IPUMS Terra:
Integrating Detailed Population Characteristics and Environmental Data

Panel contribution to the Population-Environment Research Network Cyberseminar,
"People and Pixels Revisited"
(20-27 February 2018)
https://populationenvironmentresearch.org/cyberseminars/10516

By Tracy Kugler
IPUMS, University of Minnesota, USA

IPUMS Terra (originally known as Terra Populus or TerraPop for short) was created to lower technical barriers to joining census-based population data and pixel-based environmental data. At the time the original People and Pixels workshop was convened, research utilizing linked population and remote sensing data was limited to a handful of practitioners. These researchers employed highly specialized technical skills and addressed a variety of challenges in linking data from the different domains. In the intervening decades, GIS tools, a key enabler of these types of linkages, have become more powerful and easier to use. At the same time, the volume of readily available data in both the population and environmental domains has expanded exponentially. In conjunction with these advancements, more researchers have become interested in the possibilities of linked population and environmental data. However, technical challenges posed by the different data structures used within different scientific domains remain, and there are still relatively few researchers who are readily able to work with both types of data.

IPUMS Terra aims to foster more widespread use of linked population and environmental data by taking the heavy lifting of data linking and transformation off the hands of individual researchers. IPUMS Terra utilizes location-based integration to transform among three data structures: individual-level population microdata, area-level data describing places defined by geographic boundaries, and pixel-based raster data. The IPUMS Terra data access system provides a workflow through which users first select the data structure in which they would like to receive their output data. They then select data of interest from one or more of the source data structures and make basic decisions about how transformations are conducted. The system processes the requested data, handling any necessary transformations, and delivers a customized, integrated dataset in the user's chosen data structure.

The transformations performed by IPUMS Terra depend on the output data structure and the types of input data selected (Kugler et al. 2015; Ruggles et al. 2016). Users may choose to receive area-level output data (describing administrative units) and include both census-based aggregate population data and raster data on land cover or climate as inputs. IPUMS Terra then transforms the raster data by calculating summary measures (e.g., % area of forest cover, mean annual precipitation) for the administrative units described in the aggregate population data. The output dataset includes a record for each administrative unit with variables for both population and environmental characteristics. Alternatively, users may choose to receive microdata output. In this case, they may choose individual- and household-level variables as well as contextual variables derived from area-level or raster data. IPUMS Terra would then calculate any necessary raster summaries and attach the contextual variables to the individual records based on the administrative unit in which the individual lives. Conversely, IPUMS Terra can distribute values from area-level data across the pixels located within each administrative unit to produce raster output. As described in further detail below, the transformation is based on a uniform distribution assumption, so all cells in each unit receive the same value..

The IPUMS Terra Data Collection

IPUMS Terra includes a rich collection of population data, drawing on the IPUMS International collection. IPUMS International currently includes microdata from 85 countries, with multiple census years are available for many countries. In total, the collection consists of 672 million person records from 301 censuses. Each person record includes the individual's responses to all of the questions asked in the census, covering topics such as demographics, education, employment, and housing characteristics. For each census, IPUMS International and IPUMS Terra include a sample (typically 5-10%) of the households enumerated.

The inclusion of all characteristics for each individual makes microdata extraordinarily flexible. Researchers can construct individual-level models addressing specific combinations of characteristics. Researchers can also aggregate data along any set of characteristics (e.g., all unmarried females of a certain age, ethnicity, and educational attainment). This flexibility in tabulation helps avoid ecological fallacies that may arise from using published census tables. For example, published tables may show that a district has a high proportion of people in poverty and a high proportion of elderly people, suggesting that elderly people are likely to be living in poverty. But unless a cross-tabulation of poverty against age exists, that conclusion is uncertain. With access to microdata, a researcher interested in the particular intersection between poverty and age could construct the required cross-tabulation. Furthermore, person records are organized into households, enabling investigation of household characteristics and family relationships.

IPUMS Terra has also pre-tabulated a series of area-level variables from the microdata. These variables summarize population characteristics, such as percent unemployment, literacy rate, and percent of households with electricity, for each administrative unit. Thanks to the harmonization work conducted by IPUMS International to make microdata variables comparable across time and place, these pre-tabulated variables are consistent across the various censuses (something that is decidedly not the case in tables published by each country). For countries not included in IPUMS International, IPUMS Terra incorporates basic area-level data on total population and population by sex from published census tables.[1]

On the pixels side of the data collection, IPUMS Terra includes data on land cover, land use, and climate. Classified land cover datasets include Global Land Cover 2000 and the MODIS IGBP product. GLC2000 includes 23 land cover classes at 30 arc second (~1 km) resolution for the year 2000 (Fritz et al. 2003). MODIS IGBP includes 17 classes at 500 meter resolution for annual time steps between 2001 and 2012 (Friedl et al. 2010). Land use data from the Global Landscapes Initiative include the area used for cropland and pasture, and the harvested area and yield for 175 individual crops (Monfreda, Ramankutty, and Foley 2008; Ramankutty et al. 2008; Ramankutty and Foley 1998). Climate data cover both long-term averages and time series data.

[1] An on-going project is processing published census tables covering the full range of population characteristics for these countries, as well as published agricultural census data tables from around the world.

The WorldClim dataset captures average precipitation, temperature, and derived variables over the 1950-2000 timeframe at 30 arc second resolution (Hijmans et al. 2005). The Climate Reference Unit Time Series dataset includes temperature, precipitation, vapor pressure, cloud cover, and derived variables at 0.5 degree resolution for monthly time steps from 1901-2013 (Harris, Jones, and Osborn 2014).

Transformations between microdata, area-level data, and raster data are supported by IPUMS Terra's collection of geographic administrative unit boundary data. For most countries, IPUMS Terra provides first- and second-level administrative unit boundaries. In some cases, national statistical or cartographic offices have provided administrative unit boundary data. In other cases, geographic boundaries from existing international collections (GADM, SALB, GAUL) have been adjusted to match the set of units present in census data, including tracing changes in administrative units back through time to match historical census data (UN Geographic Information Working Group 2011; Food and Agriculture Organization of the United Nations 2014; Hijmans et al. 2011). IPUMS Terra provides both the boundaries as they existed at the time of each census, as well as harmonized boundaries in which units have been aggregated so that boundaries are stable over time (Kugler, Manson, and Donato 2017). For boundaries linked to microdata (including area-level data tabulated from microdata), IPUMS International's agreements with national statistical offices require that identified geographic units contain at least 20,000 people in order to protect the confidentiality of respondents. IPUMS Terra therefore combines administrative units that have populations of less than 20,000 people with neighboring units.

Research Utilizing IPUMS Terra

IPUMS Terra's incorporation of a rich set of detailed population characteristics readily combined with environmental data complements many on-going efforts to tie people to pixels through the production of gridded population density surfaces. IPUMS Terra facilitates research on the complex, bi-directional interactions between people and the environment. Research based on IPUMS Terra data has investigated the relationship between climate change and migration, exposure to pollution among different immigrant groups (Bakhtsiyarava and Nawrotzki 2017), and the variations in vulnerability to climate variability determined by socio-economic characteristics (Adamo et al. 2014). These investigations typically rely on detailed population characteristics only available in extensive census data, such as household-level migration behavior combined with other variables such as age, household structure, and education that may influence migration decisions (Nawrotzki and Bakhtsiyarava 2016; Nawrotzki and DeWaard 2016; Nawrotzki et al. 2016; Hunter et al. 2014; Nawrotzki, Schlak, and Kugler 2016).

What's Next for IPUMS Terra?

IPUMS Terra is in a period of transition. An NSF grant under the DataNet program supported the development of the existing data collection, applications, and services. We are now looking ahead and evaluating new directions and potential enhancements to the IPUMS Terra system. Potential enhancements include incorporation of additional environmental data through links to Google Earth Engine (GEE), expanded connections to other IPUMS data products for microdata tabulation and attachment of contextual variables, and more sophisticated transformations from area-level to raster data.

The land use, land cover, and climate data currently available through IPUMS Terra have been ingested into the IPUMS Terra database, a labor-intensive and costly process. In order to expand the scope of environmental data available, we are considering tapping into the rich collection of data (and data transformation functionality) available through GEE. The GEE collection includes a variety of classified land cover products, several high-frequency climate datasets, elevation data, night lights, and more. The high-frequency (8-day, daily, and even hourly) datasets are both particularly exciting and particularly challenging. Transforming these data for integration with IPUMS Terra's population data would require both spatial and temporal summarization and the selection of metrics with which to perform temporal summarization.

IPUMS Terra currently incorporates microdata from IPUMS International, calculating area-level tabulations from these microdata and enabling attachment of contextual variables to them. IPUMS also hosts many other population data products, encompassing a broad array of census and survey data (www.ipums.org). Other IPUMS products provide U.S. microdata from decennial censuses, the American Community Survey, the Current Population Survey (labor force characteristics), the National Health Interview Survey, and the Medical Expenditures Panel Study; U.S. aggregate census data at multiple geographic levels back to 1790; and microdata from the international Demographic and Health Surveys, focusing primarily on women's health and reproductive behavior. As an organization, IPUMS is embarking on a concerted effort to develop application programming interfaces (APIs) throughout its products. This effort will facilitate connections across products and could enable the application of IPUMS Terra's transformation functionality to additional data. Microdata from other products could be tabulated and transformed into raster data, or area-level summaries of raster data could be attached to extracts from other products.

IPUMS Terra currently transforms area-level data into raster data based on a uniform distribution assumption. That is, each pixel within an administrative unit is assigned the same value. For rates, such as percent unemployed, the value for the unit as a whole is assigned to each pixel in the unit. For counts, the unit total is divided by the number of pixels, and the quotient is assigned to each pixel. Under the initial project, we explored the possibility of using dasymetric reallocation, informed by ancillary data such as elevation, land cover, night lights, and road networks. The IPUMS Terra dasymetric reallocation model would be a mid-level model, between the light touch applied for the Gridded Population of the World (Center for International Earth Science Information Network (CIESIN) 2016) datasets and heavily modeled products such as WorldPop (www.worldpop.org). Any of the characteristics in the IPUMS Terra population data could potentially be distributed across pixels using such a model.

We welcome any feedback you may have regarding these or other enhancements to IPUMS Terra. Which enhancements would you find most valuable? Which specific data are of the most interest with respect to each enhancement? What types of research could be enabled by these or other enhancements?

References

Adamo, S. B., C. A. Fitch, T. Kugler, and E. Doxsey-Whitfield. 2014. "Social Vulnerability and Climate Variability in Southern Brazil: A TerraPop Case Study." In *American Geophysical Union, Fall Meeting 2014*, abstract id. GC41B-id. 0544. http://adsabs.harvard.edu/abs/2014AGUFMGC41B0544A.

Bakhtsiyarava, Maryia, and Raphael J. Nawrotzki. 2017. "Environmental Inequality and Pollution Advantage among Immigrants in the United States." *Applied Geography* 81 (April). Pergamon: 60–69. doi:10.1016/J.APGEOG.2017.02.013.

Center for International Earth Science Information Network (CIESIN). 2016. "Gridded Population of the World, Version 4." Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). doi:10.7927/H4NP22DQ.

Food and Agriculture Organization of the United Nations. 2014. "Global Administrative Unit Layers (GAUL)." http://www.fao.org/geonetwork.

Friedl, M.A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. 2010. "MODIS Collection 5 Global Land Cover: Algorithm Refinements and Characterization of New Datasets, 2001-2012, Collection 5.1 IGBP Land Cover, Boston University, Boston, MA, USA."

Fritz, Steffen, Etienne Bartholomé, Alan Belward, Andrew Hartley, Hans-Jürgen Stibig, Hugh Eva, Philippe Mayaux, et al. 2003. "Global Land Cover 2000 [Dataset]." http://bioval.jrc.ec.europa.eu/products/glc2000/products.php.

Harris, I., P. D. Jones, and T. J. Osborn. 2014. "Updated High-Resolution Grids of Monthly Climatic Observations - the CRU TS3.10 Dataset" 34 (3): 623–42. doi:10.1002/joc.3711.

Hijmans, Robert J, S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology* 25: 1965–78. http://www.worldclim.org/.

Hijmans, Robert J, Julian Kapoor, John Wieczorek, Nel Garcia, Aileen Maunahan, Arnel Rala, and Alex Mandel. 2011. "Global Administrative Areas - GADM v2 Global Shapefile." http://biogeo.ucdavis.edu/data/ gadm2/gadm_v2_shp.zip.

Hunter, Lori M., Raphael Nawrotzki, Stefan Leyk, Galen J. Maclaurin, Wayne Twine, Mark Collinson, and Barend Erasmus. 2014. "Rural Outmigration, Natural Capital, and Livelihoods in South Africa." *Population, Space and Place* 20 (5): 402–20. doi:10.1002/psp.1776.

Kugler, Tracy A., D.C. Van Riper, S.M. Manson, D.A. Haynes Ii, J. Donato, and K. Stinebaugh. 2015. "Terra Populus: Workflows for Integrating and Harmonizing Geospatial Population and Environmental Data." *Journal of Map and Geography Libraries* 11 (2). doi:10.1080/15420353.2015.1036484.

Kugler, Tracy A, Steven M Manson, and Joshua R Donato. 2017. "Spatiotemporal Aggregation for Temporally Extensive International Microdata." *Computers, Environment and Urban Systems*, May. doi:10.1016/j.compenvurbsys.2016.07.007.

Monfreda, Chad, Navin Ramankutty, and Jonathan A Foley. 2008. "Farming the Planet: 2. Geographic Distribution of Crop Areas, Yields, Physiological Types, and Net Primary Production in the Year 2000." *Global Biogeochemical Cycles* 22 (1). American Geophysical Union: 1–19. doi:10.1029/2007GB002947.

Nawrotzki, Raphael J., and Maryia Bakhtsiyarava. 2016. "International Climate Migration: Evidence for the Climate Inhibitor Mechanism and the Agricultural Pathway." *Population, Space and Place*, May. doi:10.1002/psp.2033.

Nawrotzki, Raphael J., and Jack DeWaard. 2016. "Climate Shocks and the Timing of Migration from Mexico." *Population and Environment* 38 (1). Springer Netherlands: 72–100. doi:10.1007/s11111-016-0255-x.

Nawrotzki, Raphael J., Daniel M. Runfola, Lori M. Hunter, and Fernando Riosmena. 2016. "Domestic and International Climate Migration from Rural Mexico." *Human Ecology* 44 (6). Human Ecology: 687–99. doi:10.1007/s10745-016-9859-0.

Nawrotzki, Raphael J., A. M. Schlak, and Tracy A. Kugler. 2016. "Climate, Migration, and the Local Food Security Context: Introducing Terra Populus." *Population and Environment* 38 (2): 164–84.

Ramankutty, Navin, Amato T Evan, Chad Monfreda, and Jonathan A Foley. 2008. "Farming the Planet: 1. Geographic Distribution of Global Agricultural Lands in the Year 2000." *Global*

Biogeochemical Cycles 22 (1). American Geophysical Union: 1–19. doi:10.1029/2007GB002952.

Ramankutty, Navin, and Jonathan A Foley. 1998. "Characterizing Patterns of Global Land Use: An Analysis of Global Croplands Data." *Global Biogeochemical Cycles* 12 (4). Washington, DC: American Geophysical Union, c1987-: 667–85. doi:10.1029/98GB02512.

Ruggles, Steven, Tracy A Kugler, Catherine A Fitch, and David C Van Riper. 2016. "Terra Populus: Integrated Data on Population and Environment." In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. Atlantic City, NJ: IEEE.

UN Geographic Information Working Group. 2011. "SALB: Second Level Administrative Boundaries." www.unsalb.org.