GeoJournal (2007) 69:103–117
DOI 10.1007/s10708-007-9105-9

LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics

Budhendra Bhaduri · Edward Bright · Phillip Coleman · Marie L. Urban

Published online: 29 September 2007

Abstract High-resolution population distribution data are critical for successfully addressing important issues ranging from socio-environmental research to public health to homeland security, since scientific analyses, operational activities, and policy decisions are significantly influenced by the number of impacted people. Dasymetric modeling has been a well-recognized approach for spatial decomposition of census data to increase the spatial resolution of population distribution. However, enhancing the temporal resolution of population distribution poses a greater challenge. In this paper, we discuss the development of LandScan USA, a multi-dimensional dasymetric modeling approach, which has allowed the creation of a very high-resolution population distribution data both over space and time. At a spatial resolution of 3 arc seconds (~90 m), the initial LandScan USA database contains both a nighttime residential as well as a baseline daytime population distribution that incorporates movement of workers and students. Challenging research issues of disparate and misaligned spatial data and modeling to develop a database at a national scale, as well as model verification and validation approaches are illustrated and discussed. Initial analyses indicate a high degree of locational accuracy for LandScan USA distribution model and data. High-resolution population data such as LandScan USA, which describes both distribution and dynamics of human population, clearly has the potential to profoundly impact multiple domain applications of national and global priority.

Keywords Census · Daytime population distribution · Dasymetric modeling · High-resolution population · LandScan · Population dynamics

B. Bhaduri (✉) · E. Bright · P. Coleman · M. L. Urban
Oak Ridge National Laboratory, MS 6017, PO Box 2008, Oak Ridge, TN 37831-6017, USA
e-mail: bhaduribl@ornl.gov

Introduction

High-resolution population distribution data are essential for successfully addressing critical issues ranging from socio-environmental research to public health to homeland security (Dobson et al. 2000; Bhaduri et al. 2002, 2005; Chen 2002; Hay et al. 2005; Sutton et al. 2001). Commonly available population data, collected through modern censuses, are constrained both in space and time and do not capture the population dynamics as functions of space and time. From a spatial perspective, census data are limited by census accounting units (such as blocks), and there often is great uncertainty about spatial distribution of residents within those accounting units. This is particularly true in suburban and rural areas, where the population is dispersed to a greater degree than in urban areas. For the US, the source for

population data is the US Census Bureau, which reports population counts by census blocks (smallest polygonal unit), block groups (aggregated blocks), and tracts (aggregated block groups). At the highest resolution (block level), a uniform population distribution is assumed and the population values are typically an attribute of the block (polygon) centroids. Similarly, population values for block groups and tracts are reported at the centroids of the block group and tract polygons. In geospatial analyses, these points are used to represent the population of a census polygon. For example, calculation of travel time to health care providers considers these centroids as the origins and/or destinations for travel. For exposure and risk analyses, these centroids often serve as "receptor" points for calculating exposure or dosage from any dispersed agent.

In common practice, census data are intersected with buffers of influence (such as those from emission sources) using two primary approaches to quantify population at risk:

(a) Tally the entire population (if the centroid is inside the buffer) or zero population (if the centroid is outside the buffer)
(b) An area weighted population accounting approach (based on the ratio of the areas of the polygon included in and excluded from the buffer).

The first approach aggregates the entire population of an area to a single (point) geolocation. The second approach assumes the entire population of an area to be uniformly distributed over that area. In fact, non-uniform distribution of human population is quite obvious from simple visual observation of any landscape. From a temporal perspective, the resolution of census information is typically at anywhere between 1 and 10 year cycles. The spatial granularity of information for the decennial census is higher compared to the yearly updates. This can be logically explained by the original motivation for developing a census for social and economic planning activities aimed at medium to long-term solutions over a number of years. Consequently a general geographic assessment of population at relatively large time intervals, described through their residential locations, was adequate to address such planning processes (US Census Bureau 2000). However, with pressing needs for finer temporal resolution population distribution data for consequence assessment of natural and technological disaster events, usage of traditional census counts, represented as a "nighttime residential" population, in a daytime event simulation is irrational. Because of this uncertainty, there is significant potential to misclassify people with respect to their location, for example pollution sources, and consequently it becomes challenging to differentiate environmental exposure in specific sub-populations. These limitations, to a large degree, can be overcome by developing population data with a finer resolution in both space and time at sub-census levels. Geodemographic data at such scales will represent a more realistic non-uniform distribution of population.

Background

Spatial decomposition of census data

Spatial decomposition of census population estimates has been well studied over the last few decades. A number of interpolation and decomposition methods have been developed to address this issue with census (polygonal) population data. Among such approaches, areal weighting, pycnophylactic interpolation, dasymetric mapping, and various smart interpolation techniques are agreeably the most well recognized and widely discussed in literature. Areal weighted interpolation is the simplest approach and implies an assumption of uniform distribution of population. In this method, a regular grid is intersected with the census polygon and each grid cell is assigned a value based on the proportion of the polygon contained in each cell (Goodchild and Lam 1980; Flowerdrew and Green 1992; Goodchild et al. 1993). Pycnophylactic interpolation extends areal weighting methodology by applying a smoothing function to the raster cell values, with the weighted average of its nearest neighbors, iteratively while preserving the total population count of the polygon (Tobler 1979). The result of such interpolation develops a continuous population surface which disagrees with the noticeable discontinuous nature of population distribution. Dasymetric modeling is comparable to areal interpolation but utilizes ancillary spatial data to augment the interpolation process. The ancillary spatial data are at a finer spatial

resolution, and the variability and spatial discontinuity in their values enable an asymmetric and discontinuous allocation of population (Wright 1936; Langford and Unwin 1994; Eicher and Brewer 2001; Mennis 2003). Land cover/land use is the best example in this respect (Monmonier and Schnell 1984; Reibel and Agrawal 2006) where different land cover or land use categories for each cell can be used as a weighting function for population distribution such as urban areas which will have a higher weight than forested areas. Smart interpolation, in principle, is a multidimensional adaptation of the dasymetric model where the allocation refinement results from multiple ancillary data sources which are at a finer resolution compared to the population polygon (Langford and Unwin 1994; Cohen and Small 1998). In general, the utility of such interpolation techniques for enhancing the spatial resolution of population distribution data at local scales have been demonstrated (Sleeter and Wood 2006; Sleeter 2007) but few attempts have been made to create a model that is scalable from local to global. In fact, there are only a few well-known models that have been successfully employed to develop global population data sets, namely the Gridded Population of the World (GPW), the Global Rural Urban Mapping Project (GRUMP), and LandScan Global Population database. GPW is a product of simple areal weighting interpolation and GRUMP is derived through a simple dasymetric modeling. On the other hand, LandScan is structurally a multidimensional dasymetric model. However, as discussed in later sections of this paper, the model is not just limited to pre-determined spatial operations among input data variables. It involves a significant level of analyst intervention to validate input data and modeling parameters, as well as to improve precision of the model output based on local knowledge. Mennis and Hultgren (2006) have recently referred to this type of modeling approach as "Intelligent" dasymetric modeling.

Temporal resolution of population distribution

Human population distribution behaves as a function of both space and time. However, the spatial aspect of population distribution has received the most attention of the interpolation methods. Population distribution, as it directly relates to various human activities, can be functionally described by the various demographic groups representing those activities (Fig. 1). Mobility of population from their residences results from temporary relocation to places of daytime activities that include places of education (schools, colleges, and universities), employment, businesses (shopping, post offices, restaurants, and others), or recreational areas (parks,

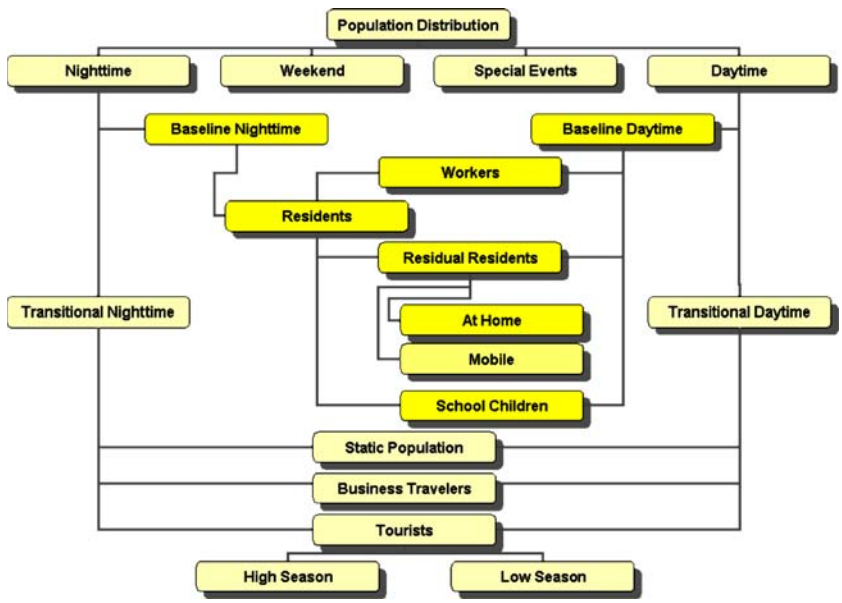

Fig. 1 Population distribution model components. LandScan USA version 1.0 follows the baseline nighttime and daytime population distribution including the static population

museums, and other tourist attractions) during the day (Quinn 1950; US Census Bureau 2000). In general population distributions of an area can be conceptually described as:

$$\begin{aligned} \textit{Nighttime Population} &= \textit{Nighttime Residential Population} + \textit{Nighttime Workers} + \textit{Tourists} + \\ &\quad \textit{Business travelers} \; (+ \textit{Static Population}) \end{aligned} \tag{1}$$

and,

$$\begin{aligned} \textit{Daytime Population} &= \textit{Workers} + \textit{School children} + \textit{Tourists} + \textit{Business travelers} + \textit{Residual} \\ &\quad \textit{Nighttime Residential Population} \; (+ \textit{Static Population}) \end{aligned} \tag{2}$$

It follows that, quantitative estimates of temporal population distribution involve two distinct elements; the identification of activity locations such as businesses, schools, and other recreational activities, and the second addresses the identification and distribution of the mobile population that are at those locations. From a modeling perspective, it is easier to gather data on the activity locations as static geographic features that are commonly captured in public and commercial databases for various infrastructures, or can be derived from remote sensing based land cover data, high resolution satellite and aerial photographs, or state and local government data (Forster 1985; Harvey 2002a, b). It is extremely challenging to quantify the number and nature of the mobile population that comprehensively captures the net displacement of residential population during the daytime or nighttime. Although, detailed population movement data sets may be available for selected local communities or even urban areas, they are not available at a national scale. In fact, the US Census Bureau's compilation of Journey to Work data are the only readily available and nationally consistent data set for the US that describes people's movement from residences to employment locations. Consequently, the US Census Bureau's estimate of daytime population based on the 2000 Census only reflect populations based on travel to work (US Census Bureau 2000). Similarly, it does not limit the work related commuting to specific hours. All worker-related travel, irrespective of what time of the day it occurs, has been used to derive these estimates of daytime population.

Modeling population dynamics

The temporal dynamics of population are well realized (Quinn 1950) and understood but very few attempts have been made to incorporate such temporal variability in a population distribution model and database. The LandScan Global model and database is the earliest example where the diurnal change in population distribution due to employment was originally captured (Dobson et al. 2000, 2003). The LandScan Global model assigns non-zero likelihood factors to areas where the land use indicates a non-residential, work related activity (such as commercial, industrial, and agricultural). Consequently, the LandScan Global Database represents an "ambient" or average population distribution over a 24 h period. In 2000, Oak Ridge National Laboratory extended the LandScan Global Population distribution model and developed the LandScan USA model, a very high resolution and scalable population distribution model for the US. At a 3 arc-second spatial resolution, it is not only the highest resolution national population distribution data ever produced, but also the first to isolate and capture the diurnal population dynamics individually. Thus the LandScan USA database contains a nighttime residential as well as a daytime population distribution data set. Subsequent to LandScan USA, the only other known effort to develop an analogous model has been at the Los Alamos National Laboratory (McPherson and Brown 2004; McPherson et al. 2006) that developed nighttime residential and daytime distribution data at a 250 m resolution. In 2000, the original prototype for LandScan USA was developed for the US Environmental Protection Agency (USEPA) and Department of Energy (DOE) for 25 southeastern counties in Texas. In the following years, the model was

extended to other areas in the US; first to the 133 larger urban areas for the Defense Threat Reduction Agency (DTRA) and then for the 99 counties in Iowa for the National Cancer Institute (NCI). Development of a national database was formally initiated for the Department of Homeland Security in 2005 and LandScan USA version 1.0 was completed in 2006 that covers the United States and Puerto Rico.

Methodology

As discussed earlier, LandScan Global and LandScan USA can be considered, at one level of abstraction, as multi-dimensional dasymetric or smart interpolation models. Since their inceptions, these have been evolving models with incremental development of spatially refined population distribution algorithms resulting from the availability of newer or higher quality input data sets to the model parameters. For example, the LandScan USA model was initially developed around publicly available input data sets but later some commercial data sets were utilized for higher spatial accuracy and greater information content. The general methodology and specific implementations (including LandScan) of both dasymetric and smart interpolation techniques are very well illustrated and documented in the literature (Dobson et al. 2000; Eicher and Brewer 2001; Mennis 2003; Mennis and Hultgren 2006). Thus, in this paper, we describe the general motivation and modeling principles behind LandScan USA in terms of capturing the temporal dynamics of population, particularly in light of the input data sets utilized for the LandScan USA version 1.0 database (Table 1).

The overall motivation behind LandScan USA is to develop a nationally to globally scalable population distribution model that adequately represents the nighttime and daytime population distributions at a very high spatial resolution of 3 arc-seconds. Thus our goal has been to select input data variables for the model based on the availability of national data sets and not to create model parameters based on the potential availability of data sets at the state and local levels. However, higher quality input data for established model parameters are often available from state and local agencies and these are utilized to enhance the quality of the data for those locations.

The LandScan USA model is essentially composed of two different components: one addressing the nighttime population distribution and the other addressing the population dynamics leading to a daytime population distribution. Further, in version 1.0 we have only accounted for the nighttime residential population and the baseline daytime population. The static population in an area is represented by the prison population and is accounted for in both components. However, since school aged children are located at K-12 school locations and colleges and universities are populated with expected student population, LandScan USA version 1.0 is representative of a regular weekday when academic institutions are in session as opposed to a weekend day. This daytime population distribution includes no business travelers or tourists, but are being included for future releases of the data set. Figure 2 shows an example of LandScan USA nighttime and daytime population distributions.

Nighttime population distribution

Estimating residential population distribution

Census population data serve as the nucleus of the LandScan USA model and the model is resolved to each census block with the goal being to maintain the integrity of the Census Bureau data at the block level. A census block is divided into finer grid cells (1 arc-second or 30 m) and the total population for the block is then allocated to the cells with weights proportional to the calculated likelihood (population coefficient) of being populated. Relative weights are empirically assigned to each cell for a number of data layers and all weights assigned from different data layers are combined to develop a cumulative weight for each cell (in a i, j matrix of cells) as follows:

$$W_{Cell\,i,j} = LC_{i,j} \times PR_{i,j} \times PRR_{i,j} \times S_{i,j} \times LM_{i,j} \times PRKS_{i,j} \times SCH_{i,j} \times PRSN_{i,j} \times ARPT_{i,j} \times WTR_{i,j}$$

where *LC*, Weight for Land Cover; *PR*, Weight for proximity to roads; *PRR*, Weight for proximity to rail roads; *S*, Weight for slope factor; *LM*, Weight for landmark polygon feature; *PRKS*, Weight for parks, *SCH*, Weight for K-12 schools, *PRSN*, Weight for prisons; *ARPT*, Weight for Airports; and *WTR*, Weight for water bodies.

Table 1 General description of input data utilization within the LandScan USA model

Input data	Source	Source date	Database	Night time	Day time	Data type
Population	US Census	1990, 2000	Block level Census	✔	✔	Table
		2000	Tract to Tract Journey to Work	✔	✔	Table
		2004	County Estimates		✔	Table
	Bureau of Labor Statistics	2004	County Workforce		✔	Table
Transportation	TeleAtlas	2004	Roads	✔		Vector
			Railroads	✔	✔	
			Airports	✔		
Cultural	TeleAtlas	2004	Large Area Landmarks	✔	✔	Vector
			Water	✔	✔	
			Recreation Areas	✔	✔	
			Retail Centers		✔	
Administrative boundary	TeleAtlas	2004	Counties	✔	✔	Vector
			Tracts	✔	✔	
			Zip Codes		✔	
			Census Blocks	✔	✔	
Land cover	USGS	1992, 2001	National Land Cover Data (NLCD)	✔	✔	Raster
Elevation	NGA	2001	DTED1 Elevation Data	✔	✔	Raster
Employment	Business Info Solutions	2004	InfoUSA		✔	Vector
Education	NCES	2004	National Center for Educational Statistics	✔	✔	Vector
	ESRI			✔	✔	
	TeleAtlas			✔	✔	
Jails and prisons	Federal Bureau of Prisons		http://www.bop.gov/	✔	✔	Table
	Prisoner Life.com		http://www.prisonerlife.com/prisondirectory.cfm	✔	✔	
	Prisoners of the Census		http://www.prisonersofthecensus.org/resources/prisons2000.html	✔	✔	
	State and County Websites			✔	✔	
	National Jail Census					
Imagery	Google Earth and Maps		http://www.maps.google.com	✔	✔	Image
	MSN Virtual Earth		http://www.local.live.com/	✔	✔	
	Yahoo Maps		http://www.maps.yahoo.com/	✔	✔	

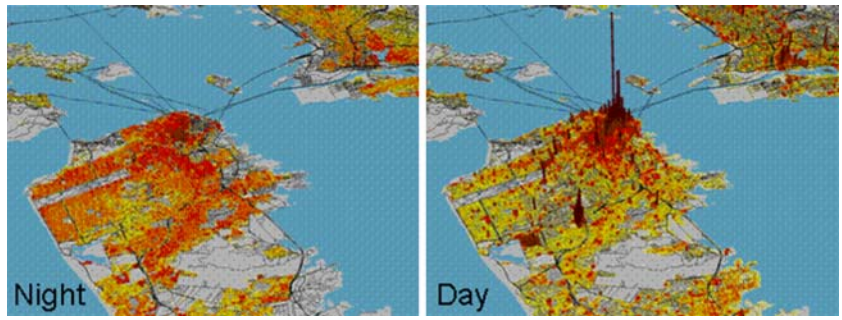

Fig. 2 LandScan USA nighttime and daytime population distributions for San Francisco, California

Once the individual cumulative cell weights are derived, these are combined and weighed with respect to the total population of the block to develop a block level population (or likelihood) coefficient as follows:

$$PC_{Block} = \frac{Total\ Population_{Block}}{\sum_{1}^{n} W_{Cell\,i,j}}$$

where *PC*, Population Coefficient and *N*, Number of LandScan USA cells describing the block.

Subsequently, the total population for that block is then allocated to each cell weighted by the calculated likelihood (population coefficient) of being populated as shown below:

$$Population_{Cell\,i,j} = PC_{Block} \times W_{Cell\,i,j}$$

For version 1.0, the decennial (2000) census block data was interpolated with the Census Bureau's estimated county totals for July 1, 2004 to be representative of a 2004 timeframe. Rather than simply increase or decrease all the blocks within an entire county at the same county level percent change for 2004, we calculated the percentage of population change for each tract within a county between 1990 and 2000, taking into account that sometimes tract boundaries mismatch between the two census years. This percentage for the tract was then prorated to the blocks within each tract, and finally the block numbers were normalized in order to sum to the Census Bureau's estimated July 2004 county total. Simply prorating the inter-census growth across a county would add proportionally more people to the blocks containing the greatest population (in all likelihood these blocks would already be fully developed). The goal was to locate the areas of the county that were experiencing growth in the preceding years and continue that growth from 2000 to 2004. This particular approach is spatially less explicit, but a better alternative to the assumption of uniform population change. Currently we are exploring new ways to extrapolate the growth within a county using new land cover or parcel data.

In the LandScan USA model, each census block is characterized using the land cover data to estimate the individual percentages of urban (residential, commercial, and industrial classes) and non-urban (agricultural, forests, and other classes) along with the census block population and number of housing units. Based on these evaluations, each census block is allocated to a sub-model that uses a specific allocation algorithm that relates such characterizations to cultural and settlement geographic understandings (Table 2). Additional spatial data for transportation (roads and railroads), physiography (water and steep slope); cultural landmarks (such as businesses, religious institutions, and schools) are then used to reclassify the LandScan USA grid cells to different likelihood values for human habitation.

The block level population counts are considered as a "control" to allocate the population within each block; i.e. the sum of the estimated populations in each cell from a block is constrained to equal the census block population, and all calculations are performed at the 1 arc-second level. Then, the cells are aggregated up to 3 arc-seconds resolution for the final output to account for spatial data geo-location and misalignment errors.

Estimating static (prison) population distribution

Initially prison locations are compiled from the National Jail Census and Tele Atlas database. These

Table 2 Sub models used within the LandScan USA model

Sub-model	Model type/ characteristic	Population presence	Population/housing unit	Population/ 1 sec cell	Population/ developed 1 sec cell	Percent developed	Percent water
1	Water/Wetlands	Yes	NA	NA	NA	NA	100
2	Rural	Yes	Low-Medium	Low-Medium	NA	None	<100
3	Dense Urban	Yes	High	High	High	High	<100
4	Suburban	Yes	Low-Medium	Low-Medium	Low-Medium	Medium	<100
5	Atypical Population Densities	Yes	Very High	Very High	Very High	Varies	<100
6	No Population	No	NA	NA	NA	NA	<100
7	Suburban-Questionable LC	Yes	Low-Medium	Low-Medium	Low-Medium	Low	<100
8	Questionable Vector Input	Yes	NA	NA	NA	NA	<100

locations have been further refined by visual identification, and correctional database websites (Table 1).

Daytime population distribution

$$Daytime\ Population = Nighttime\ Population + Daytime\ incoming\ population - Daytime\ outgoing\ population$$

Census block data do not explicitly report prison population, but indicate the possibility of such a facility with high population counts and low to zero housing

Deriving a quantitative estimate from the above qualitative expressions involves further analyses of population data which can be represented as: [1]

$$\begin{aligned} Daytime\ Population = {} & Nighttime\ Residential\ Population - Workersleavingduringtheday + \\ & Workingmovinginduringtheday - Schoolchildrenleavingduringthe \\ & day + Schoolchildrenmovinginduringtheday\ (+ Touristsvisiting \\ & duringtheday + Business\ travelerscomingintothearea) \end{aligned}$$

numbers. Such blocks are identified for further analysis. The initial static population distributions for those blocks are visually verified with high-resolution imagery and topographic maps. If a prison location is confirmed, its outline is digitized to cover the facility and grounds and the block population is updated. The prison population is then verified through various correctional database websites including the Federal, State, County and other confirmed database websites. In some cases, a prison or correctional facility location is found in an adjacent block and the population for that block is updated.

Although this may not be the most accurate or comprehensive representation, we consider this to be the expression leading to the best available daytime population estimates.

Estimating worker distribution

Workers are estimated at the block level by using a combination of a top-down and bottom-up approach.

[1] Not included in LandScan USA version. 1.0

The total number of workers for a given county from the Bureau of Labor Statistics (BLS) is prorated to the tract worker total reported by the Census Bureau's tract-to-tract worker flow table. The workers for each block are also estimated using the total worker population reported by the InfoUSA database (Table 1) for each block. Finally, these block estimates are then prorated to the final block worker counts using the tract totals. There are many instances where businesses are geocoded to zip code centroids rather than to specific addresses. In those situations, special measures are taken to distribute workers to blocks within the zip code that contain a high percentage of developed land cover with zero population (i.e. likely commercial or industrial areas).

Estimating K-12 school children and higher education students distribution

For distributing students at school, the census numbers for the demographic group between 5 through 17 and half of the 18 years age group are considered to be representative of the K-12 age population and it is assumed that all school children go to schools within the county they live. Due to a lack of readily available data at the time of model development, among the school age population 1.5% were assumed to be home-schooled and 3% were assumed to be sick, delinquent, and etc. These were subtracted from the total number of school-aged children. The sum of K-12 enrollment values in the National Center for Education Statistics (NCES) dataset is also subtracted to obtain the number of students available to be dispersed to the remaining NCES schools. If individual school enrollment data are available, the appropriate numbers of enrolled students are distributed to the specific school locations. The schools with missing enrollments are assigned weight values of 1 for elementary schools, 2 for combined and middle schools, and 3 for high schools. These weights were designed to reflect the hierarchical scaling of several elementary schools into a single middle school, and several middle schools comprising a high school. The difference between the number of children attending school in the county and the total number of students from the enrollment data are proportionally distributed to schools with missing enrollments using the weights assigned to elementary, middle, and high schools. Students attending colleges, universities and other post secondary institutions and living away from home in dormitories, apartments, and houses are typically accounted for in the census and are reflected in the nighttime model output. During the daytime, the numbers of students attending colleges or universities are found using the NCES and Tele Atlas databases for post-secondary institutions and are distributed to their corresponding locations. Where campus boundary information is available, high-resolution imagery is utilized to disaggregate population to appropriate settlement spaces (such as buildings).

Iterative data refinement process

For both nighttime and daytime population distributions, understanding the impacts that input spatial data anomalies (e.g. geolocation errors, misclassifications, and data currency) may have on the model output, a visual verification and modification process is employed to improve the spatial precision of the population distribution. Each county is evaluated by a GIS analyst by comparing the model output to high-resolution imagery and checked for obvious discrepancies. During this process, the analyst may make corrections to the locations of point features (e.g. schools or businesses) or make areal modifications to the population distribution (e.g. decreasing the distribution in a parking lot and increasing that of the adjoining apartment building). All modifications are captured as an additional input layer for the model and the population distribution model for that county is run again. These iterations continue until the analyst is satisfied with the spatial fidelity of the model output within permissible limits of time and budget constraints. For example, using high-resolution imagery, school locations are verified and subsequently the NCES database website is consulted to verify and update school enrollments. Often, some schools are found to have been closed and such school population is deleted from the daytime population distribution.

Results and discussions

It is important to realize that there is an element of subjectivity in dasymetric modeling or smart

interpolation as this approach assigns empirical weighing factors to individual data layers and to different data sub-categories within each data layer. This human element of the modeling process implies an inherent variability in the model results as different analysts independently attempts to create a model for the same area. Such variability will primarily be reflected in the predictability of an unpopulated cell based on the assumption of suitability for human habitation. For example, one version of a model may assume no residential population in agricultural areas where another may assume some residential population in those areas. This can be attributed to the signatures of smaller human settlements within larger agricultural areas that may not be detected due to an omission error in the land cover data as they are derived from remotely sensed satellite images. The more expected variability will result from using different "weighting scales" for individual data sub-categories. For example, weights to different land cover classes can be assigned in numerous ways as long as the assumptions about the relative or mutual relationship among the different sub-classes in terms of relative likelihood of population in those respective classes are preserved and consequently forces differentiating population counts to those respective sub-classes. In the real world, such spatial variability is quite expected resulting from the physiographic, socioeconomic, and cultural disparities across geographic scales. Thus the model fidelity is likely heavily influenced by modification of model parameters and assumptions based on "local" knowledge of the area being modeled. In practice, it is almost impossible to validate any of these assumptions and the predicted population counts at the individual cell level. However, the ability of a model to predict populated and unpopulated areas can be evaluated by comparing the model results with a data set that illustrates human settlements at the same or higher resolution than the input data used in the model.

LandScan USA verification and validation

True validation of any spatial model can only be achieved through ground truthing or validation. This is impractical for a dasymetric model such as LandScan USA for several reasons. First, ground validation will essentially repeat the data collection process as done in the general census. Other than the enormous resource constraints, this process is implausible due to the sensitivity associated with individual privacy issues. However, it may be possible to assess the model fidelity by comparing the results with detailed raw census data accessible through a secure process at the Census Bureau and should be addressed in future research. At present we address the verification and validation process and data integrity assessment in the following ways.

Quality assurance of input data sets

Input data sets are evaluated for possible errors or anomalies through automated algorithms as well as through manual verification. This minimizes the introduction of errors and inaccuracies into the modeling process. Assessments of population and land cover data sets are achieved through mutual comparison. Using the census block population and the land cover areas classified as residential, population densities are derived to evaluate large deviations from a normally expected range to identify possible data errors such as mis-tabulation of census data. The number of housing units and average household size from the census data as well as high-resolution satellite imagery are utilized to investigate data anomalies. For example, often the land cover data are older than the census population data and possible urbanization (as verified by imagery) helps explain a higher than expected population density. In some cases, an unusual population density is explained by the presence of static (prison) population which can be verified by the prisons database and high-resolution imagery. Similarly, an extensive spatial accuracy assessment is performed for other input data sets (schools, prisons, work and business locations) using census data and visual interpretation of orthophotographs.

Model resolution

By resolving the model at a census block resolution, we attempt to spatially restrict any possible irregularities resulting from model assumptions within the

block boundary and eliminating impacts to larger enumeration areas.

Cross validation with census data for assessing spatial integrity

In lieu of validation based on field collected and statistically robust data sets, a regression based correlation analysis based test can be designed to test the spatial integrity of LandScan USA relative to census data. This type of correlation analysis is inherently spurious since LandScan USA data is being correlated with census data which is used in the LandScan USA model. However, the objective of this correlation analysis is to illustrate that should a user choose to use LandScan USA data at the census block (or higher) level, the total population will be virtually equivalent, and hence imply the same level of user confidence for both data sets. Since the original distribution is developed and normalized exactly to the total block population at 1 arc-second and subsequently aggregated up to 3 arc-seconds resolution; an anticipated consequence of such spatial disaggregation-based modeling is a potential mensuration error. Comparison of block, block group, tract, or county population estimated from intersecting LandScan USA (with the corresponding boundaries) and those reported from the Census Bureau can provide an understanding of any measurement differences during the spatial decomposition process. Based on the distribution of the US counties with respect to their median block size and census population count (Fig. 3), Ellis County, Oklahoma and Los Angeles County, California were chosen as representatives of rural and urban landscapes. For these two counties, an evaluation of the census blocks with different LandScan USA sub-models (Table 2) criteria also attest to such rural and urban characteristics (Fig. 4). LandScan USA nighttime or residential data were intersected using census block boundaries rasterized to 3 arc-second cells, and the resulting LandScan USA population counts were compared against the adjusted census block populations. Results indicate a significant level of correlation between LandScan USA and the rasterized census block populations (Table 3) with correlation coefficients of 0.87 and 0.93 for Ellis and Los Angles Counties respectively (Fig. 5a, b). When a similar analysis is performed with the 3110 counties of the 48 contiguous

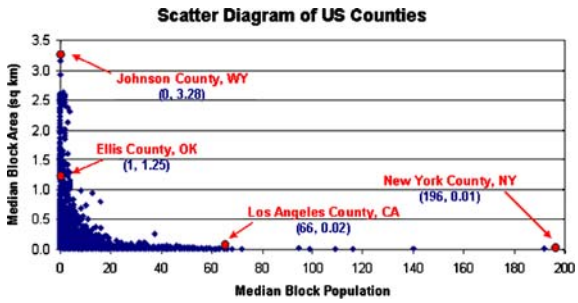

Fig. 3 Characterization of the US counties based on the median block population vs. the median block area. The urban counties tend towards a smaller block size (New York County, New York) whereas the rural counties tend towards a larger block size (Johnson County, Wyoming)

US (i.e. intersecting LandScan USA county boundaries and comparing with the adjusted census population counts), the results indicate a high degree of correlation for the LandScan USA model (Fig. 5c).

Cross validation using imagery

A goal for LandScan USA is to provide additional spatial accuracy of population distributions at the sub-census level. Although a numerical accuracy assessment of LandScan USA data at the individual cell level is impossible (i.e. validate the number of people predicted for each cell), it is possible to assess the locational accuracy and precision of the model and data. With very high resolution orthophotographs, 964 geocoded house locations were compared with an earlier version of LandScan USA data in Iowa (Cai et al. 2006). Using a spatial sensitivity filter of 90 m, the analysis indicated 72.5% accuracy in predicting populated cells over house locations and 99% accuracy in predicting unpopulated areas. Increasing the sensitivity filter to 180 m dramatically increases the accuracy level to 88%. It should be noted that the earlier version of LandScan USA used in this assessment was developed with the TIGER (Topologically Integrated Geographic Encoding and Referencing system) roads data set which has a relatively large spatial error. Though this analysis has not been repeated with the LandScan USA version 1.0 data, it will likely indicate a very satisfactory locational accuracy levels because it was developed with much higher quality data (including roads).

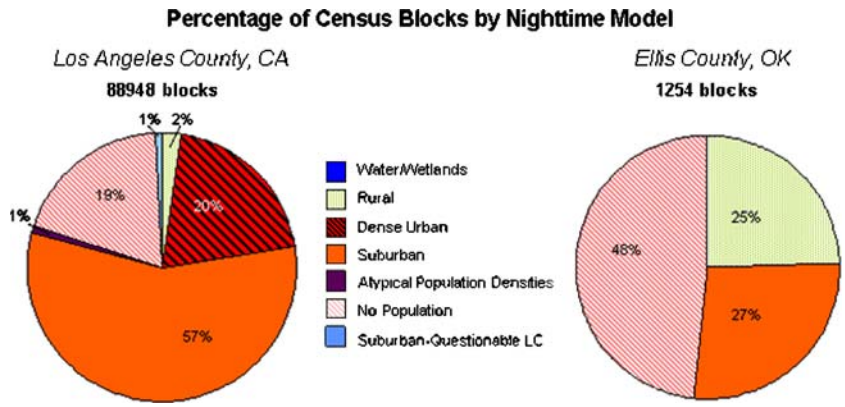

Fig. 4 Census blocks are modeled individually based upon their unique characteristics. The block sub-model frequency difference is readily apparent between an urban and rural county, as shown in Los Angeles County, California and Ellis County, Oklahoma

Table 3 Regression analyses output by block for a rural and an urban county; and by county for all counties within the contiguous United States

Spatial unit	Census population		LandScan USA population		
Ellis Co, OK	3,996		4,001		
Los Angles, CA	9,871,506		9,871,490		
Spatial unit	No. census blocks		Coefficient	R^2	p-value
Ellis Co, OK	1,254				
		Constant	0.396	0.812	0
		LandScan	0.873		
Los Angles, CA	82,948				
		Constant	1.228	0.929	0
		LandScan	0.972		
Contiguous US	3,109				
		Constant	11.576	0.999	0
		LandScan	1.000		

Applications of LandScan USA

High resolution population data serve as the nucleus to numerous application domains of national and global significance ranging from homeland security to transportation planning to socio-environmental studies. With daytime and nighttime (residential) population distributions, LandScan USA potentially magnifies the utility of high resolution population data across such a broad range of applications. Among all, disaster and consequence management, public health, and socioeconomic analysis are the three areas where the impacts are immediate and most significant. The unpredictable nature of technological and natural disasters put a large number of "unwarned" populations at risk. LandScan USA has become an integral part of homeland and national security through emergency preparedness and response including rapid risk assessment, evacuation planning, and relief delivery. Exposure analysis for public health and socio-environmental (environmental justice) studies can have tremendous benefits from LandScan USA data. In spatial epidemiology and disease (cancer) mapping, the utility of LandScan USA has been well illustrated (Cai et al. 2006). It is realized that activity based and time specific high resolution population distribution data will be of great advantage for socioeconomic (research and commercial) applications for evaluating the potential for Location Based Services (LBS) such as access to healthcare or coverage for wireless and cellular phones. Detailed discussions of the

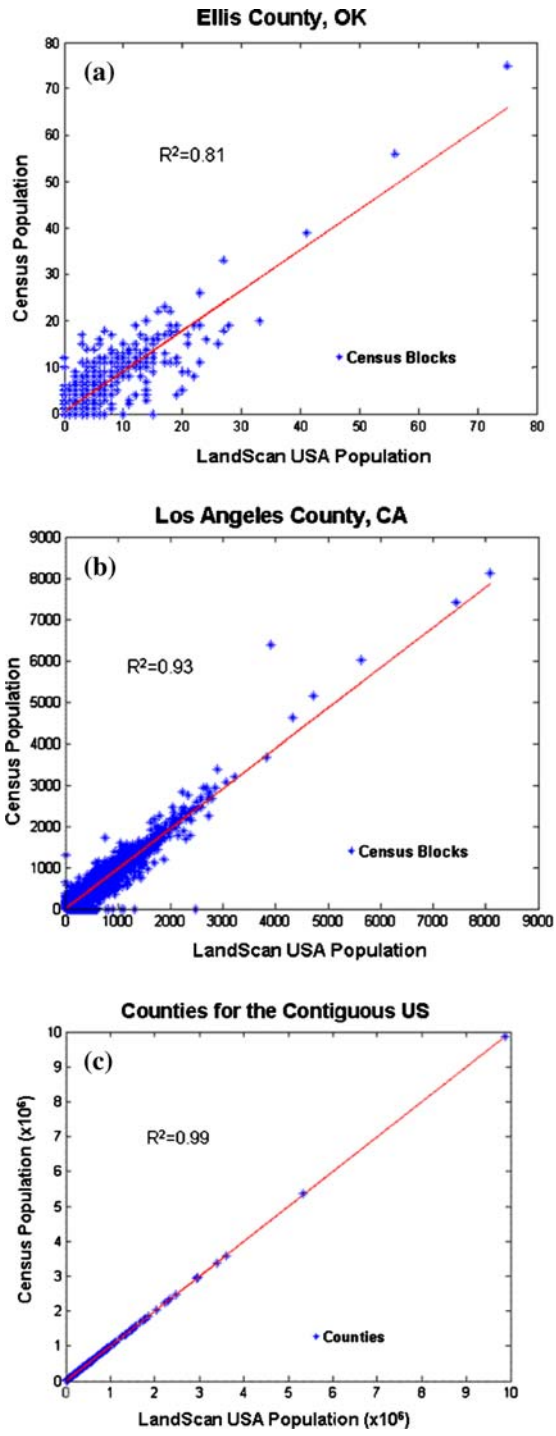

Fig. 5 Regression analyses of a rural county (Ellis County, Oklahoma), an urban county (Los Angeles, California), and all counties of the contiguous United States depicting the error induced by aggregating the population distribution cells from 1-arc second to 3-arc seconds

applications of LandScan model and data can be found elsewhere (Bhaduri 2007; Bhaduri et al. 2002, 2005; Chen 2002; Dobson et al. 2000; Hay et al. 2005; Sutton et al. 2001).

Future research

Geospatial and temporal dynamics of population are complex social processes. Consequently, effective characterization of such population dynamics requires development of high resolution spatial and temporal models that adequately capture social complexity and its influence on human movement patterns. As the resolution of available spatial data increases (for example parcel level data are now being collected and distributed by most state and local governments), it is logically possible to increase the resolution of population distribution models to the corresponding resolution. However, characterization of population with respect to functional space, such as indoor and outdoor population, will be an important aspect to investigate. Understanding and modeling the temporal resolution is a more complicated issue as the periodicity in the definition of a temporal resolution can greatly vary. The (average) representation of temporal resolution can range from a simple average nighttime and daytime distribution to hourly (and finer), weekly, monthly, seasonal, and yearly time frames. However, such approaches need to be high (spatial and temporal) resolution data driven so that the impacts of weather, climate, seasons, and special population (cultural, social, and political) events are adequately accounted for in the average distribution. At present, LandScan USA represents only an average working day population distribution. However, LandScan USA general framework is a computationally intensive approach where the goal is to develop population distribution and dynamics models at the highest possible spatial and temporal resolutions and then aggregate the results to derive the best possible average representations over larger time periods. For example, developing an average workday population distribution could be developed from hourly representations during such a day. LandScan USA is an ongoing research program and most of the research and development issues identified here are being

addressed within the scope of the research program. Substantial updates to the nighttime and daytime population distribution databases are expected in subsequent releases of LandScan USA.

Conclusions

Utilizing the increasing availability of national geospatial data sets including high resolution imagery, the LandScan USA model extends the existing paradigm of simple dasymetric modeling of census data through an innovative spatial data modeling approach. Integration of multiple high resolution indicator data sets, such as land cover, roads, cultural landmarks, educational institutions, and business activity locations, combined with human intelligence through analyst intervention allows efficient resolution enhancement in both spatial and temporal dimensions. The ability to incorporate activity-based information provides an unprecedented opportunity to design and develop a nationally consistent model that illustrates not only nighttime or residential population distribution, but also the mobility and dynamics of different demographic groups. The LandScan USA database (version 1.0) has been developed for the entire US and this initial release contains nighttime and baseline daytime population distributions at 3 arc-seconds resolution. Nighttime distribution covers residential and baseline daytime covers mobility of workers and students. Static (prison) population is included in both distributions. Transient population (business travelers and tourists) are not included in this version and will be included in the subsequent release of the database. Qualitative and quantitative verification and validation of the modeling parameters and quality assessment analysis demonstrate a high degree of precision and locational accuracy for the LandScan USA model and database. Current research efforts address the coupling of transportation modeling framework with population distribution data to develop population distribution scenarios at even finer time intervals (for example, hourly). Very high-resolution population databases, such as LandScan USA are imminently expected to enhance the current fidelity of spatial analysis, modeling, and decision support activities in application domains across the areas of homeland security, emergency preparedness and response, socio-environmental studies, and public health and consequently allow evaluation of existing policy.

Acknowledgements The authors would like to acknowledge the ongoing financial support for the development of LandScan and LandScan USA models and databases from the Department of Defense and the Department of Homeland Security and past financial support from the Department of Energy, the US Environmental Protection Agency and the National Cancer Institute. Our efforts continue to benefit from significant contributions from some of the best and brightest student research associates, whose efforts in data search, acquisition, modeling, and validation allow us to develop the LandScan USA database. Such tireless contributions from Nagendra Singh, Lauren Patterson, Aaron Myers, Pamela Dalal, Neal Feierabend, Aarthy Sabesan, Patrick Hagge, and Allan Jolly are thankfully acknowledged. We would also like to thank other members of the Geographic Information Science and Technology group for their periodic insights and contributions to this work. This manuscript has benefited through comments from a number of internal and external reviewers and the authors acknowledge their valuable insights.

References

Bhaduri, B. (2007). Population distribution during the day. In S. Shekhar & H. Xiong (Eds.), *Encyclopedia of GIS*, Springer, December 2007 (print edition ISBN 978-0-387-30858-6).

Bhaduri, B., Bright, E., & Coleman, P. (2005). *Development of a high resolution population dynamics model*. Paper presented at Geocomputation 2005, Ann Arbor, Michigan; http://www.geocomputation.org/2005/Abstracts/Bhaduri.pdf.

Bhaduri, B., Bright, E., Coleman, P., & Dobson, J. (2002). LandScan: Locating people is what matters. *Geoinformatics, 5*(2), 34–37.

Cai, Q., Rushton, G., Bhaduri, B., Bright, E., & Coleman, P. (2006). Estimating small-area populations by age and sex using spatial interpolation and statistical inference methods. *Transactions in GIS, 10*(4), 577–598.

Chen, K. (2002). An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing, 23*(1), 37–48.

Cohen, J., & Small, C. (1998). Hypsographic demography: The distribution of human population by altitude. *Proceedings of the National Academy of Science, 95*(24), 14009–14014.

Dobson, J., Bright, E., Coleman, P., & Bhaduri, B. (2003). LandScan 2000: A new global population geography. In V. Mesev (Ed.), *Remotely-sensed cities* (pp. 267–279). London: Taylor & Francis, Ltd.

Dobson, J., Bright, E., Coleman, P., Durfee, R., & Worley, B. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering & 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, 125–138.

Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *Annals of Regional Science, 26*, 67–78.

Forster, B. C. (1985). An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing, 6*(1), 139–151.

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

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

Harvey, J. T. (2002a). Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing, 23*, 2071–2095.

Harvey, J. T. (2002b). Population estimation models based on individual TM pixels. *Photogrammetric Engineering and Remote Sensing, 68*, 1181–1192.

Hay, S. I., Noor, A. M., Nelson, A., & Tatem, A. J. (2005). The accuracy of human population maps for public health application. *Tropical Medicine and International Health, 10*, 1073–1086.

Langford, M., & Unwin, D. (1994). Generating and mapping population density surfaces within a geographical information system. *Cartography Journal, 31*(1), 21–26.

McPherson, T., & Brown, M. (2004). *Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities*. *Symposium on Planning, Nowcasting, and Forecasting in the Urban Zone*. Paper presented at the American Meteorological Society Annual Meeting, Washington.

McPherson, T. N., Rush, J. F., Khalsa, H., Ivey, A., & Brown, M. J. (2006). *A day-night population exchange model for better exposure and consequence management assessments* Paper presented at the 6th Annual Meeting of the Urban Environment American Meteorological Society, Atlanta.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *Professional Geographer, 55*(1), 31–42.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194.

Monmonier, M. S., & Schnell, G. A. (1984). Land-use and land-cover data and the mapping of population density. *The International Yearbook of Cartography, 24*, 115–121.

Quinn, J. (1950). The daytime population of the central business district of Chicago. Review by Breese, Gerald W. *American Sociological Review, 15*(6), 827–828.

Reibel, M., & Agrawal, A. (2006). *Areal interpolation of population counts using pre-classified land cover data.* Paper presented at the 2006 Population Association of America Annual Meeting.

Sleeter, R. (2007). *Dasymetric mapping for estimating populations exposed to natural disasters*. Paper presented at the 2007 Annual Meeting of the American Association of Geographers, California.

Sleeter, R., & Wood, N. (2006). *Estimating daytime and nighttime population density for coastal communities in Oregon*. Paper presented at the 44th Urban and Regional Information Systems Association Annual Conference, British Columbia.

Sutton, P., Roberts, C., Elvidge, D., & Baugh, K. (2001). Census from heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing, 22*, 3061–3076.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association, 74*(367), 519–530.

U.S. Census Bureau, Population Division, Journey to Work and Migration Statistics Branch (2000) *Census 2000 PHC-T-40; Estimated daytime population and employment-residence ratios: Technical notes on the estimated daytime population*. Retrieved April 20, 2007, from (http://www.census.gov/population/www/socdemo/daytime/daytimepoptechnotes.html).

Wright, J. (1936). A method of mapping densities of population: With Cape Cod as an example. *Geographical Review, 26*(1), 103–110.