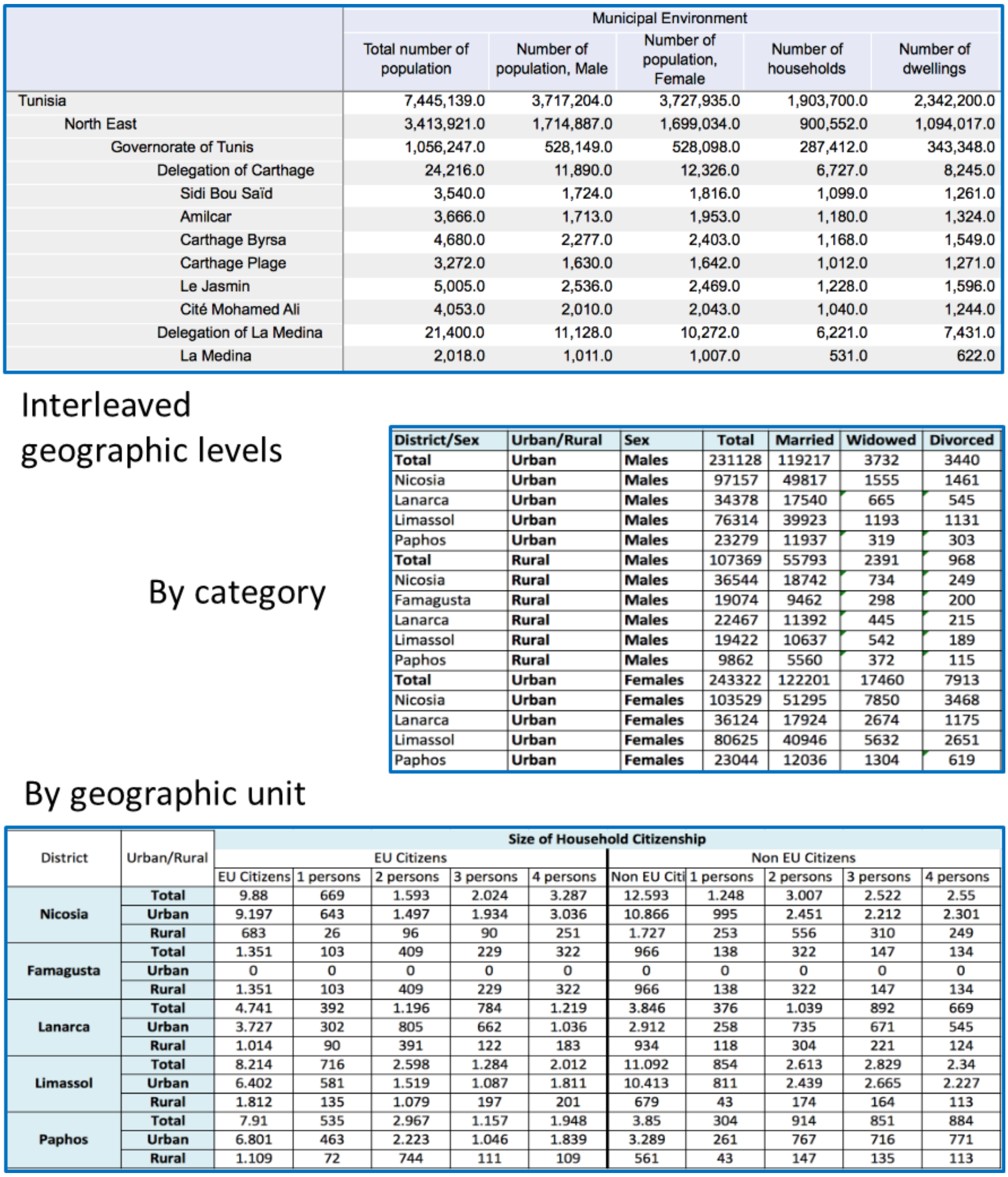
Capturing these population and agricultural census tables and associated metadata will also serve an important preservation mission. Data are easily lost (Sobek *et al.*, 2011). Technology changes and human error can lock data away in outdated data formats and separate the data files from the documentation necessary to understand and use the information. Data found on NSO websites now may not be found there in the future. We must protect the world's statistical heritage by preserving these data and making them usable in modern formats, discoverable by researchers, and interoperable with other data.

4. Methods

We describe progress on three major work components: (1) developing table markup, metadata capture, and data intake tools; (2) creation of tools and workflows for metadata enhancement and standardization; and (3) and development of data dissemination systems. The sections that follow describe each of these components in turn. Note that much of this is a work in progress.

4.1 Table markup, metadata capture, and data intake

A core geocomputational challenge for big population-environment data include development of architectures and services that facilitate cataloging, processing, and integration of heterogeneous spatial data with complicated attribute and temporal dimensions. In particular, a key challenge of aggregate data is heterogeneity in how the published tables are structured. In general, the contents of table cells are counts of individuals, households, or other entities in a geographic unit that have some characteristic or combination of characteristics. The organization of geographic units and characteristics varies widely from country to country and, within a country, table to table. Table columns typically represent variables or categories of characteristics, such as sex, marital status, or citizenship. Table rows may represent either geographic units or additional variables or categories, and can be arranged in a variety of ways. A single table may include geographic units at many levels of a parent-child hierarchy, and when geographic units and population characteristics are both represented, rows may be grouped either by category or by geographic unit (Figure 2).


Interleaved geographic levels
By category
By geographic unit

	Municipal Environment				
	Total number of population	Number of population, Male	Number of population, Female	Number of households	Number of dwellings
Tunisia	7,445,139.0	3,717,204.0	3,727,935.0	1,903,700.0	2,342,200.0
North East	3,413,921.0	1,714,887.0	1,699,034.0	900,552.0	1,094,017.0
Governorate of Tunis	1,056,247.0	528,149.0	528,098.0	287,412.0	343,348.0
Delegation of Carthage	24,216.0	11,890.0	12,326.0	6,727.0	8,245.0
Sidi Bou Saïd	3,540.0	1,724.0	1,816.0	1,099.0	1,261.0
Amilcar	3,666.0	1,713.0	1,953.0	1,180.0	1,324.0
Carthage Byrsa	4,680.0	2,277.0	2,403.0	1,168.0	1,549.0
Carthage Plage	3,272.0	1,630.0	1,642.0	1,012.0	1,271.0
Le Jasmin	5,005.0	2,536.0	2,469.0	1,228.0	1,596.0
Cité Mohamed Ali	4,053.0	2,010.0	2,043.0	1,040.0	1,244.0
Delegation of La Medina	21,400.0	11,128.0	10,272.0	6,221.0	7,431.0
La Medina	2,018.0	1,011.0	1,007.0	531.0	622.0

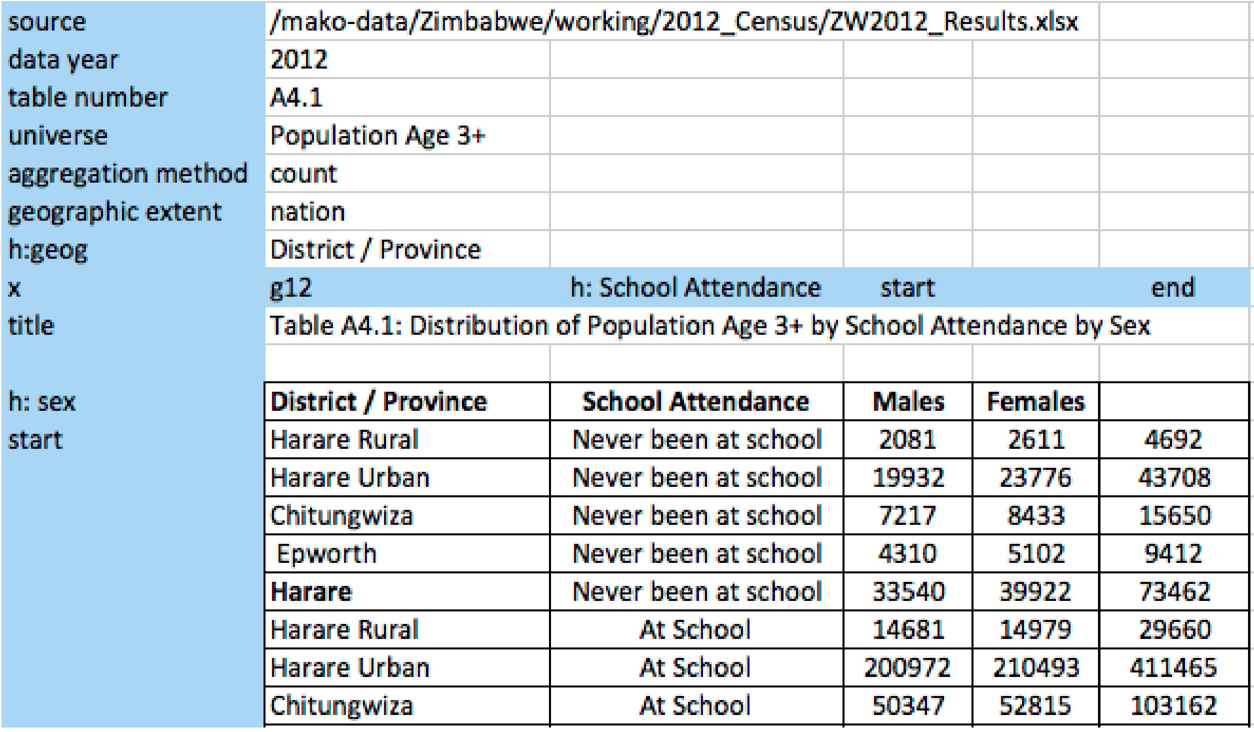
District/Sex	Urban/Rural	Sex	Total	Married	Widowed	Divorced
Total	**Urban**	**Males**	231128	119217	3732	3440
Nicosia	**Urban**	**Males**	97157	49817	1555	1461
Lanarca	**Urban**	**Males**	34378	17540	665	545
Limassol	**Urban**	**Males**	76314	39923	1193	1131
Paphos	**Urban**	**Males**	23279	11937	319	303
Total	**Rural**	**Males**	107369	55793	2391	968
Nicosia	**Rural**	**Males**	36544	18742	734	249
Famagusta	**Rural**	**Males**	19074	9462	298	200
Lanarca	**Rural**	**Males**	22467	11392	445	215
Limassol	**Rural**	**Males**	19422	10637	542	189
Paphos	**Rural**	**Males**	9862	5560	372	115
Total	**Urban**	**Females**	243322	122201	17460	7913
Nicosia	**Urban**	**Females**	103529	51295	7850	3468
Lanarca	**Urban**	**Females**	36124	17924	2674	1175
Limassol	**Urban**	**Females**	80625	40946	5632	2651
Paphos	**Urban**	**Females**	23044	12036	1304	619

District	Urban/Rural	Size of Household Citizenship									
		EU Citizens					Non EU Citizens				
		EU Citizens	1 persons	2 persons	3 persons	4 persons	Non EU Citi	1 persons	2 persons	3 persons	4 persons
Nicosia	**Total**	9.88	669	1.593	2.024	3.287	12.593	1.248	3.007	2.522	2.55
	Urban	9.197	643	1.497	1.934	3.036	10.866	995	2.451	2.212	2.301
	Rural	683	26	96	90	251	1.727	253	556	310	249
Famagusta	**Total**	1.351	103	409	229	322	966	138	322	147	134
	Urban	0	0	0	0	0	0	0	0	0	0
	Rural	1.351	103	409	229	322	966	138	322	147	134
Lanarca	**Total**	4.741	392	1.196	784	1.219	3.846	376	1.039	892	669
	Urban	3.727	302	805	662	1.036	2.912	258	735	671	545
	Rural	1.014	90	391	122	183	934	118	304	221	124
Limassol	**Total**	8.214	716	2.598	1.284	2.012	11.092	854	2.613	2.829	2.34
	Urban	6.402	581	1.519	1.087	1.811	10.413	811	2.439	2.665	2.227
	Rural	1.812	135	1.079	197	201	679	43	174	164	113
Paphos	**Total**	7.91	535	2.967	1.157	1.948	3.85	304	914	851	884
	Urban	6.801	463	2.223	1.046	1.839	3.289	261	767	716	771
	Rural	1.109	72	744	111	109	561	43	147	135	113

Figure 2: Heterogeneous table structures that can vary between countries and between tables within countries

While tables may be structured in many ways, all tables include a common set of elements, and each numerical value in a table can be described with a common set of metadata. We have developed a markup framework that enables trained researchers to quickly identify the metadata for each table and mark the location of key elements for further automated processing. The markup process is designed to minimize manual manipulation because such manipulation is time consuming, error prone, and difficult to document. After conversion from PDF to Excel, worksheets are preprocessed with a Python script that unmerges any merged cells and inserts rows and columns to provide space for markup.

Researchers then manually markup the table. First, key pieces of information are extracted into a standard sheet header section, including data year, table universe, and the geographic extent of the table (Figure 3). Next, researchers add tags indicating the location of standard elements . The tags also provide a brief description of the element (e.g., ‘g12’ in Figure 3 marks the names of first- and second-level geographic units). If elements such as data year or aggregation method vary by column, the tag is moved from the sheet header to mark the row containing the corresponding per-column information. For most tables, markup takes only a few minutes per table, representing an efficient hybrid of human interpretive skills and automated data capture and manipulation long central to geocomputation.



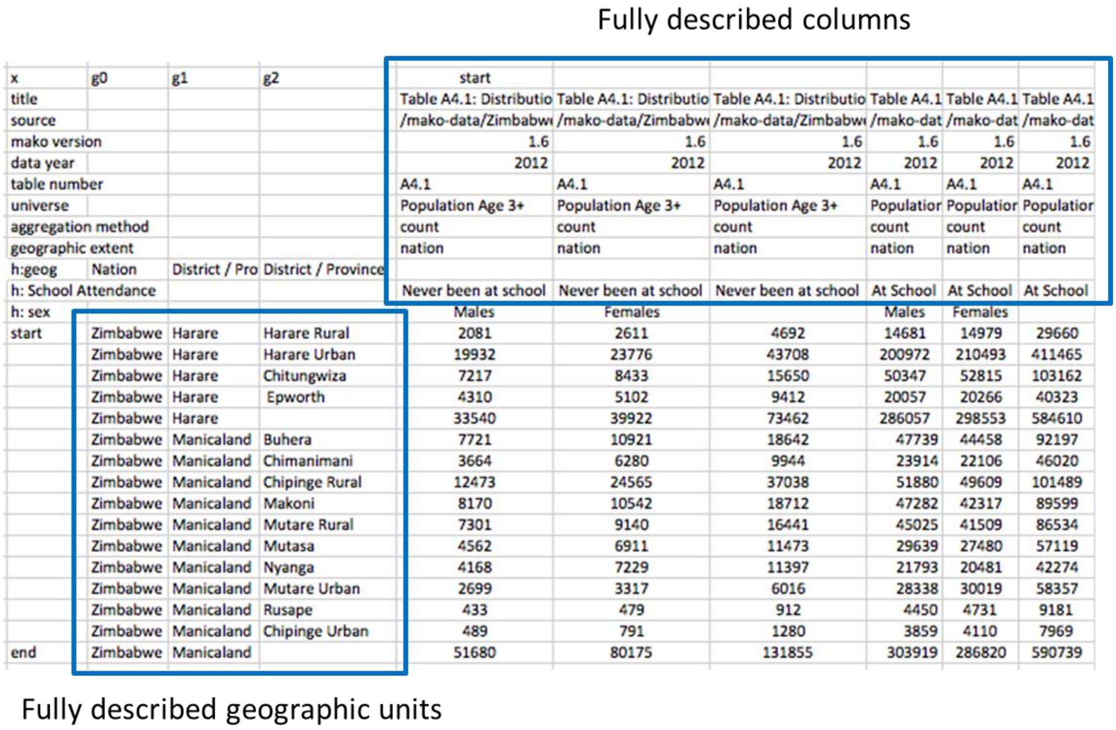
source	/mako-data/Zimbabwe/working/2012_Census/ZW2012_Results.xlsx				
data year	2012				
table number	A4.1				
universe	Population Age 3+				
aggregation method	count				
geographic extent	nation				
h:geog	District / Province				
x	g12	h: School Attendance	start		end
title	Table A4.1: Distribution of Population Age 3+ by School Attendance by Sex				
h: sex	**District / Province**	**School Attendance**	**Males**	**Females**	
start	Harare Rural	Never been at school	2081	2611	4692
	Harare Urban	Never been at school	19932	23776	43708
	Chitungwiza	Never been at school	7217	8433	15650
	Epworth	Never been at school	4310	5102	9412
	Harare	Never been at school	33540	39922	73462
	Harare Rural	At School	14681	14979	29660
	Harare Urban	At School	200972	210493	411465
	Chitungwiza	At School	50347	52815	103162

Figure 3: Marked up table with sheet header and tags indicating location of key elements

Once the sheet has been marked up, another automated tool processes it into a standardized CSV file. In the standardized CSV, each row and column is fully described with a complete set of metadata. Each column is described by all elements in the sheet header and the category or variable descriptions from the column headers in the original table. On the rows, categorical headers are separated from geographic unit names and pulled into a new column. In addition, the hierarchical structure of the geographic units is fully specified, with a column for each level of geography.

These operations are supported by a geographic dictionary that lists all of the unit names in their hierarchical relationships, and in turn may be tied to the IPUMS Terra’s comprehensive collection of spatiotemporal multilevel administrative geographic units. We have drawn in data sources including the United Nations Second Administrative Level Boundaries (SALB), Global Administrative Unit

Layers (GAUL), and Global Administrative Areas (GADM) database (Hijmans *et al.*, 2011). We inventoried, analyzed, and corrected these sources to provide multilevel administrative historic geographic boundaries. Automated tools are then able to intake these standardized tables into a database that captures the characteristics of each row and column. Within the database, tables are linked to previously captured information about the census itself, such as the name of the agency that conducted it and the official census day of record. Linkages can also be made between the geographic levels described in each table and shapefiles representing the boundaries of the units (Figure 4).



Fully described columns

x	g0	g1	g2	start					
title				Table A4.1: Distributio	Table A4.1: Distributio	Table A4.1: Distributio	Table A4.1	Table A4.1	Table A4.1
source				/mako-data/Zimbabw	/mako-data/Zimbabw	/mako-data/Zimbabw	/mako-dat	/mako-dat	/mako-dat
mako version				1.6	1.6	1.6	1.6	1.6	1.6
data year				2012	2012	2012	2012	2012	2012
table number				A4.1	A4.1	A4.1	A4.1	A4.1	A4.1
universe				Population Age 3+	Population Age 3+	Population Age 3+	Populatior	Populatior	Populatior
aggregation method				count	count	count	count	count	count
geographic extent				nation	nation	nation	nation	nation	nation
h:geog	Nation	District / Pro	District / Province						
h: School Attendance				Never been at school	Never been at school	Never been at school	At School	At School	At School
h: sex				Males	Females		Males	Females	
start	Zimbabwe	Harare	Harare Rural	2081	2611	4692	14681	14979	29660
	Zimbabwe	Harare	Harare Urban	19932	23776	43708	200972	210493	411465
	Zimbabwe	Harare	Chitungwiza	7217	8433	15650	50347	52815	103162
	Zimbabwe	Harare	Epworth	4310	5102	9412	20057	20266	40323
	Zimbabwe	Harare		33540	39922	73462	286057	298553	584610
	Zimbabwe	Manicaland	Buhera	7721	10921	18642	47739	44458	92197
	Zimbabwe	Manicaland	Chimanimani	3664	6280	9944	23914	22106	46020
	Zimbabwe	Manicaland	Chipinge Rural	12473	24565	37038	51880	49609	101489
	Zimbabwe	Manicaland	Makoni	8170	10542	18712	47282	42317	89599
	Zimbabwe	Manicaland	Mutare Rural	7301	9140	16441	45025	41509	86534
	Zimbabwe	Manicaland	Mutasa	4562	6911	11473	29639	27480	57119
	Zimbabwe	Manicaland	Nyanga	4168	7229	11397	21793	20481	42274
	Zimbabwe	Manicaland	Mutare Urban	2699	3317	6016	28338	30019	58357
	Zimbabwe	Manicaland	Rusape	433	479	912	4450	4731	9181
	Zimbabwe	Manicaland	Chipinge Urban	489	791	1280	3859	4110	7969
end	Zimbabwe	Manicaland		51680	80175	131855	303919	286820	590739

Fully described geographic units

Figure 4. Standardized structure fully describing columns and geographic areas

4.2 Metadata enhancement and standardization

Having metadata in a database enables queries that can be used to verify and further document data and conduct additional metadata standardization. For example, queries could identify the rows that describe the children of each parent geographic unit. That information can then be used to verify that the sum of the counts for the child units matches the reported value for the parent unit. Similar cross-checks can be performed across categories. For example, in a table containing counts of married males, married females, and total married persons, the sum of married males and married females should match the count of total married persons.

Following Data Documentation Initiative (DDI) constructs, population characteristics are described in terms of dimensions and categories (Blank and Rasmussen, 2004). Dimensions generally relate to a particular question on a census enumeration form and are concepts such as marital status, type of water supply to a household, or land use. The "h:" tags added during markup typically describe dimensions. We use established international standards to construct a consistent set of dimensions to apply across all census datasets, which will enable users to easily find data from many countries that relate to their interests. Relevant standards include the Principles and Recommendations for Population and Housing Censuses, the World Programme for the Census of Agriculture guidelines, and Eurostat's concepts and definitions database.

In order to speed the markup process, dimensions are not standardized during markup. Researchers simply mark each worksheet with "h:" tags that describe their interpretation in the context of that sheet. After the metadata are in the database, queries and reports listing the dimension tags and associated category labels across multiple tables assist with standardization. For example, a query could identify all the categories associated with h: tags containing "school" or "education" and those labels could then be standardized.

4.3 Data dissemination

The international aggregate census data processed under this project will be disseminated through a new IPUMS data product. This new product will complement our existing U.S. microdata product (IPUMS USA), U.S. aggregate data product (IPUMS NHGIS), and international microdata product (IPUMS International). We expect that the new international aggregate data will attract a new segment of users who are especially interested in international agricultural census and small area data. These users will join a community of IPUMS users that already numbers in the tens of thousands.

The data will be delivered through both a web application and via API. The web application will allow users to browse and select tables through a graphical interface and download their selected tables as CSV files with accompanying codebooks. The API will allow users to request data for seamless integration into data analysis packages, scripts, or web maps. Finally, we continue development of GIS-compatible boundary files of the first- and second-level subnational administrative units for many countries. We will utilize efficient processes and tools that we have already developed to construct any additional boundary files required (Kugler *et al.*, 2015). These join these IPUMS-Terra data finder that allows creation of a customized data subsets, of data, TerraClip to provide map extracts, and TerraScope for data visualization.

5. Conclusion and future steps

IPUMS Terra is combines geocomputation and spatial data infrastructure to develop and deliver data, analysis, and visualization for human-environment systems. It is expanding to include aggregate census and agricultural census data. Doing so requires advances in data manipulation and workflow management approaches to transform and standardize data and metadata for tens of thousands of data tables. Given the sheer volume of material, processing and documentation will extend over a period of several years. We are prioritizing tables that fit within the constraints of our initial markup and metadata capture tools and then building out both these tools and more complex tables. We are also prioritizing the needs of users, who have identified recent population censuses from countries not participating in IPUMS International and recent agricultural censuses from all available countries. In cases where individual countries have exceptionally large numbers of tables, we focus on topics most frequently represented in the data collection. We will continue to gather feedback from users through surveys, direct engagement, and usage statistics to identify countries and topics that are in greatest demand.

6. Acknowledgments

This work is supported in part by the National Institutes of Health supported Minnesota Population Center (R24 HD041023), National Spatiotemporal Population Research Infrastructure (2R01HD057929-11), NSF OCI: Terra Populus: A Global Population/Environment Data Network (0940818), National Science Foundation Sociology and Data Infrastructure Programs (SMA-1738369), and the University of Minnesota Office of the Vice President for Research. The authors gratefully acknowledge the assistance of the editor and anonymous reviewers. Responsibility for the opinions expressed herein is solely that of the authors.

7. References

Abrahart, R. J. and See, L. M. (2014) *GeoComputation*. CRC Press.

Batty, M. (2017) ‘Geocomputation’, *Environment and Planning B*. SAGE Publications Sage UK: London, England, 44(4), pp. 595–597.

Blank, G. and Rasmussen, K. B. (2004) ‘The data documentation initiative: the value and significance of a worldwide standard’, *Social Science Computer Review*, 22(3), pp. 307–318.

Bloom, D. E. (2011) ‘7 billion and counting’, *Science*. American Association for the Advancement of Science, 333(6042), pp. 562–569.

Brunsdon, C. and Singleton, A. (2015) *Geocomputation: A Practical Primer*. Thousand Oaks, California: Sage.

Doxsey-Whitfield, E. *et al.* (2015) ‘Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4’, *Papers in Applied Geography*. Taylor & Francis, 1(3), pp. 226–234.

Ehrlich, P. R., Kareiva, P. M. and Daily, G. C. (2012) ‘Securing natural capital and expanding equity to rescale civilization’, *Nature*, 486(7401), pp. 68–73.

Haynes, D. *et al.* (2015) ‘High performance analysis of big spatial data’, in *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*. doi: 10.1109/BigData.2015.7363974.

Haynes, D., Manson, S. M. and Shook, E. (2017) ‘Terra Populus’ architecture for integrated big geospatial services’, *Transactions in GIS*, 21(3). doi: 10.1111/tgis.12286.

Hendriks, P. H. J., Dessers, E. and Van Hootegem, G. (2012) ‘Reconsidering the definition of a spatial data infrastructure’, *International journal of geographical information science*. Taylor & Francis, 26(8), pp. 1479–1494.

Hijmans, R. *et al.* (2011) *Global Administrative Areas (GADM v2)*. Available at: http://biogeo.ucdavis.edu/data/gadm2/gadm_v2_shp.zip.

Holm, P. *et al.* (2013) ‘Collaboration between the natural, social and human sciences in global change research’, *Environmental science & policy*. Elsevier, 28, pp. 25–35.

JMGL (2016) ‘Best Paper of the Year Award for Volume 11’, *Journal of Map & Geography Libraries*.

Routledge, 12(2), p. 228. doi: 10.1080/15420353.2016.1187017.

Kugler, T. A. *et al.* (2015) ‘Terra Populus: Workflows for integrating and harmonizing geospatial population and environmental data’, *Journal of Map & Geography Libraries*. Taylor & Francis, 11(2), pp. 180–206. doi: 10.1080/15420353.2015.1036484.

Longley, P. A. *et al.* (1998) *Geocomputation: A Primer*. Chichester, UK: John Wiley and Sons.

Millett, L. I. and Estrin, D. L. (2012) *Computing Research for Sustainability*. Washington D. C.: National Academies Press.

Rosa, E. A. *et al.* (2010) *Human footprints on the global environment: Threats to sustainability*. Cambridge, MA: MIT Press.

Ruggles, S. (2014) ‘Big microdata for population research’, *Demography*. Springer, 51(1), pp. 287–297.

Sobek, M. *et al.* (2011) ‘Big data: large-scale historical infrastructure from the Minnesota Population Center’, *Historical methods*. Taylor & Francis, 44(2), pp. 61–68.