2015 IEEE 15th International Conference on Data Mining Workshops

Terra Populus

Integrated Data on Population and Environment

Steven Ruggles, Tracy A. Kugler, Catherine A. Fitch and David C. Van Riper
Minnesota Population Center
University of Minnesota
Minneapolis, Minnesota, 55414
Email: ruggles@umn.edu, takugler@umn.edu, fitch@umn.edu, vanriper@umn.edu

***Abstract*—*Terra Populus*, part of National Science Foundation's DataNet initiative, is developing organizational and technical infrastructure to integrate, preserve, and disseminate data describing changes in the human population and environment over time. A large number of high-quality environmental and population datasets are available, but they are widely dispersed, have incompatible or inadequate metadata, and have incompatible geographic identifiers. The new Terra Populus infrastructure enables researchers to identify and merge data from heterogeneous sources to study the relationships between human behavior and the natural world.**

Keywords*—*data integration; population; environment

I. INTRODUCTION

Terra Populus (TerraPop) enables research, learning, and policy analysis by providing integrated spatiotemporal data describing people and the environment in which they live. The project is developing technical and organizational infrastructure to integrate, preserve, and disseminate data describing population and environment on a global scale over the past two centuries, including data on human population characteristics, land use, land cover, and climate change.

The past five decades have seen remarkable changes in the characteristics and spatial distribution of human population. Population has more than doubled, and increasing urbanization and international migration have altered spatial distributions. Similarly, world per-capita gross domestic product roughly doubled, but with large inequalities among countries and populations [1-4]. At the same time, fertility rates have begun to decline, bringing about a shift in the age composition of the world's population [5, 6].

The extraordinary levels of global demographic and economic growth since the 1950s have had ominous consequences: alarming environmental degradation, resource depletion, and climate change [2, 3]. In just the last 50 years, food and water consumption roughly tripled, alongside a four-fold increase in the use of fossil fuels. Global land resources, biodiversity, and "ecosystem goods and services" are experiencing rapidly increasing pressures [7, 8]. The average global temperature has gone up 0.74˚ Celsius over the past century, and is now rising at an accelerating pace; predictions for temperature increase over the next century range from 1.1˚ to 6.4˚. Sea levels are rising, and the oceans are growing more acidic. New precipitation patterns—including increased precipitation in high latitudes and decreased rainfall in subtropical regions—are becoming more pronounced [9]. Deforestation and pollution are compounding the direct effects of global warming and contributing to the destruction of ecosystems and decline of biodiversity [8, 9].

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—are direct consequences of population growth and economic development. Conversely, environmental change has profound implications for demographic behavior. Flooding, erosion of coastal areas, destruction of ecosystems, and drought and degradation of water supplies at lower latitudes all have consequences for human populations, such as food scarcity, increased armed conflict, and mass migration.

Our understanding of the interactions of population and environment has been hampered by the dearth of internationally comparable data. While high-quality data are available describing both the human population and the environment, data from the two realms are not commonly used together. TerraPop provides population data closely integrated with data on the environment that will allow researchers to describe the unfolding transformation of human and ecological systems.

Despite the need to study humans and the environment as coupled systems, there are relatively few large-scale efforts to make available the integrated data necessary to support such investigations. Existing efforts tend to focus on two extremes of the integration spectrum. On one end are projects working to build federated data catalogs encompassing multiple domains, such as the DataNet project DataONE [10]. While federated catalogs can encompass large quantities of data and make it easier to discover data from different domains, they do little to enable analyses drawing on data from different sources in different structures. On the other end of the spectrum are data fusion projects, which conduct in-depth processing and analysis to combine data from across domains, creating new datasets. The suite of agricultural lands datasets produced by the Global Landscapes Initiative [11-14] is an example of data fusion. These datasets draw on information from both agricultural censuses and satellite imagery to delineate and characterize land used for agricultural purposes. Data fusion projects tightly integrate data from multiple sources, but the

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.204

techniques incorporated are tailored to particular datasets and are not easily transferrable to other applications.

In contrast to federated catalogs and data fusion projects, TerraPop is situated at the middle of the integration spectrum. TerraPop combines the data discovery advantages of federated catalogs with the tight integration of data fusion, providing broadly applicable tools for integrating data from multiple sources. TerraPop has created and continues to grow a curated catalog of high-quality datasets describing both population and environment. TerraPop also provides mechanisms to blend data from any of the sources in its collection, facilitating analyses incorporating both population and environmental aspects. By creating a unified framework for locating, analyzing, and visualizing the world's population and environment in time and space, TerraPop provides unprecedented opportunities for investigating the agents of change, assessing their implications for human society and the environment, and developing policies to meet future challenges.

II. Source Data

The TerraPop framework is designed to accommodate any kind of geographically- and spatially-referenced data sources. Our initial work focuses on three data formats: microdata, area-level data, and raster data. Microdata are structured as records of individuals and households. Area-level, also called vector data, are structured as records of places defined by boundaries. Raster data are structured as values arranged in a spatial grid. Researchers are often familiar with one or two of these data types but rarely are able to manage all three. TerraPop will allow researchers to easily obtain integrated population and environmental data in formats conducive to their work.

We are building a curated collection of high-quality population and environmental data across these data formats. Our initial work has focused on five types of data that have a significant temporal dimension; much of the data span the past five decades, and some reach back before the nineteenth century:

1) Census and survey microdata describing the social and economic characteristics of individuals and their families and households (microdata)
2) Aggregate census and survey data, describing the population characteristics of places (area-level)
3) Ecosystem characteristics, economic indicators and health information describing places (area-level)
4) Remote-sensing data describing land cover and other environmental characteristics (raster)
5) Climate data describing temperature, precipitation, and other climate-related variables interpolated from weather stations (raster)

A. Microdata

Microdata provides individual-level information describing characteristics such as age, sex, and occupation for large samples or complete census enumerations. The core of TerraPop's population microdata is the Integrated Public Use Microdata Series (IPUMS), the world's largest collection of spatiotemporal population data. IPUMS was the first data integration system driven by structured metadata. The system uses a data warehousing approach to transform data from heterogeneous sources into a single view schema. The IPUMS project receives data from national statistical offices, which are cleaned, processed, and integrated across time and place before being incorporated into TerraPop [15-17].

By making thousands of population databases interoperable, IPUMS demonstrated the feasibility of large-scale data integration. The system presently provides individual-level data on 859 million people from 765 censuses and surveys of 82 countries spanning the period from 1703 to 2011 [18-20]. By 2018 IPUMS will disseminate about 2 billion records [21].

The individual records are nested within families and households, with full information about the interrelationships of the members of each residential group. For every person, we have information about geographic location and economic activities. In most cases, the data also cover educational attainment, literacy, fertility history, child mortality, migration and place of former residence, marital status and consensual unions, disabilities, water supply, sewage, characteristics of the building (floor, roof, etc.), and a host of other characteristics.

Due to the detailed and potentially sensitive nature of microdata, protections are in place to ensure confidentiality. Basic protections include deidentifying records, bottom- and top-coding continuous variables, combining small groups in categorical variables, and limited random swapping of records across geographic units. Another key protection is restricting the release of fine-scale geographic information. Agreements between IPUMS and national statistical offices typically stipulate that publicly available geographic unit codes will encompass populations of at least 20,000 people.

B. Area-level data

Area-level data provide summary characteristics for a polygon corresponding to an administrative or statistical area such as a county or a census tract or for another geographic unit such as a watershed. Area-level data are available for both population and environmental characteristics.

Because they are aggregated, area-level data lack the detail and flexibility of microdata, but they offer other advantages. Most important, area-level data can often provide greater geographic detail than microdata, as they pose fewer confidentiality concerns. This advantage is especially salient in developed countries, which often limit microdata geographic detail to places with very large populations (such as greater than 100,000 people). Moreover, area-level sources often provide variables not available from the microdata. Some sources provide data only at the area-level, such as agricultural censuses, which include data related to land use. Finally, area-level data are the only source of population data available for countries that have not joined IPUMS.

The current version of TerraPop includes area-level population data from a variety of sources. TerraPop incorporates selected data from the National Historical

Geographic Information System (NHGIS), the most extensive collection of small-area population data for the United States [22]. Like the IPUMS projects, NHGIS cleans, processes, and integrates data over time before they are incorporated into TerraPop. We have also tabulated IPUMS microdata to generate a defined set of area-level variables for countries participating in IPUMS. For countries that have not joined IPUMS, we have harvested published data from the websites of national statistical offices. Fig. 1 shows the availability of population data in both microdata and area-level formats by finest available geographic level.

The aggregate data from international statistical offices present a significant harmonization challenge. The set of published tables varies widely country to country, from nothing more than a simple population by sex table to complex collections of hundreds of tables covering dozens of topics. Furthermore, tables are published in heterogeneous structures that cannot be easily ingested into a common database.

To address the challenge of describing the range of tables and variables and making them discoverable through the TerraPop interface, we have adopted an NHGIS-based model of classifications and categories. Classifications describe the characteristics represented by a variable, such as sex and marital status. Categories describe the subsets within a classification included in a variable, for example, female and married. Any single variable may have multiple classifications and/or categories assigned. Classifications can then be used as filters or search terms for identifying variables of interest. Classifications and categories help to unify the wide variety of tables at a conceptual level.

To handle the technical challenge of heterogeneous table structures, we are developing a suite of Python tools utilizing the Pandas library. The tools convert Excel worksheets containing the original table structures into CSV files in a standard structure with one row per geographic unit, one variable per column, and geographic units at the same level (e.g., states or counties) contained within the same file. The CSV files are then easily ingested into the TerraPop database.

An initial survey of the collection of tables revealed that nearly all of the tables could be categorized into a handful of structural families. Each family is handled by a Python tool tailored to tables of that general structure. The tools are parameterized by a series of rules that define the specifics of a particular table or set of tables, and each structure family has an associated set of rules.

After converting the tables to a standard format, we perform checks for internal consistency. These checks include ensuring that the sum of counts over a set of geographic units matches the count for the parent unit and similarly that the sum of counts across a set of categories matches the total universe. Because the data have been vetted and published by national statistical offices, the initial quality is generally high and these checks rarely fail. In cases of inconsistencies, the source data are checked. If the problem is present in the source data, we do not often have the information necessary to correct the error, unless its source is clearly apparent (e.g., the presence of an extraneous zero).

C. *Raster data*

Raster datasets divide a geographical area into grid cells and provide a value (e.g. average annual temperature) for each cell. Satellite imagery and data derived from it, such as land cover, are typically in raster format. Data from point measurements, such as climate stations, may also be interpolated to produce raster data that covers the entirety of a geographical area. Data from environmental models, especially for climate, are typically in raster format. TerraPop currently delivers three land use/land cover datasets and one climate

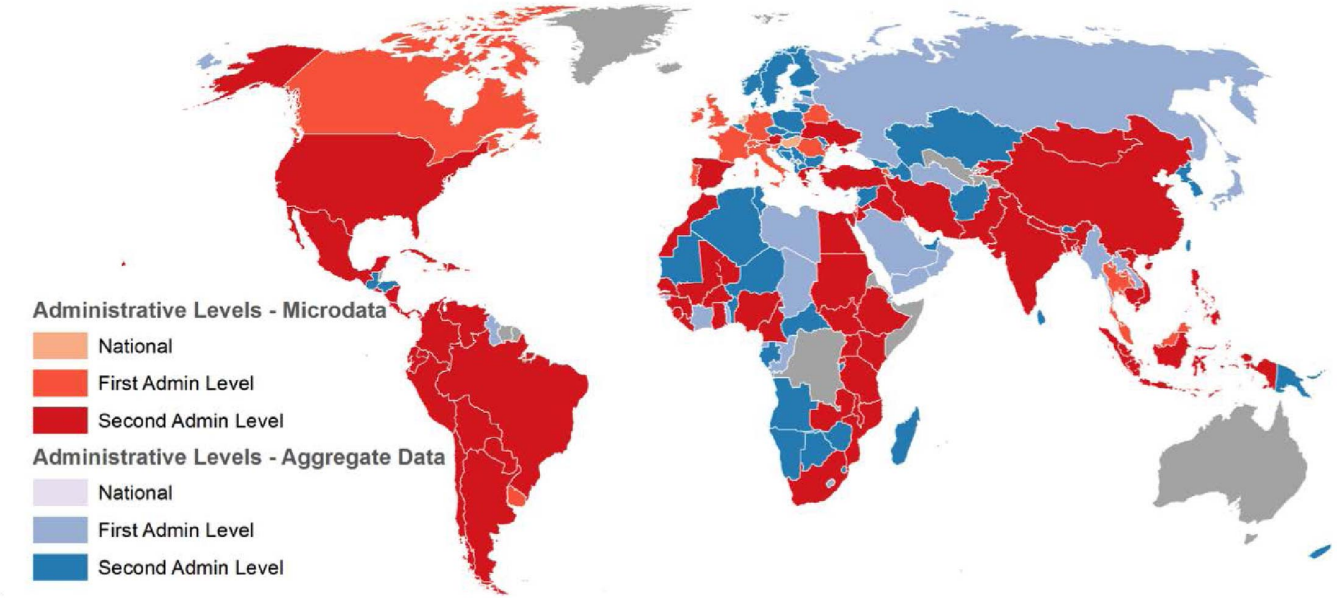

Fig. 1. TerraPop Population Data Availability

dataset. In the next release we will add an additional climate dataset.

TerraPop currently incorporates land use/land cover data from the Global Landscapes Initiative's (GLI) agricultural lands datasets, the Global Land Cover 2000 dataset, and the MODIS Land Cover Type product (MCD12Q1). The GLI agricultural lands datasets are the world's largest collection of spatiotemporal data on agricultural acreage and yields. They include information on harvested area and yields for 175 crops in the year 2000 and cropland and pasture data back to 1700. These data are derived from census-based land-use and land-cover records in combination with remotely sensed data, enriching both forms of data [11-14]. The Global Land Cover 2000 [23] dataset is based on SPOT-4 vegetation imagery and uses a globally standardized 22-class classification system based on the Food and Agriculture Organization (FAO) Cover Classification System [24]. It provides global coverage with approximately 1-kilometer resolution. The MODIS Land Cover Type product includes annual time steps of land cover classified from MODIS satellite data over the 2001-2012 time period [25]. The data are at 500 meter resolution, and the collection includes five different classification systems ranging from nine to 17 classes. (TerraPop currently includes only the 17-class IGBP classification. Other classification systems will be included in future releases.)

TerraPop currently includes climate data from WorldClim and will be adding Climate Reference Unit-Time Series (CRU-TS) data in the next release. Both datasets are derived from climate station data interpolated to generate global surfaces. The WorldClim dataset includes long-term temperature and precipitation averages by calendar month over the 1950-2000 timeframe [26]. It also includes a series of bioclimatic variables as long-term averages with an annual basis. WorldClim data are at 1-kilometer resolution. The CRU-TS data include temperature and precipitation data for each month from 1901 to 2013 at half-degree resolution [27]. TerraPop has also calculated annual bioclimatic variables parallel to those present in the WorldClim dataset from the CRU-TS data.

We perform relatively little cleaning and harmonization of raster datasets prior to ingest into the TerraPop database. TerraPop's curated collection consists of datasets that the scientific community has used extensively and determined to be of high quality. While the various datasets are provided in different projections and resolutions, reprojecting and changing the resolution of raster data is prone to introducing errors [28-30]. TerraPop therefore opts to retain the data in their native projection and resolution. The most significant processing of raster data performed by TerraPop is to merge datasets provided as individual tiles into seamless global rasters.

III. Location-Based Data Integration

TerraPop allows investigators to design customized datasets that incorporate information from multiple sources and data formats. TerraPop alleviates much of the data processing burden associated with handling data from diverse sources, in multiple formats, and making connections across datasets, allowing scientists to focus on the data analysis relevant to their research questions. TerraPop combines data derived from the three different formats—microdata, area data, and raster data—and delivers them to investigators in any of the three formats.

The ability to integrate data across the three basic formats relies on spatial data delineating the boundaries of administrative and statistical units. Microdata records include codes identifying the unit in which the individual lives and area-level data include similar codes identifying the unit described by the record. These codes can be used to link microdata and area-level data and connect both types of data to real-world locations through boundary data. The boundaries can then be overlaid on raster data to link all three data formats. However, working with spatial data involves a number of challenges. For example, data derived from statistical agencies and administrative sources often use different spatial units, and those units often change over time. Similarly, different raster datasets use varying projections and resolutions.

TerraPop allows scientists to bypass these challenges and easily construct datasets that combine data derived from all three formats and that are delivered in the format of their choice:

(1) Microdata describing the characteristics of individuals, families, and households merged with variables derived from area-level data and raster data describing the characteristics of the area in which they live

(2) Area-level data describing population and environmental characteristics of spatial units, including summaries of variables from both microdata and raster data

(3) Raster data that include estimates of both environmental and population attributes derived from microdata and area-level data

Fig. 2 illustrates the concept of data integration across data formats. Unlike the data warehouse approach used by IPUMS, TerraPop integrates data on demand.

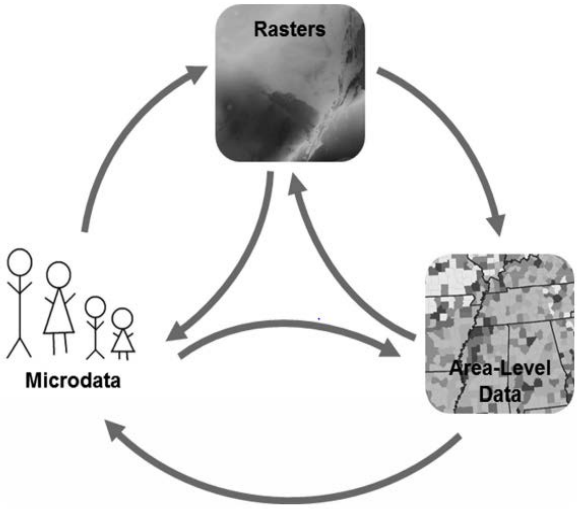

Fig. 2. TerraPop Data Integration

Area-level data are the key to transforming data across formats. Linkages and transformations are performed at the level of geographic units because TerraPop's population data do not include point locations of individual households. Both microdata and raster data can be transformed to area-level data by summarizing these more granular data types to describe geographic units. Area-level data, either from native area-level sources or summarized from raster data or microdata, may then be attached to microdata by matching the codes of the geographic units. Conversely, area-level data, either native or tabulated from microdata, may be distributed spatially across the grid cells in each geographic unit to create raster data.

To convert microdata to area-level data, we must tabulate the population with a characteristic or combination of characteristics within each geographic area. For example, we might count the number of girls age 5-9 who attend school in each administrative district. This tabulation is currently conducted offline for a defined set of area-level variables using software written in Java. The software utilizes definitions of variables in YML format that specify the microdata variables (e.g. age, sex, school attendance) and values (e.g. > 4 and < 10, female, yes) to be counted. Variable definitions may also include universes (e.g. persons age 15-65) and describe proportional variables (e.g. the proportion of the labor force that is unemployed). Since IPUMS-I microdata is stored as fixed-width flat files, variables are specified by their start location and width within the file.

We are in the process of improving the performance of the tabulator so that it can be used in an on-demand manner. The high-speed tabulator will allow users to specify their own area-level variable definitions, which can then be incorporated into customized datasets within the TerraPop interface. Implementation of threading on the existing tabulator software has improved performance approximately four-fold. The next stage of development will involve porting the microdata into the Parquet distributed column store format and adapting the tabulator software to work within a Spark environment.

Converting raster data to area-level data is a similar process: we must tabulate the grid cells within each geographic unit to derive a useful summary measure. For example, we might count the number of cells with forest cover in a geographic unit to derive the percentage of the unit that is forested. The computations are currently performed within TerraPop's PostgreSQL/PostGIS database. PostGIS has the advantage of being able to store and operate on both raster and vector (e.g., boundary polygons) data types. Spatial queries identify the grid cells that fall within each geographic unit and perform summary operations on the set of grid cell values. Because spatial queries are computationally intensive, Common Table Expressions (CTEs) are used to improve efficiency. The results of a spatial query are stored as a CTE for use in subsequent queries, reducing the need to perform multiple spatial queries.

We are also in the process of improving the performance of raster summarization. We have explored several potential parallelization approaches, with limited success. Most existing platforms for parallelizing PostgreSQL do not support geospatial sharding. The most efficient means of parallelizing raster summarization operations is to operate independently on individual geographic units (e.g., handle Minnesota independently of Arizona). Doing so requires that the data for each geographic unit are contained within a single core, while data for other units are contained on other cores, making geospatial sharding a critical component to this type of approach. Other spatially-aware parallel platforms exist, but they tend to be structured to support either raster or vector data, not both. Currently, the most promising platform is the Stado extension for PostgreSQL. Stado includes some geospatial sharding functionality for vector data. We have efforts underway to implement raster support within Stado and extend geospatial sharding to raster data. We are also exploring alternate methods of storing data and structuring queries to improve performance within non-parallel PostGIS.

Data in area-level format can be transformed to raster data by distributing values to grid cells. The simplest method of distributing values is to assume a uniform distribution within each geographic unit. For ratio-like values (e.g. population density), the value for the geographic unit is simply assigned to all the cells in geographic area. For count-like variables (e.g., total population), the area-level value is divided by the number of cells in the geographic unit, and the quotient is assigned to each cell. These operations are performed by queries within the project's PostGIS database. The queries may use any raster dataset in the TerraPop collection to serve as a template for redistribution, so that the transformed data are spatially aligned to the grid cells in the original raster data.

TerraPop is also developing more sophisticated methods incorporating data from a variety of sources to make informed estimates of how characteristics are spatially distributed across units. In these methods, known as dasymetric mapping, ancillary data are used to identify populated and non-populated zones within each geographic unit or a range of zones with varying levels of population density [31]. In the simpler binary populated/non-populated case, population is assigned only to the populated zones. Rather than dividing the population (or population subset) by the total number of cells in the unit, it is divided by the number of cells in the populated zone. Cells within the populated zone are then assigned the value of the quotient, while cells in the non-populated zone receive values of 0. Weighting schemes based on empirical or heuristic estimates may be used to differentially divide population among multiple zones. For example, urban areas may receive a greater proportion of the population than agricultural areas. Ancillary data used by TerraPop to determine population zones include land cover, roads, and nightlights.

Area-level variables, including those derived from raster data, can be attached to microdata records. For example, a researcher could augment microdata records of individuals with variables describing the total agricultural output and average annual precipitation for the county in which the individual lives.

IV. GEOGRAPHIC DATA

Data integration across formats in TerraPop hinges on geography. A key element of the integration processes is a global and accurate map of administrative units that can be

linked to the codes found in census data for current and historical time periods. Unfortunately, such a map was not previously available. Boundary data are particularly sparse for small geographic units and for historical time periods.

Accordingly, one of our most challenging tasks is developing a digital map of census units. We are building on three freely available administrative unit boundary data sets: the Food and Agriculture Organization's Global Administrative Unit Layers (GAUL), the United Nations Second Administrative Level Boundaries (UNSALB), and the Global Administrative Areas (GADM) database [32-34]. Each of these datasets has strengths and weaknesses. We also obtain shapefiles directly from national statistical and mapping agencies. When digital shapefiles are not available, as is often the case for historical time periods, we obtain paper maps. A collection of published census volumes held by the U.S. Census Bureau's International Division has proven to be invaluable, because the units represented in the maps are directly related to the sets of units described in the census data.

Boundaries digitized from print sources are aligned to the vertices in a reference shapefile (typically from a more recent year). Fig. 3 illustrates the process. The dotted boundaries were digitized from a PDF image depicting Brazilian municipios in 1980, and the solid lines derive from a shapefile of Brazil municipios in 2000. The shapefile boundaries are more accurate than the digitized boundaries. The hash marked areas illustrate automated processes for integrating the boundaries of 1980 and 2000. In the lower left, a unit that had not changed is simply copied from the 2000 shapefile. In the north, three year-2000 units are merged to create a larger 1980 unit. In the east, two 1980 units were rearranged to create four units in 2000. These units require manual processing, so the TerraPop software copies the year-2000 units and flags them for manual editing.

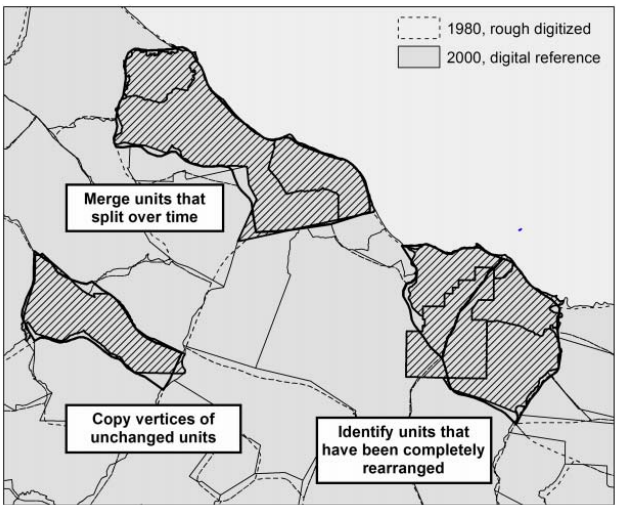

Figure 3. Brazil 1980 and 2000 Municipios: Harmonized Shapefiles

Once an administrative unit shapefile has been digitized from a map image or acquired from another source, we link the

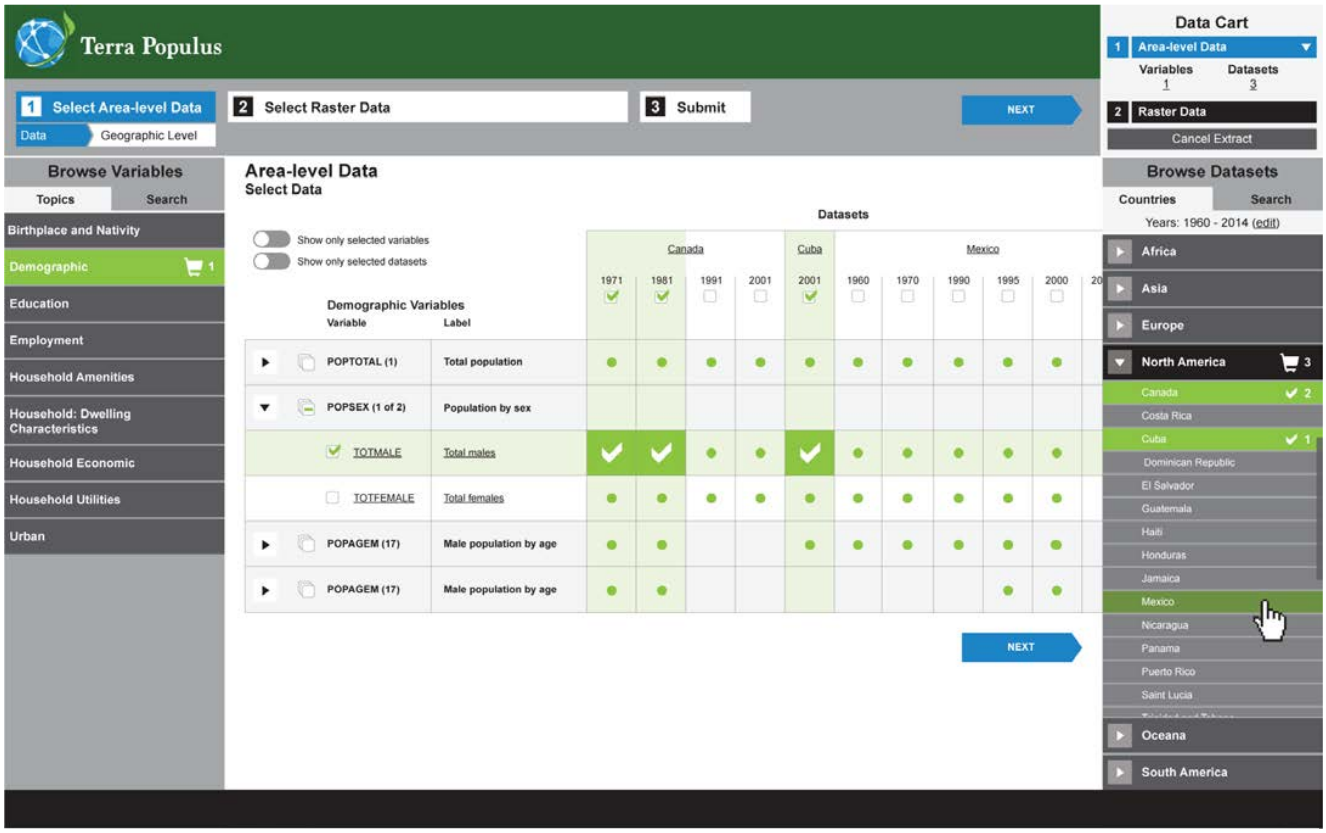

Figure 4. TerraPop Dissemination System: Area-level Data Selection

units represented in the map to units represented in census data, adjusting the map as needed based on ancillary sources so that the set of units in the census data is kept as close as possible to the original census data source. We conduct extensive research to ensure that all units in the census data are represented as accurately as possible in our final boundary files. In addition to shapefiles linked to codes in a specific census year, we also create harmonized shapefiles, with boundaries that are stable over time to enable analysis of change. Harmonized shapefiles are created with a minimum aggregation algorithm, which combines units that have experienced boundary changes with other units participating in the change until a stable set of boundaries has been reached. For boundary files linked to microdata, units with populations of less than 20,000 are regionalized with neighboring units to comply with confidentiality protections. For further details on our geographic boundary processing see [35].

V. Dissemination

The TerraPop web-based dissemination system allows users to select and integrate data in an information-rich environment where metadata browsing is fully integrated with data selection. Users can locate and select variables from hundreds of datasets, merge data derived from any of the three formats, and export the data in the format needed for their particular analyses.

A prototype data access system (https://beta.terrapop.org) has been available since May 2014. This system provides the core functionality of transformations between raster and area-level data and of attaching area-level data to microdata. We are continually improving the performance of the transformation operations. Based on lessons learned from the prototype interface, we are rolling out a redesigned user interface in stages during fall 2015 (https://data.terrapop.org). The prototype site will remain available in parallel until all functionality has been implemented in the new interface. During 2016 we will be developing on-the-fly tabulation, allowing users to create customized area-level variables from microdata. This functionality will complete the data transformation cycle.

We illustrate the data integration process in the new interface in Figs. 4-7. This example shows the workflow for producing a dataset that includes data from area-level and raster sources integrated to create an area-level file for analysis. The first step is to select the output data format. With the output format selected up-front, the workflow is tailored to

Fig. 4 shows the area-level data selection screen. The rows in the central table represent specific variables available for inclusion in the data extract. There may be thousands of such variables, and users can narrow the range of choices by selecting a particular topic (e.g. demographic characteristics), or they can do an open-ended search for variables. In many cases, variables are arranged in groups similar to published census tables; for example, the table of population by sex contains two cells describing the number of men and women in each geographic unit. Users may select entire groups, or they can "open" the group and select only the particular categories required for their analysis. The columns in Fig. 4 represent different countries and census years. Particular countries, regions, and time periods can be filtered using the tools on the right of the screen. Datasets are selected using the checkboxes at the top of each column.

Users may access extensive metadata while browsing. Dataset-level documentation—including source information, provenance, and sample density (if relevant)—can be obtained by clicking on the dataset name. Variable-level documentation can be obtained by clicking on the variable name; this includes codes (if applicable), universe, questionnaire text, and analysis of comparability across time and space.

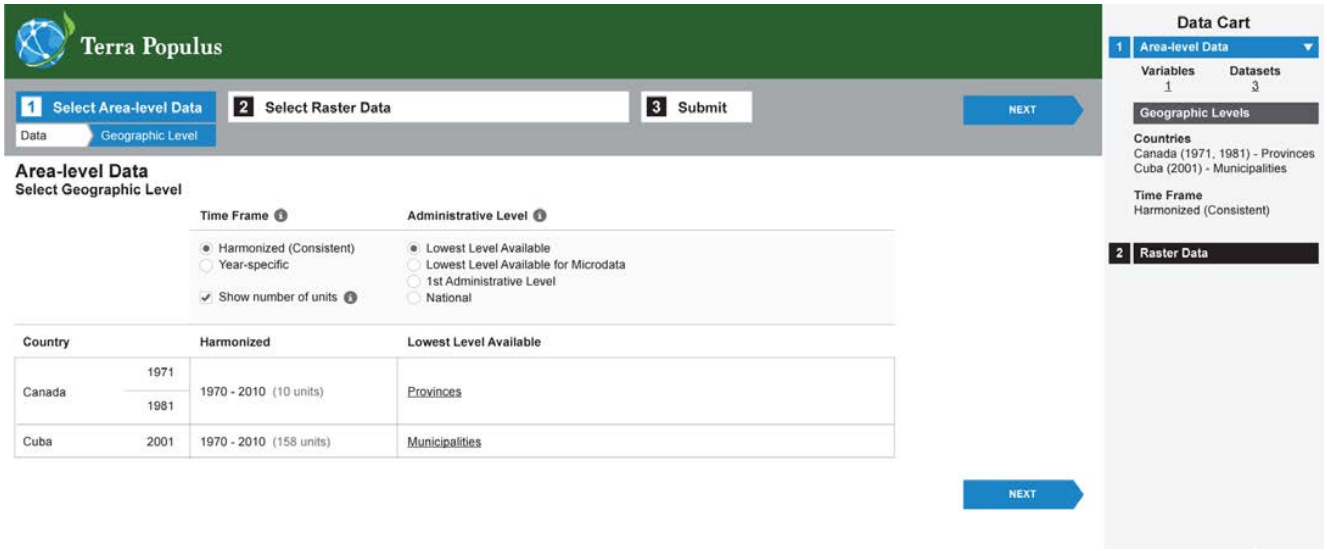

Figure 5. TerraPop Dissemination System: Geographic Selection

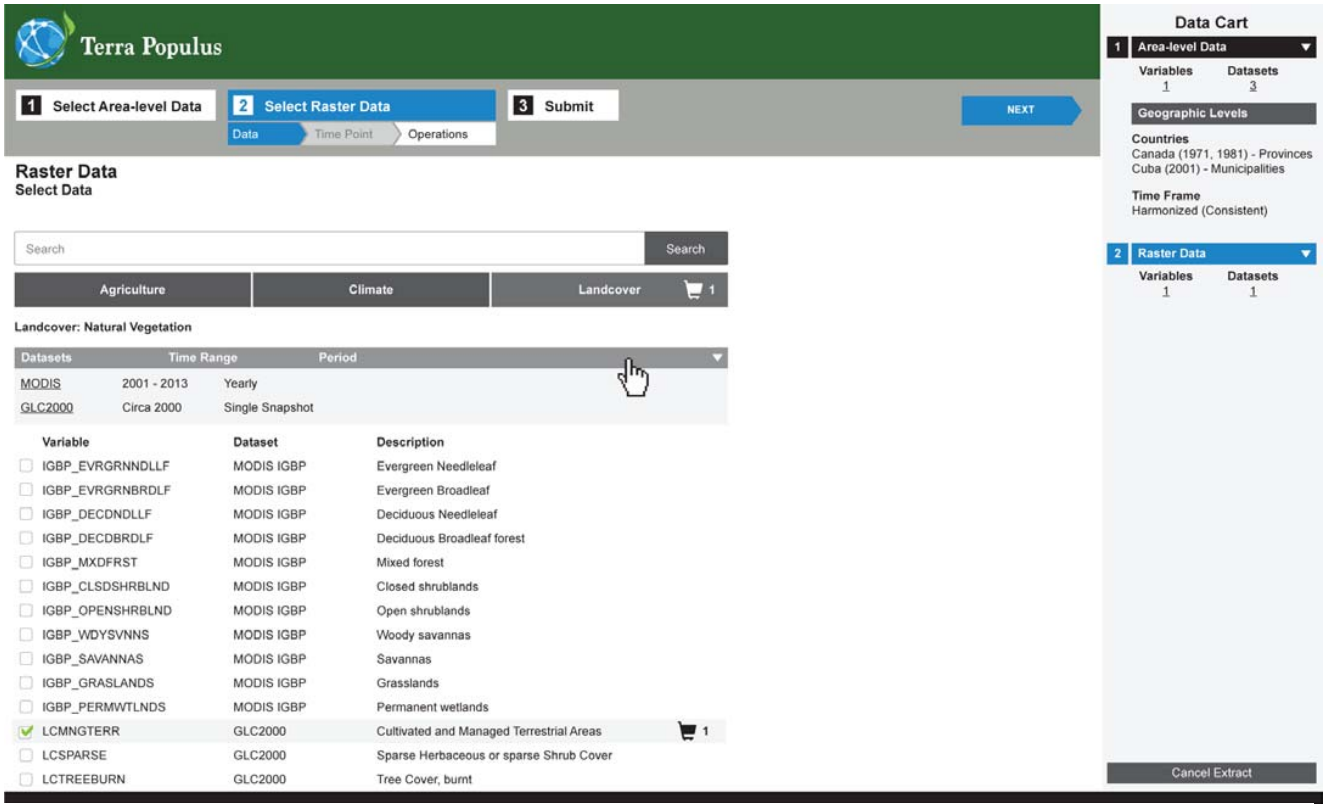

Figure 6. TerraPop Dissemination System: Raster Data Selection

The interior of the grid in Fig. 4 displays variable availability by dataset. Selection is performed at the level of variables and datasets. For each selected dataset, the output will include all selected variables that are available in that dataset.

After selecting area-level variables and datasets, the user selects the geographic levels that will constitute the rows in their output. Fig. 5 shows the interface for this selection. The user makes choices between harmonized and year-specific and among the types of administrative levels. These choices are then applied to the datasets they selected in the previous step, and information about the corresponding geographic levels is displayed in the table in the lower portion of the screen. Additional details about the geographic levels available for each country may be accessed by clicking the links in the administrative level column of the table.

Fig. 6 shows the raster data selection screen. Like area-level variables, raster variables are organized into topics or may be searched. Unlike area-level variables, which may occur in many different country-year census datasets, raster variables are typically unique to a single dataset. When a topic is selected, variables available within the topic are listed, and the dataset each variable appears in is identified. As with area-level data, documentation about each variable and dataset can be obtained by clicking on the name of the variable or dataset. For raster data, dataset-level documentation includes information about the spatial reference system, primary source data, and a processing summary. Variable level documentation includes a more detailed variable description, data type and units, and summary statistics of the values.

For output formatted as area-level data, raster variables are summarized to the units in the selected geographic levels and incorporated in the same output table as the native area-level data. Operations used to summarize raster variables are selected on the screen shown in Fig. 7. Each combination of raster variable and operation will create an area-level variable column in the output table. The summary operations that may be applied to a raster variable depend on the variable's data type, so raster variables are organized by dataset and data type. Users may apply the same operation(s) to all variables within a dataset/data type, or they may expand the data types to select operations for individual variables.

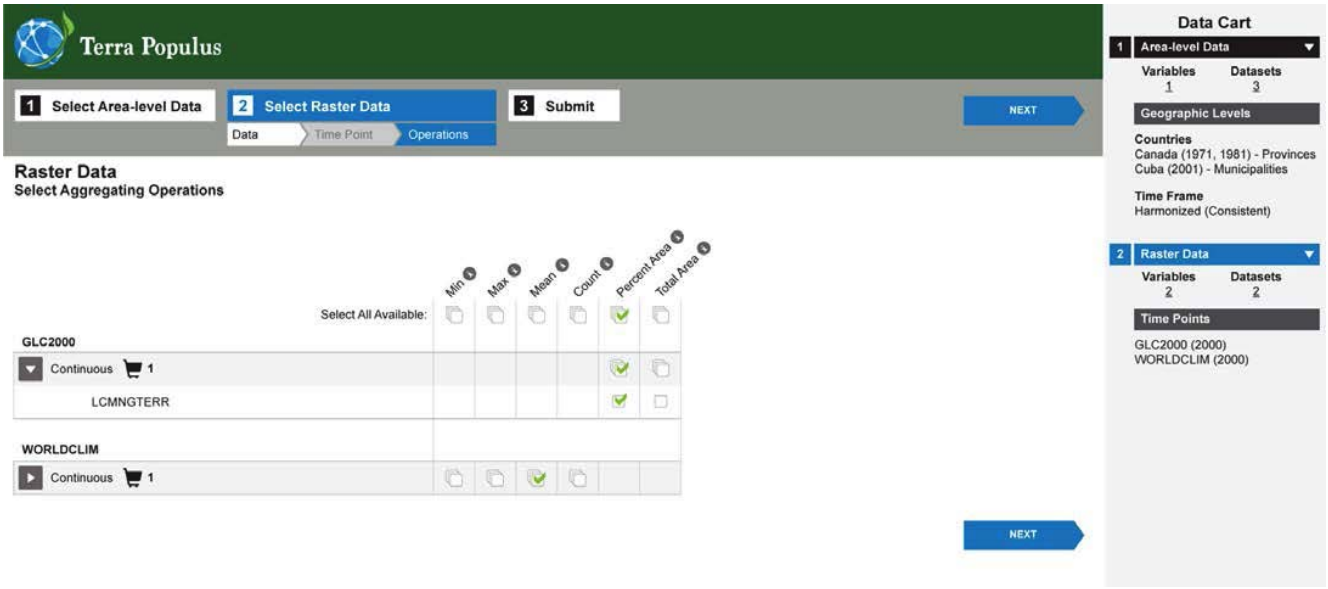

Figure 7. TerraPop Dissemination System: Raster Operations

	A	B	C	D	E	F
1	GEOLEV1_LABEL	GEOLEV1	TOTMALE_GEOLEV1_CA1971A	TOTMALE_GEOLEV1_CA1981A	ANPRECIP_mean_1_GEOLEV1	LCMNGTERR_percent_area_areal_1_GEOLEV1
2	British Columbia	124059	1100900	1365250	764.8	0.0678
3	Alberta	124048	828200	1143600	453.6	1.1157
4	Saskatchewan	124047	470900	486000	426.4	2.1658
5	Manitoba	124046	494700	506600	476.2	0.6061
6	Ontario	124035	3843600	4246650	729.2	0.4198
7	Quebec	124024	2996700	3172200	773.4	0.0875
8	New Brunswick	124013	319600	346000	1104.5	0.1796
9	Nova Scotia	124012	396300	419700	1335.6	0.1670
10	Prince Edward Is, Yukon, NW Territories, & Nunavut	124011		96950	233.9	0.0020
11	Newfoundland	124010	266000	285800	985.9	0.0000

Figure 8. TerraPop Data Extract

A portion of the extract produced by the workflow described in Figures 4-7 is shown in Figure 8. The figure shows the data table for Canada, included in the extract as a CSV. The data table includes columns for the geographic unit names and codes, the total male population for each of the years requested, the mean annual precipitation, and the percent area cultivated and managed in agriculture. This table may be joined to the shapefile of Canadian provinces, also included with the extract, based on the geographic unit codes.

Although dissemination of integrated TerraPop data is primarily through the data access system, we are also developing an application programming interface (API) that will facilitate direct access to the data, metadata, and integration tools. Functionality will include providing variable-dataset availability information, detailed metadata on particular objects, available operations and transformations given a set of input variables and datasets, and submission of fully defined extract queries.

VI. Conclusion

The profound significance of the scientific questions that integrated population and environmental data can answer is the central motivation for creating TerraPop. TerraPop has created a unique international reference collection for investigating changes in the human-environment system by integrating global data collections on land use and climate change with major global collections of population data. The data collection and integration tools are a powerful resource for understanding the causes and consequences of the cataclysmic transformations in the human population and environment that are reshaping the planet. By providing richly detailed data spanning the globe over multiple decades, TerraPop represents a unique laboratory for developing and testing social, economic, ecological, and climate models. Accordingly, TerraPop has the potential to serve thousands of researchers from dozens of disciplines, including transportation and environmental engineering, biology, geography, ecology, sociology, climate science, history, economics, public affairs, epidemiology, and demography.

TerraPop is also advancing information and computer science. Organizing, archiving, and making global datasets interoperable and easily accessible requires both advanced information technology and the expertise of the domain scientists who produce and use the data. The challenges derive not only from the large scale of the data collections but also from their complexity. The population and environmental data are multi-scale 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 are

engaging at the leading edge of computer and information science and develop new technologies and processes.

ACKNOWLEDGMENT

TerraPop is supported by the National Science Foundation (ACI-0940818), and relies on research infrastructure provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R24HD041023).

REFERENCES

[1] F. Ferreira and M. Ravallion, "Poverty and Inequality: The Global Context," in The Oxford Handbook on Economic Inequality, W. Salverda, B. Nolan, and T. Smeeding, Eds. Oxford: Oxford University Press, 2011, pp. 599-638.

[2] World Bank, World Development Indicators [website]. Available: http://data.worldbank.org/indicator.

[3] B. O'Neill, C.F. MacKellar, and W. Lutz, *Population and Climate Change*. Cambridge: Cambridge University Press, 2001.

[4] W. Lutz and A. Goujon, "The World's Changing Human Capital Stock: Multi-State Population Projections by Educational Attainment," *Population and Develop. Review*, vol. 27, no. 2, pp. 323-339, June 2001, DOI: 10.1111/j.1728-4457.2001.00323.x.

[5] National Research Council Panel on Research Agenda and New Data for an Aging World, *Preparing for an Aging World: The Case for Cross-National Research*. Washington D.C.: National Academy Press, 2001.

[6] United Nations, Department of Economic and Social Affairs, Population Division, "World Population Prospects: The 2006 Revision, Highlights," New York, United Nations, Working Paper No. ESA/P/WP.202, 2007.

[7] J.A. Foley et al, "Global Consequences of Land Use," *Science*, vol. 309, no. 5734, pp. 570-574, 22 July 2005, DOI:10.1126/science.1111772.

[8] A. Fischlin et al., "Ecosystems, Their Properties, Goods and Services," in *Climate Change 2007: Impacts, Adaptation and Vulnerability* (Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change), M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden, and C.E. Hanson, Eds. Cambridge: Cambridge University Press, 2007. https://www.ipcc.ch/report/ar4/.

[9] IPCC, "Summary for Policymakers," in *Climate Change 2007: The Physical Science Basis* (Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change), S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, Eds. Cambridge: Cambridge University Press, 2007. https://www.ipcc.ch/report/ar4/.

[10] W.K. Michener et al., "Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences," *Ecological Informatics*, vol. 11, pp. 5-15, Sept. 2012, doi: 10.1016/j.ecoinf.2011.08.007.

[11] N. Ramankutty and J.A. Foley, "Characterizing Patterns of Global Land Use: An Analysis of Global Croplands Data," *Global Biogeochemical Cycles*, vol. 12, no. 4, pp. 667-685, 1998.

[12] N. Ramankutty, and J.A. Foley, "Estimating Historical Changes in Global Land Cover: Croplands from 1700 to 1992," *Global Biogeochemical Cycles*, vol. 13, no. 4, pp. 997-1027, 1999.

[13] N. Ramankutty et al., "Farming the Planet: 1. Geographic Distribution of Global Agricultural Lands in the Year 2000," *Global Biogeochemical Cycles*, vol. 22, no. 1, Mar. 2008, doi:10.1029/2007GB002952.

[14] C. Monfreda, N. Ramankutty, and J.A. Foley, "Farming the Planet: 2. Geographic Distribution of Crop Areas, Yields, Physiological Types, and Net Primary Production in the Year 2000," *Global Biogeochemical Cycles*, vol. 22, no. 1., Mar. 2008, doi:10.1029/2007/GB002947.

[15] S. Ruggles, J. D. Hacker, and M. Sobek, "Order out of Chaos: The Integrated Public Use Microdata Series," *Historical Methods*, vol. 28, no. 1, pp. 33–39, 1995.

[16] A. Esteve and M. Sobek, "Challenges and methods of international census harmonization," *Historical Methods*, vol. 36, no. 2, pp. 66–79, 2003.

[17] S. Ruggles et al., "The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata," *J. of Demographic Economics*, vol. 81, no. 2, pp. 203-216, 2015, doi:10.1017/dem.2014.6.

[18] S. Ruggles et al.,, *Integrated Public Use Microdata Series: Version 6.0* [Machine-readable database]. Minneapolis: University of Minnesota, 2015. Available: http://usa.ipums.org.

[19] Minnesota Population Center, *Integrated Public Use Microdata Series, International: Version 6.3* [Machine-readable database]. Minneapolis: University of Minnesota, 2014. Available: http://international.ipums.org.

[20] North Atlantic Population Project and Minnesota Population Center, *NAPP: Complete Count Microdata*, Version 2.0 [Computer Files]. Minneapolis: Minnesota Population Center [distributor], 2008. Available: http://www.nappdata.org.

[21] S. Ruggles, "Big microdata for population research," *Demography*, vol. 51, no. 1., pp. 287-297, 2014.

[22] Minnesota Population Center, *National Historical Geographic Information System* [Machine-readable database]. Minneapolis: University of Minnesota, 2009. Available: http://www.nhgis.org/.

[23] E.M. Bartholome and A.S. Belward, "GLC2000: A new approach to global land cover mapping from Earth Observation data," *Int. J. of Remote Sensing*, vol. 26, no. 9, pp. 1959-1977, 2005.

[24] A. Di Gregorio and L. J. M. Jansen, *Land cover classification system: classification concepts and user manual: LCCS*. Rome, Food & Agriculture Org., No. 8, 2005.

[25] M.A. Friedl et al., "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of the Environment*, vol. 83, no. 1, pp. 287–302, 2002, doi:10.1016/929 S0034-4257(02)00078-0.

[26] R. J. Hijmans et al., "Very high resolution interpolated climate surfaces for global land areas," *Int. J. of Climatology*, vol. 25, no. 15, pp. 1965-1978, 2005, doi: 10.1002/joc.1276.

[27] I. Harris, et al., "Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset," *Int. J. of Climatology*, vol. 34, no. 3, pp. 623–642, 2014, doi: 10.1002/joc.3711

[28] D.R. Steinwand, J.A. Hutchinson, and J.P. Snyder, "Map Projections for Global and Continental Data Sets and an Analysis of Pixel Distortion Caused by Reprojection," *Photogrammetric Engineering and Remote Sensing*, vol. 61, no. 12, pp. 1487-1497, 1995.

[29] J.C. Seong and E.L. Usery, "Assessing Raster Representation Accuracy Using a Scale Factor Model," *Photogrammetric Engineering and Remote Sensing*, vol. 67, no. 10, pp. 1185-1191, 2001.

[30] E.L. Usery et al., "Projecting Global Datasets to Achieve Equal Areas," *Cartography and Geographic Information Science*, vol. 30, no. 1, pp. 69-79, 2003, doi: 10.1559/152304003100010956.

[31] J. B. Holt and H. Lu, "Dasymetric Mapping for Population and Sociodemographic Data Redistribution," in U*rban Remote Sensing Synthesis and Modeling in the Urban Environment*, X. Yang, ed. Hoboken: John Wiley & Sons, Ltd, 2011, pp. 195–210, doi:10.1002/9780470979563.ch14.

[32] Food and Agriculture Organization of the United Nations, Global Administrative Unit Layers [GAUL], 2014. Available: http://www.fao.org/geonetwork.

[33] R. Hijmans et al., Global Administrative Areas - GADM v2 global shapefile, 2011. Available: http://biogeo.ucdavis.edu/data/gadm2/gadm_v2_shp.zip.

[34] UN Geographic Information Working Group, SALB: Second Level Administrative Boundaries, 2011. Accessed www.unsalb.org, Feb. 2012 (no longer available).

[35] T.A. Kugler et al., "Terra Populus: Workflows for Integrating and Harmonizing Geospatial Population and Environmental Data," *J. of Map & Geography Libraries*, vol. 11, no. 2, pp. 180-206, 2015, doi: 10.1080/15420353.2015.1036484.