https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Allocating people to pixels: A review of large-scale gridded population data products and their fitness for use

Stefan Leyk[1*], Andrea E. Gaughan[2, 10], Susana B. Adamo[3], Alex de Sherbinin[3], Deborah Balk[9], Sergio Freire[5], Amy Rose[4], Forrest R. Stevens[2], Brian Blankespoor[8], Charlie Frye[7], Joshua Comenetz[6], Alessandro Sorichetta[10], Kytt MacManus[3], Linda Pistolesi[3], Marc Levy[3], Andrew J Tatem[10]

[1]Department of Geography, University of Colorado Boulder, Boulder, CO 80309, U.S.A.
[2]Department of Geography and Geosciences, University of Louisville, KY, 40292, U.S.A.
[3]CIESIN, Columbia University, Palisades, NY, 10964, U.S.A.
[4]Human Dynamics Group, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, U.S.A.
[5]European Commission, Joint Research Centre (JRC), Ispra, Italy
[6]U.S. Census Bureau, Washington, D.C., 20233, U.S.A.
[7]Environmental Systems Research Institute, Redlands, CA 92373
[8]Development Research Group, World Bank, Washington, D.C. 20433, U.S.A.
[9]CUNY Institute for Demographic Research, and Marxe School of Public And International Affairs, Baruch College, City University of New York, 10010, USA
[10]WorldPop, School of Geography and Environmental Sciences, University of Southampton, Southampton, SO17 1BJ, UK

[*] *Correspondence to*: Stefan Leyk (stefan.leyk@colorado.edu)

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Abstract. Population data represent an essential component in studies focusing on human-nature interrelationships, disaster risk assessment and environmental health. Several recent efforts have produced global and continental-extent gridded population data which are becoming increasingly popular among various research communities. However, these data products, which are of very different characteristics and based on different modeling assumptions, have never been systematically reviewed and compared which may impede their appropriate use. This article fills this gap and presents, compares and discusses a set of large-scale (global and continental) gridded datasets representing population counts or densities. It focuses on data properties, methodological approaches and relative quality aspects that are important to fully understand the characteristics of the data with regard to the intended uses. Written by the data producers and members of the user community, through the lens of the "fitness for use" concept, the aim of this paper is to provide potential data users with the knowledge base needed to make informed decisions about the appropriateness of the data products available in relation to the target application and for critical analysis.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Attribution Table. Gridded population data collections described in this review article, years covered, digital object identifiers and reference links. This review covers sources and versions available as of May 2019.

Data collection	Year(s)	Population Themes	Digital Object Identifier (doi)	Reference Link
Gridded Population of the World (GPWv4.11)	2000; 2005; 2010; 2015; 2020	Persons	10.7927/H4JW8BX5	http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
		UN WPP-adj.	10.7927/H4PN93PB	
		Pop. Density	10.7927/H49C6VHW	
		UN WPP-adj.	10.7927/H4F47M65	
Global Rural Urban Mapping Project (GRUMPv1)	1990; 1995; 2000	Persons	10.7927/H4VT1Q1H	http://sedac.ciesin.columbia.edu/data/collection/grump-v1
		Pop. Density	10.7927/H4R20Z93	
LandScan Global Population Database (Landscan Global)	annual: 2000–2016	Persons	N/A; data download at:	https://landscan.ornl.gov/
WorldPop	2000-2020	Persons	10.5258/SOTON/WP00645	www.worldpop.org
Global Human Settlement Layer - Population (GHS-POP)	1975; 1990; 2000; 2015	Persons	http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a	http://ghsl.jrc.ec.europa.eu/ghs_pop.php
World Population Estimate (WPE)	2013	Persons	10.13140/RG.2.2.18213.14565	https://sites.google.com/ciesin.columbia.edu/popgrid/find-data/esri
	2015	Persons	10.13140/RG.2.2.16160.79367	
		Pop. Density	10.13140/RG.2.2.14857.70248	
	2016	Persons	10.13140/RG.2.2.12996.48007	
		Pop. Density	10.13140/RG.2.2.21568.58885	
History Database of the Global Environment (HYDE) Population Grids v3.2	10,000 BC - 2015	Persons	10.17026/dans-25g-gez3	https://themasites.pbl.nl/tridion/en/themasites/hyde/download/index-2.html
High Resolution Settlement Layer (HRSL)	2015	Persons	N/A; data download at	https://ciesin.columbia.edu/data/hrsl/
European GHS Population Grid (GHS-POP-EUROSTAT)	2011	Persons	http://data.europa.eu/89h/jrc-ghsl-ghs_pop_eurostat_europe_r2016a	http://data.jrc.ec.europa.eu/dataset/jrc-ghsl-ghs_pop_eurostat_europe_r2016a
Gridded Population Mapping (Demobase)	1998-present	Persons	N/A; data download at:	https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/demobase.html

5

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

1 Introduction

The distribution and density of human population continues to be a critical component to measuring, mapping and understanding human-environment interrelationships, identifying populations at risk of infectious diseases or disasters, and informing management and policy decisions from local to global level initiatives (e.g. Wesolowski et al. 2014, Simarro et al. 2011, McDonald et al. 2011, Jones et al. 2008, McGranahan et al. 2007, Doocy et al. 2007). The traditional form of collecting population data is through a census or registry, and those population counts can be spatially linked to boundary datasets representing enumeration areas (the most basic unit of collected census data) or administrative units in a Geographic Information System (GIS). More recently, an increasing use of fully georeferenced censuses has made building-level mapping more feasible in some countries. However, census data vary substantially across countries with regard to quality, the number and size of enumerated areas, the frequency of data collection and the level of confidentiality depending on detail. The size of census units also varies significantly within countries between rural and urban areas. Thus, to be useful for many analytical purposes, substantial efforts are required to harmonize such enumerated data (de Sherbinin 2017). Since Tobler's "World population in a grid of spherical quadrilaterals" (Tobler et al. 1997) and Liverman et al.'s "People and Pixels" (Liverman et al. 1998), the benefits of gridded population data have been acknowledged. As a consequence, the scientific community has increasingly invested in ways to create global georeferenced data products that help overcome the inconsistencies in census-derived national population data and facilitate their integration with other gridded geospatial datasets such as, for example, remote sensing data products. This article, a product of the POPGRID Data Collaborative (POPGRID 2018), describes the variety of gridded population data products that have been created over the past 20 years and is an effort to aid users in better understanding the nature of these products, their qualities and forms of appropriate uses.

There is high demand for modeled gridded population datasets particularly in countries with less detailed or infrequent censuses. These datasets, for example, support land use and urban planning (Dong et al. 2017), measurement of economic development (Nordhaus 2006, Roberts et al. 2017), transportation infrastructure management (Iimi et al. 2016), resource allocation strategies (Islam et al. 2006), disaster risk mitigation, management and reduction (Ehrlich et al. 2018a, Aubrecht et al. 2016, Gunasekera et al. 2015, Mondal and Tatem 2012, Taramelli et al., 2010), climate change research (Blankespoor, Dasgupta and Lange 2017, Dasgupta et al. 2011), sampling design for household surveys (Blankespoor et al. 2018, Thomson et al. 2017), public health campaigns and assessments (Hay et al. 2004, Weber et al. 2018) and sustainable resource management (Koch et al. 2008, Parish et al. 2012) among many other applications1[1]. International frameworks for development and sustainability depend on the availability of population data, which are commonly used as a denominator in calculating different metrics and indicators. Such frameworks include the Sustainable Development Goals (SDGs), the Sendai Framework for Disaster Risk Reduction, the UNFCCC Paris Agreement, and the United Nations New Urban Agenda, to mention just a few.

The field has seen advances at multiple levels. First, the spatial resolution of underlying census data available for geoprocessing, along with the standards for producing such data (United Nations 2009), has improved dramatically in many countries since the creation of the earliest gridded population data products such as the Gridded Population of the World version 1 (Tobler et al. 1995). Second, significant progress has been made through advances in information extraction and classification of populated land area from remote sensing data at various resolutions (Wardrop et al. 2018). The increased availability and spatial granularity of remotely sensed information about the topography, vegetation and land cover has been critical to improve the identification of such places that are potentially inhabited and even the estimation of counts of people living there (Frye et al. 2018, Nieves et al. 2017, Pesaresi et al. 2013). Third, the combination of access, increased computing power, and greater spatial accuracy in ancillary datasets has provided the basis for methodological advances to redistribute

[1] The list of citations here are just a few of hundreds of possibilities that could have been identified. This paper does not aim to be a complete review of the literature or applications of all usages of the gridded data products under review.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

census-enumerated population counts to grid cells at continental and global scales with high accuracy (e.g., Freire et al. 2018) and to create time series of population estimates that can be used to fill in data gaps between national census surveys that are commonly taken at decadal intervals (Figure 1) (e.g., WorldPop & CIESIN 2018).

As a result of these recent developments, there are now several global and continental gridded population data sets that are based on different modeling approaches and input data layers. As might be expected, there are similarities but also important differences among these products, and yet to date there has neither been a systematic review of these various approaches, nor a comparison of the corresponding outputs. This represents a serious gap in the literature as these differences can easily lead to misunderstandings or inappropriate use of population grids. The objective of the paper is to fill that gap by helping guide users in forms of appropriate, uncertainty-aware use of the available global gridded population datasets in different application areas. Such an assessment is necessary as knowledge of underlying approaches and input data can inform about what each gridded product actually measures. For example, the exposure of a target population to disasters requires a population grid that 1) covers the area of interest, 2) provides a meaningful analytical unit (i.e., the size of the grid cell), 3) warrants the temporal currency needed relative to the time of interest, and 4) estimates the correct target population.[2] This example demonstrates why applying population grids is not trivial; grids have different characteristics that may affect the accuracy and precision of an analysis but also their suitability in a given context.

The above aspects together provide the essential components to assess the fitness for use of a data product in the context of relative data quality (Tayi and Ballou 1998). Fitness for use is a concept that has often been used to assess the appropriateness of a given spatial dataset for an intended purpose (Agumya and Hunter 1999, de Bruin, Bregt and Ven 2001, Devillers et al. 2007). Here, this concept will be applied to guide a growing user community in making informed decisions regarding the most appropriate dataset(s) for their intended use by better understanding the characteristics of the available different data products that also include the modeling assumptions behind them. Spatial, thematic and temporal accuracy play a key role in formalizing fitness for use. However, the multidimensionality of accuracy in the case of population grids is further driven by the nature and heterogeneity of the input population data, the use and characteristics of ancillary data involved and the methodological framework applied to redistribute population counts to grid cells. All these factors will be systematically explored in this article.

This review targets researchers and applied users in the geospatial, demographic, environmental and land use research communities with diverse needs. Section 2 begins with a brief history of population gridding, and Section 3 provides an introduction to the data products of interest herein, and summarizes the approaches behind the most recently released global gridded population datasets. Section 4 looks at commonalities and differences in methods applied and ancillary data used, and Section 5 provides a discussion for general guidance on the fitness for use of the different products and illustrates these aspects by visualizing the products in direct comparison. Finally, we identify future avenues of work and needed investments in Section 6.

2 People in pixels: Background and historical development

In the past, mapping population typically entailed linking tabulated population statistics to "vector features", such as points (for example, geographic coordinates indicating city centers) and/or polygons (most notably, administrative units or census enumeration areas). Beginning in the 1990s, a new approach to mapping population distributions emerged, which was to

[2] In the production of gridded population data, the underlying census data are accepted as demographically accurate. While demographers concern themselves with such issues as age-heaping (Myers, 1993) or completeness of registrations or census-samples (e.g., Potter and Ordóñez,, 1976) at the national and first-order administrative level, to the extent that such problems exist (perhaps to an even greater degree) in the fine-grain, underlying spatially-refined data, these issues are inherited into the gridded products.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

convert population data from irregular vector formats to gridded surfaces composed of regular, standardized grid cells or pixels (e.g., Tobler et al. 1995, Balk et al. 2006, Thomson et al. 2017).

The impetus to grid population data arose soon after the first GIS software packages were developed, and as the spatially-oriented research community began to use a growing number of gridded biophysical and geophysical data products. Regular grids represented an efficient and consistent data storage format, and the move to gridded data - already in use by the climatological modeling community - was reinforced by the growing array of remote sensing data products that began to appear in the 1970s and 1980s. By gridding population, researchers were able to more easily integrate population count and density data with biophysical data to better understand spatial distributions and components of socio-environmental systems. Furthermore, by decoupling the data from their original administrative boundaries, populations could then be easily aggregated to different units of interest (e.g., watersheds or climate zones) for spatial and statistical analysis (Balk et al. 2009).

Early efforts to grid populations include an African population grid for UNEP's Global Atlas of Desertification (Deichmann and Eklundh 1991), the NASA Goddard Institute for Space Studies' Global Distribution of 1984 Population Density at 1°×1 ° Resolution (Fung et al. 1991), and Tobler's pycnophylactic method (Tobler et al. 1997), which resulted in the first version of Gridded Population of the World in 1995. These early approaches spread populations evenly across grid cells within input census units, with adjustment effects applied (in the case of the pycnophylactic method) at the unit boundaries. One inherent problem of these early modeled outputs is the existence of aggregation effects that often lead to analytical challenges, as described in the next paragraph. Two concomitant changes helped to partially overcome this inherent problem: First, improvement in the spatial resolution of the underlying population data, and increased computation capacity to use higher-resolution data, have reduced the impact of this problem for many applications. Second, as methods and data availability have progressed, researchers also sought to improve the spatial resolution of population estimates by reallocating populations using ancillary datasets, a spatial refinement strategy known as dasymetric mapping (Semenov-Tian-Shansky 1928, Wright 1936), in combination with different statistical methods (e.g., Wu et al. 2005). Both dasymetric and statistical techniques continue to play an important role in gridded population mapping (Mennis 2009), as discussed below. In addition to such spatial refinement strategies, ongoing efforts also focus on improving the temporal coverage and temporal resolution as well as increasing the variety of population characteristics mapped.

While the development of consistent, comparable grids is what makes gridded data products so useful, there are some important implications that need to be addressed, as should be the case for any geospatial data. Population is not randomly distributed and therefore the allocation and representation of populations will always be subject to aggregation effects. These effects have been described in the geography literature as the Modifiable Areal Unit Problem (MAUP) (Openshaw and Taylor 1981). According to MAUP, the level of aggregation -- in this case the census unit or administrative level -- and the shape of the reporting units can affect the analysis in ways that are difficult to predict. MAUP is manifested in the flawed assumption of homogeneity of population distributions across census reporting units. The spatial resolution of a gridded population dataset determines the output analytical unit and thus will have implications due to these same aggregation effects after transitioning population counts from vector boundaries to grid cells. In other words, these aggregation-related problems of enumerating data are not eliminated but are propagated into a different data structure through the creation of gridded population data.

As one of the most persistent problems in geographical analysis, MAUP-related research has made significant progress to better understand the sensitivity of analytical results due to changing aggregation levels using synthetic and real-world data (Amrhein 1995, Steel and Holt 1996, Flowerdew et al. 2001, Pawitan and Steel 2006, Wong 2009, Arbia and Petrarca 2011, Maclaurin et al. 2015). However, because of this sensitivity, it's important to recognize that MAUP affects the fitness for use of data products for specific analyses in which the spatial precision of population locations is critical. Other implications that affect the quality of population grids have been reported by the data producers including temporal differences of input and ancillary variables as well as the kind of population that is mapped (daytime, night-time, residential, ambient, etc.). While

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

these quality aspects are important to help the user community by guiding general applications, the impact of these aspects on the fitness for use of the data products for specific applications is difficult to measure and not well understood.

3 Current data products, characteristics and availability

This section summarizes several global data products including the Center for International Earth Science Information Network (CIESIN)'s Gridded Population of the World (GPWv4.11) and Global Rural Urban Mapping Project (GRUMPv1); The European Commission Joint Research Centre (JRC) and CIESIN's Global Human Settlement Population Layer (GHS-POP); Oak Ridge National Laboratory's LandScan; ESRI's World Population Estimate (WPE); and WorldPop's WorldPop datasets. We also reference The History Database of the Global Environment (HYDE) as a gridded data product representing a long-term historical context (i.e. ~12,000 years). While the focus of this review is on global population grids, we also discuss a number of country and regional/continental grids, including Facebook and CIESIN's High Resolution Settlement Layer (HRSL), JRC's European GHS Population Grid, and the U.S. Census Bureau's country grids (Demobase). Owing to space constraints, we omit gridded population projections such as those developed by Jones and O'Neill (2016). Detailed summaries of the similarities and differences in these data products are summarized in Table 1. Extended data documentation and visual comparison tools (tables and map services) are available through the POPGRID website (www.popgrid.org).

3.1 Global population data production efforts

Gridded Population of the World version 4 (GPW4) is a data collection consisting of gridded data products on total population counts and densities and other key demographic variables, globally at a nominal spatial resolution of 1km using the World Geodetic System (WGS84) as geographic reference system (Doxsey-Whitfield et al. 2015). GPW4 includes estimates for the years 2000, 2005, 2010, 2015, and 2020 respectively. Additionally, GPW4 includes vector point data representing the centroids of input census enumeration units, and gridded data on land and water area estimates, national identifiers, and data quality metrics. GPW4 employs a uniform allocation approach to disaggregate population which is based purely on the land area of a given pixel. The Mean Input Administrative Area can be used as a data quality metric to provide users with guidance as to the effective local resolution of original input population data. Because the size and extent of input census geographies is highly variable, within and across countries, the scale at which GPW4 data should be analyzed differs by region. For example, in the USA, where Census blocks are the primary input units, highly localized analysis is appropriate, whereas the coarse input geographies of Libya require aggregations to provincial scales for analysis. Detailed documentation and metadata including nominal resolution and sources of input data are provided. These data are freely accessible and downloadable at: http://sedac.ciesin.columbia.edu/data/collection/gpw-v4 .

The **Global Rural Urban Mapping Project, v1 (GRUMP)** data collection builds on GPW with the explicit aim to capture urban locations and populations and to distinguish those from surrounding rural areas. The collection, also in WGS84, consists of global data sets that indicate urban settlement points and grids of urban extents, as well as population count and density grids that are lightly modeled taking the urban location information into account (Balk et al. 2005, Balk 2009). Using the stable-city lights data from the National Oceanic and Atmospheric Administration (Elvidge et al., 1997), GRUMP was the first global database to render urban areas spatially and connect those locations with estimates of population. At a nominal spatial resolution of 1 km, and normalized to the years 2000, 1995, and 1990, these data are freely available at https://sedac.ciesin.columbia.edu/data/collection/grump-v1. Although newer night time light time-series data are now available (e.g., Elvidge et al. 2017), for a variety of reasons, updates to this exact data product are not presently expected. This is partly due to the fact that the time-series does not extend as far back as other possible settlement input layers, and that more recent night-lights can be better put to use as an independent proxy for economic activity rather than urban location.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

The **Global Human Settlement Population Grid (GHS-POP)** depicts the distribution and density of the total population as the number of people per grid cell (250m spatial resolution) in World Mollweide equal-area projection (EC 2015). Residential population estimates (counts) per smallest census units available, used also by CIESIN GPWv4 for the years of interest, are disaggregated to grid cells, directly (linearly) proportional to the ratio of built-up areas within a cell to the total cell surface (Freire et al. 2016, 2018). Global mapping of built-up areas was performed through the Global Human Settlement Layer (GHSL) project using Landsat imagery collections for nominal epochs 1975, 1990, 2000 and 2014 (Pesaresi et al. 2013, 2016a, 2016b). The GHSL approach is grounded on the concept that buildings and their agglomerations (i.e., settlements) are nowadays the main visible and direct manifestation of human presence (and activity) on the Earth's surface. GHS-POP aims to constitute a detailed and consistent time series of population distributions that is based on reproducible methods for sustainable data production (Melchiorri et al. 2019) and can be used in policy support in numerous domains (Ehrlich et al. 2018b). These grids are created using open and free input data and are also freely accessible and downloadable at: https://ghslsys.jrc.ec.europa.eu/ghs_pop.php.

Oak Ridge National Laboratory's **LandScan Global** represents an ambient (average day/night) population distribution in a 30 arc-second (~1 km) resolution grid using the World Geodetic System (WGS84) for spatial reference (Dobson et al. 2000). LandScan uses census and other geographic data, as well as remote sensing imagery in a multivariate dasymetric modeling framework to disaggregate census counts within administrative boundaries (Dobson et al. 2003, Bhaduri et al. 2002). The final product displays a combination of locally adaptive models tailored to match input geographies and different environmental conditions in countries and regions. The modeling approach, defined as "smart interpolation," uses subnational level census counts for each country and primary geospatial input or ancillary datasets, including land cover, roads, slope, urban areas, village locations, and high resolution image classifications; all of which are key indicators of population distributions. Based upon the spatial data and the socioeconomic and cultural understanding of an area, cells are preferentially weighted for the possible occurrence of population during the course of a day. Within each country, the population distribution model calculates a "likelihood" coefficient for each cell and applies the coefficients to the census counts, which are employed as control totals for respective areas. The total population for that area is then allocated to each cell proportionally to the calculated population coefficient to compute counts of ambient or average day/night population. LandScan Global is available for download free of charge to the educational community at https://landscan.ornl.gov/.

Esri's **World Population Estimate (WPE)** represents another global dasymetric re-distribution of human population. Initiated in 2014, WPE is produced at the Environmental Systems Research Institute (ESRI) by apportioning population data enumerated within the most detailed census data available for each country to raster cells using a raster model representing the footprint of human settlement (Frye et al. 2018). The footprint of human settlement is produced using various ancillary data layers. First, base scores are derived through the combination of a 30-meter resolution global classified land cover dataset (MacDonald Dettwiler and Associates (MDA) 2017), road intersection points (HERE 2019, OpenStreetMap Foundation (OSMF) 2015), and populated place points from GeoNames (GeoNames 2013). The base scores are augmented with texture scores derived from 15-meter resolution Landsat 8 panchromatic images using a rugosity (i.e., terrain roughness) model (Jenness 2004). The base scores are used to allocate population to WPE cells to create gridded representations of estimates of population counts, population density (number of persons per square kilometer), the likelihood of settlement, as well as confidence scores. The population counts and density grids have a spatial resolution of 150 meters and are referenced through the WGS84 geographic coordinate system (Frye et al. 2018). WPE is the only commercial product described, available through https://www.arcgis.com/home/item.html?id=92d3005feb84428a8f85160f2451ec63.

The **WorldPop program** produces a variety of demographic gridded data products at the global and country scales (Tatem et al. 2017). Initiated in October 2013, the WorldPop project replaces and merges the regional AfriPop (Linard et al. 2012),

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

AsiaPop (Gaughan et al. 2013) and AmeriPop (Sorichetta et al. 2015) population mapping projects. The main method for producing WorldPop products is a weighted dasymetric approach that relies on a random forest model (Breiman, 2001) to produce a predictive weighting layer for dasymetrically redistributing population counts into 3 arc-seconds grid cells (~100m at the equator) in the Geographic projection WGS84 (Stevens et al. 2015). Individual country outputs from the WorldPop project provide an open access, transparently documented archive of spatial demographic datasets for many regions in the world including Central and South America, Africa and Asia to support development, disaster response and health applications (Gaughan et al. 2013, Stevens et al. 2015, Sorichetta et al. 2016). In addition, the WorldPop project produces a standardized, temporally and spatially consistent set of gridded products at the global scale. These freely available datasets include the input population data and covariates used in model prediction, annual gridded population count datasets structured by 36 age/sex classes from 2000 to 2020, and grid cell area estimates that can be used to derive gridded population density datasets (Lloyd et al., 2017). All data can be downloaded from www.worldpop.org.

The **History Database of the Global Environment (HYDE)** is an internally consistent combination of historical population estimates and allocation algorithms with time-dependent weight maps for land use (Klein Goldewijk et al. 2010, 2011 and 2017). Population is represented by maps of total, urban and rural population, population density and built-up area at a spatial resolution of 5 min longitude/latitude. HYDE covers the time period from 10,000 before Common Era (BCE) to 2015 Common Era (CE). For the period after 1950, the underlying input data is based on 1950-2015 population estimates from the United Nations World Population Prospects (2008 Revision). All data can be downloaded from https://doi.org/10.17026/dans-25g-gez3.

3.2 National and regional/continental population data production efforts

It is imperative for a review of existing global population data products to also reference production efforts at national, regional or continental scales that often make use of more detailed input data but are based on similar methodological frameworks. Such country- and regional-level products are often produced for specific purposes, which may influence the decision rules applied for their creation. Often these data products are based on more up-to-date ancillary and input population data and thus may provide pointers for future global population data creation once those ancillary data could become available worldwide.

For example, Facebook Connectivity Lab and CIESIN's **High Resolution Settlement Layer (HRSL)** provides estimates of human population distribution at a resolution of 1 arc-second (approximately 30 m) for the year 2015. Machine learning techniques are used to identify potentially populated areas (settlement) using very high resolution satellite imagery. Proportional allocation is then applied to redistribute population from recent census data onto grid cells identified as settlement extent (Tiecke 2016, Tiecke et al. 2017). This data production effort was driven mostly by Facebook's interest in locating people in remote areas of developing countries such as Burkina Faso, Ghana, Haiti and Sri Lanka who may be in need of internet access and is available from: https://www.ciesin.columbia.edu/data/hrsl/.

Developed by the European Commission for the purpose of producing the most detailed population grid for policy analysis and support, **the European Global Human Settlement (GHS) population grid** represents the distribution and density of total residential population, expressed as the number of people per grid cell (100 m spatial resolution) in equal-area projection (LAEA ETRS89) for 43 countries and territories in 2011. Intelligent dasymetric mapping (Mennis and Hultgren 2006) was employed in order to disaggregate best-available census data for each country (vector grids or census tracts) to built-up areas as mapped by the European Settlement Map 2016 (Ferri et al. 2014, Florczyk et al. 2016), and weighted by enhanced land use/cover data from a refined Corine Land Cover map where available (Freire and Halkia 2014). For eight countries, population grids were originally modeled at 10m spatial resolution and then aggregated to 100 m grid cells. This data product is freely accessible and downloadable at: http://data.jrc.ec.europa.eu/dataset/jrc-ghsl-ghs_pop_eurostat_europe_r2016a.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

The U.S. Census Bureau has invested in efforts to provide data on population patterns by administrative areas and grid cells for various regions with a focus on improving the availability of detailed population maps (1998-present) in regions likely in need of humanitarian relief and disaster assistance from external partners (U.S. Census Bureau 2018). Data inputs include census data from every country and territory that conducts a census, demographic surveys, maps of administrative boundaries from national and international mapping agencies, high- and medium-resolution satellite imagery, and a range of ancillary layers such as land cover, road networks, and elevation. The U.S. Census Bureau has developed gridded **Demobase** population maps at 100m resolution for selected countries including Haiti, Pakistan, and Rwanda (e.g., Azar et al. 2010, Azar et al. 2013), as well as maps of subnational population by age and sex within administrative areas. Both Demobase gridded data and administrative-area based subnational datasets are freely accessible and downloadable via links at:
https://www.census.gov/programs-surveys/international-programs/about/global-mapping.html.

3.3 Data availability

The above described data products and their characteristics including the underlying population concept, method, resolution, points in time, the source for national-level population statistics used as well as reference links to access the data can be found in Table 1. All of these population grids are open access except two that have some restrictions. The different data producers host the data in different ways, typically using internal servers and data repositories. Summaries and links to the various data repositories can be found at www.popgrid.org, facilitating access to, documentation and comparison of different data products. As mentioned before, the user can also find visual comparison tools (outputs as tables and through map services) that provide effective ways to perform visual analytics and identify differences in patterns of population distributions exhibited by the different data products.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Table 1. Detailed characteristics of the datasets described in this review. See also https://www.popgrid.org/popgrid_files/popgrid-data-comparison-tables_0.pdf .

Dataset	Source	Population concept	Method	Spatial Resolution	Year(s) Represented	National level population totals	Distribution Policy	Reference Link
GLOBAL POPULATION GRIDS								
Gridded Population of the World (GPWv4.11)	Center for International Earth Science Information Network (CIESIN), Columbia University	De jure / de facto	GPW1: pycnophylactic; GPW2,3,4: areal weighting	1 km (v4)	2000; 2005; 2010; 2015; 2020	2 versions: 1) official country census totals; 2) country totals adjusted to United Nations Population Division (UNPD) estimates and projections	Open access	http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
Global Rural Urban Mapping Project (GRUMPv1)	CIESIN, Columbia University; International Food Policy Research Institute, The World Bank, Centro Internacional de Agricultural Tropical	De jure / de facto	Dasymetric	1 km	1990; 1995; 2000	UNPD estimates and projections	Open access	http://sedac.ciesin.columbia.edu/data/collection/grump-v1
LandScan Global Population Database (Landscan Global)	Oak Ridge National Laboratory (ORNL)	Ambient (day-time)	Smart interpolation	30 arc-seconds	annual releases 2000–2016	US Census Bureau	Open for research; commercial use at cost	https://landscan.ornl.gov/
WorldPop	WorldPop, University of Southampton	De jure / de facto	Statistical / dasymetric	100 m	2000-2020	2 versions: 1) Country-official estimates, and 2) UNPD estimates and projections	Open access	www.worldpop.org
Global Human Settlement Layer - Population (GHS-POP)	European Commission Joint Research Centre (JRC) and CIESIN, Columbia University	De jure / de facto	Dasymetric refinement, proportional to built-up density	250 m	1975; 1990; 2000; 2015	UNPD estimates and projections	Open access	http://ghsl.jrc.ec.europa.eu/ghs_pop.php
World Population Estimate (WPE)	Environmental Systems Research Institute (ESRI)	Combined (de jure, defacto, estimates)	Dasymetric Redistribution (Smart)	250 m 250 m 150 m	2013, 2015, and 2016	Country-official estimates with 134 countries processed further by M. Bauer Research GmbH	Free to ArcGIS Users	https://sites.google.com/ciesin.columbia.edu/popgrid/find-data/esri
History Database of the Global Environment (HYDE) Population Grids v3.2	Netherlands Environmental Assessment Agency (PBL)	De jure / de facto population	Dasymetric mapping using historical population, cropland, pasture data, satellite data	5 arc-min (ca. 10km)	10,000 BC - 2015	United Nations World Population Prospects (2008 Revision) after 1950	Open access	https://themasites.pbl.nl/tridion/en/themasites/hyde/download/index-2.html
REGIONAL/ CONTINENTAL POPULATION GRIDS								
High Resolution Settlement Layer (HRSL)	Facebook Connectivity Lab and CIESIN	De jure / de facto population	Binary Dasymetric	30 m (1 arc-second)	2015	Country-official estimates of more than 30 countries	Open access	https://ciesin.columbia.edu/data/hrsl/
European GHS Population Grid (GHS-POP-EUROSTAT)	European Commission Joint Research Centre (JRC)	De jure / de facto population	Intelligent dasymetric mapping	100 m	2011	Country-official estimates	Open access	http://data.jrc.ec.europa.eu/dataset/jrc-ghsl-ghs_pop_eurostat_europe_r2016a
Gridded Population Mapping (Demobase)	U.S. Census Bureau	de jure population	Statistical / dasymetric	100 m	Depends on country (1998-present)	U.S. Census Bureau International Data Base and national censuses	Open access	https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/demobase.html

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

4. Putting people in places: Key methods and ancillary data

4.1 Ancillary data

The products included in this comparative review are the outcomes of different data integration approaches to produce gridded population distribution datasets based on different techniques of refinement, zonal statistics, reallocation or inter- and extrapolation. Different ancillary data have been used in slightly different ways to create different population models. All ancillary data have in common that they exhibit some kind of relationship to population that can be exploited in population redistribution models to increase the accuracy of population estimates. These relationships may be of correlative nature, based on empiric rules or even binary. While the literature on population modeling and dasymetric mapping has described a variety of such ancillary variables, the data that can be used in national, regional and global population grid production has to be available consistently for large extents, and for different points in time, thus limiting the choices for researchers and data producers. One important class of ancillary data is that of urban land use area or human settlements detections. Figure 1 provides an overview of this type of ancillary data available at different points in time including satellite images (Landsat, MODIS), land cover products, settlement layers (e.g., GHSL or the Global Urban Footprint (GUF+)) in relation to commonly available census data. This overview highlights apparent temporal offsets between input population data and some ancillary data but also high temporal resolution of satellite data that provides the basis for the derivation of abundant ancillary variables.

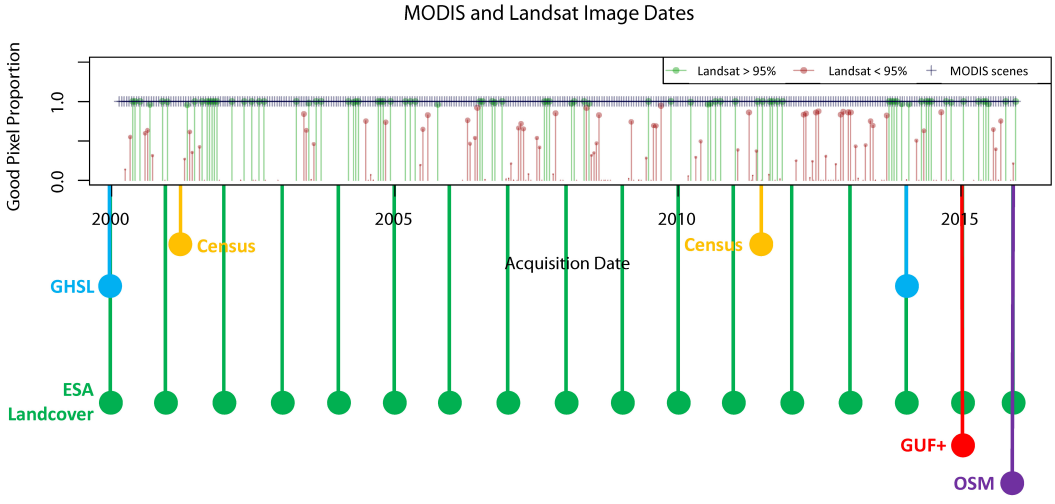

Figure 1. Identification of different ancillary data that inform spatial and temporal interpolation approaches to create gridded population data across scales of interest. Temporal fidelity in the Landsat (30m resolution; with varying proportions of cloud-free area, labeled as Good Pixel Proportion) and MODIS (250m resolution) sensors are shown in relation to typical points in time for censuses alongside several derived ancillary data products such as the European Space Agency (ESA) annual land cover data (300m resolution), and the Global Human Settlement Layer (38m resolution) at various publication dates. Also noted are OpenStreetMap data, potentially useful for more contemporary time periods, representing vector-based information that is increasingly explored as a possible ancillary data source.

Table 2 summarizes the input variables, including these land-use type and other ancillary data, used to create the different products (also available at https://www.popgrid.org/compare-data); as described earlier, Table 1 provides additional information on the modeling methods used.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Table 2. Summary of input variables used in modeling gridded population, globally.

Gridded Population Dataset	Population	Ancillary Data Layers								
		Roads	Land Cover	Built structures	Cities or Urban areas	Night-time lights	Infrastructure	Environmental data[b]	Protected areas[a]	Water bodies
GPW	x								[a]	x
GRUMP	x				x	x			[a]	x
LandScan	x	x	x	x	x		x	x	x	x
GHS-POP	x			x						
WPE	x	x	x		x					x
WorldPop	x	x	x	x	x	x	x	x	x	x
HYDE 1950-	x							x		x

[a] *Protected areas were not masked out, but national statistical offices often assign no data or 0 (zero) to protected areas;*

[b] *climate, topography, elevation*

4.2 Methods for population redistribution

Understanding the fundamentals of the different data integration approaches is an important aspect in evaluating the fitness of any given dataset for specific uses or cases. The process of gridded population mapping requires reallocation of spatial data from "source" units into "target" units, which can result in disaggregation or aggregation depending on the spatial detail and resolution of the input and output data, and can be done through different approaches including various forms of areal interpolation and statistical modeling.

Areal weighting techniques (the simplest form of areal interpolation, also known as proportional reallocation) evenly re-distribute source data into target grid cells based on proportions of overlap with no ancillary data input informing the process (Goodchild and Lam 1980, Mennis and Hultgren 2006) (Figure 2). The source input data may be census-based or other administrative data and the target grid cell represents a spatial unit which could be generally larger or smaller than the source units. An assumption associated with this approach is that there is uniform redistribution from the source units to target cells that overlap with the source units. This assumption is a gross simplification as population distributions are not uniform, but the approach is computationally efficient and simple in creating spatially-explicit and globally consistent population estimates. Such products are well suited for informing policy-making efforts that do not require fine spatial resolution (Doxsey-Whitfield et al. 2015), or for performing correlation analyses in which endogeneity issues are excluded (e.g., Cohen & Small 1998). An example of this approach would be the Gridded Population of the World (GPWv4).

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Figure 2. Schematic illustration of different types of techniques for population redistribution or allocation from source to target grid cells.

When ancillary data informs the redistribution from source area to target cell, the technique is referred to as **dasymetric mapping** (Semenov-Tian-Shansky 1928, Wright 1936, Eicher and Brewer 2001, Mennis 2003, Mennis and Hultgren 2006). The ancillary variables, often produced and available at finer spatial detail than the input population data, can be used to develop weighting schemes for reallocating population from the source area to target units depending on existing or assumed relationships between the two. Ancillary variables can include land cover, topography, land-use zones, street networks, remote sensing data and more (for details and more examples see Zandbergen and Ignizio 2010, Nieves et al. 2017; for an overview see Mennis 2009). For example, redistributing population from a source area (e.g., a census tract) that includes built or developed parts along with forest and agricultural land uses will more heavily weight the built area in redistributing population counts because it is more likely that these areas are populated (Mennis and Hultgren 2006, Bhaduri et al. 2014). Dasymetric approaches vary in the allocation method applied, ranging from binary dasymetric (Figure 2, Eicher and Brewer 2001) to 'intelligent' dasymetric mapping (Mennis and Hultgren 2006) to statistical approaches (Leyk et al. 2013, Nagle et al. 2014). However, these methods have in common that they rely on existing relationships between population (e.g., provided by the input census data) and ancillary information (e.g., land cover) that can be exploited to redistribute population to different geometric units with higher accuracy.

Several **statistical modeling** approaches have been described in the literature that blur the line and can be viewed as another means of population estimation, traditionally focusing on the problem of small area estimation (e.g., Birkin and Clarke 1988, Wong 1992, Bogaert 2002) or as a type of dasymetric refinement (e.g., Mrozinski and Cromley 1999, Nagle et al. 2014) (Figure 2). Statistical estimations rely on correlating population counts through regression (Mennis 2009) and can be based on various types of predictive variables, derived from ancillary data layers such as density or length of streets (Reibel and Bufalino 2005), or remotely-sensed data to inform the estimation process (Harvey 2002, Wu et al. 2005).

More recently, various **hybrid approaches** that often rely on machine learning techniques or ensemble prediction, and dasymetric redistribution, have shown promising results. In a hybrid approach, first, a statistical model estimates a population

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

density layer. The estimated population densities provide a weighting that is then constrained by a dasymetric process in order to redistribute total population counts aggregated within source units to target cells. Such weighted dasymetric approaches have shown promising results when compared to other techniques for producing gridded population maps (Stevens et al. 2015, Sorichetta et al. 2015, Reed et al. 2018).

Figure 3 illustrates, using a region in Kenya, how different ancillary data layers, typically used for population redistribution including roads, land cover, protected areas and topography (Figure 3b-e) affect the resulting population distribution (Figure 3f). Different methods described above will employ these variables in different ways and operate under varying assumptions, and thus result in different estimates. Thus, there are expected relationships and trends that can be observed for most population grids. For example, low road density of roads, rough topography and high elevations, the presence of protected area and non-urban land cover are commonly related to low population densities. However, Figure 4 illustrates remarkable differences between the population distributions of the data products described in this review for a larger area in Kenya, highlighting the importance of informing the user about critical aspects and characteristics of the different data layers. Note that in panel A population counts (not density) are rendered per irregularly shaped level-5 census unit. In panels C-H, population is rendered per grid cell. Note that the grid cell size is not the same across the panels and is specific to each data product. For each panel, however, the grid cells have the same extent and can be interpreted as population densities.

It is important to acknowledge error accompanying the estimation results from such redistribution approaches. This includes uncertainty associated with the original census, the areal aggregation of both the input census data and the ancillary data products (Wu et al. 2005), and the model used to estimate statistical relationships (Nagle et al. 2014, Sinha et al. 2019). Recent research has increasingly stressed the complexity of uncertainty in such applications as well as the difficulty to carry out validation due to the lack of reference data (Mennis and Hultgren 2006, Zandbergen and Ignizio 2010). Therefore, error assessments tend to appear mostly in studies in data-rich settings.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

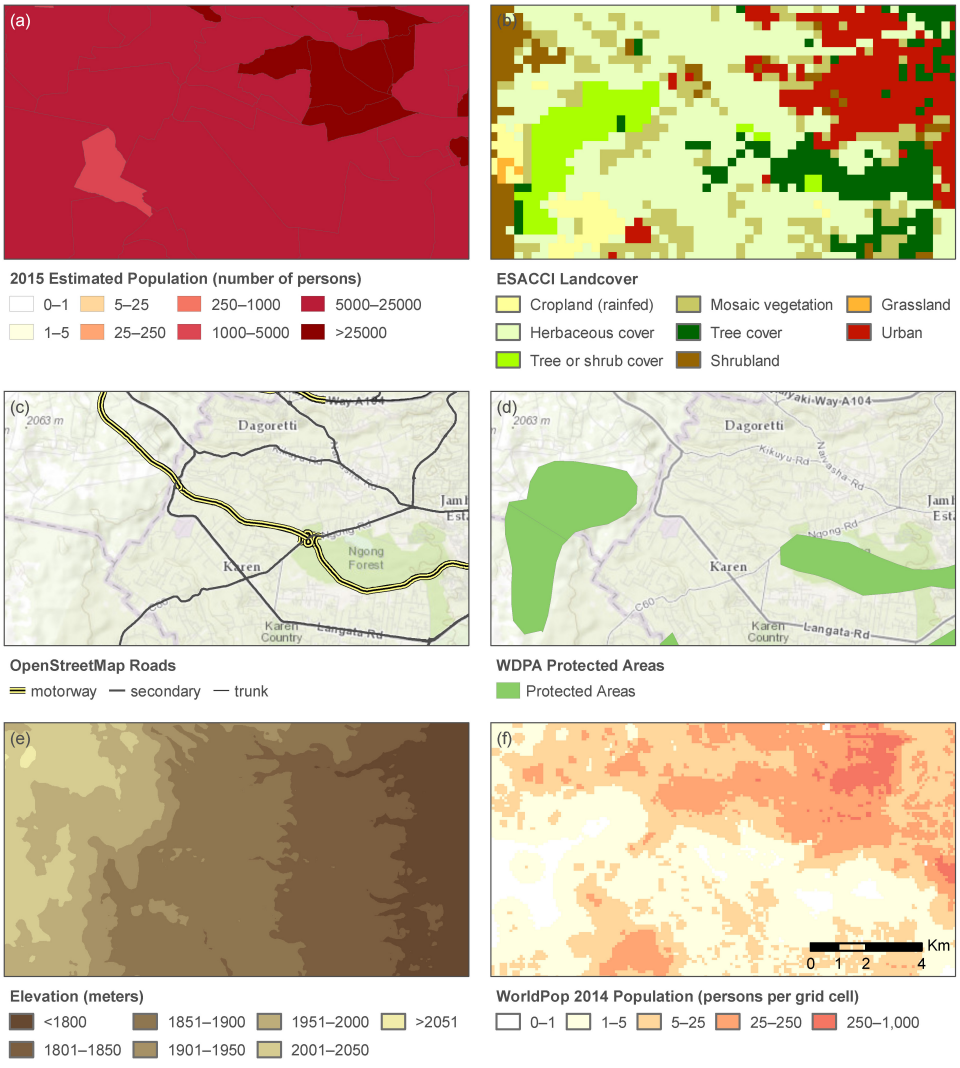

Figure 3. A schematic illustration of refinement effects of ancillary data layers on census population data to create gridded population grids at fine spatial resolution for a small study area near Nairobi, Kenya: (a) Kenya National Bureau of Statistics, Population and Housing Census 2009, level 5 population units (Center for Development and Environment, Kenyan Atlas Project) as input, (b) European Space Agency (ESA) Climate Change Initiative (CCI) Land Cover 2015 (300 m resolution), (c) OpenStreetMap major roads, (d) World Database on Protected Areas (March 2019 Release), (e) Viewfinder Panoramas 3 Arc seconds Digital Elevation Model, (f) WorldPop 2014 Population Count (100 m resolution) as one exemplary population grid created.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

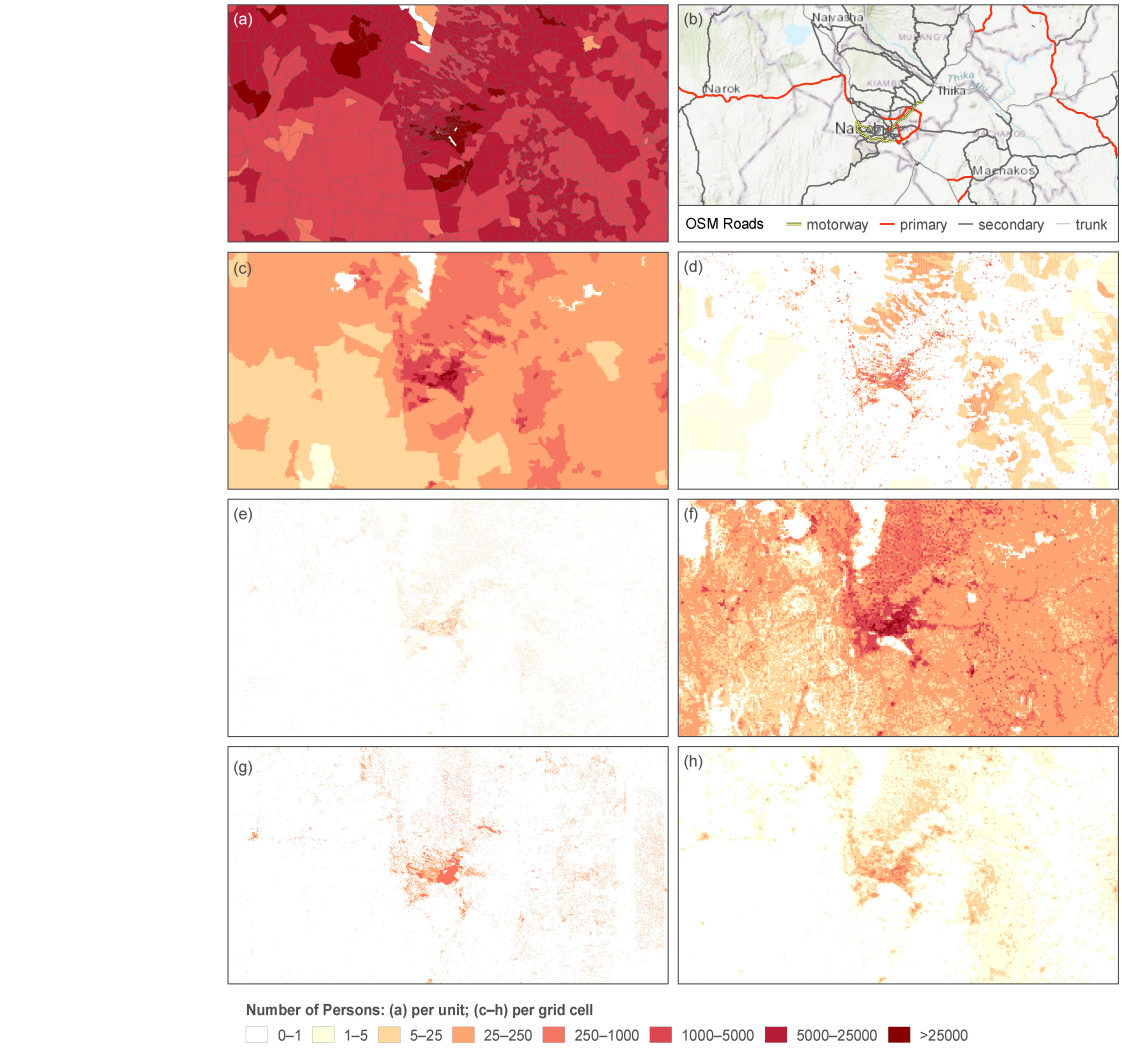

Figure 4: Illustration of population input, exemplary ancillary data and different outcome data for a larger region around Nairobi, Kenya: (a) Kenya National Bureau of Statistics, Population and Housing Census 2009, level 5 population units (Center for Development and Environment, Kenyan Atlas Project), (b) basemap with roads and topography as ancillary data, (c) Gridded Population of the World version 4 Revision 10, UN Adjusted 2015 Population Count (1 km), (d) Global Human Settlement Layer 2015 Population Count (250 m), (e) High Resolution Settlement Layer 2015 Population Count (30 m), (f) Landscan 2015 Population Count (1 km), (g) Esri World Population Estimate 2016 Population Count (150 m), (h) WorldPop 2014 Population Count (100 m).

The persistent challenges with modeling and validating gridded population datasets especially in data-poor regions has driven more recent initiatives that focus on modeling gridded population from the ground up, relying on micro-census data and geostatistical covariates in a statistical modeling framework (Wardrop et al. 2018). Such techniques, in the absence of reliable or recent census data, leverage advances in computational and statistical frameworks along with increased spatial fidelity of remotely sensed products and advances in global positioning system (GPS)-enabled field survey techniques to produce gridded population surfaces. This type of approach is considered complementary to more traditional, census enumeration-based efforts.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

5. Different populations or different data? A Fitness-for-use perspective

The process of creating gridded population products redistributes population estimates from census or administrative areas to grid cells, conditional on where human populations and settlements may be located. The nature, quality and accuracy of the input population data, the characteristics of the output gridded population dataset, the properties of the ancillary data used and the implications of the methodological approach applied for population allocation and redistribution are all important determinants of spatial data quality in general (FGDC 1998, Guptill and Morrison 2013) but also help to shed light on the relative data quality of each of the population grids described in this review. While data quality and its reporting in standardized metadata has been the focus of much research in the last decades, the discussion of relative quality or fitness for use of spatial data has received less attention (see Devillers et al. 2007, Devillers et al. 2010, Ivánová et al. 2013). Since the described population grids show fundamental differences, the question whether a data product is fit for a given purpose is of high relevance. Thus, in this section, we discuss several determinants (not an exhaustive list) that aid the data user in the assessment of the data product's fitness for use relative to the target application. We briefly discuss data-related aspects including scale, currency and semantics, as well as modeling and processing-related implications for uncertainty. We address them separately, but the reader may be reminded that all those relative quality aspects have to be understood interrelated as one can affect all others. We will also address the problem of validation of large-scale population grids.

5.1 Data aspects of relative quality

The **accuracy** of the input census/population data and ancillary data includes thematic, spatial and temporal accuracies, which contribute to the level of uncertainty of the final data product. For this reason, the user needs to consider and understand what kind of data are input to a certain data production process. For census data, the completeness of coverage, the margin of error (if sampled), the time period the census is taken and the positional accuracy of the boundaries are measures that can be used but might not be always known and the data need to be used with caution. This kind of knowledge is important to reflect when using population grids in a given region (e.g., Tatem 2014). With regard to the ancillary data, needless to say, the quality of the final population grid depends on the quality of the ancillary data used for redistributing population counts. Apart from the existence and strength of the assumed relationship between population and ancillary variable (Nieves et al. 2017), the accuracy of these spatial layers themselves is critical for the accuracy of gridded population estimates. For example, the classification accuracy of built-up or developed land layers that are used to redistribute census counts to different regions tends to be lower in rural than in urban settings (Wickham et al. 2013; Leyk et al. 2014 and 2018), but can also vary across larger regions and countries. The quality of remotely-sensed ancillary data also depends heavily on the characteristics of the instrument (optical daytime, optical night-time, or radar) and the processing algorithm (e.g., Small et al. 2005, Potere et al. 2009, Pesaresi et al. 2016b, Esch et al. 2017). Such differences propagate through to strongly influence the accuracy of the final population data product and may cause over- or underestimations in different subregions. Knowledge of such issues would be critical for the data user if population estimates in different regions are compared with each other. Due to the nature of the input and ancillary data, these accuracies translate into aspects of scale, currency and semantics critical for evaluating the fitness for use of the final population grids as discussed below.

Scale: Since input data are typically enumerated counts, issues due to spatial aggregation including the MAUP (Openshaw 1983), as the geographical manifestation of the ecological fallacy (Piantadosi et al. 1988, Waller and Gotway, 2004), are one of the main sources of the "unknown." Differences in granularity of the input (census) data across different regions or countries must be taken into account since the same population redistribution model may perform very differently under different circumstances due to possible scale effects. In using the final population grids, the grid cell, defined by the spatial resolution (aka cell size), would often be assumed to define the analytical scale (Montello 2001, Cao and Lam 1997). The user would often attempt to model a certain process or phenomenon of interest but often there is a mismatch between this 'operational' scale (e.g., Montello 2001, Maclaurin et al. 2015) and the analytical scale. However, it is imperative for the user to understand

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

that due to the difference between input population data (i.e., source unit) and output grid cell (i.e., target unit) granularity this assumption may be fundamentally flawed and result in either oversampling or generalization. For example, if an analysis is intended to be conducted at the neighborhood scale, population estimates provided in grid cells of 150m or 250m appear to represent meaningful target units. If these input data were at the census block or tract level the grid cell size would represent an appropriate proxy and can be used as a valid analytical unit at the intended target scale. If, however, the input data originated from large administrative units (e.g., districts or county level source units) there would be a significant offset between input and output. In such cases, the user would face a higher risk of using oversampled population estimates that might result in higher degrees of local inaccuracy. To complicate matters, if ancillary data are used to redistribute population (e.g., to built-up portions of the source unit) based on existing relationships, such scale-related problems may be mitigated to some degree. In addition, variation in how a model is trained or the units selected to build the estimation model will influence the final gridded distribution (Sinha et al. 2019). For example, if census data from one region or country is very coarse, a model built based on finer-resolution data from a neighbouring region and then applied to the region of interest can be more accurate (Gaughan et al. 2015). Thus, scale effects are inherent to each of the described population grids at different degrees, and represents a geographically varying characteristic depending on the granularity of the input data, the strength of the associations between population and ancillary data, and the resolution of the output data. These effects need to be interpreted in the context of the target scale of the intended analysis.

The **currency** of the data represents another important issue. In a few instances, underlying census data are old (e.g., in Haiti) or the period between censuses is more than 10 years. While some of the ancillary data are more or less constant over the near term (e.g., water bodies and permanent ice), there may also be *temporal mismatches* between population data and any of the intrinsically time-varying ancillary data layers. For example, it may be unknown whether a given built-up land grid cell has been developed at the time the census has been taken. Such temporal offsets may be critical if the assumption for the intended application necessitates a high degree of temporal agreement (currency). This form of uncertainty is difficult to handle and can be further complicated by differences across regions and countries. In response to this, few efforts (e.g., WorldPop and GHS-POP) ensure the use of temporally implicit or invariant ancillary data in the modeling process (Gaughan et al. 2016). However, even under those conditions, there might still be underlying issues for projecting forward/backward from census data for a target year of interest. The user is well advised to understand the gridded population estimates as approximations over a period of time and avoid flawed assumptions of high currency in a given analysis.

Semantics: As mentioned before, what the population modeled represents can be very different among data products. This meaning can even be different within one product if, e.g., the census input data account for different population concepts or population groups in different regions or countries. For example, the population estimate might refer to *de-jure* (or legal, typically closer to night time) populations vs. *de-facto* (or present, closer to daytime) populations and using the one over the other product would possibly result in dramatically different results. The user has to be aware that data on resident populations as provided by censuses is itself a convention, whose distribution never occurs at any moment in time (*de jure* census population) or if it does occur (*de facto*, location at the time of the census) that distribution may not be representative of a different situation or in medium/long term (i.e., a year): the concept of usual residence. Most of the global population data products use a night-time / usual residence (*de jure*) concept, or mostly rely on underlying data that use a *de jure* concept, with LandScan being the notable exception. Thus, the user is well advised to get informed about the meaning of the populations modeled in the population grid in question to avoid such misinterpretations, as indicated in Table 1. The aspects of scale mismatch described above can further add to semantic differences since due to such aggregation effects, different populations may be modeled. Thus, these implications have to be understood by the user, spatially and semantically, and caution is advised when interpreting analytical results.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

5.2 Processing- and model-related implications of uncertainty

Regardless of the approach of choice employed for data production, all efforts described in this review do carry out some form of data conversion (e.g., vector-to-raster) and data integration (re-allocation or resampling). Any such, data processing step will propagate uncertainty in some way and have consequences for the quality of the outcome data and the subsequent analyses, depending on the input data quality as described above. For example, if large census units (e.g., counties or districts) in vector format are converted to grid cells (rasterization) of fine spatial resolution (e.g., 150m), while there is a clear scale effect to be addressed (see above), the resulting population estimates may differ dramatically for different redistribution models applied that may or may not use ancillary data. The data user needs to be aware that existing uncertainty is not eliminated by applying certain models or integrating different data sources. However, through the process of data integration we may be able to improve the accuracy based on spatial refinement strategies such as dasymetric modeling (Mennis 2009).

GPW, GHS-POP and HYDE do not employ statistical methods to produce their grids, and thus traditional metrics of uncertainty are not available. Because fine resolution inputs reduce errors of aggregation, GPW reports the number of input units per country used in the gridding process. Nevertheless, errors may persist in countries with highly variable input units. For example, Sahelian countries that have finely resolved units for densely populated areas but very coarse units for sparsely populated regions.

The specific model applied to re-allocate population counts and densities, which can be empirical or statistical, will always have some error. When the modelling process is statistical or hybrid such as in the case of the WorldPop, Landscan and WPE, estimates of such errors (e.g., standard error of regression coefficients, prediction error) can be derived as a by-product of the modeling process. To fully understand the quality of a population grid, the error of the applied model needs to be evaluated. Highly accurate ancillary data are not useful if the relationship to population is weak or the model applied inappropriate, and thus the model predictions are unreliable (e.g., low R squared, or deviance explained). Such prediction errors are often assessed in comparison to alternative models but are hard to quantify in the absence of validation data. To complicate matters, the same model might perform very differently in different geographies or under different environmental conditions, an effect known as spatial non-stationarity or spatial variation of the target relationships (Fotheringham et al. 1996). Such variations will further affect the model predictions if left unaccounted.

5.3 Validation challenges

Validation of population data has always been a challenge, simply because validation data at fine resolutions are rarely available and even when available, may exist at different time periods or confidentiality rules may limit their use in order to not expose individual and household level information. Access to such confidential data is only possible with special permission or sworn status and even then, often the demographic data are only a sample of the whole population. These challenges can be very different between countries and thus a validation that may be possible in one country does not necessarily translate to another location. While a true validation of the gridded output remains a challenge, it is possible to internally test the accuracy of the modeling approaches (Gaughan et al. 2015, Sorichetta et al. 2015, Reed et al. 2018). Such an assessment can be done when different levels of census input data are available for use in a model. The approach leverages the coarser level data in different modelling approaches and then compares the gridded outputs to the finer level census data to determine how well and plausible populations were distributed.

Validating ancillary data may have its own challenges. However, the existence of new, more detailed reference data in some regions (e.g., parcel data, crowdsourcing data) has helped to make progress in evaluating land cover data and built-up land layers, which is key to most of the described population grids (See et al. 2015, Leyk et al. 2018, Leyk and Uhl 2018). In general, depending on the level of land development and land use patterns in the region of interest, different products may serve the intended purpose, differently.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

6. Fit for use or not fit? Concluding remarks and future work

The different critical elements described above all have some impact on the fitness for use as a measure of relative data quality. Despite the importance of data quality, it does not receive the attention it deserves, in part because comparative measures may be difficult to conceive, derive or quantify. Furthermore, such assessment also importantly depends on the application of interest. The different aspects above have to be seen in context and considered interrelated. Different analytical and data processing steps such as conversions or data integration do not cause isolated uncertainties but through all those steps uncertainty can be propagated and thus becomes difficult to control and account for.

Whether or not data are fit for an intended use is not based on standardized measures nor is it well understood as to what the concept of fitness for use actually entails. Often it is at the discretion of the data user to decide whether the use of a given data product is appropriate or not, particularly in the age of open public data and open science. Based on the above discussion, there are a few guidelines that, in general, can help a user make informed decisions related to the fitness of a given data product for their intended use:

(1) **How important is spatial refinement of the population grid to be used?** In the last 20 years, considerable attention has been paid to the spatial refinement of gridded population estimates. Some applications such as estimation of populations at risk of seaward natural hazards benefit substantially from these improvements. Other applications such as some climate scenario modeling do not require such finely resolved data as information on the general spatial distribution of population at moderate resolution would be sufficient.

(2) **Which data product is appropriate and effective for urban population analysis?** Closely related to the above concern, if the aim of the analysis is to examine urban population distribution as opposed to rural population, one would be better off using a data set for which information on human settlements or urban extents has been used in the modeling. Urban land tends to be concentrated and can be clearly distinguished from the surrounding areas in remote sensing images and thus settlement data products (or other measures of urban extent) are effective in spatially refining population data along an urban gradient that will most likely improve the spatial precision of resulting estimates (though there is always the possibility that they will over concentrate population in built-up locations). In contrast, data products that do not include such refinements tend to underestimate urban population. Data with extremely high resolution may mitigate such effects even if no settlement data were involved in the data production.

(3) **What is the target population for the question at hand?** Questions aimed at understanding long-term population change are likely to be well served by the use of population grids that represent night-time (de-jure), residential population. In other instances, however, for example on emergency response, one may need to know where populations are during the day-time or rely on an ambient population concept and thus would be better served by data products that incorporate that concept in the modeling process.

(4) **Is the population grid being used to model other outcomes?** If so, and if that outcome may be one of the ancillary variables (or one closely linked to it) used in the production of the population grid, one needs to select a population grid that is not endogenous to the question at hand. For example, if the goal of the analysis is to estimate changes in built-up area using population as one of the explanatory variables, then one must not use GHS-POP (which uses built-up area). This necessitates that users become familiar with the ancillary data used in the production of the population grids, and even those used, perhaps, in training data sets but not actually the modelling.

(5) **Analysing change over time?** If one's goal is to examine change over time in population distributions, one of the data products representing multiple years is most suitable. However, there may be differences in how different grids have been generated for specific years. In order to analyze change in population distributions, ideally a data product built from data layers representing the respective time period, would be preferred. Data which represents only one time period and applies, say, national-level growth rates to derive data distributions for earlier time periods, would be less amenable.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

(6) **How have these data sets been used previously?** Some of data providers make available citation lists of publication from the providers' team or the broader user community that may provide some guidance for the novice user. For example, GPW, WorldPop, LandScan, and others provide such lists and are extremely useful as a collection of common applications in which those data have been used. Based on such lists, the user can explore whether prior use of the data products appears to be appropriate with regard to target applications and how these scenarios compare to their own study.

This review is an attempt to shed light on underlying data considerations to raise the awareness of relative data quality concerns related to the described population data products. The data user community is encouraged to consider the described quality aspects and metadata carefully, before making decisions about any given data product's fitness for the target application. This can include the full assessment of the above aspects, the use of metadata as well as sensitivity analysis including running an analysis at different spatial resolutions or the comparison of analytical results using different population grids (see e.g., Mondal and Tatem 2012; Tatem et al. 2011) to understand and quantify the sensitivity of the study results.

There has been significant progress in the spatial rendering of population and related characteristics in the past 20 years, but persistent challenges remain. We depend on existing population grids that are created using ancillary data to provide hints for where people live or spend time. In an ideal world, the research community would also have access to detailed building footprint and height data for all structures, and know whether these structures are residential or commercial, if indeed they are occupied at all to pair with population data. Future work will help to close these gaps by employing new high-resolution satellite technology as well as more reliable population surveys. This includes new and improved nighttime lights products (e.g. Visible Infrared Imaging Radiometer Suite (VIIRS) with respect to DMSP-OLS Nighttime Lights), that have been already successfully tested in urban mapping applications (Elvidge et al. 2017), and settlement data production (GHSL, DLR WFS, Digital Globe, etc.) to further refine the available population grids.

Members of the POPGRID data collaborative are investing work in a number of emerging areas, including future population projections (Jones and O'Neill 2016), population projections incorporating climate change impacts (Rigaud et al. 2018), near-real-time population modeling (Bharti et al. 2015), mobility mapping, population dynamics (Deville et al. 2014), increased temporal resolution (Batista et al. 2018) and working directly with national statistical offices to improve the spatial accuracy of census products (www.grid3.org). These efforts often make use of novel data streams such as cell phone call detail records or social media data, or best practices in data collection using mobile devices. Finally, future work will be dedicated to improving the accuracy of population estimates, particularly in rural regions, where the reliability of existing data products is limited to date.

Author contribution

SL, AG, SA, AdS, DB, SF, AT, MP, KM and ML conceptualized this review article and formulated its vision. SL, AG, SA, DB and AdS structured the manuscript and developed contents of the various sections. SL, in collaboration with AG, SA and AdS, drafted the manuscript, with contributions from all co-authors. DB, AR, FRS, BB, CF, JC, AS, SF and KM wrote and revised data product descriptions and provided insights on data product characteristics and underlying procedures. All authors have read and revised the manuscript. AG and LP created the figures.

Competing interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Acknowledgements

POPGRID has been supported by seed funding from the Columbia University Earth Institute's Cross-Cutting Initiatives and the Bill & Melinda Gates Foundation. It is an element of the Group on Earth Observations (GEO) Human Planet Initiative (HPI) and is exploring linkages with key sustainable development data organizations and networks. SL is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD066613. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SL and DB are supported by the US National Science Foundation award #1416860 to the City University of New York, the Population Council, the National Center for Atmospheric Research and the University of Colorado at Boulder. AEG, FRS, AS and AJT are supported by the funding from the Bill & Melinda Gates Foundation Investment ID OPP1134076, and SA and AD by Gates Foundation Investment ID OPP1177328. AEG and FRS are also supported through the NASA Land Cover and Land Use Change Program and the NASA GEO-Human Planet Program.

References

Agumya, A. and Hunter, G. F.: A Risk-Based Approach to Assessing the 'Fitness for Use' of Spatial Data, URISA JOURNAL, 11, 33-44, 1999.

Amrhein, C. G.: Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations, ENVIRON PLANN A, 27, 105-119, doi: 10.1068/a270105, 1995.

Arbia, G. and Petrarca, F.: Effects of MAUP on spatial econometric models, LETTERS IN SPATIAL AND RESOURCE SCIENCES, 4, 173, doi: 10.1007/s12076-011-0065-9, 2011.

Aubrecht, C., Gunasekera, R., Ungar, J., and Ishizawa, O.: Consistent yet adaptive global geospatial identification of urban–rural patterns: The iURBAN model, REMOTE SENS ENVIRON, 187, 230-240, doi: 10.1016/j.rse.2016.10.031, 2016.

Azar, D., Engstrom, R., Graesser, J., and Comenetz, J.: Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data, REMOTE SENS ENVIRON, 130, 219-232, doi: https://doi.org/10.1016/j.rse.2012.11.022, 2013.

Azar, D., Graesser, J., Engstrom, R., Comenetz, J., Leddy, R. M., Schechtman, N. G., and Andrews, T.: Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti, INT J REMOTE SENS, 31, 5635-5655, doi: 10.1080/01431161.2010.496799, 2010.

Balk,D.: More than a name: Why is Global Urban Population Mapping a GRUMPy proposition?" In P. Gamba and M. Herold, (eds.) Global Mapping of Human Settlement: Experiences, Data Sets, and Prospects, New York: Taylor and Francis, pp. 145-161, 2009.

Balk, D., M. R. Montgomery, G. McGranahan, and M. Todd, "Understanding the Impacts of Climate Change: Linking Satellite and Other Spatial Data with Population Data ," In G. Martine, J.M. Guzman, G. McGranahan, D. Schensul, and C. Tacoli (editors), Population Dynamics and Climate Change, New York: United Nations Population Fund and International Institute for the Environment and Development, pp. 206-217, 2009.

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., and Nelson, A.: Determining Global Population Distribution: Methods, Applications and Data, ADV PARASIT, 62, 119-156, doi: https://doi.org/10.1016/S0065-308X(05)62004-0, 2006.

Balk, D., F. Pozzi, G. Yetman, U. Deichmann, and A. Nelson. 2005. "The distribution of people and the dimension of place: Methodologies to improve the global estimation of urban extents." International Society for Photogrammetry and Remote Sensing Proceedings of the Urban Remote Sensing Conference, Tempe, AZ, March 2005.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Batista e Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., and Schiavina, M.: Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources, TOURISM MANAGE, 68, 101-115, doi: https://doi.org/10.1016/j.tourman.2018.02.020, 2018.

Bhaduri, B., Bright, E., Coleman, P., and Dobson, J.: LandScan: Locating people is what matters, GEOINFORMATICS, 5, 34-37, 2002.

Bhaduri, B. L., Bright, E. A., Rose, A. N., Liu, C., Urban, M. L., and Stewart, R. N.: Data Driven Approach for High Resolution Population Distribution and Dynamics Models, Savannah, GA, 2014.

Bharti, N., Lu, X., Bengtsson, L., Wetter, E., and Tatem, A. J.: Remotely measuring populations during a crisis by overlaying two data sources, INT HEALTH, 7, 90-98, doi: 10.1093/inthealth/ihv003, 2015.

Birkin, M. and Clarke, M.: Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples, ENVIRON PLANN A, 20, 1645-1671, doi: 10.1068/a201645, 1988.

Blankespoor, B., Dasgupta, S., and Lange, G.-M.: Mangroves as a protection from storm surges in a changing climate, AMBIO, 46, 478-491, doi: 10.1007/s13280-016-0838-x, 2017.

Blankespoor, B., Kilic, T., Murray, S., and Wild, M.: Can remote sensing data complement or even replace the current sampling frames in household surveys in developing countries?, Washington, DC, 2018.

Bogaert, P.: Spatial prediction of categorical variables: the Bayesian maximum entropy approach, STOCH ENV RES RISK A, 16, 425-448, doi: 10.1007/s00477-002-0114-4, 2002.

Breiman, L.: Random Forests, MACH LEARN, 45, 5-32, doi: 10.1023/a:1010933404324, 2001.

Cao, C. Y. and Lam, N.: Understanding the scale and resolution effects in remote sensing and GIS. In: Scale in remote sensing and GIS, Quattrochi, D. A. and Goodchild, M. F. (Eds.), Lewis Publishers, Boca Raton, FL, 1997.

Cohen, J. E., and Small, C.: Hypsographic demography: The distribution of human population by altitude, P NATL ACAD SCI USA, 95, 14009-14014, 10.1073/pnas.95.24.14009, 1998.

Dasgupta, S., Laplante, B., Murray, S., and Wheeler, D.: Exposure of developing countries to sea-level rise and storm surges, CLIMATIC CHANGE, 106, 567-579, doi: 10.1007/s10584-010-9959-6, 2011.

de Bruin, S., Bregt, A., and Ven, M. v. d.: Assessing fitness for use: the expected value of spatial data sets, INT J GEOGR INF SCI, 15, 457-471, doi: 10.1080/13658810110053116, 2001.

de Sherbinin, A.: Remote Sensing and Socioeconomic Data Integration: Lessons from the NASA Socioeconomic Data and Applications Center. In: Integrating Scale in Remote Sensing and GIS, Quattrochi, D., Wentz, E., Lam, N. N., and Emerson, C. (Eds.), CRC Press, Boca Raton, FL, 2017.

Deichmann, U. and Eklundh, L.: Global digital datasets for land degradation studies: a GIS approach, United Nations Environment Programme, Nairobi, 1991.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J.: Dynamic population mapping using mobile phone data, Proceedings of the National Academy of Sciences, 111, 15888-15893, doi: 10.1073/pnas.1408439111, 2014.

Devillers, R., Bédard, Y., Jeansoulin, R., and Moulin, B.: Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data, INT J GEOGR INF SCI, 21, 261-282, doi: 10.1080/13658810600911879, 2007.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., and Shi, W.: Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities, T GIS, 14, 387-400, doi: 10.1111/j.1467-9671.2010.01212.x, 2010.

Dobson, J., Bright, E., Coleman, P., and Bhaduri, B.: LandScan: a global population database for estimating populations at risk. In: Remotely-Sensed Cities, Mesev, V. (Ed.), Taylor & Francis, London, 2003.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., and Worley, B. A.: LandScan: A Global Population Database for Estimating Populations at Risk, PHOTOGRAMM ENG REM S, 66, 849-857, 2000.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Doocy, S., Gorokhovich, Y., Burnham, G., Balk, D., and Robinson, C.: Tsunami mortality estimates and vulnerability mapping in Aceh, Indonesia, AM J PUBLIC HEALTH, 97 Suppl 1, S146-S151, doi: 10.2105/AJPH.2006.095240, 2007.

Dong, N., Yang, X., Cai, H., and Xu, F.: Research on Grid Size Suitability of Gridded Population Distribution in Urban Area: A Case Study in Urban Area of Xuanzhou District, China, PLOS ONE, 12, e0170830, doi: 10.1371/journal.pone.0170830, 2017.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., and Baptista, S. R.: 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, 1, 226-234, doi: 10.1080/23754931.2015.1014272, 2015.

Ehrlich, D., Melchiorri, M., Florczyk, A. J., Pesaresi, M., Kemper, T., Corbane, C., Freire, S., Schiavina, M., and Siragusa, A.: Remote Sensing Derived Built-Up Area and Population Density to Quantify Global Exposure to Five Natural Hazards over Time, REMOTE SENS-BASEL, 10, 1378, doi: doi:10.3390/rs10091378, 2018a.

Ehrlich, D., Kemper, T., Pesaresi, M., and Corbane, C.: Built-up area and population density: Two Essential Societal Variables to address climate hazard impact, ENVIRON SCI POLICY, 90, 73-82, doi: https://doi.org/10.1016/j.envsci.2018.10.001, 2018b.

Eicher, C. L. and Brewer, C. A.: Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation, CARTOGR GEOGR INF SC, 28, 125-138, doi: 10.1559/152304001782173727, 2001.

Elvidge, C. D., Baugh, K., Kihn, E., Kroehl, H., and Davis, E.: Mapping city lights with nighttime data from the DMSP operational linescan system, PHOTOGRAMM ENG REM S, 63, 727−734, 1997.

Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., and Ghosh, T.: VIIRS night-time lights, INT J REMOTE SENS, 38, 5860-5879, doi: 10.1080/01431161.2017.1342050, 2017.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., and Strano, E.: Breaking new ground in mapping human settlements from space – The Global Urban Footprint, ISPRS J PHOTOGRAMM, 134, 30-42, doi: https://doi.org/10.1016/j.isprsjprs.2017.10.012, 2017.

FGDC (Federal Geographic Data Committee): Geospatial Positioning Accuracy Standards, part 3: National standard for spatial data accuracy. Subcommittee for Base Cartographic Data, 25p. 1998.

Ferri, S., Syrris, V., Florczyk, A., Scavazzon, M., Halkia, M., and Pesaresi, M.: A new map of the European settlements by automatic classification of 2.5m resolution SPOT data, Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13-18 July 2014, 1160-1163.

Florczyk, A. J., Ferri, S., Syrris, V., Kemper, T., Halkia, M., Soille, P., and Pesaresi, M.: A New European Settlement Map From Optical Remotely Sensed Data, IEEE J SEL TOP APPL, 9, 1978-1992, 10.1109/JSTARS.2015.2485662, 2016.

Flowerdew, R., Geddes, A., and Green, M.: Behaviour of Regression Models under Random Aggregation. In: Modelling scale in geographical information science, Tate, N. and Atkinson, P. (Eds.), Wiley, Chichester, 2001.

Fotheringham, A. S., Charlton, M., and Brunsdon, C.: The geography of parameter space: an investigation of spatial non-stationarity, INT J GEOGR INF SYST, 10, 605-627, doi: 10.1080/02693799608902100, 1996.

Freire, S. and Halkia, M.: GHSL application in Europe: Towards new population grids, Krakow, Poland, 2014.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., and Mills, J.: Development of New Open and Free Multi-Temporal Global Population Grids at 250 m Resolution, Helsinki, Finland, 2016.

Freire, S., Schiavina, M., Florczyk, A. J., MacManus, K., Pesaresi, M., Corbane, C., Borkovska, O., Mills, J., Pistolesi, L., Squires, J., and Sliuzas, R.: Enhanced data and methods for improving open and free global population grids: putting 'leaving no one behind' into practice, INT J DIGIT EARTH, 1-17, doi: 10.1080/17538947.2018.1548656, 2018.

Frye, C., Wright, D. J., Nordstrand, E., Terborgh, C., and Foust, J.: Using Classified and Unclassified Land Cover Data to Estimate the Footprint of Human Settlement, DATA SCIENCE JOURNAL, 17, doi: 10.5334/dsj-2018-020, 2018.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Fung, I., John, J., Lerner, J., Matthews, E., Prather, M., Steele, L. P., and Fraser, P. J.: Three-dimensional model synthesis of the global methane cycle, J GEOPHYS RES-ATMOS, 96, 13033-13065, doi: 10.1029/91jd01247, 1991.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., and Tatem, A. J.: High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015, PLOS ONE, 8, e55882, doi: 10.1371/journal.pone.0055882, 2013.

Gaughan, A. E., Stevens, F. R., Linard, C., Patel, N. N., and Tatem, A. J.: Exploring nationally and regionally defined models for large area population mapping, INT J DIGIT EARTH, 8, 989-1006, 10.1080/17538947.2014.965761, 2015.

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., Ye, X., Linard, C., Hornby, G. M., Hay, S. I., Yu, H., and Tatem, A. J.: Spatiotemporal patterns of population in mainland China, 1990 to 2010, SCIENTIFIC DATA, 3, 160005, doi: 10.1038/sdata.2016.5, 2016.

GeoNames: http://www.geonames.org/.

Goodchild, M. and Lam, N.: Areal interpolation: a variant of the traditional spatial problem, GEOPROCESSING, 1, 297-312, 1980.

Gunasekera, R., Ishizawa, O., Aubrecht, C., Blankespoor, B., Murray, S., Pomonis, A., and Daniell, J.: Developing an adaptive global exposure model to support the generation of country disaster risk profiles, EARTH-SCIENCE REVIEWS, 150, 594-608, doi: https://doi.org/10.1016/j.earscirev.2015.08.012, 2015.

Guptill, S.C., Morrison, J.L. (eds.): Elements of spatial data quality. Elsevier; 2013 Oct 22.

Harvey, J. T.: Estimating census district populations from satellite imagery: Some approaches and limitations, INT J REMOTE SENS, 23, 2071-2095, doi: 10.1080/01431160110075901, 2002.

Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M., and Snow, R. W.: The global distribution and population at risk of malaria: past, present, and future, LANCET INFECT DIS, 4, 327-336, doi: 10.1016/S1473-3099(04)01043-6, 2004.

HERE: https://www.here.com/products/mapping/map-data.

Iimi, A., Ahmed, A. K. F., Anderson, E. C., Diehl, A. S., Maiyo, L., Peralta Quiros, T., and Rao, K. S.: New rural access index : main determinants and correlation to poverty, World Bank Group, Washington, D.C., 2016.

Islam, M. S., Oki, T., Kanae, S., Hanasaki, N., Agata, Y., and Yoshimura, K.: A grid-based assessment of global water scarcity including virtual water trading, WATER RESOUR MANAG, 21, 19, doi: 10.1007/s11269-006-9038-y, 2006.

Ivánová, I., Morales, J., de By, R. A., Beshe, T. S., and Gebresilassie, M. A.: Searching for spatial data resources by fitness for use, J SPAT SCIS, 58, 15-28, doi: 10.1080/14498596.2012.759087, 2013.

Jenness, J. S.: Calculating landscape surface area from digital elevation models, WILDLIFE SOC B, 32, 829-839, 2004.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P.: Global trends in emerging infectious diseases, NATURE, 451, 990, doi: 10.1038/nature06536, 2008.

Jones, B. and O'Neill, B. C.: Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways, ENVIRON RES LETT, 11, 084003, doi: 10.1088/1748-9326/11/8/084003, 2016.

Klein Goldewijk, K., Beusen, A., Doelman, J., and Stehfest, E.: Anthropogenic land use estimates for the Holocene – HYDE 3.2, EARTH SYST SCI DATA, 9, 927-953, doi: 10.5194/essd-9-927-2017, 2017.

Klein Goldewijk, K., Beusen, A., and Janssen, P.: Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1, HOLOCENE, 20, 565-573, doi: 10.1177/0959683609356587, 2010.

Klein Goldewijk, K., Beusen, A., van Drecht, G., and de Vos, M.: The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years, GLOBAL ECOL BIOGEOGR, 20, 73-86, doi: 10.1111/j.1466-8238.2010.00587.x, 2011.

Koch, J., Schaldach, R., and Köchy, M.: Modeling the impacts of grazing land management on land-use change for the Jordan River region, GLOBAL PLANET CHANGE, 64, 177-187, doi: https://doi.org/10.1016/j.gloplacha.2008.09.005, 2008.

Leyk, S., Nagle, N. N., and Buttenfield, B. P.: Maximum Entropy Dasymetric Modeling for Demographic Small Area Estimation, GEOGR ANAL, 45, 285-306, doi: 10.1111/gean.12011, 2013.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Leyk, S., Ruther, M., Buttenfield, B. P., Nagle, N. N., and Stum, A. K.: Modeling residential developed land in rural areas: A size-restricted approach using parcel data, APPL GEOGR, 47, 33-45, doi: https://doi.org/10.1016/j.apgeog.2013.11.013, 2014.

Leyk, S. and Uhl, J. H.: HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years, SCIENTIFIC DATA, 5, 180175, doi: 10.1038/sdata.2018.175, 2018.

Leyk, S., Uhl, J. H., Balk, D., and Jones, B.: Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States, REMOTE SENS ENVIRON, 204, 898-917, doi: https://doi.org/10.1016/j.rse.2017.08.035, 2018.

Linard, C. and Tatem, A. J.: Large-scale spatial population databases in infectious disease research, INT J HEALTH GEOGR, 11, 7, doi: 10.1186/1476-072x-11-7, 2012.

Liverman, D., Moran, E., Rindfuss, R., and Stern, P. (Eds.): People and Pixels: Linking Remote Sensing and Social Science, National Research Council. The National Academies Press, Washington, DC, 1998.

Lloyd, C. T., Sorichetta, A., and Tatem, A. J.: High resolution global gridded data for use in population studies, SCIENTIFIC DATA, 4, 170001, doi: 10.1038/sdata.2017.1, 2017.

MacDonald Dettwiler and Associates (MDA): https://landscape6.arcgis.com/arcgis/rest/services/World_Land_Cover_30m_BaseVue_2013/ImageServer

Maclaurin, G., Leyk, S., and Hunter, L. M.: Understanding the combined impacts of aggregation and spatial non-stationarity: The case of migration-environment associations in rural South Africa, TRANSACTIONS IN GIS: TG 19, 877-895, doi: 10.1111/tgis.12134, 2015.

McDonald, R. I., Green, P., Balk, D., Fekete, B. M., Revenga, C., Todd, M., and Montgomery, M.: Urban growth, climate change, and freshwater availability, P NATL ACAD SCI USA, 108, 6312-6317, doi: 10.1073/pnas.1011615108, 2011.

McGranahan, G., Balk, D., and Anderson, B.: The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones, ENVIRON URBAN, 19, 17-37, doi: 10.1177/0956247807076960, 2007.

Melchiorri, M., Pesaresi, M., Florczyk, A. J., Corbane, C., and Kemper, T.: Principles and Applications of the Global Human Settlement Layer as Baseline for the Land Use Efficiency Indicator—SDG 11.3.1, ISPRS INT J GEO-INF, 8, 96, doi: doi:10.3390/ijgi8020096, 2019.

Mennis, J.: Dasymetric Mapping for Estimating Population in Small Areas, GEOGRAPHY COMPASS, 3, 727-745, doi: doi:10.1111/j.1749-8198.2009.00220.x, 2009.

Mennis, J.: Generating Surface Models of Population Using Dasymetric Mapping, PROF GEOGR, 55, 31-42, doi: 10.1111/0033-0124.10042, 2003.

Mennis, J. and Hultgren, T.: Intelligent Dasymetric Mapping and Its Application to Areal Interpolation, CARTOGR GEOGR INF SC, 33, 179-194, doi: 10.1559/152304006779077309, 2006.

Mondal, P. and Tatem, A. J.: Uncertainties in Measuring Populations Potentially Impacted by Sea Level Rise and Coastal Flooding, PLOS ONE, 7, e48191, doi: 10.1371/journal.pone.0048191, 2012.

Montello, D. R.: Scale in geography. In: International Encyclopedia of the Social and Behavioral Sciences, Smelser, N. J. and Baltes, B. (Eds.), 2001.

Mrozinski, R. D. and Cromley, R. G.: Singly- and Doubly-Constrained Methods of Areal Interpolation for Vector-based GIS, T GIS, 3, 285-301, doi: 10.1111/1467-9671.00022, 1999.

Myers, R. J.: Errors and bias in the reporting of ages in census data. In: Readings in population research methodology. Volume 1: Basic Tools, Bogue, D. J., Arriaga, E. E., Anderton, D. L., and Rumsey, G. W. (Eds.), Social Development Center, Chicago, Illinois, 1993.

Nagle, N. N., Buttenfield, B. P., Leyk, S., and Speilman, S.: Dasymetric Modeling and Uncertainty, ANN ASSOC AM GEOGR, 104, 80-95, doi: 10.1080/00045608.2013.843439, 2014.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Nieves, J. J., Stevens, F. R., Gaughan, A. E., Linard, C., Sorichetta, A., Hornby, G., Patel, N. N., and Tatem, A. J.: Examining the correlates and drivers of human population distributions across low- and middle-income countries, J R SOC INTERFACE, 14, 20170401, doi: doi:10.1098/rsif.2017.0401, 2017.

Nordhaus, W. D.: Geography and macroeconomics: New data and new findings, P NATL ACAD SCI USA, 103, 3510-3517, doi: 10.1073/pnas.0509842103, 2006.

Openshaw, S.: The modifiable areal unit problem, Geo Books, Norwick [Norfolk], 1983.

Openshaw, S. and Taylor, P. J.: The modifiable areal unit problem. In: Quantitative Geography: A British View, Wrigley, N. and Bennett, R. (Eds.), Routledge and Kegan Paul, London, 1981.

OpenStreetmap Foundation (OSMF): https://www.openstreetmap.org.

Parish, E. S., Kodra, E., Steinhaeuser, K., and Ganguly, A. R.: Estimating future global per capita water availability based on changes in climate and population, COMPUT GEOSCI, 42, 79-86, doi: https://doi.org/10.1016/j.cageo.2012.01.019, 2012.

Pawitan, G. and Steel, D. G.: Exploring a Relationship Between Aggregate and Individual Levels Spatial Data Through Semivariogram Models, GEOGR ANAL, 38, 310-325, doi: 10.1111/j.1538-4632.2006.00688.x, 2006.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S. C., Halkia, S., Julea, A. M., Kemper, T., Soille, P., and Syrris, V.: Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014, Joint Research Center, 2016a.

Pesaresi, M., Ehrlich, D., Florczyk, A. J., Freire, S., Julea, A., Kemper, T., and Syrris, V.: The global human settlement layer from landsat imagery, 10-15 July 2016b, 7276-7279.

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M. A., Ouzounis, G. K., Scavazzon, M., Soille, P., Syrris, V., and Zanchetta, L.: A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results, IEEE J SEL TOP APPL, 6, 2102-2131, doi: 10.1109/JSTARS.2013.2271445, 2013.

Piantadosi, S., Byar, D. P., and Green, S. B.: The ecological fallacy, AM J EPIDEMIOL, 127, 893-904, doi: 10.1093/oxfordjournals.aje.a114892, 1988.

POPGRID: https://www.popgrid.org/, https://www.popgrid.org/compare-data, http://sedac.ciesin.columbia.edu/mapping/popgrid/. 2018

Potere, D., Schneider, A., Angel, S., and Civco, D. L.: Mapping urban areas on a global scale: which of the eight maps now available is more accurate?, INT J REMOTE SENS, 30, 6531-6558, doi: 10.1080/01431160903121134, 2009.

Potter, J. E. and Ordóñez, M. G.: The Completeness of Enumeration in the 1973 Census of the Population of Colombia, POPUL INDEX, 42, 377-403, doi: 10.2307/2734458, 1976.

Reed, F. J., Gaughan, A. E., Stevens, F. R., Yetman, G., Sorichetta, A., and Tatem, A. J.: Gridded Population Maps Informed by Different Built Settlement Products, DATA, 3, 33, 2018.

Reibel, M. and Bufalino, M. E.: Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems, ENVIRON PLANN A, 37, 127-139, doi: 10.1068/a36202, 2005.

Rigaud, K. K., de Sherbinin, A., Jones, B., Bergmann, J., Clement, V., Ober, K., Schewe, J., Adamo, S., McCusker, B., Heuser, S., and Midgley, A.: Groundswell: Preparing for Internal Climate Migration, The World Bank, Washington, DC, 2018.

Roberts, M., Blankespoor, B., Deuskar, C., and Stewart, B.: Urbanization and Development: Is Latin America and the Caribbean Different from the Rest of the World?, World Bank Group, Washington, DC, 2017.

See, L., Fritz, S., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., Duerauer, M., Sturn, T., Karner, M., Kraxner, F., and Obersteiner, M.: Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki, TECHNOL FORECAST SOC, 98, 324-335, doi: https://doi.org/10.1016/j.techfore.2015.03.002, 2015.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

Semenov-Tian-Shansky, B.: Russia: Territory and Population: A Perspective on the 1926 Census, GEOGR REV, 18, 616-640, doi: 10.2307/207951, 1928.

Simarro, P. P., Cecchi, G., Franco, J. R., Paone, M., Fèvre, E. M., Diarra, A., Ruiz Postigo, J. A., Mattioli, R. C., and Jannin, J. C.: Risk for Human African Trypanosomiasis, Central Africa, 2000–2009., EMERG INFECT DIS, 17, 2322-2324, doi: https://dx.doi.org/10.3201/eid1712.110921, 2011.

Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., and Tatem, A. J.: Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling, COMPUT ENVIRON URBAN, 75, 132-145, doi: https://doi.org/10.1016/j.compenvurbsys.2019.01.006, 2019.

Small, C., Pozzi, F., and Elvidge, C. D.: Spatial analysis of global urban extent from DMSP-OLS night lights, REMOTE SENS ENVIRON, 96, 277-291, doi: https://doi.org/10.1016/j.rse.2005.02.002, 2005.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J.: High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020, SCIENTIFIC DATA, 2, 150045, doi: 10.1038/sdata.2015.45 https://www.nature.com/articles/sdata201545#supplementary-information, 2015.

Sorichetta, A., Bird, T. J., Ruktanonchai, N. W., zu Erbach-Schoenberg, E., Pezzulo, C., Tejedor, N., Waldock, I. C., Sadler, J. D., Garcia, A. J., Sedda, L., and Tatem, A. J.: Mapping internal connectivity through human migration in malaria endemic countries, SCIENTIFIC DATA, 3, 160066, 10.1038/sdata.2016.66 https://www.nature.com/articles/sdata201666#supplementary-information, 2016.

Steel, D. G. and Holt, D.: Rules for Random Aggregation, ENVIRON PLANN A, 28, 957-978, doi: 10.1068/a280957, 1996.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J.: Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data, PLOS ONE, 10, e0107042, doi: 10.1371/journal.pone.0107042, 2015.

Taramelli, A., Melelli, L., Pasqui, M., and Sorichetta, A.: Modelling risk hurricane elements in potentially affected areas by a GIS system, GEOMAT NAT HAZ RISK, 14, 349-373, doi: DOI: 10.1080/19475705.2010.532972, 2010.

Tatem, A. J.: WorldPop, open data for spatial demography, SCIENTIFIC DATA, 4, 170004, doi: 10.1038/sdata.2017.4, 2017.

Tatem, A. J.: Mapping the denominator: spatial demography in the measurement of progress, INT HEALTH, 6, 153-155, doi: 10.1093/inthealth/ihu057, 2014.

Tatem, A. J., Campiz, N., Gething, P. W., Snow, R. W., and Linard, C.: The effects of spatial population dataset choice on estimates of population at risk of disease, POPUL HEALTH METRS, 9, 4-4, doi: 10.1186/1478-7954-9-4, 2011.

Tayi, G. K. and Ballou, D. P.: Examining data quality, COMMUN ACM, 41, 54-57, doi: 10.1145/269012.269021, 1998.

Thomson, D. R., Stevens, F. R., Ruktanonchai, N. W., Tatem, A. J., and Castro, M. C.: GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data, INT J HEALTH GEOGR, 16, 25, doi: 10.1186/s12942-017-0098-4, 2017.

Tiecke, T.: Open population datasets and open challenges. https://code.fb.com/connectivity/open-population-datasets-and-open-challenges/, 2016.

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., and Dang, H.-A. H.: Mapping the world population one building at a time., Retrieved from https://arxiv.org/abs/ 1712.05839, 2017.

Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K.: The Global Demography Project, National Center for Geographic Information and Analysis, 1995.

Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K.: World population in a grid of spherical quadrilaterals, INTERNATIONAL JOURNAL OF POPULATION GEOGRAPHY, 3, 203-225, doi: doi:10.1002/(SICI)1099-1220(199709)3:3<203::AID-IJPG68>3.0.CO;2-C, 1997.

United Nations. Handbook on Geospatial Infrastructure in Support of Census Activities New York: United Nations, Department of Economic and Social Affairs, Statistics Division, Series F 103, 2009.

https://doi.org/10.5194/essd-2019-82
Preprint. Discussion started: 6 June 2019
© Author(s) 2019. CC BY 4.0 License.

U.S. Census Bureau International Programs: https://www.census.gov/programs-surveys/international-programs/about/global-mapping.html, last access: 12/11/2018.

Waller, L. and Gotway, C.: Applied Spatial Statistics for Public Health Data, Wiley, Hoboken, NJ, 2004.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., and Tatem, A. J.: Spatially disaggregated population estimates in the absence of national population and housing census data, P NATL ACAD SCI USA, 115, 3529-3537, doi: 10.1073/pnas.1715305115, 2018.

Weber, E. M., Seaman, V. Y., Stewart, R. N., Bird, T. J., Tatem, A. J., McKee, J. J., Bhaduri, B. L., Moehl, J. J., and Reith, A. E.: Census-independent population mapping in northern Nigeria, REMOTE SENS ENVIRON, 204, 786-798, doi: https://doi.org/10.1016/j.rse.2017.09.024, 2018.

Wesolowski A, B. C., Bengtsson L, Wetter E, Lu X, Tatem AJ. 2014 Sep 29 . Edition 1. : Commentary: Containing the Ebola Outbreak – the Potential and Challenge of Mobile Network Data, PLOS CURRENT OUTBREAKS, Sep 29, doi: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e., 2014.

Wickham, J. D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., and Wade, T. G.: Accuracy assessment of NLCD 2006 land cover and impervious surface, REMOTE SENS ENVIRON, 130, 294-304, doi: https://doi.org/10.1016/j.rse.2012.12.001, 2013.

Wong, D.: The modifiable areal unit problem (MAUP). In: The SAGE handbook of spatial analysis, Fotheringham, A. and Rogerson, P. (Eds.), SAGE, Los Angeles, 2009.

Wong, D. W. S.: The Reliability of Using the Iterative Proportional Fitting Procedure, PROF GEOGR, 44, 340-348, doi: 10.1111/j.0033-0124.1992.00340.x, 1992.

WorldPop (School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Département de Géographie, Université de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University: Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00###, 2018.

Wright, J. K.: A Method of Mapping Densities of Population: With Cape Cod as an Example, GEOGR REV, 26, 103-110, 1936.

Wu, S.-s., Qiu, X., and Wang, L.: Population Estimation Methods in GIS and Remote Sensing: A Review, GISCI REMOTE SENS, 42, 80-96, doi: 10.2747/1548-1603.42.1.80, 2005.

Zandbergen, P. A. and Ignizio, D. A.: Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates, CARTOGR GEOGR INF SC, 37, 199-214, doi: 10.1559/152304010792194985, 2010.