The LandScan Global Population Distribution Project: Current State of the Art and Prospective Innovation

Amy N. Rose[1]*, Eddie Bright[1]

[1]Computational Sciences and Engineering Division, Oak Ridge National Laboratory

*Corresponding Author: Computational Sciences and Engineering Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831-6017. Email:rosean@ornl.gov.

Abstract

Advances in remote sensing, dasymetric mapping techniques, and the ever-increasing availability of spatial datasets have enhanced global human population distribution databases. These datasets demonstrate an enormous improvement over the conventional use of choropleth maps to represent population distribution and are vital for analysis and planning purposes including humanitarian response, disease mapping, risk analysis, and evacuation modeling. Dasymetric mapping techniques have been employed to address spatial mismatch, but also to develop finer resolution population distributions in areas of the world where subnational census data are coarse or non-existent. One such implementation is the LandScan Global model which provides a 30 arc-second global population distribution based on ancillary datasets such as land cover, slope, proximity to roads, and settlement locations. This work will review the current state of the LandScan model, future innovations aimed at increasing spatial and demographic resolution, and situate LandScan within the landscape of other global population distribution datasets.

Introduction

LandScan is now in its fifteenth year of production, a milestone which warrants a long overdue examination of this dataset; including its origins, data inputs, ongoing innovations, and future pathways. Although LandScan is a widely distributed dataset, a frequent critique is the lack of documentation of and transparency into the process by which the data is developed each year. Over the course of the last 15 years, LandScan has undergone numerous changes with regard to input data sources. Furthermore, the basic methodology used to produce LandScan continues to be refined based on the increased availability of good quality national level datasets, as well as new and innovative imagery processing techniques. This paper highlights the input data issues and availability, recent advances in the LandScan modeling process, and the plans for the future. Future requirements and the appropriate strategy to meet those needs are critical; the necessity of population distribution datasets is apparent through the continued funding and production of such efforts. Often the similarities and differences between each of these efforts is questioned and so this paper examines other datasets that are frequently likened to LandScan.

High resolution population data is a necessary requirement for research and analysis in two primary contexts: prevention and response situations. For prevention, the requirement focuses on developing capabilities to improve the response to critical events over time and increase the resilience of local populations. For response situations, there is often a shorter-term need for information that can aid in estimating damage and providing relief to affected areas (NRC 2007, 109). The LandScan population distribution database is explicitly developed to address both of these contexts by helping to identify populations at risk, whether from natural or man-made events. The goal is not to specifically quantify a certain number of people in a particular geographical location, but rather to provide a more realistic population distribution for consequence assessments than is afforded strictly by census counts.

The uses of LandScan data are varied and extensive. LandScan has significantly enhanced the utility and impact of various applications in areas including counter-terrorism, homeland security, emergency planning and management, consequence analysis, epidemiology, exposure analysis, and urban sprawl detection. National and international organizations including the United Nations (UN), the World Health Organization (WHO), the Food and Agricultural Organization (FAO), and several federal agencies in the U.S. and other countries currently employ this data in their analyses. In all, Oak Ridge National Laboratory (ORNL) has distributed LandScan Global population data to over 750 different organizations spread across the world (Bhaduri et al. 2002).

Background

Population data is vital for an array of analysis and planning purposes and is a basic component of humanitarian response efforts (NRC 2007). Unfortunately, those analyses are often hampered by the reality that population source data (e.g. census information) rarely conforms to the spatial extent of analysis regions (Goodchild, Anselin, and Deichmann 1993). To mitigate the problem of disparate spatial data assimilation, a surface representing population distribution is often employed for spatial analyses (Fotheringham and Rogerson 1993). Numerous methodologies have been devised to create population

surfaces including simple areal interpolation (Tobler 1979), allometric and regression models (Lo and Welch 1977), and dasymetric models (Wright 1936; Deichmann 1996; Eicher and Brewer 2001; Mennis and Hultgren 2006). The following section discusses these methods.

Population Distribution Methods

Bracken and Martin (1989) developed surface representations of demographic data for census enumeration areas in the United Kingdom using a centroid-kernel method. Using this point-based method, population is assigned to the centroids of the enumeration areas. A moving kernel window is positioned over each centroid and population is allocated to cells falling within the window based on weights of other centroids falling within the window. This method uses a weighting based on a distance decay function so that closer centroids have more weight than those farther away. While this does allow for some areas of the resulting surface to contain zero population, it also assumes that population density decreases further away from the centroid.

The largest issue with a point-based interpolation method is that generally the total value within each source enumeration area is not preserved. To address this, areal interpolation methods were developed (Goodchild and Lam 1980; Flowerdew and Green 1992; Goodchild, Anselin, and Deichmann 1993; Fisher and Langford 1996). Simple areal interpolation uses the geometric properties of the source area to determine the proportion of the total value that will be allocated to that area. Essentially, an area-weighted function is used where the area proportion serves as a weight for population distribution. However, the assumed homogeneity of an area, especially with regard to population distribution is problematic (Lam 1983).

The dasymetric method was originally developed by Benjamin Semenov-Tian-Shansky in the early 1900's in his proposal to produce the 'Dasymetric Map of European Russia' (Petrov 2012). However, this fact is often overlooked in literature, owing instead the popularization of this method to J.K. Wright by way of his population density mapping of Cape Cod (Wright 1936). Wright felt that choropleth maps, with their uniform distribution, did not provide a realistic representation of population within enumeration units, so he used his local knowledge to develop a method to refine the population densities (Fisher and Langford 1996). In this method, partitions are iteratively created to disaggregate general zones to detailed zones of population density while preserving the population counts of the original zones. Building on this methodology, more robust dasymetric methods have been developed recently. Since population is commonly related to other factors including land use and transportation networks, it stands to reason that those features would be important inputs in the development of representative population distribution surfaces. These so-called "intelligent" dasymetric methods use ancillary variables to guide the redistribution of the population (Flowerdew, Green, and Kehris 1991; Flowerdew and Green 1992; Mennis and Hultgren 2006). Figures 1 and 2 below show an example of the dasymetric mapping process.

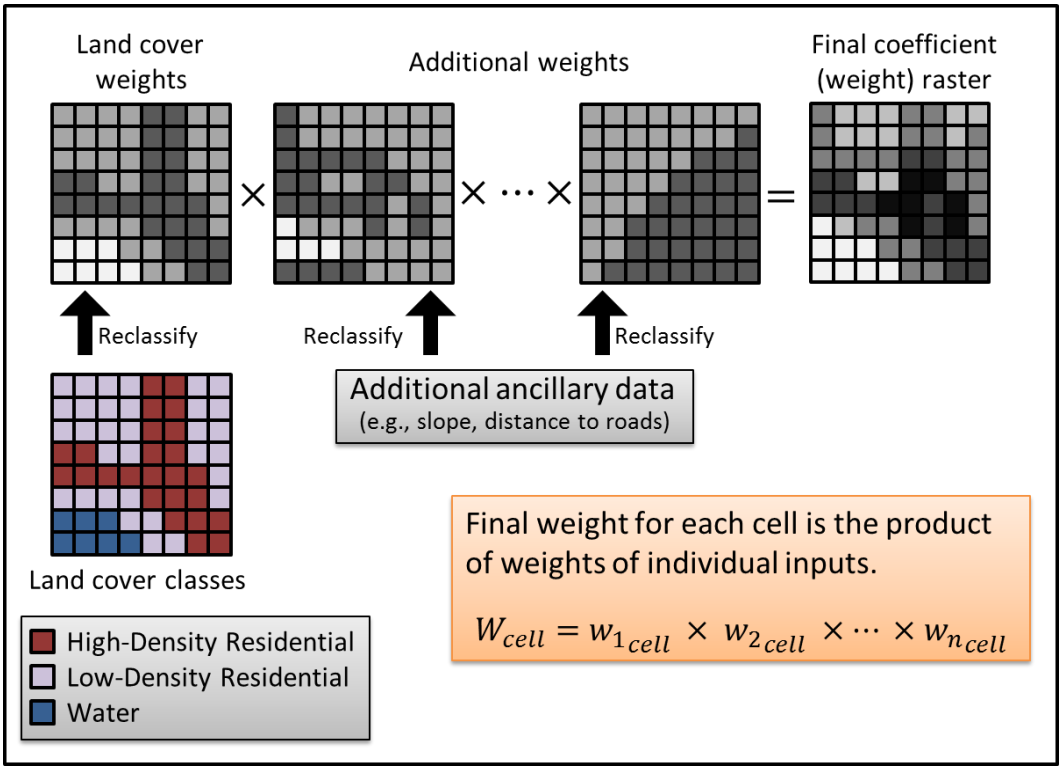

Figure 1. Creating a coefficient surface using ancillary variables.

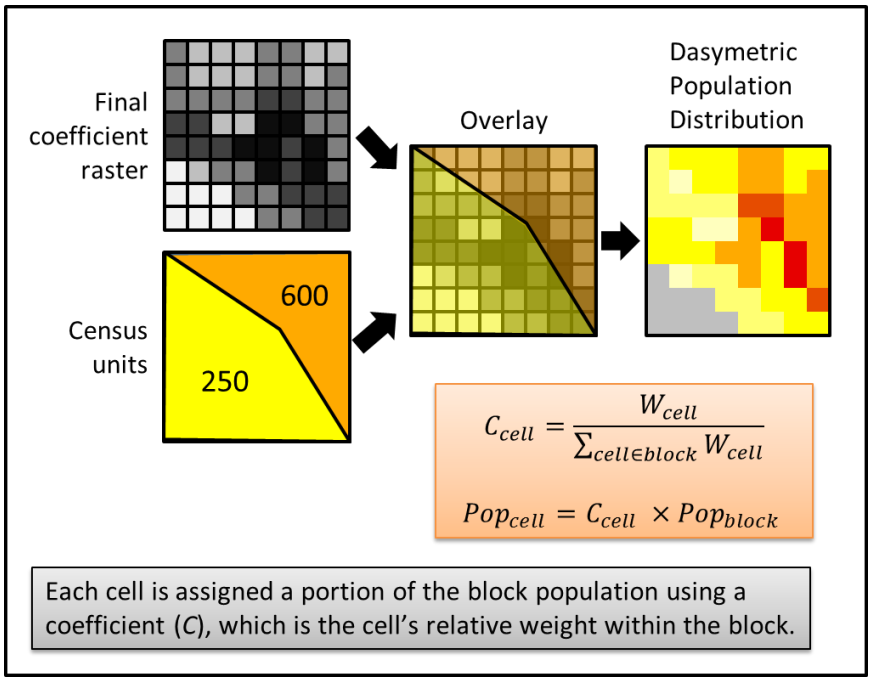

Figure 2. Using the coefficient surface to redistribute population.

Methodological Considerations

The application of a particular population allocation approach should consider not just methodology, but also the theoretical issues associated with implementing spatiotemporal representations of real world phenomena. When developing population distribution surfaces, identifying what phenomena we are attempting to measure is critical since the degree of geographical detail at which phenomena occur, the data we use to measure it, and the models we use to represent it can greatly vary. Population occurs at one scale, input data is available at a different scale, and the output population distribution at yet another scale. How can all this information be integrated and reconciled in order to most accurately characterize human activity space?

The foundation of population distribution methods is the reliance on spatial information that either directly describes population at or within a given place, or provides a reasonable proxy for human activity. At the most basic level, point or areal features with an associated population count are required. Additional features including land cover, slope, or transportation infrastructure that can be used to inform the distribution can help refine the geographical detail of the output. Regardless, uncertainty will still manifest in the output, particularly due to both spatial and temporal misalignment of input data.

The increasingly high geographical detail that is required for useful analytical output makes global population distribution datasets with a very high spatial resolution extremely desirable as input for modeling. In some areas of the world, particularly developed nations, very detailed census and spatial data is available, so that creating population distributions with high geographic detail is not difficult. However, most of the world lacks in either good quality spatial data or recent (and accurate) census data. This lack of data however should not be a barrier to producing population distribution datasets, but rather the implications of data limitations should be acknowledged so that users of population distribution datasets are aware of how these issues may affect the outcome of their own analyses.

Input Data Effects

The annual update of LandScan, as well as other global population distribution datasets, requires that subnational boundaries for every country are examined with a critical eye. Often subnational boundaries coincide with natural features or circumscribe urban areas. The spatial accuracy and precision of administrative boundaries can be evaluated by overlaying the data on high resolution geo-rectified imagery. Some boundary datasets are topologically correct but not spatially accurate, with errors commonly exceeding multiple kilometers. Often subnational administrative boundaries have been greatly generalized for cartographic purposes which may result in villages, towns, and even portions of larger cities being located into the wrong administrative region.

The resolution, or more specifically the scale and detail, of administrative unit boundaries and the population associated with them is the single most important input into a dasymetric model – or any population distribution model. The idea is that the smaller the administrative units, the more accurate any population distribution model within those units will be. This is a reasonable assumption in most cases, but it also assumes that the boundaries are spatially accurate. Furthermore, population associated with administrative boundaries is typically census or residential population counts. Using

these population counts to develop a finer resolution population distribution for small boundary units within an urban area may not accurately reflect the true human activity signature in that area since it will not account for other potential activity outside of residences.

Temporal Factors

Population distribution datasets are most useful when the date of representation aligns with the analyses being performed. Regardless of the temporal domain of the input data used to create the population distribution surface, the output data is either implicitly or explicitly capturing a snapshot in time of the actual distribution of humans on the earth's surface. The nature of population dynamics prevents any census from being truly representative of where people are located. However, population distribution methods aim to capture the population at a given time.

Universally, population distribution methods use some type of "official" population counts whether census year counts, intercensal estimates, or registry information. Given this, it's important to note that censuses can vary greatly in their regularity and level of execution (Table 1). Particularly for developing countries, even if scheduled, censuses may not be conducted regularly and in some cases have been lacking for multiple decades.

It is important to consider the implications of using outdated census information for developing any population distribution datasets. First, for many countries migration – either voluntary or forced – can have an enormous impact on the population dynamics of an area. For example, recent civil conflicts in neighboring countries Sudan (and areas that are now South Sudan), Ethiopia, and Somalia have produced an associated internal displacement, as well as cross-border displacement with each other in what essentially constituted a population swap. Of these countries, Ethiopia conducted its most recent census in 2007, Sudan in 2008 – prior to South Sudan becoming an independent state, and the last official census conducted in Somalia was in 1987. Therefore, using these numbers to develop any population distribution dataset must involve some type of validation of current conditions, and applied as corrections to either the input datasets or the output itself.

Uncertainty Assessments

Quantifying the accuracy of population distribution data has been the focus of several recent publications (Hall, Stroh, and Paya 2012; Mondal and Tatem 2012; Tatem et al. 2011), yet the reality is that the true accuracy of these datasets are difficult if not impossible to measure due not only to the lack of independent ground-truth data, but also as a result of spatiotemporal population dynamics continually taking place at a variety of scales.

Table 1. Census history of selected nations (IPC 2008).

Country or Area	Round of Population Census						
	1950 Round (1945-54)	1960 Round (1955-64)	1970 Round (1965-74)	1980 Round (1975-84)	1990 Round (1985-94)	2000 Round (1995-2004)	2010 Round (2005-2014)
Benin	--	--	--	3/20-30/79 (F)	2/15/92	2/11/02	2012 (S)
Botswana	5/7/46	4/1/64	8/31/71	8/12-26/81	8/14-23/91	8/17-26/01	2011 (P)
Burundi	--	--	--	8/15-16/79 (F)	8/15-16/90	--	8/08 (S)
Cameroon	--	--	--	4/9/76 (F)	4/14-28/87	--	11/11/05
Cape Verde	12/15/50	12/15/60	12/15/70	6/1-2/80	6/23/90	6/16-30/00	2010 (P)
Chad	--	--	--	--	4/15/93 (F)	--	--
Comoros	--	9/7/58 (F)	7/66-9/66	9/15/80	9/15/91	9/03	2013 (S)
Cote d'Ivoire	--	--	--	4/30/75 (F)	3/1/88	11/21/98	2008 (S)
Ethiopia	--	--	--	5/9/84 (F)	10/11-27/94	--	5/29/07
Gabon	--	10/8/60-5/61 (F)	6/1/69-6/70	8/1-31/80	7/1-31/93	12/31/03	2014 (S)
Gambia, The	--	4/17/63 (F)	4/21/73	4/15/83	4/15/93	4/15/03	2013 (S)
Ghana	2/1/48 (F)	3/20/60	3/1/70	3/11/84	--	3/26/00	2010 (S)
Kenya	2/25-8/23/48 (F)	8/15-16/62	8/24-25/69	8/25/79	8/24/89	8/24/99	2009 (S)
Liberia	--	4/2/62 (F)	2/1/74	2/1-14/84	--	--	3/21-27/08
Libya	7/31/54 (F)	7/31/64	7/31/73	7/31/84	--	8/95	4/15-5/1/06
Madagascar	--	--	--	1/26-8/18/75 (F)	8/1-19/93	--	2009 (S)
Morocco	--	6/18/60 (F)	7/20/71	9/3-21/82	9/4/94	9/1/04	9/14 (S)
Mozambique	9/21/50	9/5/60	12/15/70	8/1/80	--	8/1-15/97	8/1-15/07
Namibia	--	9/6/60 (F)	5/6/70	8/26/81	10/21/91	8/27-9/11/01	2011 (P)
Niger	--	--	--	10/7-11/6/77 (F)	5/10-24/88	5/20-6/18/01	2011 (S)
Nigeria	7/52-6/53	11/5-8/63	--	--	11/27-29/91		3/21/06
Rwanda	--	--	--	8/15-16/78 (F)	8/15-16/91	8/16-30/02	2012 (S)
Somalia	--	--	--	2/7-20/75 (F)	11/86-2/87	--	--
Sudan	--	7/1/55-9/2/56 (F)	4/3/73	2/1/83	4/15/93	--	4/22-5/6/08
Swaziland	5/7/46	7-8/56	5/24/66	8/25/76	8/25/86	5/12/97	5/12/07
Togo	--	--	3/1-4/30/70 (F)	11/22/81	--	--	--
Zimbabwe	--	4/10-5/20/62 (F)	4/21-5/11/69	8/18/82	8/18/92	8/18/02	2012 (S)

(F) First full modern census taken **(P) Projected based on pattern of census dates** **(S) Scheduled; not yet taken or known if taken**

Current State of the Art

More recently, dasymetric mapping techniques have been employed not just to address spatial mismatch, but also to develop finer resolution population distributions in areas of the world where subnational census data are coarse or non-existent. One such implementation is the LandScan Global dataset which provides a 30 arc-second global population distribution based on ancillary datasets such as land cover, slope, proximity to roads, and settlement locations (Dobson et al. 2000). LandScan is just one of the available high-resolution global population datasets that has been developed in recent years. Projects like the Gridded Population of the World (GPW), the Global Rural Urban Mapping Project (GRUMP), and AfriPop are all public domain datasets available for research and analysis use. Of these, AfriPop is a recent addition, initiated in 2009 with an aim of producing population distribution maps for

the whole of Africa. While each of these population distribution efforts relies on some of the same datasets as inputs to their model, they are methodologically dissimilar. The next section discusses each of these population distribution models in turn.

Gridded Population of the World

The Gridded Population of the World (GPW) project, originally produced at the National Center for Geographic Information Analysis (NCGIA) in 1995 (Tobler et al., 1995), can be considered the first noteworthy attempt to generate a consistent spatial global population dataset. GPW was developed with the purpose of providing a (nighttime) global population dataset at a scale required to perform analysis at the country or global level. GPW has been successively updated in 2000 (Deichmann et al., 2001) and 2005 (Balk and Yetman, 2004) by The Center for International Earth Science Information Network (CIESIN) at Columbia University. GPW is simple, areal-weighted redistribution (i.e. proportional allocation) of census population from their census boundaries (or administrative polygons) to a uniform quadrilateral grid. The population totals in GPW datasets are based solely on administrative boundary data and population estimates associated with those administrative units. Since that first release, successive releases of GPW have incorporated larger numbers of administrative units which are used to guide the population redistribution. However, no effort is made to model population distribution, and no ancillary data were used to predict population distribution or revise the population estimates. The only assumption made was that population is uniformly distributed within each administrative unit.

The principal drawbacks of the GPW dataset are the coarse resolution of 2.5 arc-minutes and the lack of any modeling of population distribution within administrative units. The latter is particularly significant in that the result is evenly distributed populations across any given administrative unit. This is unlikely to represent a realistic population distribution, especially within large units with significant variation in land cover characteristics. Although GPW utilizes administrative unit data at the finest resolution available, this by far is not the resolution of human settlement on the ground.

Global Urban-Rural Mapping Project

The Global Rural-Urban Mapping Project (GRUMP) began around 2004 building on the methodology employed by GPW in order to provide a framework of urban and rural area (nighttime) populations. The GRUMP population grid represents a nighttime population at a 30 arc-second resolution. GRUMP employs urban proportional reallocation based on land area using nighttime lights, and in some cases the Digital Chart of the World (DCW). Although areal weighting is used to redistribute population from polygons to grid, it uses urban polygons as well as administrative polygons and makes assumptions about the parameters of the assignment in those urban areas. GRUMP relies on accurate subnational population projections as well as the usefulness of the nighttime lights dataset. The update frequency of GRUMP is not published, and no update has been produced since the initial release of the dataset in the mid-2000s.

AfriPop

The AfriPop project was initiated in 2009 in order to provide spatial data on African (nighttime) population distributions to support epidemiological modeling. Although at the present time this is not a global population distribution dataset, related projects, AsiaPop and AmeriPop, were launched in 2012

and 2013 respectively. Initial development and release of country level population datasets has been swift, although updates to the data have not yet been undertaken. AfriPop is promoted as a 100m resolution dataset which uses an area-weighted reallocation method to distribute population counts. The AfriPop methodology combines census boundaries and population, settlement points, and land cover information in a semi-automated process to produce population distribution maps (Linard et al. 2012). Specifically for the population distribution weighting, 30 meter LandSat Enhanced Thematic Mapper (ETM) is the primary input, and there is limited use of varying scale vector data indicating populated places.

The AfriPop methodology relies on accurate subnational census data (population counts) as well as accurate open source settlement locations. However, particularly in the case of Africa, these are not always available. Similarly to GPW, AfriPop considers greater numbers (and smaller sizes) of subnational administrative areas to be a primary factor in improving the resolution of the output data. However, simply increasing the number of administrative areas will not improve the actual resolution of the output population distribution if those boundaries do not have a high level of geographical detail on their own. Since the output population distribution is fractional rather than integer, AfriPop generally has few areas of no population. That is, if an administrative unit has at least one person, that person will be divided up over the entire area. While population allocated to most areas may or may not be accurate indications of human activity space, one of the limitations of AfriPop is that it depicts population density as being greater in urban peripheral areas and villages rather than urban centers.

LandScan

Created in 1998, LandScan is a global population database depicting an ambient (24-hour average) population distribution. It was conceived as an improved resolution global population distribution database for estimating populations at risk. The LandScan methodology disaggregates subnational census information through a suite of novel and dynamically adaptable algorithms using spatial data, imagery derived spatial products, and manual corrections. LandScan exploits spatial data and imagery analysis technologies in a multi-variable dasymetric modeling approach (Dobson et al. 2000). LandScan data represents an average, or ambient, population that integrates diurnal movements and collective travel habits into a single measure (Dobson et al. 2000). Since natural or man-made emergencies may occur at any time of the day, the goal of LandScan is to develop a population distribution surface in totality, not just the locations of where people sleep. LandScan data is analogous to mapping biological habitat where the species total environment (e.g. nests, feeding areas, travel pathways, density gradients and boundary conditions) are considered (Guisan and Thuiller 2005).

Because of the ambient nature of LandScan, care should be taken with direct comparisons of LandScan data with other population distribution surfaces. Furthermore, since LandScan incorporates new spatial data and imagery into the distribution algorithms for each new version, comparing different versions of the dataset on a cell to cell basis may result to misleading conclusions. While some of the differences between LandScan versions are due to recently developed urban or suburban expansion, there are many cases where a village identified with high resolution imagery may have existed for hundreds of years, but was never represented in various spatial data products. Figure 3 illustrates the differences

between the original version of Landscan from 1998 and a later version, both of which are at a 30 arc-second resolution.

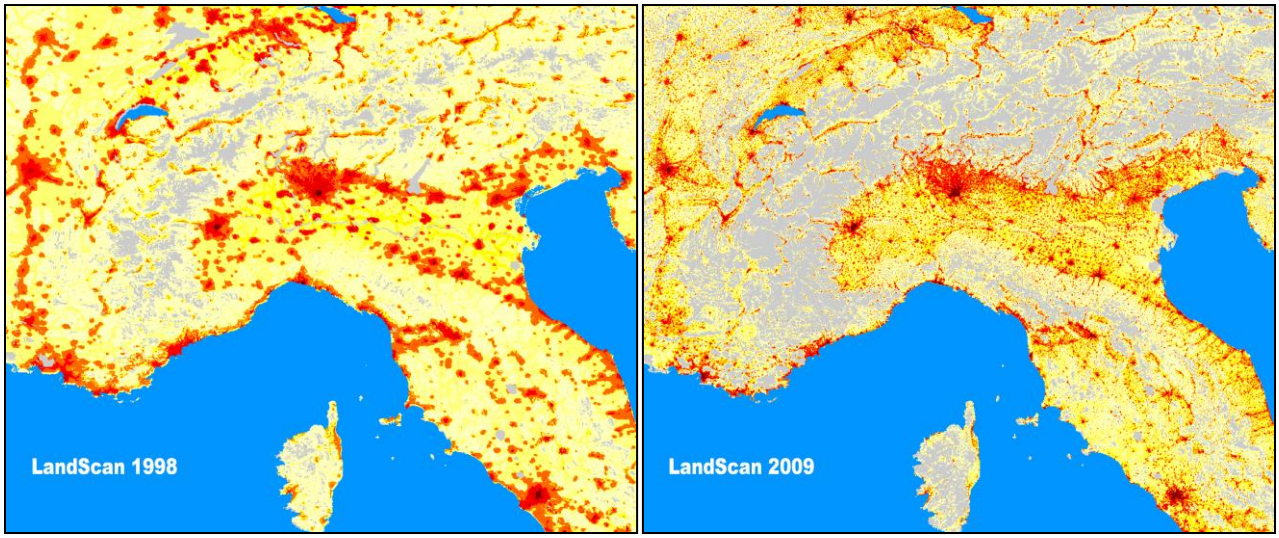

Figure 3. Evolution of LandScan.

Data and Methods

Census Data

LandScan, like other previously discussed datasets, relies on accurate subnational census data as the basis for population distribution. The LandScan development team expends significant effort each year to collect population censuses or estimates for each country from a myriad of open sources. While recent census tabulations are preferred, censuses can vary greatly in their regularity and level of execution. As such, official intercensal estimates, registration data, or other proxies may be used. Regardless of how subnational population counts are obtained, at the country level the population is always normalized to the mid-year national estimates provided through the CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/).

Once subnational population numbers have been acquired, those population counts are matched to their corresponding administrative boundaries. All LandScan data meet a pycnophylactic condition (i.e. mass preserving). The LandScan distribution algorithm uses a normalization process that ensures that the population counts distributed throughout each administrative unit will reflect exactly the number reported for that bounded area. After the initial population distribution, the procedure determines if either more or fewer people need to be added to the calculation of each administrative area in order to total to the subnational numbers. If more population is needed in an area, then population is added to the cells with the greatest likelihood coefficient; if less population is needed, then population is subtracted from the cells with the least likelihood coefficient. This procedure is repeated for each administrative area in a country. Since administrative units are used as spatial controls for population

distribution, the spatial and attribute accuracy of administrative boundaries are integral considerations for the LandScan model.

Subnational Administrative Boundaries

A required input for the development of any population distribution dataset is administrative boundaries, yet the availability of accurate digital subnational administrative boundaries remains a problem for many nations of the world. Accurate subnational boundary information has been identified as a significant need for population research and emergency response (NRC 2007). The United Nations Geospatial Information Working Group (UNGIWG) began work on a Second Administrative Level Boundary (SALB) dataset to globally map second level administrative boundaries. Due to differences in the quality of the documents and data compiled for the SALB, the spatial data layers of the database are more suitable for thematic mapping rather than precise representation or modeling. Another effort, the Global Administrative Unit Layers (GAUL) by the Food and Agriculture Organization (FAO) of the United Nations within the European Commission Food Security Programme, aims at "compiling and disseminating the most reliable spatial information on administrative units for all countries in the world" (FAO 2010). Unfortunately, many countries are not willing to share detailed subnational boundary information (Lauber 2007). Furthermore, much of the data volunteered in these repositories lack adequate spatial fidelity for accurate representations at 30 arc-second cell resolution. Numerous digital versions of subnational boundary information exist, but only the minority possesses the spatial accuracy desired for input into the LandScan model.

The annual update of LandScan requires that subnational boundaries for every country are examined with a critical eye. Often subnational boundaries coincide with natural features or circumscribe urban areas. The spatial accuracy and precision of administrative boundaries can be evaluated by overlaying the data on high resolution geo-rectified imagery. Some boundary datasets are topologically correct but not spatially accurate, with errors commonly exceeding multiple kilometers. Often subnational administrative boundaries have been greatly generalized for cartographic purposes which may result in villages, towns, and even portions of larger cities being located into the wrong administrative region. Digital representations of boundaries with inadequate detail create topological incongruities. For example, a city alongside a meandering river coincident with a boundary may be placed on the incorrect side of a generalized line segment. The census population that is associated with the city would not be reflected in the output because the erroneous representation would position the city in the adjacent administrative region. In effect, the population of one administrative region would be underrepresented and the population of the adjacent administrative region would be overrepresented. In other cases, either purposely or by accident, digital boundaries may simply be spatially incorrect.

Inconsistencies between census information and the concurrent subnational administrative boundaries are common. For many nations, subnational administrative boundaries or other census enumeration areas are dynamic. Often administrative boundaries move, merge, or change names for a variety of reasons. Recent census information may not be temporally synced to the existing boundary files. That is, the name of a region remains the same, but the area it represents may have changed. These types of discrepancies can be difficult to detect by automated quality assurance algorithms.

The administrative unit level by which the census data is distributed by the LandScan algorithm varies considerably in size and spatial precision from country to country. The number of administrative units per nation is considered in the LandScan model parameterization process. Nations with few, but very large administrative areas require a different weight in the model parameters to allocate representative populations to their appropriate locations. Generally smaller administrative boundaries lead to better population distribution – if the boundaries are spatially accurate. Small administrative areas that are spatially misplaced actually induce population distribution errors. To mitigate this, where possible, LandScan algorithms merge poor sub-province boundaries to the province level and distribute the entire province population according to the population likelihood locations determined by the model rather than constrict population distributions to incorrect locations. Very small administrative or enumeration areas equivalent to U.S. Census blocks or block groups have unintended consequences for modeling an ambient population. Since the populations associated with census tables are places of residence, commercial and industrial areas may have zero or very low populations associated with them. Thus the output would be reflective of a residential only population distribution instead of an ambient population distribution.

If they display reasonable geographic detail for mapping at 30 arc-seconds, the lowest level administrative units available are used in the weighting process, but they are not used to construct the pycnophylactic constraints. Given the intent to represent the distribution of persons, not simply the distribution of persons at night, using the lower level administrative units would not provide an accurate representation. Take for example a U.S. Census block level illustration; using the block unit which contains the Willis Tower in Chicago would yield an output population distribution with no people in it. Although this is a reasonable residential population representation, it does not in any way depict the true population activity of that block. Using higher level administrative units (larger areas) also has the benefit of smoothing out the noise in the distribution process in so that the weights are essentially a trend line.

Vector Spatial Data

Vector spatial data used in the LandScan model are chosen for their usefulness in predicting possible population distributions. The vector map (VMap) series distributed through the National Geospatial-Intelligence Agency (NGA) includes global coverage for a set of spatial features including road and rail networks, hydrologic features and drainage systems, utility networks, airports, selected elevation contours, international boundaries, populated places, and geographical names. VMap data is created by extracting spatial features from hardcopy maps at a consistent scale, converting these features into a digital representation of their locations and attributes, and finalizing the output to the Vector Product Format (VPF). The initial version of LandScan incorporated the VMap Level 0 series data. The primary source for the VMap0 data was the 1:1,000,000 scale Operational Navigation Chart (ONC) series co-produced by the military mapping authorities of Australia, Canada, United Kingdom, and the United States. The absolute horizontal accuracy of VMap0 data for all features derived from ONCs is 2,040 meters at 90 percent circular error.

At the time of the initial release of LandScan, VMap0 was the only global coverage available. Subsequent versions of LandScan incorporated NGA's VMap Level 1 data as it became available. VMap1

is a medium resolution database derived primarily from 1:250,000 Joint Operation Graphic hardcopy sources. Vector information collected from a 1:250,000 Class 1 source may be expected to be accurate to within 125 meters while certain other sources may result in accuracy of 500 meters on the ground (NGA 1995). For selected areas, LandScan employed both VMap Level 2 data (based on 1:50,000 scale maps) and Urban Vector Map UVMAP data (based on 1:7,500 – 1:25,000 scale maps). These higher resolution input data result in increased spatial precision and an increased quantity of various spatial features. However, VMAP products are not without problems. VMAP products are based on hardcopy maps created by various producers at different dates and therefore feature details are not necessarily consistent from one map to the next and features may not match at tile edges. The date of the original map compilations captured within a VMAP tile may, in some cases, be quite old, creating population distribution anomalies especially for rapidly developing urban areas. Vector data layers used in the LandScan modeling algorithms include roads, populated areas (urban boundaries), and populated points (towns and villages). Each data layer serves as an indicator of likely population locations. Spatial feature representation and accuracy is markedly improved from VMAP0 to VMAP1 to VMAP2 for all features used as inputs into the LandScan modeling algorithms.

Given their intricate spatial patterns, digital coastlines require a very high resolution to represent coastal features accurately. Coastal areas are dynamic landscapes; shorelines change and coastal islands may grow, shift positions, or disappear entirely. In some parts of the world, such as Dubai or parts of Yemen, new land is being created on the coast which is essentially pushing the coastlines seaward. Popular coastline databases may lack both the spatial resolution and the update frequency to capture these complexities, and instead, intersect current land features thereby potentially missing populated areas. For this reason, LandScan extends all coastal boundaries several kilometers seaward to ensure all shore and small island features are encapsulated within an administrative unit boundary. Instead of a vector shoreline, land cover data and imagery is used to capture populated areas along the coast. That being said, most subnational boundary data do not have enough geographic detail to allow for even finer resolution population distributions than LandScan that are currently being pursued. Errors of omission and commission are common problems with using poor quality boundary data, particularly near coastlines. Figure 4 shows an example of this in Dar es Salaam, Tanzania.

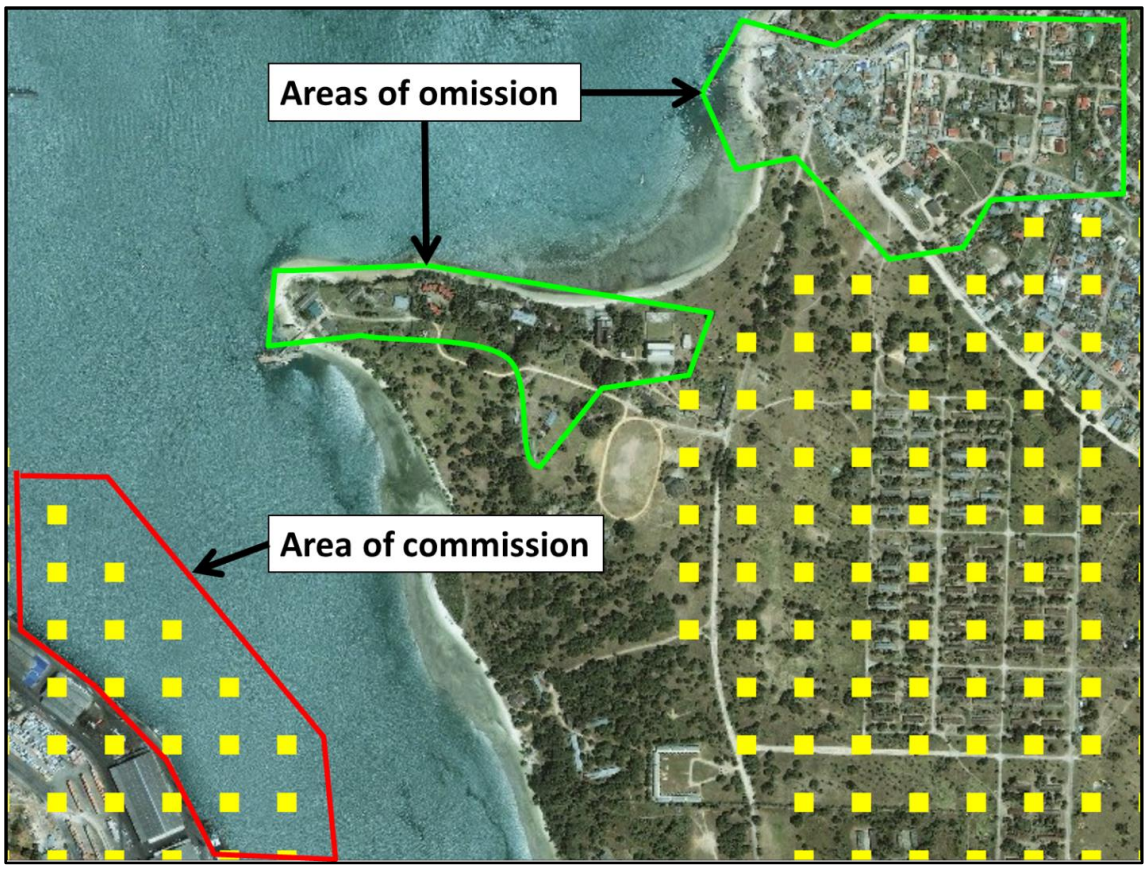

Figure 4. Example of errors of omission and commission that can arise when using subnational administrative boundaries with incorporated coastlines that lack appropriate geographic detail to develop very high resolution population distributions. Yellow squares mark the centerpoint of a 100m resolution population distribution.

Raster Spatial Data

Elevation and slope are incorporated into the LandScan modeling algorithms as most human settlements occur at low to moderate elevations on flat to gently sloping terrain (Cohen and Small 1998; Dobson et al. 2000). However, every continent has instances of settlements at high elevations. For some nations (e.g. Eritrea and Yemen) coastal regions may be particularly inhospitable for human habitation and settlements occur at higher elevations. Although on occasion humans will frequent the very highest elevations on the planet, there are elevation levels above which consistent habitation is impractical. Since settlements can occur at all but the very highest elevations, the LandScan model does not evaluate elevation other than to exclude population distributions at elevations over 18,000 ft.

Local area slope is often a good indicator of population likelihood. Picturesque hillside settlements are found in many hilly or mountainous regions of the world, but generally, very steep slopes inhibit settlements, agriculture, and industrial development which are the most likely areas of ambient population distribution. In mountainous areas, villages are normally found in valleys. However, because

cultural settlement patterns vary significantly from region to region with respect to slope tolerance and the relative availability of arable land, a general global regression-based weighting scheme is not ideal. In regions where farmland is a precious resource, the very flat areas are reserved for agriculture production and settlements are found on the periphery of nearby slopes.

The initial version of LandScan incorporated the DTED0 global elevation dataset into the population distribution model. Using DTED0, a single slope value was calculated for each cell by examining the elevation of the eight neighboring cells. Thus for approximately each square kilometer a single slope value was calculated and given a weight in the model according to the settlement patterns of each country. Calculating an average slope value for each 30 arc-second cell may mask many small, relatively flat areas within the cell that are suitable for habitation. Subsequent versions of LandScan processed NGA's Digital Terrain Elevation Data (DTED1). DTED 1 has near global land area coverage with resolution of 3 arc-seconds.

Compared to DTED0, the DTED1 data has 100 times more elevation values for each output cell. The elevation data from each DTED data tile were extracted, converted to raster format, projected to a Universal Transverse Mercator (UTM) projection to calculate slope values in appropriate units, and the calculated slope values were re-projected to the original geographic projection. Rather than average these high resolution slope values into an aggregated 30 arc-second cell slope value, a slope suitability ranking index was assigned to each 3 arc-second cell. By processing the higher resolution elevation data into discrete slope categories, cells with very high slope values do not disproportionately skew the average slope of an aggregated cell. An average of these 3 arc-second slope index values is calculated for each 30 arc-second cell. This methodology produces a slope suitability index by capturing the percentage of land that is preferable for habitation within each 30 arc-second cell. This discrete slope index technique accounts for some of the small areas with relatively gentle slope such as narrow valleys or knolls within an overall steep landscape. More recent versions of LandScan integrated the Shuttle Radar Topography Mission (SRTM1) elevation data. The SRTM data also has a resolution of 3 arc-seconds with an extent in latitude from 60 degrees North to 56 degrees South. The SRTM data was processed in a fashion similar to that used for the DTED1 data, and SRTM2 data were processed for limited areas.

Land cover is a crucial data layer for modeling population distribution (Tian et al. 2005, Mubareka et al. 2008). Signatures of anthropogenic activity are observed in all but the most remote places on earth. Dedicated land cover or land use classes such as urban, developed, residential, or even agricultural, reinforce the importance of human modifications on the landscape. LandScan analyzes each land cover dataset by comparing the data for each nation to settlement characteristics observed in coincident high resolution imagery. Relative weights assigned to different land cover types generally follow a logical progression (Urban > Agriculture > Grass/Forest > Desert > Ice/Water). Absolute weighting values are adjusted for different settlement patterns, data completeness, and spatial accuracy of the land cover data.

The values for the weights are amended considerably from one area to another due to cultural settlement patterns, economic activity, and history. For example, prime agricultural lands in the U.S.

generally have low population densities whereas principal agricultural regions in developing nations generally exhibit much higher densities. Even many medium resolution land cover databases do not capture thousands of smaller towns, rural villages, or kampongs, that have population densities rivaling that found in many developed cities. The land cover data associated with these agricultural areas receive higher population likelihood in the distribution model, not only for the missing villages, but also to represent the potential number of workers in fields.

High Resolution Imagery Analysis

Currently, image archives ingest terabytes of data per day (National Academies Press 2006). High resolution imagery is employed in every phase of the LandScan population distribution process. At the outset, high resolution imagery is used to identify settlement patterns and building characteristics, but is also used to evaluate the accuracy and precision of the different spatial data layers used in the models as well as to adapt the weighting factor for each layer in the model algorithms. Preliminary model output is superimposed on high resolution imagery to verify relative population distributions and magnitude. As new spatial data are received, iterative modifications to variable weights in the likelihood coefficient file are made and the distribution algorithms are re-calculated.

High resolution imagery is used to create or modify existing spatial data layers, especially to update or refine the land cover data related to urban boundary delineations. To speed processing of vast image archives, an automated urban boundary delineation algorithm based on texture and edge information extracted from high-resolution panchromatic images has been developed. Feature vectors using statistical features derived from gray level co-occurrence matrices (GLCM) and local edge pattern (LEP) matrices were extracted and deployed in a parallel computing environment to speed computation time (Cheriyadat et al. 2007). Additional automated extraction of urban features is accomplished by implementing a parallel algorithm for a Gabor filter based multi-resolution representation of panchromatic images to compute the texture features at different scales and orientations. The Gabor filter representation has been widely used for texture analysis as it can be shown that it minimizes the joint two-dimensional uncertainties in space and time (Vijayaraj, Bright, and Bhaduri 2007).

Positional or attribute errors and anomalies are to be anticipated in large volumes of disparate spatial data. LandScan includes a manual verification and modification process to improve the spatial precision and relative magnitude of the population distribution. Imagery analysts identify obvious population distribution errors and create an additional spatial data layer of population likelihood coefficient modifications to correct or mitigate input data anomalies. Derived land cover such as that from Thematic Mapper data can not reveal urban properties such as building densities or building heights that can be readily inferred with visual inspection using high resolution imagery. As a result, a large number of modifications are made to urban areas and urban extents.

Future Innovations

Although it may seem that dissecting the data and methodologies used in the creation of the LandScan population distribution database is simply a metadata exercise, it is an important start for understanding the limitations, as well as potential for improvement of these data, including new directions for

research. Despite the extensive use of Geographic Information Systems (GIS) to map spatial patterns and distributions of population, much less research has been undertaken with globally mapping the sociocultural properties of place including human experience, social hierarchies, and power relations. It is yet to be determined whether LandScan is an appropriate or plausible vehicle for this type of critical theoretical extension; the scale of the output alone suggests that it would be impractical, if not impossible, to foray into this realm using the current "top-down" approach. However, some recent research has been done using LandScan as the starting point for developing more complex population models. Specifically, Fernandez et al. (2010) experimented with building a credible synthetic population for Afghanistan consisting of 31 million social atoms. While the research centered on data and computational feasibility, the outcome showed promise in using social theories to model regions at the individual level.

The current spatial resolution of LandScan is appropriate for a global dataset and is more than adequate for use in national and subnational analyses. However, there is a continued effort to improve the resolution without sacrificing the availability of annual updates. To this end, work has been done in the area of automating both the extraction and characterization of human settlements in a high performance computing (HPC) environment (Cheriyadat et al. 2007). The Settlement Mapper Tool (SMT) developed by ORNL rapidly delineates and characterizes settlements using high resolution imagery. The application of this to LandScan is twofold. First, the extraction of human settlements will allow for a baseline understanding of settlement area vs. non-settled area across the globe. Second, the characterization allows a deeper dive by using low-level image features (i.e., edges, lines, textures, etc.) to characterize types of settlement structures (Graesser et al. 2012). In a remote sensing environment, neighborhoods can be classified as formal or informal using both spatial and spectral characteristics (Hofmann et al. 2008). While formal settlements tend to have larger building sizes, regular street patterns, and more homogeneous building materials, informal settlements are more likely to consist of small buildings, narrow streets, and heterogeneity of materials.

Another component of the ongoing efforts to improve the spatial, temporal, and demographic dimensions of LandScan, is the Population Density Tables (PDT). PDT is a set of national-level databases, which model density by the sociocultural activities and associated facility spaces that define normal patterns of life. Such a modeling exercise involves automated collection, and subsequent fusion of information from open, published sources including academic journals, official government statistics, websites of humanitarian organizations, corporate and university web pages, buildings databases, tourism brochures, reported field observations, and NGO reports. The result is a collection of data that provide a baseline, spatiotemporal and demographic snapshot of facility occupancy levels for nighttime and daytime periods at a higher spatial resolution. Activity spaces, which include a variety of buildings and open air facilities, are characterized by functional categories which include residential, cultural institutions, retail, commercial, and transportation spaces. Population density estimates for night and day capture a normal workweek and routinely scheduled periodic and episodic events. The PDT database also contains other population density values representing weekends, holidays, special events, and seasonal population distribution scenarios.

In light of efforts such as PDT and SMT, the advancement of population distribution models has moved beyond numerical input and toward attempting to quantify qualitative data. In the same sense that expert or local knowledge can inform intelligent dasymetric mapping, so too can the fusion of SMT and PDT type outputs produce a richer understanding of not just population distribution, but also the nature of settlement and the characteristics of the people and activities that take place within that settlement. The challenge and current focus then is to facilitate the convergence of remote sensing techniques and qualitative sociological data and methods in order to produce a richer understanding of the population and human activity of a particular area at very high spatial, temporal, and demographic resolutions.

Conclusion

What's missing from LandScan and other global population distribution datasets is the demographic depth that is most useful for analysis and policy formation. So while it can be argued that LandScan is enormously useful for addressing questions about the magnitude of population at risk in cases of natural disasters or conflict, it does not address the more complex relationships and characteristics of these populations that are tied to long term policy questions of health, education, and poverty.

The need for finer resolution spatiodemographic data for analysis is well documented in the literature. Analyses of health, accessibility, and poverty issues, of emergency response planning, and of political stability are all examples of where both population distribution and characterization must be key components, particularly in the developing world. In response to this need, work has begun to improve the demographic depth of LandScan beyond simply the age/sex breakdown that is currently available to LandScan users.

Linking demographic survey data such as the Demographic and Health Survey (DHS) to a population distribution database would unquestionably enable the use of both datasets in a wider variety of studies. Furthermore, a generalizable methodology for providing this linkage to LandScan Global rather than for specific cities or countries would provide a particularly useful dimension to the LandScan data, as well as a possible pathway for many researchers to replicate for their particular needs. To this end, current work is being undertaken to develop a prototype data fusion methodology to link DHS data to census data for a single country as a proof of concept, as well as to document challenges that arise with regard to uncertainty and validation issues.

Acknowledgements

This work was prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725.

The authors would like to thank Nicholas Nagle, who provided valuable suggestions that helped improve the quality of the manuscript. The authors would also like to thank Eric Weber for contributing Figures 1 and 2 to this manuscript.

References

Balk, D., & Yetman, G. (2004). The Global Distribution of Population: Evaluating the gains in resolution refinement. CIESIN, Columbia University, N.Y.

Bhaduri, B. L., Bright, E. A., Coleman, P. R., & Dobson, J. E. (2002). LandScan: Locating People is What Matters. *Geoinformatics, 5*(2), 34-37.

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A, 21*(4), 537-543.

Cheriyadat, A., Bright, E., Potere, D., & Bhaduri, B. (2007). Mapping of settlements in high-resolution satellite imagery using high performance computing. *GeoJournal, 69*(1-2), 119-129. doi: 10.1007/s10708-007-9101-0

Cohen, J. E., & Small, C. (1998). Hypsographic demography: The distribution of human population by altitude. *Proceedings of the National Academy of Sciences, 95*, 14009-14014.

Deichmann, U., Global Resource Information Database National Center for Geographic Information Analysis Consultative Group on International Agricultural Research (1996). *A Review of Spatial Population Database Design and Modeling*: National Center for Geographic Information and Analysis.

Deichmann, U., Balk, D., & Yetman, G. (2001). Transforming Population Data for Interdisciplinary Usages: From census to grid. CIESIN, Columbia University, NY.

Dobson, J., Bright, E., Coleman, P., Durfee, R., & Worley, B. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing, 66*(7), 849-857.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science, 28*(2), 125-138.

Fernandez, S. J., Rose, A. N., Bright, E. A., Beaver, J. M., Symons, C. T., Omitaomu, O. A., & Jiao, C. (2010, 20-22 Aug. 2010). *Construction of Synthetic Populations with Key Attributes: Simulation Set-Up While Accommodating Multiple Approaches within a Flexible Simulation Platform.* Paper presented at the 2010 IEEE Second International Conference on Social Computing (SocialCom).

Fisher, P., & Langford, M. (1996). Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping. *The Professional Geographer, 48*(3), 299-309. doi: 10.1111/j.0033-0124.1996.00299.x

Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *The Annals of Regional Science, 26*(1), 67-78. doi: 10.1007/bf01581481

Flowerdew, R., Green, M., & Kehris, E. (1991). USING AREAL INTERPOLATION METHODS IN GEOGRAPHIC INFORMATION SYSTEMS. *Papers in Regional Science, 70*(3), 303-315. doi: 10.1111/j.1435-5597.1991.tb01734.x

Fotheringham, A. S., & Rogerson, P. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Science, 7*(1), 3-19. doi: 10.1080/02693799308901936

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*, 383-397.

Goodchild, M. F., & Lam, N. S.-N. (1980). Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing, 1*, 297-312.

Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., & Bright, E. (2012). Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*(4), 1164-1176.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters, 8*(9), 993-1009. doi: 10.1111/j.1461-0248.2005.00792.x

Hall, O., Stroh, E., & Paya, F. (2012). From Census to Grids: Comparing Gridded Population of the World with Swedish Census Records. *The Open Geography Journal, 5*, 1-5.

Hofmann, P., Strobl, J., Blaschke, T., & Kux, H. (2008). Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach. In T. Blaschke, S. Lang & G. Hay (Eds.), *Object-Based Image Analysis* (pp. 531-553): Springer Berlin Heidelberg.

Lam, N. S.-N. (1983). Spatial Interpolation Methods: A Review. *The American Cartographer, 10*(2), 129-149.

Lauber, W.G. (2007). Political Geography and Emergency Relief. In *Tools and Methods for Estimating Population at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: The National Academies Press.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One, 7*(2), e31743. doi: 10.1371/journal.pone.0031743

Lo, C. P., & Welch, R. (1977). Chinese Urban Population Estimates. *Annals of the Association of American Geographers, 67*(2), 246-253.

Mennis, J., & Hultgren, T. (2006). Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science, 33*(3), 179-194.

Mondal, P., & Tatem, A. J. (2012). Uncertainties in Measuring Populations Potentially Impacted by Sea Level Rise and Coastal Flooding. *PLoS One, 7*(10), e48191. doi: 10.1371/journal.pone.0048191

Mubareka, S., Ehrlich, D., Bonn, F., & Kayitakire, F. (2008). Settlement location and population density estimation in rugged terrain using information derived from Landsat ETM and SRTM data. *International Journal of Remote Sensing, 29*(8), 2339-2357. doi: 10.1080/01431160701422247

National Geospatial-Intelligence Agency (1995). MILITARY SPECIFICATION VECTOR MAP (VMap) Level 1. http://earth-info.nga.mil/publications/specs/printed/VMAP1/vmap1.html. Last Accessed 9/27/2013.

National Research Council (2007). *Tools and Methods for Estimating Population at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: The National Academies Press.

Petrov, A. (2012). One Hundred Years of Dasymetric Mapping: Back to the Origin. *The Cartographic Journal, 49*(3), 256-264. doi: 10.1179/1743277412y.0000000001

Priorities for GEOINT Research at the National Geospatial-Intelligence Agency. (2006). The National Academies Press.

Tatem, A., Campiz, N., Gething, P., Snow, R., & Linard, C. (2011). The effects of spatial population dataset choice on estimates of population at risk of disease. *Population Health Metrics, 9*(1), 1-14. doi: 10.1186/1478-7954-9-4

Tian, Y., Yue, T., Zhu, L., & Clinton, N. (2005). Modeling population density using land cover data. *Ecological Modelling, 189*(1-2), 72-88. doi: 10.1016/j.ecolmodel.2005.03.012

Tobler, W. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association, 74*(367), 519-530.

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. (1995). The Global Demography Project (D. o. Geography, Trans.): National Center for Geographic Information and Analysis.

Vijayaraj, V., Bright, E. A., & Bhaduri, B. L. (2007, 23-28 July 2007). *High resolution urban feature extraction for global population mapping using high performance computing.* Paper presented at the Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International.

Wright, J. K. (1936). A Method of Mapping Densities of Population: With Cape Cod as an Example. *Geographical Review, 26*(1), 103-110.