Computers, Environment and Urban Systems 61 (2017) 13–23

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A high resolution population grid for the conterminous United States: The 2010 edition

Anna Dmowska, Tomasz F. Stepinski *

Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA

ARTICLE INFO

Article history:
Received 30 January 2016
Received in revised form 22 May 2016
Accepted 20 August 2016
Available online 15 October 2016

Keywords:
Dasymetric modeling
Gridded population data
Census
NLCD
Land use

ABSTRACT

Readily available high resolution data on population distribution is an important resource for monitoring human-environment interactions and for supporting planning and management decisions. Using a grid that approximates population density over the entire country seems like the most practical approach to exploring and distributing detailed population data but instead data based on census aggregation units is still the most widely used method. In this paper we describe the construction of 30 m resolution grid representing the distribution of population in 2010 over the entire conterminous United States. The grid is computed using 2010 U.S. Census block level population counts disaggregated by a dasymetric model that uses land cover (2011 NLCD) and land use (2010 NLUD) as ancillary data. Detailed descriptions of the ancillary data and dasymetric model are given. Methods of computing the grid are presented followed by an extensive assessment of model accuracy. Overall the expected value for relative error of the model is 44% which is at the lower limit of errors reported for other continental-sized, high resolution population grids. We also offer a more specific error estimate for areas with specified value of population density. Using two example areas, one highly urbanized and another rural, we demonstrate the advantages of using the gridded population data over the census block-based data. Our 30 m population grid is available for online exploration and for download from the custom-made GeoWeb application SocScape at http://sil.uc.edu.

1. Introduction

Accurate information about the human population distribution is essential for formulating informed responses to population-related social, economic, and environmental problems. Governments need precise population data to support planning for infrastructure projects (Benn, 1995; Murray, Davis, Stimson, & Ferreira, 1998; Pattnaik, Mohan, & Tom, 1998), locating public facilities (Deng, Wu, & Wang, 2010), allocating and managing of resources (Gleick, 1996; Smith, Nogle, & Cody, 2002), and for preparing responses to natural disasters (Dobson, Brlght, Coleman, & Worley, 2000; Balcik & Beamon, 2008; Maantay & Maroko, 2009; Tenerelli, Gallego, & Ehrlich, 2015). Similarly, the private sector needs population data for planning the locations of their facilities (Martin & Williams, 1992), for optimization of service delivery systems, and for risk assessments (Chen et al., 2004; Thieken et al., 2006). Reliable information about the population distribution is also essential to assess human pressure on the environment (Weber & Christophersen, 2002), for quantifying environmental impact on population (Vinkx & Visee, 2008), and for public health applications (Hay, Noor, Nelson, & Tatem, 2005).

An authoritative source of population data is the government instigated national census; in the U.S. population data is collected every 10 years by the U.S. Census Bureau (hereafter referred to as the census). The most recent census was performed in 2010. The census collects population data with the ultimate resolution of an individual household but it releases this data aggregated to fixed areal units due to privacy concerns. The smallest aggregated areal unit released by the census is the census block. Census blocks in urban areas may be as small as a city block, but they are much larger in suburban and rural areas. There are several reasons why population data aggregated to fixed administrative units is not an ideal form of information about population density.

First, it suffers from the modifiable areal unit problem (Lloyd, 2014). Second, the spatial detail of aggregated data is variable and low, except in the most densely populated urban areas. Third, there is a spatial mismatch (Voss, Long, & Hammer, 1999) between census areal units (blocks, tracts etc.) and user-desired units (for example, neighborhoods, tax zones, postal delivery zones, vegetation zones, watersheds, etc.). Finally, the boundaries of census aggregation units (particularly blocks) may change from one census to another, making the analysis of population change at high spatial resolution difficult (Holt, Lo, & Hodler, 2004; Schroeder, 2007; Ruther, Leyk, & Buttenfield, 2015).

Overall, the properties of aggregation unit-based data make it ill-suited for the spatial analysis of population-related socio-economic

* Corresponding author.
E-mail address: stepintz@uc.edu (T.F. Stepinski).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.08.006
0198-9715/

and environmental problems. Instead, the population grid has emerged as an alternative format to deliver population data. A population grid is a geographically referenced lattice of square cells with each cell carrying a population count or the value of population density at its location. Population grids are constructed from census unit-based data using either areal weighting interpolation (Goodchild & Lam, 1980; Flowerdew & Green, 1992; Goodchild, Anselin, & Deichmann, 1993) or dasymetric modeling (Wright, 1936; Langford & Unwin, 1994; Eicher & Brewer, 2001). Population grids have the following advantages: all cells have the same size, the cells are stable in time, there is no spatial mismatch problem as any partition of the study area can be rasterized to be co-registered with a population grid. In addition, if dasymetric modeling is used (see below), a population grid offers a spatial resolution superior to that of the unit-based data.

With respect to the construction of a population grid, dasymetric modeling can be described as a technique of disaggregating aggregation unit-based population data into grid cells of a higher spatial resolution using ancillary data that correlates with population density but which has a higher resolution. Sharpening population data using dasymetric modeling has been extensively studied (Petrov, 2012) with a focus on the utilization of different types of ancillary data in order to increase the accuracy of a model. The original, and still the most widely used, ancillary data are land cover/land use data (Wright, 1936; Mennis, 2003, 2009; Linard, Gilbert, & Tatem, 2011). High-resolution satellite images have recently been utilized as ancillary data to identify individual buildings (Ural, Hussain, & Shan, 2011; Lu, Im, Quackenbush, & Halligan, 2010; Lung, Lübker, Ngochoch, & Schaab, 2013). A regression analysis is able to link the area or volume of each building to the number of people in it. If available, the Light Detection and Ranging (LiDAR) data is used (Lu et al., 2010), to help establish the volume of a building. Another approach to dasymetric modeling is to use local infrastructure information, such as street density (Reibel & Bufalino, 2005; Su, Lin, Hsieh, Tsai, & Lin, 2010) or the density of points of interest (Bakillah, Liang, Mobasheri, Arsanjani, & Zipf, 2014) as ancillary data. Tax parcel data have also been used (Maantay, Maroko, & Herrmann, 2007; Kar & Hodgson, 2012; Mitsova, Esnard, & Li, 2012; Jia, Qiu, & Gaughan, 2014; Jia & Gaughan, 2016) to disaggregate census population data. Other proposed sources of ancillary data include light emission data (Briggs, Gulliver, Fecht, & Vienneau, 2007; Sridharan & Qiu, 2013) and address datasets (Zandbergen, 2011).

Despite rapid progress in developing various techniques for dasymetric modeling, the practical adoption of population grids is low. This is because the majority of potential users are only able to utilize the ready-to-use product (a population grid) rather than actually create their own. In order to increase the adoption of demographic data for spatial analysis, high resolution grids over broad geographical areas need to be available in the public domain. Such grids have been developed and made available for all countries in the European Union (Gallego, 2010; Gallego, Batista, Rocha, & Mubareka, 2011) and, through the WorldPop project (http://www.worldpop.org.uk), for countries in South and Central America, Asia and Africa (Gaughan, Stevens, Linard, Jia, & Tatem, 2013; Linard, Gilbert, Snow, Noor, & Tatem, 2012; Sorichetta et al., 2015). For the United States, the Socioeconomic Data and Application Center (SEDAC) (http://sedac.ciesin.columbia.edu/) provides 1 km resolution (250 m for selected metropolitan areas) demographic grids. However, in addition to having a rather coarse resolution, these grids are only available for the years 1990 and 2000. A higher resolution (90 m) US–wide demographic grid, presumably based on most recent census data, is under development by the Oak Ridge National Laboratory (Bhaduri, Bright, Coleman, & Urban, 2007). This project, called LandScan–USA, aims at providing both nighttime (residential) as well as daytime population densities, but it is not currently available, nor is it expected to be in the public domain once it becomes available.

Since 2014 we have been developing high resolution demographic grids for the entire conterminous US. Our goal is to develop grids that offer a significant improvement over SEDAC grids and make them available for exploration and download through our interactive web-based application SocScape (Social Landscape) at http://sil.uc.edu. The first generation of our grids (referred to as SocScape–90) were the results of sharpening SEDAC grids to 90 m resolution using dasymetric modeling with the National Land Cover Dataset (NLCD) as ancillary data (Dmowska & Stepinski, 2014). Using this approach we have developed and made available through SocScape the population grids for 1990 and 2000. However, our original approach had several shortcomings and limitations. First, it did not use original census data, instead it relied on the SEDAC grid, which, in addition to containing a number of errors and inconsistencies (Dmowska & Stepinski, 2014), was also spatially coarser than census blocks in densely populated urban areas. Second, it was limited to years 1990 and 2000 – the only years for which SEDAC published its grids.

In this paper we report on our second generation of U.S.-wide grids (referred to as SocScape–30). Our new approach differs from the previous approach in the following ways: (1) It uses dasymetric modeling to disaggregate census blocks directly, rather than disaggregating SEDAC cells. (2) It uses two ancillary datasets, the NLCD 2011 and the newly available National Land Use Dataset (NLUD2010) (Theobald, 2014). (3) The new grid has a nominal resolution of 30 m, equal to the resolution of both ancillary datasets. (4) We offer an assessment of uncertainty of the model in the form which is directly relevant to a user. SocScape–30 is calculated on the basis of 2010 Census block-level data and is available online through our GeoWeb application. Section 2 describes the datasets used for the construction of the 2010 grid and Section 3 describes our methodology for obtaining the population grid. Section 4 gives the details of our calculations, presents a quality assessment, and describes how to access the data. Section 5 uses two examples, one urban and one rural, to demonstrate the advantages of using gridded data over the census block-based data. Discussion and conclusions are given in Section 6.

2. Input data

The SocScape–30 population grid is constructed using dasymetric modeling. In the context of this paper the dasymetric modeling technique requires two types of data - areal unit-based population data, to be disaggregated to a high resolution grid, and ancillary data at the resolution of this grid.

2.1. Census data

The primary source of spatio-demographic information in the United States are decennial censuses (http://www.census.gov). The U.S. Census Bureau provides data as a series of summary text files labeled from 1 to 4 which provide information at different levels of spatial aggregation (from as small as a census block to as large as the entire U.S.). To construct SocScape–30 we used the population count for each block based on variable P1 (total population) from Summary File 1 (SF1). We used the census data distribution provided by the National Historical Geographic Information System (NHGIS) at https://www.nhgis.org/ as we deemed NHGIS distribution easier to use than the Census Bureau distribution. This is because NHGIS-distributed demographic tables and shapefiles depicting block boundaries contain identifiers expediting the task of joining population counts to shapefiles.

Sizes of shapefiles containing block boundaries and their population counts vary from 34 MB for the District of Columbia to 4037 MB for the state of California. The overall size of the block-level shapefile/population count data for the entire conterminous U.S. (11,007,989 blocks) was 39 GB. We converted the block shapefile into a 30 m resolution raster grid co-registered with the ancillary grid (see the next subsection). Grid cells constituting a given block store numerical identifier of this block. There are 8,651,157,015 cells in the grid. Block rasterization is performed in order to expedite the computation of the dasymetric

model. An unintended consequence of rasterization is the "loss" of 264,565 blocks having a size about equal or smaller than a grid cell. This constitutes about 2% of all the blocks, however, it constitutes a "loss" of only 0.035% of the population because many of these blocks are uninhabited.

2.2. Ancillary data

The role of ancillary data is twofold, first to distinguish between inhabited and uninhabited sections of each block, and second, to provide information about variations in population density within inhabited sections of each block. We use two different ancillary datasets: the National Land Cover Dataset 2011 (NLCD 2011) (Homer et al., 2015) and the National Land Use Dataset 2010 (NLUD 2010) (Theobald, 2014). Both NLCD 2011 and NLUD 2010 grids have a resolution of 30 m and are co-registered with the grid of rasterized blocks. All three grids are in the Albers Conical Equal Area (EPSG 5070) projection so each grid cell has approximately the same area.

Within the conterminous U.S. each 30 m cell in NLCD 2011 is assigned one of possible 16 land cover classes: open water (11), perennial ice/snow (12), developed, open space (21), developed, low intensity (22), developed, medium intensity (23), developed high intensity (24), barren land (31), deciduous forest (41), evergreen forest (42), mixed forest (43), shrub (52), grassland (71), pasture (81), crops (82), woody wetlands (90), herbaceous wetlands (92). The numbers in brackets are the numerical labels of land cover classes. The problem with using the NLCD as an ancillary variable for dasymetric modeling is that its classes are based on surface spectral properties and thus cannot distinguish between populated buildings and unpopulated buildings as well as other impervious surfaces.

To have better information about populated vs. unpopulated areas we also utilize the NLUD 2010. Each cell in NLUD 2010 is assigned one of 79 land use classes (see Table 1 in Theobald, 2014). These classes are divided into five broad categories: water, built-up, production, recreation, and conservation. A built-up category is further subdivided into residential and non-residential (commercial, industrial, institutional, transportation) subcategories. In principle, NLUD is superior to NLCD as an ancillary variable for dasymetric modeling because it relates more directly to population density. Therefore, it could be argued that the dasymetric model should be constructed exclusively on the basis of NLUD. However, NLUD contains artifacts due to fact that it is a compilation of many different datasets of varying quality and form. Thus, for the purpose of our dasymetric model we utilize NLUD 2010 only to distinguish between inhabited and uninhabited areas - a distinction that is the most problematic while using NLCD. Thus, we reclassify NLUD into two categories: uninhabited (water, built-up commercial, built-up industrial, built-up institutional (except nursing homes), built-up transportation, mining area, and general and developed parks) and inhabited (all other classes).

Information from the census blocks, NLCD 2010 and the reclassified NLUD 2010 is combined to assign to each grid cell one of 6 possible ancillary classes: uninhabited (6), inhabited, vegetation (5), inhabited, high intensity (4), inhabited, medium intensity (3), inhabited low intensity (2), and inhabited, open space (1). The number in brackets are our numerical labels for these classes.

The process of assigning these classes is illustrated by the decision tree shown in Fig. 1. Each cell is subjected to a hierarchy of predicate statements to decide its ancillary class. At the first node the cell is assigned to class 6 (uninhabited) if the block to which it belongs has population count of zero. At the second node the cell is assigned to class 6 if the cell has NLCD label 11 (open water), 12 (perennial ice/snow), or 31 (barren land). The third node tests whether the cell is assigned to inhabited or uninhabited categories according to our reclassification of NLUD. If the cell has an "inhabited" category, the fifth node assigns it to one of 5 inhabited classes (1 - open space, 2 - low intensity, 3 - medium intensity, 4 - high intensity, 5 - vegetation) based on its NLCD labels as shown in Fig. 1. If the cell has an "uninhabited" category then the fourth node assigns it to class 6 unless all cells in a given block are assigned to the "uninhabited" category which is in the conflict with the fact that the block has a nonzero population count (follow the tree back to the first node). As the census information is considered more reliable, the cell is assigned its ancillary class based on its NLCD labels as shown in Fig. 1.

Fig. 2 illustrates the construction of the ancillary data layer using six adjacent blocks in Cincinnati, OH as an example. A satellite image of the blocks is given for reference (panel A). Panel B shows the NLCD map; it can be checked alongside the image for accuracy. Panel C shows the division of the area into inhabited and uninhabited sections according to the NLUD. Finally, panel D shows the map of ancillary classes we derived from the NLCD and NLUD using the decision tree in Fig. 1.

3. Dasymetric model

The key step when constructing a dasymetric model is the establishment of the relationship between ancillary variables and population density. A significant body of literature exists on the different methods used to establish such a relationship. Their review is beyond the scope of this paper. However, for the most relevant background we refer the reader to the descriptions of models used in other projects whose aim was to construct large scale population grids: Gallego et al. (2011) discussed various models tested while computing the 100 m grid for countries in the European Union, and Stevens, Gaughan, Linard, and Tatem (2015) discussed a model used to compute the 100 m grids for the WorldPop project.

In our model the relationship between the ancillary variable (6 classes resulting from combining NLCD and NLUD information) and population density is in the form of characteristic values of population density for each ancillary class (Mennis & Hultgren, 2006). These values are established by sampling the population density in blocks selected from the entire U.S. which are (almost) homogeneous with respect to their ancillary classes. For ancillary classes 1 to 4 we set the block homogeneity threshold to 90% and for ancillary class 5 we set the block homogeneity threshold to 95%.

Table 1 summarizes the block samples. The second column gives the count of homogeneous blocks with respect to each ancillary class, the third column gives the population count in homogeneous blocks, and the fourth column gives the total area of homogeneous blocks. The next three columns (fifth to seventh) give different estimates of characteristic population density (in people/km^2) for each ancillary class. Values in the fifth column are calculated by dividing the entire

Table 1
Characteristic population densities for six ancillary classes.

Ancillary class	Blocks count	Population	Area [km ^{2}]	Mean density	Median density	Max. prob	d [%]
Uninhabited (6)	4,576,562	0	3,001,721	0	0	0	0
Inhabited, open space (1)	47,198	701,428	902	778	850	801	4.21
Inhabited, low intensity (2)	208,171	5,571,086	2511	2219	2051	1874	11.98
Inhabited, medium intensity (3)	143,543	7,847,399	1648	4761	4060	3346	25.73
Inhabited, high intensity (4)	30,887	2,966,009	276	10,743	6389	1076	58.05
Inhabited vegetation (5)	753,140	11,795,302	2,153,420	5	6	6	0.03

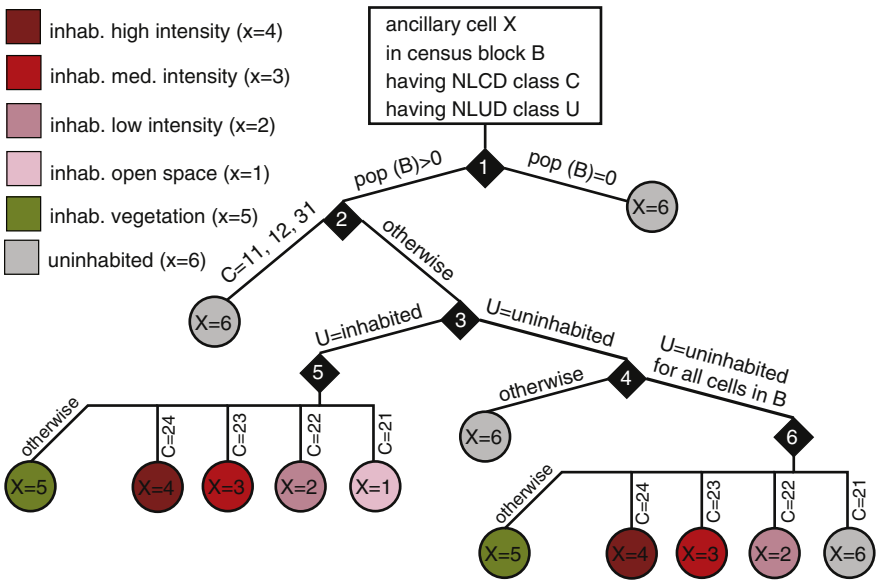

Fig. 1. Decision tree showing the process of assigning a cell's ancillary class on the basis of its NLCD and NLUD classes and the population count in the block to which it belongs.

population in a given sample by the total area of all blocks in this sample. Values in the sixth column are the medians of population densities of individual blocks in a given sample. Values in the seventh column correspond to the maximum probability of a given sample probability distribution function. Only for class 4 (inhabited, high intensity) do the three estimates of population density vary considerably.

Fig. 3 shows the probability distribution functions for values of population density in each sample of homogeneous blocks. From these distributions it is clear that the density values are broadly distributed. Despite this broadness, the shapes of density distribution functions for ancillary classes 1,2,3, and 5 indicate the existence of characteristic values – values of maximum probability which correspond to the values

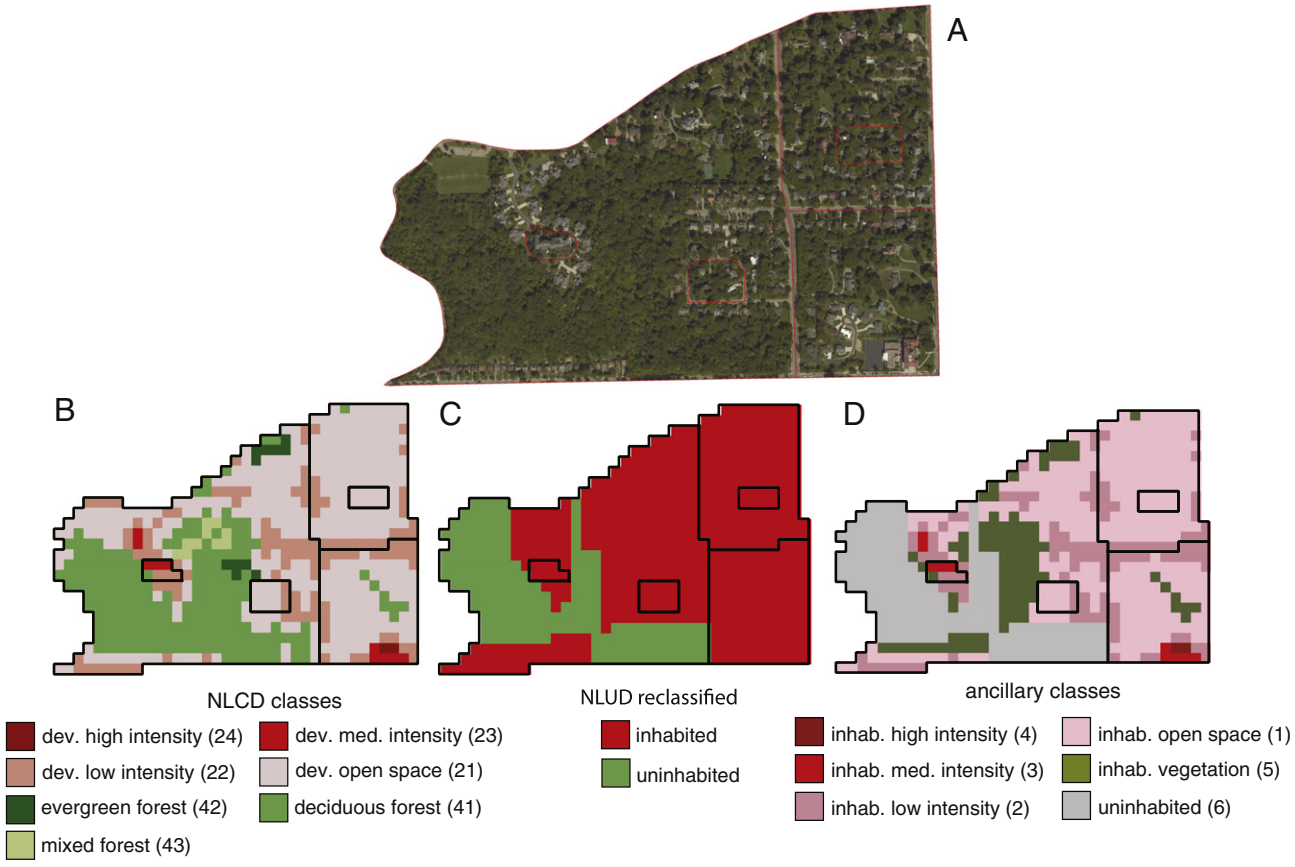

Fig. 2. Construction of the ancillary layer using area consisting of six adjacent blocks in Cincinnati, OH as an example. (A) A satellite image showing the surface masked to the spatial extent of the area; six constituent blocks are indicated by orange lines. (B) Spatial distribution of NLCD classes over the area. (C) Spatial distribution of reclassified NLUD classes over the area. (D) Spatial distribution of final ancillary classes over the area.

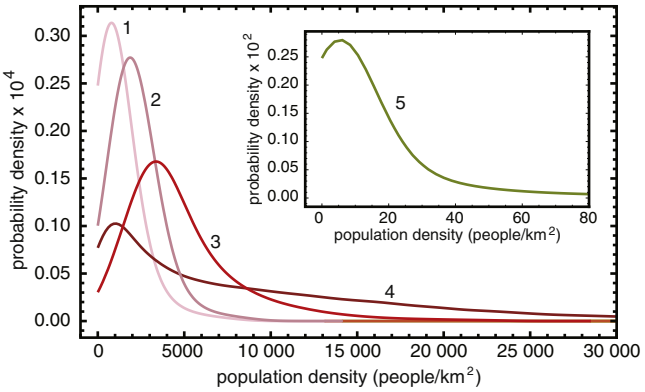

Fig. 3. Probability distribution functions of population density values in samples of blocks which are homogeneous with respect to ancillary class. The distribution for ancillary class 5 is shown as an inset because of its vastly different scale of density values. The colors of distribution curves and the numerical labels of the curves correspond to ancillary class labels (see legend of Fig. 2).

of a sample's mean as well as its median. However, this is not the case for ancillary class 4 (inhabited, high intensity) for which the probability peak (albeit a small one) occurs at a much smaller density value than the sample average – there is simply no good characteristic value of population density associated with this ancillary class.

We have chosen to use the values in the fifth column (average method) of Table 1 as sampled characteristic population densities and denoted them by the symbols p_i, where $i = 1 \ldots 6$ are numerical labels of ancillary classes. The last column in Table 1 gives the values of relative population density coefficients d_i

$$d_i = \frac{p_i}{p_1 + \ldots + p_6} \times 100, i = 1, \ldots 6 \quad (1)$$

The relative population density coefficients could be interpreted as percentages of a given block population assigned to portions of the block covered by corresponding ancillary classes in a hypothetical case when the block area is divided equally between all ancillary classes.

According to this model the relative population density in areas belonging to the high intensity class is about twice the relative density of areas belonging to the medium intensity class, which, in turn is about twice the relative density of the low intensity class. The relative density of the open space class is about three times lower than that of the low intensity class. Finally, the relative density of the vegetation class is about two orders of magnitude lower than that of the open space class.

A dasymetric model disaggregates population within a block to its constituent cells in proportion to a relative population density coefficient corresponding to the ancillary class associated with a cell. Denoting a given block by label x, where $x = 1, \ldots, 11{,}007{,}989$, and denoting a given cell in this block by label j, we calculate the weight $W_{x,j}$ associated with this cell as:

$$W_{x,j} = \frac{d_{x,j}}{\sum_{k=1}^{6} A_{x,k} d_k} \quad (2)$$

where $d_{x,j}$ is the value of the relative population density coefficient in cell j of block x determined by the label of the cell's ancillary class and $A_{x,k}$ is an area (in units of cells) within cell x associated with ancillary class k. The denominator in Eq. (2) is to ensure that weights over all cells in the block add to 1. The population count in cell j of block x is given as:

$$pop_{x,j} = W_{x,j} \times pop_x \quad (3)$$

where pop_x is the population in block x and $pop_{x,j}$ is the population in cell j of block x. From the properties of the weights it is clear that the sum of populations of all constituting cells is equal to the population of the block. Thus, our dasymetric model has a pycnophylactic (mass-preserving) property. Note that $pop_{x,j}$ is not an integer number and, in many areas, it is smaller than 1. Thus, referring to $pop_{x,j}$ as a population count is inappropriate. Instead, $pop_{x,j}$ should be referred to as a population density with units of people/area. Calculated values of $pop_{x,j}$ are given in units of people/900 m ^{2} and our downloadable data (see Section 4.3) are given in these units. For the purpose of illustration in this paper, and for the online population density map, we recalculated the values to the more customary units of people/km^2 by multiplying each cell value by 9×10^{-4} which is the area of a cell in km^2.

Fig. 4 demonstrates the use of the dasymetric model using the 6 blocks introduced in Section 2.2 as an example. Panel A shows spatial distribution of population density inferred from block data alone; in the absence of any additional information population density over each block is assumed constant and equal to the number of people in the block divided by the area of the block. Panel B shows the spatial distribution of weight values calculated on the basis of ancillary data (Eq. (2)) and panel C shows the spatial distribution of population density after disaggregation by the dasymetric model (Eq. (3)). According to the model, five out of six blocks in this example are characterized by a heterogeneous distribution of population. The grid-based map (panel C) offers more specific information about where inhabitants of these blocks live. The spatial precision of the grid-based map can be visually checked using a high resolution image (Fig. 2A).

4. U.S.–wide population grid

Although dasymetric modeling is a well established method and is relatively straightforward to apply, its application to the high resolution disaggregation of continental-scale areas requires an efficient computational algorithm and the ability to handle big datasets. The major challenge for the calculation of a 30 m dasymetric model of population density for the entire conterminous United States is the size of input and output data. Another challenge is to provide an intuitive and convenient means of reviewing and accessing the grid data.

4.1. Computation of population grid

Our computational task was to disaggregate over 11 millions census blocks into over 8 billion 30 m grid cells, and to do this in a reasonable time on a relatively modest computer with Intel 3.4GHz, 4-core processor and 16 GB of memory running the Linux operating system. We use a combination of two open source software packages: GRASS 7.0 (Neteler & Mitasova, 2007) which is a geographical information system platform, and R (R Development Core Team, 2008) which is a programming language and software environment. In addition, we also use the spgrass6 (Bivand, 2007) package that allows for the efficient transport of data between GRASS and R.

Computation consists of several steps that follow the methodology described in Sections 2 and 3: data pre-processing (in GRASS), calculation of the population grid (in R), and post-processing (in GRASS). The input for the pre-processing step are vector data containing the boundaries of census blocks together with an attribute table as well as ancillary datasets (NLCD and NLUD) for each county separately. Pre-processing performs block rasterization and computes a single ancillary dataset from NLCD + NLUD datasets (see Fig. 1). At the end it imports the data (organized by county) to R. The pre-processing step takes 37 h. Actual calculation of the grid using the dasymetric model (see Section 3) is performed in R and consists of establishing weights and disaggregating the block population according to these weights. This step takes 6 h. The post-processing step consists of exporting the grid (organized by county) to GRASS and joining grids for separate counties into one grid for the entire conterminous U.S. This step takes 18 h.

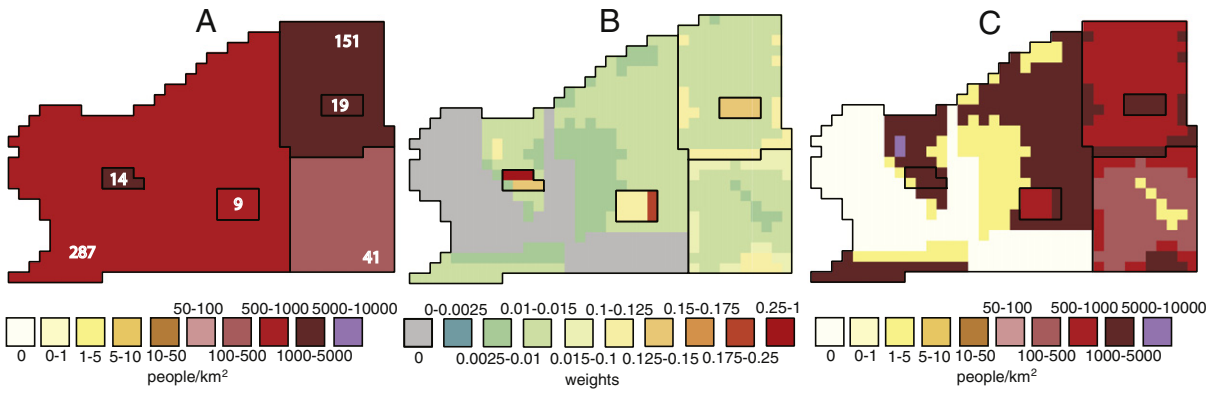

Fig. 4. Demonstration of dasymetric modeling using an area consisting of six adjacent blocks in Cincinnati, OH as an example. (A) Map of population density using only block-level data. Numbers indicate population counts for each block. (B) Spatial distribution of weight values. (C) Map of population density using grid data calculated using a dasymetric model.

Altogether, the entire computation takes 61 h. However, the grid of weights is stored and can be reused for disaggregation of other block-level variables which are related to population, such as sex, age, and race. Thus, for example, to obtain a U.S.–wide grid of the population of African-Americans we would start from weight we have already calculated and perform only disaggregation and post-processing.

4.2. Assessment of accuracy

The accuracy of the grid can be assessed directly only if ground truth data is available. In this context the ground truth data would consist of certifiable population counts within aggregation units smaller than those used in the dasymetric model. For example, the EU population grid (Gallego, 2010; Gallego et al., 2011) was obtained by disaggregating population in communes – relatively large areal aggregation units with areas much larger than 1 km^2. In this case, ground truth data in the form of 1 km^2 population grids was available for several countries in the EU. On the other hand, census blocks – the areal aggregation units we chose to disaggregate – are the smallest aggregation units available for the U.S., thus, we lack any sub-block population ground truth data to assess the accuracy of our grid directly.

Instead, we assess the accuracy of our method by calculating an additional grid based on the disaggregation of larger units – census block groups – and comparing the population of the resultant grid aggregated to blocks with the population of the blocks as given by the census. This method was used previously, including most recently by Jia et al. (2014) and Stevens et al. (2015).

Let's consider a given block group consisting of M blocks. We denote the population of m-th block, as given by block level data, by pop_m^{GT}, where the superscript GT indicates ground truth. We denote the population of m-th block as obtained from group-based dasymetric model by pop_m^{DM}, where the superscript DM indicates dasymetric model. Jia et al. (2014) following Eicher and Brewer (2001) used two quantities to assess the accuracy of their methods:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(pop_m^{GT} - pop_m^{DM}\right)^2} \tag{4}$$

and

$$\mathrm{CV} = \frac{\mathrm{RMSE}}{\frac{1}{M}\sum_{m=1}^{M} pop_m^{GT}} \tag{5}$$

where RMSE is the root square mean error and CV is the coefficient of variance. RMSE expresses the deviation (in the number of people) of the model from the ground truth in absolute terms, whereas CV expresses this deviation relative to the population. Both quantities pertain to a single block group. Calculating the mean (or median) of these quantities over all block groups in the region covered by the grid assesses the overall accuracy of the grid.

We first calculate the statistics of RSME and CV for our grid in order to compare them to the analogous statistics calculated by Jia et al. (2014) for their 30 m resolution population grids covering Alachua county in the state of Florida. They calculated two grids (disaggregated from 2010 census blocks) using two different ancillary datasets: NLCD, and 2010 tax parcel data. Parcel data is considered better ancillary data for dasymetric modeling than land cover because it relates more directly to population count (Jia et al., 2014). However, it is only available (in various degrees of completeness) for nineteen states in the U.S. (Stage & VonMeyer, 2006) and thus cannot constitute a base for a U.S.–wide population grid.

The first row in Table 2 shows the mean value of RMSE, the mean value of CV, and the median of CV, respectively. These statistics were calculated over all block groups in the conterminous U.S. using our dasymetric model. For the remaining three rows in Table 2 statistics were calculated only over block groups in Alachua county. The second row shows values calculated using our model, the third row shows values calculated using Jia et al. (2014) model utilizing NLCD as an ancillary variable, and the fourth row shows values calculated using the Jia et al. (2014) model utilizing tax parcel data as an ancillary variable. Statistics for Alachua county indicate that our grid (restricted to Alachua county) has higher accuracy than the Jia et al. (2014) model based on NLCD despite using values of characteristic population densities derived from nationwide block samples. This can be attributed to our more advanced dasymetric model which utilizes NLCD as well as NLUD, and to having six ancillary classes instead of three. In comparison to our model the Jia et al. (2014) model based on parcel data has a slightly higher value of ⟨RMSE⟩, a slightly lower value of ⟨CV⟩, and a lower value of median CV value. Thus, the population grid (within the Alachua county) constructed using tax parcels as ancillary data has somewhat higher accuracy than our grid constructed using a combination of NLCD and NLUD but this difference is small. Assessment of our grid over the entire conterminous U.S. yields superior accuracy in comparison with the part restricted to Alachua county and in comparison with both grids constructed by Jia et al. (2014).

Table 2
Accuracy assessment using RMSE and CV.

Grid	Mean RMSE	Mean CV	Median CV
Conterminous U.S.	43.17	0.97	0.78
Alachua	63.12	1.29	1.21
Alachua NLCD	73.26	1.36	1.30
Alachua parcels	63.96	1.20	0.88

Although the accuracy measures discussed above are useful for the comparison of different dasymetric models, they do not indicate to a user the degree of accuracy that can be expected from a grid. If a user utilizes the grid to estimate the population count of a sub-block areal unit, what is an uncertainty of this estimation? To provide this information we assume that uncertainty indicators calculated for blocks using the dasymetric model obtained by disaggregation of block groups are valid for sub-block units when using a dasymetric model which disaggregate blocks.

We calculate the statistics of relative errors over all inhabited blocks in the conterminous U.S. Note that due to the way we construct our ancillary classes (see Fig. 1) uninhabited blocks have zero population error in our model and could be excluded from statistics. The relative error δ_m of the population count for block m is the absolute error divided by the magnitude of the ground truth value.

$$\delta_m = \frac{|pop_m^{GT} - pop_m^{DM}|}{pop_m^{GT}} \quad (6)$$

The value of δ_m has a simple interpretation – if multiplied by 100 it expresses the overestimation or underestimation of the number of people as a percentage of the actual population of the block. As the distribution of the values of δ_m over all inhabited blocks is skewed, the usual statistical indicators (mean and standard variation) are not providing useful information, instead a median (equal to 0.44) provides a robust estimation of the "expected" value of δ_M and the median absolute deviation (equal to 0.4) provides a robust estimation of spread in the values of δ_m around the expected value. Thus, when using our grid to estimate a population in a sub-block area, a user can expect, on average, 44% uncertainty in the population count.

We can provide further information on the uncertainty of grid-based population estimations by calculating a two-dimensional histogram (shown in Fig. 5) of blocks with respect to their population density and relative error. Note that both population density (horizontal axis) and relative error (vertical axis) are shown in logarithmic scale due to the orders of magnitude variations in their values. The histogram has a final number of bins with each bin represented by a small square in Fig. 5. The color of a bin carries information on the number of blocks having the population density and relative error as indicated by the bin's coordinates. Note that the block count legend is also logarithmic, 83% of all blocks are in the three top block counts categories indicated by pink, darker red, and lighter red colors.

One way to use Fig. 5 to obtain information about population count uncertainty is to first select a value of population density for an area of interest. Values of block counts along the vertical line corresponding to the selected value of density combine into distribution of error values for these blocks. The peak of this distribution is located at the expected value of an error for blocks in this population density range and a deduced shape of the distribution informs about spread of error values around the expected value. Thus, assuming that histogram obtained for blocks is also valid for sub-block areas a user interested in an area having population density of about 3000 people/km ^{2} finds that the error is between 0.23 and 0.45. On the other hand, a user interested in areas having population density of about 100 people/km ^{2} finds that the error is most likely to be about 0.67.

4.3. Data access

We provide convenient access to the 2010 population grid via SocScape (Social Landscape) – a GeoWeb application designed for exploration and data distribution of population density and racial diversity grids. SocScape is available at http://sil.uc.edu. Upon launching SocScape shows a background map of the United States and a menu to select data. When "Population density 2010" is selected from the menu a map of population density appears categorized to eleven bins represented by different colors (the map legend is accessible from the navigation panel). It is important to differentiate between the map of population density, which is intended only for online exploration, and actual (not categorized) data that can be downloaded once a user decides on an area of interest.

Data in the population grid corresponds to population density and has units of people per cell. It can be downloaded by an area of interest. To download the data a user has to zoom into a general area of interest and select the download tool from the navigation panel (see Fig.6). When the download tool appears the "Population density 2010" data layer needs to be selected. Next, the user indicates a specific region of

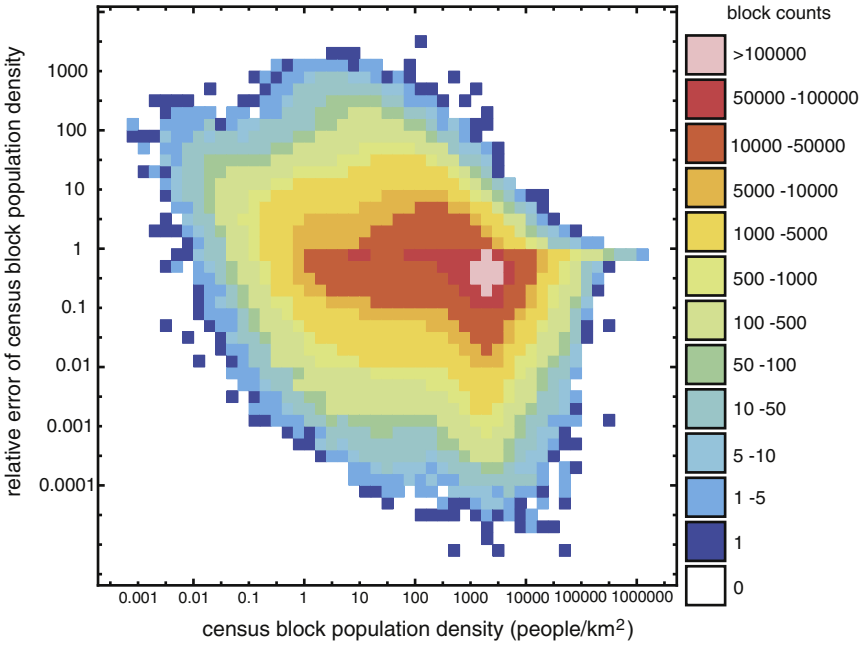

Fig. 5. Two-dimensional histogram of all inhabited blocks with respect to their population density and the value of relative error between its modeled and actual populations. Number of blocks in each bin of the histogram is color-coded according to the legend. Note logarithmic scales for all variables.

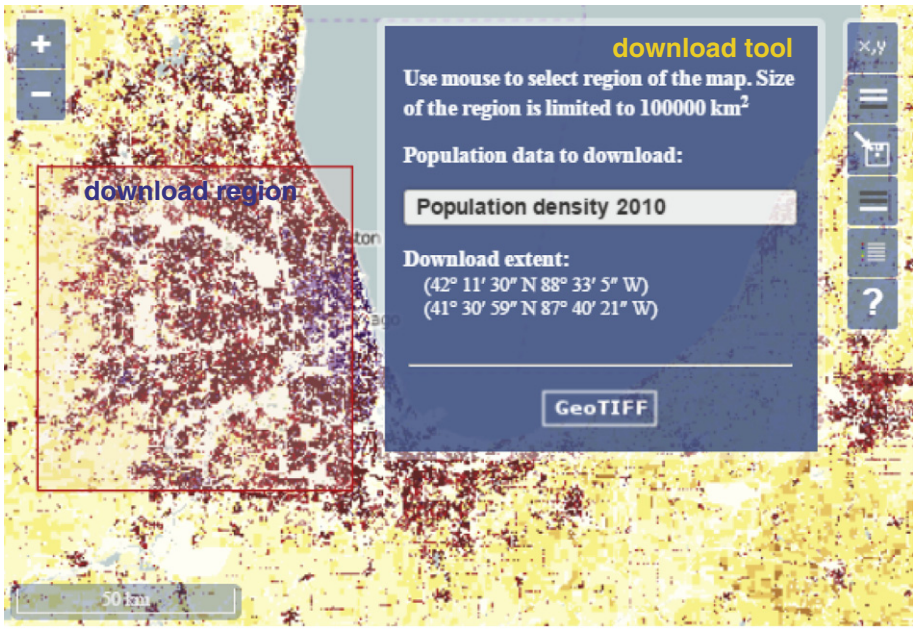

Fig. 6. Screenshot of SocScape after population density data layer has been selected, the map zoomed to show the area around Chicago, IL. The download tool is selected from the navigation panel, and download region is indicated by a user.

interest by dragging the mouse to draw a rectangle. Pressing on the GeoTIFF button in the download tool will start the download. The data is provided in geotiff format.

5. Grid-based versus blocks-based population maps

What differences can one expected when mapping population using population density (an intensive variable defined on a regular grid) versus a population aggregated to units (an extensive variable defined on census blocks). To point out differences between the two approaches we selected two locations as examples, first – a highly urbanized area of San Francisco CA, and second – a rural area centered on the Lake Loramie State Park, OH.

Fig. 7A shows a map of population density in San Francisco calculated from census blocks; density is uniform over each block as it was calculated by dividing the population in a block by its area. Block boundaries are not shown because the size of the blocks in this area are very small and the lines representing their boundaries would obscure the map at the figure's level of resolution. Fig. 7B shows a map of population density as represented by our grid. Overall, the two maps are similar but there some differences that can be summarized into two categories: (a) the block-based map is more consolidated and thus appears to have less detail, and (b) the grid-based map shows uninhabited areas which appear as inhabited on the block-based map.

The first difference stems directly from the fact that the dasymetric model disaggregates blocks into sub-block cells whose densities may vary. In general this results in superior spatial resolution for the grid. However, we need to keep in mind the uncertainty of the dasymetric model. Consider areas immediately south and north of Golden Gate Park (a prominent rectangle elongated along the west-east axis). Inspection of these areas using a high resolution image or map (for example those available in Google Maps) reveals that they are characterized by a grid layout of streets with houses filling the grid. This is not

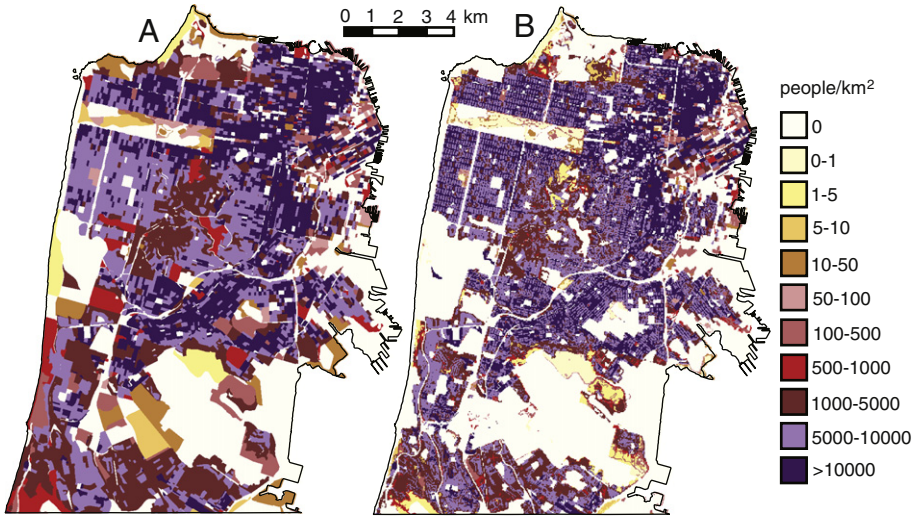

Fig. 7. Comparison of population maps for San Francisco (CA). (A) The block-based map. (B) The grid-based map. Boundaries of census blocks are not show for clarity.

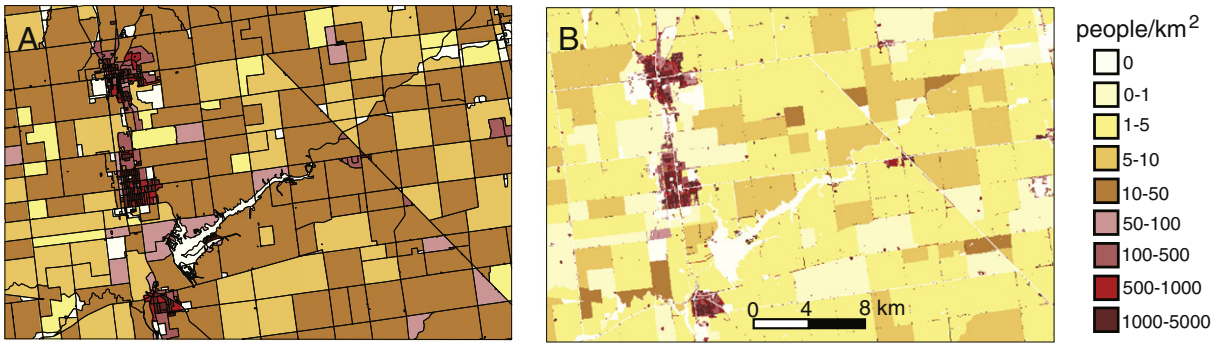

Fig. 8. Comparison of population maps for Lake Loramie State Park (OH) site. (A) The block-based map; block boundaries are shown as black lines. (B) The grid-based map.

captured by the block-based map because the blocks contain both houses and streets. With 30 m resolution one would expect that the grid-based map would show a white-purple pattern corresponding to streets (uninhabited) and houses (inhabited with high density). Instead, we observe a pattern consisting of purple (high population density) and light-purple (smaller population density) colors which is due to the fact that the NLCD in these areas does not recognize narrow streets and interprets the Landsat image as a mosaic of high and medium intensity developed land cover classes; the NLUD also do not delineate streets in these areas. Thus, although the grid correctly indicates the heterogeneity of population density in these areas, the spatial scale of streets is too small for them to be recognized as uninhabited areas. Still, there is some gain in information with respect to the block-based map.

The second difference is a direct result of the modifiable areal unit problem, entire blocks in the block-based map are marked by colors corresponding to low or medium values of population density even if they are mostly uninhabited but have small portions occupied by housing. Inspection of a high-resolution image or map reveals that most areas on the block-based map shown in yellow, light brown, or red colors are really predominantly uninhabited.

Fig. 8A shows a map of population density in a rural site centered on the Lake Loramie State Park, OH calculated from census blocks; density is uniform over each block as it was calculated by dividing the population in a block by its area. A high resolution image of this site (Google Earth) reveals the presence of Lake Loramie (center) and three small towns (the western part of the site), but most of the site is an agricultural landscape crossed by a grid-like network of secondary roads. Farmhouses are located predominantly at intersections of the roads, leaving most of the land uninhabited.

The block-based map does not reflect the real distribution of population because it assigns homogeneous density to predominantly uninhabited blocks. As a result most of the map is shown in a brown color corresponding to a population density of 10–50 people/km ^{2}. This is a relatively small value of density but still the reality is that these areas are uninhabited except at the locations of individual farmhouses. Fig. 8B shows the map of population density as represented by our grid. This map does not completely eliminate the inaccurate impression about the character of population distribution at this site but it significantly alleviates the problem by concentrating population in farmhouses and along the roads leaving the rest of the countryside either uninhabited or with a negligibly small density of population. Fig. 8B must be magnified in order to see small red dots indicating the population concentration along the secondary roads and at individual locations.

6. Discussion and conclusions

The purpose of the work presented in this paper, was to deliver the best possible nationwide population grid that can be constructed using readily available public domain data. The resultant product, which we call SocScape–30, is a 30 m resolution grid carrying values of residential (nighttime) population densities in 2010. The grid is freely distributed through the web using the SocScape GeoWeb application. The availability of this resource should make population data more accessible and thus more utilized.

We decided to combine land cover (NLCD) and land use (NLUD) datasets into a single ancillary variable to guide dasymetric modeling – the technique on which our grid is based. It may be asked why we did not utilize more ancillary datasets such as tax parcel data, road network data, the density of points of interest, topography, light emission etc. There are three reasons behind our choice. First, unlike other parts of the world, in the U.S. available land cover and land use datasets are highly reliable and the smallest census units (blocks) are small making additional ancillary information redundant or unnecessary. Second, apart from the road network (which we claim is redundant when NLUD is utilized) land cover and land use grids are the only readily available ancillary data that are consistent over the entire U.S. Tax parcel data – potentially useful ancillary information – is only available for some states (see Section 4.2) and each state or even county releases this data in its own format making the consistency of tax parcel data an issue. Our uncertainty estimates (Section 4.2), admittingly performed only for a single county, indicate that using ancillary data based on the combination of NLCD and NLUD yields a grid that is only slightly less accurate than a grid based on disaggregation using tax parcel data. Point of interest data is equally inconsistent on the continental scale. Third, land cover and land use datasets are available in a convenient grid format making data pre-processing more efficient and free from potential artifacts.

We use a relatively simple dasymetric model instead of seemingly more advanced models based on supervised machine learning (Stevens et al., 2015) because our model uses only a single ancillary variable. The model is based on nation-wide statistics rather than on a series of local statistics which could potentially capture a non-stationarity of the relationship between population density and an ancillary variable (Lo, 2008; Gallego, 2010; Schroeder & VanRiper, 2013). We selected this model from among the three we have calculated. The other two models attempted to address the non-stationarity issue. For the second model we divided the U.S. into five zones following the United States Department of Agriculture (USDA) Rural-Urban Continuum Codes for U.S. counties: rural areas, small town, micropolitan areas, metropolitan statistical areas (MSAs), and MSAs with population > 1,000,000 people. For each zone separately we calculated a dasymetric model as described in Section 3. As expected, the values of relative population density coefficients vary somewhat from one zone to another but the accuracy of such model, as measured by the mean value of CV calculated over the entire conterminous U.S., is not higher than that of our default model. In addition, the zoned model contains artifacts as the population density is discontinuous at the boundaries between the zones. In our third model we fitted characteristic values of population density separately for each county or census tract (like in Gallego

(2010) who deployed such a technique for Nomenclature of Territorial Units for Statistics (NUTS) – very large territorial regions) but we found that tracts or counties are too small to have a statistically valid sample of homogeneous blocks. Thus, overall, we deemed the default model to be the best choice for our purpose.

We have performed a comprehensive assessment of the accuracy of the method used to obtain our grid (Section 4.2). We estimate an overall relative error to be 44%, which may appear to be large. However, it is at the lower limit of errors estimated for other methods used to create state-of-the-art population grids. The error of the EU 100 m grid (Gallego et al., 2011) is estimated to be between 40% and 105% depending on the dasymetric model used and the country for which the assessment was done. The error for the WorldPop population grids is estimated to be between 39% and 91% for the newest models (Stevens et al., 2015) or between 46% and 120% for older models (Gaughan et al., 2013) depending on the country. It is important to note that each project uses a slightly different methodology to assess accuracy, but conclusions are similar – dasymetric model has an expected uncertainty of about 40–100%. We also provide the means for a more specific estimation of expected error (Fig. 5) based on additional knowledge about the population density of the area of interest. The largest source of error stems from the blurred relationship between land cover classes and population density. Using land use classes would help with this problem but the quality of U.S.-wide land use data is not sufficient to be utilized in a role other than for delineation of uninhabited areas. It is important to realize that the expected error estimates give the difference between a predicted and an actual number of people in a area of interest. It does not comment on the spatial precision of delineation between inhabited and uninhabited areas. In our grid this delineation follows the land use data and is expected to be fairly accurate.

As we mentioned in Section 4.1 weights (Eq. (2)) calculated on the basis of our model can be used to disaggregate other census block-level (SF1) variables which represent population segments; examples include sex, age and race. Such a disaggregation will not be able to account for possible sub-block heterogeneities in the proportion of each population segment to the total population, this will remain fixed throughout each block, but it will give a more accurate spatial location of each population segment by keeping it away from uninhabited areas and making an adjustment in line with the sub-block overall population density. One immediate application is the construction of a 2010 version of diversity maps like those introduced in Dmowska and Stepinski (2014) and analyzed in Dmowska and Stepinski (2016) for years 1990 and 2000.

Future plans call for recalculation of the 1990 and 2000 editions of U.S. population grids from 90 m resolution obtained by disaggregating SEDAC 1 km grid (Dmowska & Stepinski, 2014) to 30 m resolution using the technique presented in this paper. The major promise of having grids for various years is the ability to assess spatio-temporal population change. For such grids not to lead to discovery of spurious change they must be constructed using compatible ancillary datasets. Our plans call for us to make available through SocScape not only the best possible population grid for 2010 as described in this paper – but also for compatible, but not necessarily the best possible – population grids for 1990, 2000, and 2010 for the purpose of analyzing spatio-temporal change.

Acknowledgments

This work was supported by the University of Cincinnati Space Exploration Institute.

References

Bakillah, M., Liang, S., Mobasheri, A., Arsanjani, J. J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest International. *Journal of Geographical Information Science*, *28*, 1940–1963.

Balcik, B., & Beamon, B. M. (2008). Facility location in humanitarian relief. *International Journal of Logistics*, *11*(2), 101–121.

Benn, H. P. (1995). Synthesis of transit practice 10: Bus route evaluation standards. *Tech. rep.*. Washington, DC: Transit Cooperative Research Program, Transportation Research Board, National Research Council.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, *69*(1–2), 103–117.

Bivand, R. (2007). Using the R – GRASS interface: Current status. *OSGeo Journal*, *1*, 36–38.

Briggs, D. J., Gulliver, J., Fecht, D., & Vienneau, D. M. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment*, *108*(4), 451–466.

Chen, K., McAneney, J., Blong, R., Leigh, R., Hunter, L., & Magill, C. (2004). Defining area at risk and its effect in catastrophe loss estimation: A dasymetric mapping approach. *Applied Geography*, *24*, 97–117.

Deng, C., Wu, C., & Wang, L. (2010). Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information. *International Journal of Remote Sensing*, *31*(21), 5673–5688.

Dmowska, A., & Stepinski, T. F. (2014). High resolution dasymetric model of U.S demographics with application to spatial distribution of racial diversity. *Applied Geography*, *53*, 417–426.

Dmowska, A., & Stepinski, T. F. (2016). Mapping changes in spatial patterns of racial diversity across the entire United States with application to a 1990–2000 period. *Applied Geography*, *68*, 1–8.

Dobson, J. E., Brlght, E. A., Coleman, P. R., & Worley, B. A. (2000). LandScan: A global population database for estimating populations at risk*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.

Gallego, F. J. (2010). A population density grid of the European Union. *Population and Environment*, *31*(6), 460–473.

Gallego, F., Batista, F., Rocha, C., & Mubareka, S. (2011). Disaggregating population density of the European Union with CORINE land cover. *International Journal of Geographical Information Science*, *25*(February 2015), 2051–2069.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PloS One*, *8*(2), e55882.

Gleick, P. H. (1996). Basic water requirements for human activities: Meeting basic needs. *Water International*, *21*(2), 83–92.

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

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

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

Holt, J. B., Lo, C. P., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, *31*(2), 103–121 (31(2), 103–121).

Homer, C. G., Dewitz, J. A., Yang, L., Jin, S., Danielson, P., Xian, G., ... Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, *81*(5), 345–354.

Jia, P., & Gaughan, A. E. (2016). Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Applied Geography*, *66*, 100–108.

Jia, P., Qiu, Y., & Gaughan, A. E. (2014). A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. *Applied Geography*, *50*, 99–107.

Kar, B., & Hodgson, M. E. (2012). A process oriented areal interpolation technique: A coastal county example. *Cartography and Geographic Information Science*, *39*(1), 3–16.

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

Linard, C., Gilbert, M., Snow, R. V., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PloS One*, *7*, e31743.

Linard, C., Gilbert, M., & Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, *76*, 525–538.

Lloyd, C. D. (2014). The modifiable areal unit problem. *Exploring spatial scale in geography*. Chichester, UK: John Wiley & Sons, Ltd.

Lo, C. P. (2008). Population estimation using geographically weighted regression. *GIScience & Remote Sensing*, *45*(2), 131–148.

Lu, Z., Im, J., Quackenbush, L., & Halligan, K. (2010). Population estimation based on multisensor data fusion. *International Journal of Remote Sensing*, *31*(21), 5587–5604.

Lung, T., Lübker, T., Ngochoch, J. K., & Schaab, G. (2013). Human population distribution modelling at regional level using very high resolution satellite imagery. *Applied Geography*, *41*, 36–45.

Maantay, J., & Maroko, A. (2009). Mapping urban risk: Flood hazards, race, & environmental justice in New York. *Applied Geography*, *29*(1), 111–124 ((2009): 29 (1)).

Maantay, J. A., Maroko, A. R., & Herrmann, C. (2007). Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science*, *34*(2), 77–102.

Martin, D., & Williams, H. C. (1992). Market-area analysis and accessibility to primary health-care centres. *Environment and Planning A*, *24*(7), 1009–1019.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, *55*(1), 31–42.
Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass*, *3*(2), 727–745.
Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, *33*(3), 179–194.
Mitsova, D., Esnard, A. M., & Li, Y. (2012). Using enhanced dasymetric mapping techniques to improve the spatial accuracy of sea level rise vulnerability assessments. *Journal of Coastal Conservation*, *16*(3), 355–372.
Murray, A. T., Davis, R., Stimson, R. J., & Ferreira, L. (1998). Public transportation access. *Transportation Research Part D: Transport and Environment*, *3*(5), 319–328.
Neteler, M., & Mitasova, H. (2007). *Open source GIS: A GRASS GIS approach* (3rd edition). New York: Springer.
Pattnaik, S. B., Mohan, S., & Tom, V. M. (1998). Urban bus transit route network design using genetic algorithm. *Journal of Transportation Engineering*, *124*(4), 368–375 ((1998): 124(4)).
Petrov, A. (2012). One hundred years of dasymetric mapping: Back to the origin. *Cartographic Journal*, *49*(3), 256–264.
R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Reibel, M., & Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, *37*, 127–139.
Ruther, M., Leyk, S., & Buttenfield, B. P. (2015). Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. *GIScience & Remote Sensing*, *52*(2), 158–178 (52(2), 158–178).
Schroeder, J. P. (2007). Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis*, *39*(3), 311–335.
Schroeder, J. P., & VanRiper, D. C. (2013). Because Muncie's densities are not Manhattan's: Using geographical weighting in the expectation–maximization algorithm for areal interpolation. *Geographical Analysis*, *45*(3), 216–237 (45(3), 216–237).
Smith, S. K., Nogle, J., & Cody, S. (2002). A regression approach to estimating the average number of persons per household. *Demography*, *39*(4), 697–712.
Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data*, *2*, 150045.
Sridharan, H., & Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis*, *45*(3), 238–258 ((2013): 45 (3)).
Stage, D., & VonMeyer, N. (2006). An assessment of parcel data in the United States – 2005 survey results. *Tech. rep.*. Federal Geographic Data Subcommittee on Cadastral Data.
Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, *10*(2), e0107042.
Su, M. D., Lin, M. C., Hsieh, H. I., Tsai, B. W., & Lin, C. H. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *Science of the Total Environment*, *408*(20), 4807–4816.
Tenerelli, P., Gallego, J. F., & Ehrlich, D. (2015). Population density modelling in support of disaster risk assessment. *International Journal of Disaster Risk Reduction*, *13*, 334–341 ((2015): 13).
Theobald, D. M. (2014). Development and applications of a comprehensive land use classification and map for the US. *PloS One*, *9*(4), e94628.
Thieken, A. H., Müller, M., Kleist, L., Seifert, I., Borst, D., & Werner, U. (2006). Regionalisation of asset values for risk analyses. *Natural Hazards and Earth System Sciences*, *6*(2), 167–178.
Ural, S., Hussain, E., & Shan, J. (2011). Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation*, *13*(6), 841–852.
Vinkx, K., & Visee, T. (2008). Usefulness of population files for estimation of noise hindrance effects. *ICAO committee on aviation environmental protection. CAEP/8 modelling and database task force (MODTF). 4th meeting. Sunnyvale, USA* (pp. 20–22).
Voss, P. R., Long, D. D., & Hammer, R. B. (1999). When census geography doesn't work: Using ancillary information to improve the spatial interpolation of demographic data. *Tech. rep.*. Center for Demography and Ecology, University of Madison-Wisconsin.
Weber, N., & Christophersen, T. (2002). The influence of non-governmental organisations on the creation of Natura 2000 during the European Policy process. *Forest Policy and Economics*, *4*(1), 1–12.
Wright, J. K. (1936). A method of mapping densities of population with Cape Cod as an example. *Geographical Review*, *26*(1), 103–110.
Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, *15*(s1), 5–27.