High-Resolution Population Grids for the Entire Conterminous United States

Anna Dmowska and Tomasz F. Stepinski

Abstract To have a more complete awareness of the global environment and how it changes, remotely sensed data pertaining to the physical aspects of the environment need to be complemented by broad-scale demographic data having high spatial resolution. Although such data are available for many parts of the world, there are none available for the United States. Here we report on our ongoing project to develop high-resolution (30–90 m/cell) population/demographic grids for the entire conterminous United States and to bring them into the public domain. Two different, dasymetric modeling-based approaches to disaggregation of block-level census data into a fine grid are described, and resulting maps are compared to existing resources. We also show how to utilize these methods to obtain racial diversity grids for the entire conterminous United States. Our nationwide grids of population and racial diversity can be explored using the online application SocScape at http://sil.uc.edu.

Keywords Population grids · Dasymetric modeling · Demographic data

1 Introduction

Quick and convenient access to high-resolution data for the spatial distribution of population is needed for a wide range of analyses related to resource management, facility allocation, land-use planning, natural hazards, and environmental risk (Dobson et al. 2000; Chen et al. 2004), disaster relief/mitigation (Bhaduri et al. 2002), and socio-environment interactions (Weber and Christophersen 2002). Widely available population and demographic data are a result of aggregating census information into arbitrary areal units to ensure privacy. Consequently, maps showing the

A. Dmowska · T.F. Stepinski (✉)
Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH, USA
e-mail: stepintz@uc.edu

A. Dmowska
e-mail: dmowskaa@ucmail.uc.edu

D.A. Griffith et al. (eds.), *Advances in Geocomputation*,
Advances in Geographic Information Science,
DOI 10.1007/978-3-319-22786-3_4

spatial distribution of population are in vector form, resulting in an unrealistic level of homogeneity (lack of spatial resolution) in the population distribution. Moreover, these data are inconvenient to use, especially when a study area covers several different administrative regions along which the data are organized (e.g., a metropolitan area that stretches over two or more states).

Given the limitations of areal units-based population data, the focus of investigation has shifted to populations grids that are capable of depicting the population distribution at a higher resolution. These grids also are convenient to use and support means of algorithmic analysis that are not available for vector-based maps. Population grids are obtained using the principle of dasymetric mapping (Wright 1936)—a procedure that subdivides areal units into a regular grid of cells using ancillary information that can serve as a proxy for a more accurate population distribution.

Disaggregation by means of dasymetric modeling is well established and relatively straightforward (Langford and Unwin 1994; Eicher and Brewer 2001; Mennis 2003; Qiu and Cromley 2013), with a majority of applications being applied to small, local study areas for which detailed, local ancillary information—for example, parcels (Jia et al. 2014), buildings (Dong et al. 2010), addresses (Reibel and Bufalino 2005)—is available. However, we are interested in using dasymetric modeling to obtain a continental-scale, high-resolution population grid because it can serve a much bigger number of potential users. Using dasymetric modeling on such a large scale is a technically demanding task due to its needs to handle very large datasets and to develop computationally efficient algorithms. Recognizing a need for broad-coverage population data at a high spatial resolution, population grids have been produced for countries within the European Union, Africa, South America, and Asia (see Table 1).

Until recently, the only public domain gridded population data for the United States (U.S.) were census grids developed by the Socioeconomic Data and Application Center (SEDAC) (http://sedac.ciesin.columbia.edu). SEDAC grids have a number of shortcomings: (1) they have a relatively coarse 1 km resolution (250 m

Table 1 Availability of broad-scale, high-resolution population grids

Project	Region	Resolution	Availability
WorldPop[a]	S. America, Africa, Asia	100 m	http://www.worldpop.org.uk
E.U. pop. grid[b]	Europe	100 m	http://www.eea.europa.eu/
Australian pop. grid	Australia	1000 m	http://www.abs.gov.au
SEDAC-USA	United States	1000/250 m	http://sedac.ciesin.columbia.edu
LandScan-USA[c]	United States	90 m	Not available
SocScape[d]	United States	90/30 m	http://sil.uc.edu/

[a]Tatem et al. (2007), Linard et al. (2012), Gaughan et al. (2013), [b]Gallego (2010), Gallego et al. (2011), [c]Bhaduri et al. (2007), [d]Dmowska and Stepinski (2014)

resolution for selected metropolitan areas), (2) they are a product of simple areal weighting interpolation (Goodchild et al. 1993) rather than disaggregation using dasymetric modeling, and (3) they are available only for 1990 and 2000. The Oak Ridge National Laboratory developed (Bhaduri et al. 2007) the LandScan USA—a 90 m population grid covering the US LandScan uses advanced dasymetric modeling utilizing a multitude of ancillary datasets and comes in two versions, nighttime (density of residential population) and daytime (density of workplace population). However, LandScan 90 is neither in the public domain nor is it commercially available, so it cannot be utilized by the broader scientific community.

Here we report on our ongoing project to bring high-resolution population grids of the entire conterminous United States to the public domain. We have developed two approaches to achieve this goal. Our first approach (hereafter referred to as generation-1 or Gen-1) yields 90 m nationwide population grids for 1990 and 2000. This approach is based on sharpening already existing SEDAC grids using the National Land Cover Dataset (NLCD) (http://www.mrlc.gov/) as ancillary data. Because SEDAC is not producing grids based on the 2010 census, we have decided to change our approach to one that is computationally much more demanding but can be applied to 2010 census data as well as to the data from previous censuses. Our second approach (hereafter referred to as generation-2 or Gen-2) yields 30 m United States-wide population grids for 2010. This approach disaggregates census blocks directly using NLCD 2011 and the newly available, 30 m resolution, United States-wide National Land Use Dataset (NLUD 2010) (Theobald 2014) as ancillary datasets.

For these grids to be easily previewed, we have developed a web-based application called SocScape (Social Landscape) (http://sil.uc.edu). It is designed to facilitate exploration of population density over the entire conterminous United States for 1990, 2000, and 2010. In addition, it can be used to explore the geographical distribution of racial diversity for 1990 and 2000 (the map based on 2010 data is in the development phase). High-resolution racial diversity maps are by-products of our population disaggregation method.

2 Data and Methods

Both the Gen-1 and Gen-2 methods use census data. Gen-2 diaggregates ~11 millions census blocks directly, whereas Gen-1 disaggregates the SEDAC grids. In addition, both methods use the NLCD as ancillary data; but Gen-2 also uses the NLUD 2010 data.

2.1 The Gen-1 Disaggregation Method

Gen-1 calculates 90 m population grids from pre-existing 1 km SEDAC grids. SEDAC grids are a product of simple areal weighting interpolation from 1990 and 2000 census block data; no ancillary data have been used in the process of their creation. We sharpen SEDAC grids from 1,000/250 m/cell to 90 m/cell using land cover (1992 and 2001 editions of the NLCD) as ancillary data. This method was initially selected because it is less computationally demanding than disaggregating directly from census blocks. Computational efficiency comes from working exclusively with grids.

Land cover has been used because it is the only ancillary information that has uniform quality across the entire United States. However, the main editions of 1992 and 2001 NLCD have different land cover legends and cannot be used to produce population grids that can be compared between 1990 and 2000. Because such a comparison is one of our goals, we instead use the NLCD 1992/2001 Retrofit Land Cover Change Product (Fry et al. 2009). This product provides compatible land cover classifications for 1992 and 2001, but at the cost of reducing the number of land cover categories to only eight classes that constitute Level I of the Anderson classification scheme (Anderson et al. 1976). We further reclassify the eight-class land cover data into just three classes: urban, vegetation, and uninhabited area. The Gen-1 dasymetric model uses these three-classes of land cover as its ancillary data.

For details of Gen-1 disaggregation, see Dmowska and Stepinski (2014). Briefly, the population in each SEDAC cell is redistributed to 90 m cells using weights calculated from land cover composition within each 90 m cell and the average population density of the three land cover classes. In metropolitan statistical areas (MSAs), in addition to average population density in land cover classes, we use information about the distribution of population from the finer, 250 m/cell, SEDAC-MSA grids. The population in each 90 m cell is determined by multiplying the population count in a SEDAC cell by a weight specific to a given 90 m cell.

2.2 The Gen-2 Disaggregation Method

Population maps produced by our Gen-1 method offer an efficient means of improving upon the SEDAC grids (see the comparison in Sect. 3). However, because SEDAC has no population grids for 2010, and because SEDAC grids for 1990 and 2000 contain some inconsistencies, we decided to change our disaggregation algorithm to a direct disaggregation from census blocks. This approach requires larger hardware resources and is less computationally efficient, but the outcome justifies the extra effort. In addition, a new ancillary resource—a 30 m/cell land use map (NLUD 2010) over the entire conterminous United States—became available (Theobald 2014) in 2014, and we incorporate it into our Gen-2 method. The Gen-2 method consists of several steps: (1) preprocessing of U.S Census data, (2) preprocessing of ancillary data, (3) sampling population density over different land cover/use classes, and (4) calculating weights and the redistribution of population.

The 2010 United States Census block-level data consist of two components: shapefiles (TIGER/Line Files) with geographical boundaries of ~11 millions blocks, and summary text files that list population data for each block. In the census data preprocessing step, we first join the boundaries shapefile with the data from the summary file to form a vector file. In a key step, which departs from conventional dasymetric modeling, the vector files are rasterized to a 30 m resolution grid. Note that all 30 m cells belonging to a single block initially have the same values. The 30 m resolution was selected because it is the resolution of ancillary datasets; having all datasets as co-registered grids of the same resolution improves the computational efficiency of the disaggregation algorithm.

The major problem with using land cover (NLCD 2011) as ancillary information is that it may not differentiate correctly between inhabited and uninhabited impervious areas. We use land use data (NLUD 2010) to make this differentiation. In the ancillary data preprocessing step, we combine information from NLCD 2011 and NLUD 2010 to define six land cover/use classes: developed open space, developed low intensity, developed medium intensity, developed high intensity, vegetation, and uninhabited. Following Mennis and Hultgren (2006), representative population density for each of the six land cover/use classes is established using a set of blocks (selected from the entire United States) having relatively homogeneous land cover (90 % for developed classes, and 95 % for vegetation classes). Within each block class, weights are calculated using class abundances in the block and their representative population densities (Mennis 2003). Finally, the population in each (rasterized) block is redistributed to its constituent cells using the weights. Note that once the weights are calculated, they can be used not only for disaggregation of total population, but also for disaggregation of segments of populations defined by any attribute for which block-level data are available. In particular, these procedure can be used to disaggregate race-specific populations. This procedure preserves the block-level value of a ratio between different races because no race-specific ancillary information is used.

The Gen-2 algorithm redistributes ~11 millions census blocks over 8 billion grid cells. The output grid file size is 139 GB. To keep computational cost under control, calculations were performed for each state separately, and results were joined into a single map for the entire United States. These calculations were done using a computer with an Intel 3.4 GHz, 4-cores processor and 16 GB of memory running the Linux operating system. All calculations were performed using Python scripts written for GRASS GIS 7.0 software. The total time of Gen-2 calculations was 66 h.

3 Results

The best way to examine population density grids generated by our Gen-1 and Gen-2 algorithms is to explore them using the SocScape web application. SocScape is a computerized map application that allows a population density map to be overlaid on a base layer of either a street map or an aerial image. It works on desktops,

laptops, and mobile devices. It supports downloading of data in either GeoTiFF or PNG formats. Note, however, that only classified data (as seen in the application) can be downloaded. The numerical data (people counts per cell) are too large to be distributed via SocScape, but are available upon request from the authors. Here we use three examples to demonstrate how our grids compare with other available population density resources. The first example compares different population density maps/grids in an urban setting (Cincinnati, Ohio). The second example compares different population density maps/grids in a rural setting (Somerset, Ohio). Finally, the third example shows how our method can be used to produce high-resolution racial diversity grids.

Figure 1 compares different population density grids in a portion of the Cincinnati (Ohio) metropolitan region. This site captures a region around the Ohio River, with Cincinnati located north of the river and Kentucky located south of the river. The industrial transportation corridor (that includes railroad tracks and Interstate 71/75) runs through the middle of the site, from the Ohio River northward. To the west of this corridor are residential neighborhoods, and to the east is the downtown area. For reference, Fig. 1a shows a satellite image of the site, and Fig. 1b shows a land cover map (NLCD 2011) of the site. Five population density maps are compared: census block-based (Fig. 1c), SEDAC 1 km grid (Fig. 1d), SEDAC 250 m grid (available for the Cincinnati MSA) (Fig. 1e), our Gen-1 90 m grid (Fig. 1f), and our Gen-2 30 m grid (Fig. 1g). All grids have the same legends, and the land cover legend is also shown for reference.

Because sizes of census blocks are small in heavily populated areas, a block-based map (Fig. 1c) offers more actual details in the downtown area than the other maps do, except for our Gen-2 grid (Fig. 1g). Outside the downtown, where blocks are large and they may include parks and other sparsely inhabited areas, the block-based map still offers advantages over the SEDAC 1 km grid but not over the other grids. The SEDAC 1 km grid (Fig. 1d) captures only the most basic features of the population distribution. The presence of a river and of an industrial transportation corridor cannot be deduced from this map. The SEDAC 250 m grid (Fig. 1e) shows some additional details. The river and the industrial transportation corridor are delineated; however, parks are not distinguished from built-up areas. Our Gen-1 grid (Fig. 1f) offers a significant improvement over the SEDAC 250 m grid. The river and the industrial transportation corridor are well delineated and parks and green spaces are distinguished from built-up areas. Comparing the Gen-1 grid with the block-based map, however, reveals that some industrial, uninhabited areas are shown as inhabited because some NLCD classes contain residential and non-residential buildings. Also, the resolution of our Gen-1 map in the downtown area is not as good as it is in the block-based map. Based on a comparison with a satellite image (Fig. 1a) and a land cover map (Fig. 1b), as well as our familiarity with the site, we conclude that our Gen-2 grid (Fig. 1g) offers the best depiction of population density in this site. Utilizing land use ancillary data results in a proper delineation of uninhabited areas and the parks, and forested areas also are correctly shown as having low population density. Finally, the resolution in the downtown area is as good as, or better than, it is in the block-based map.

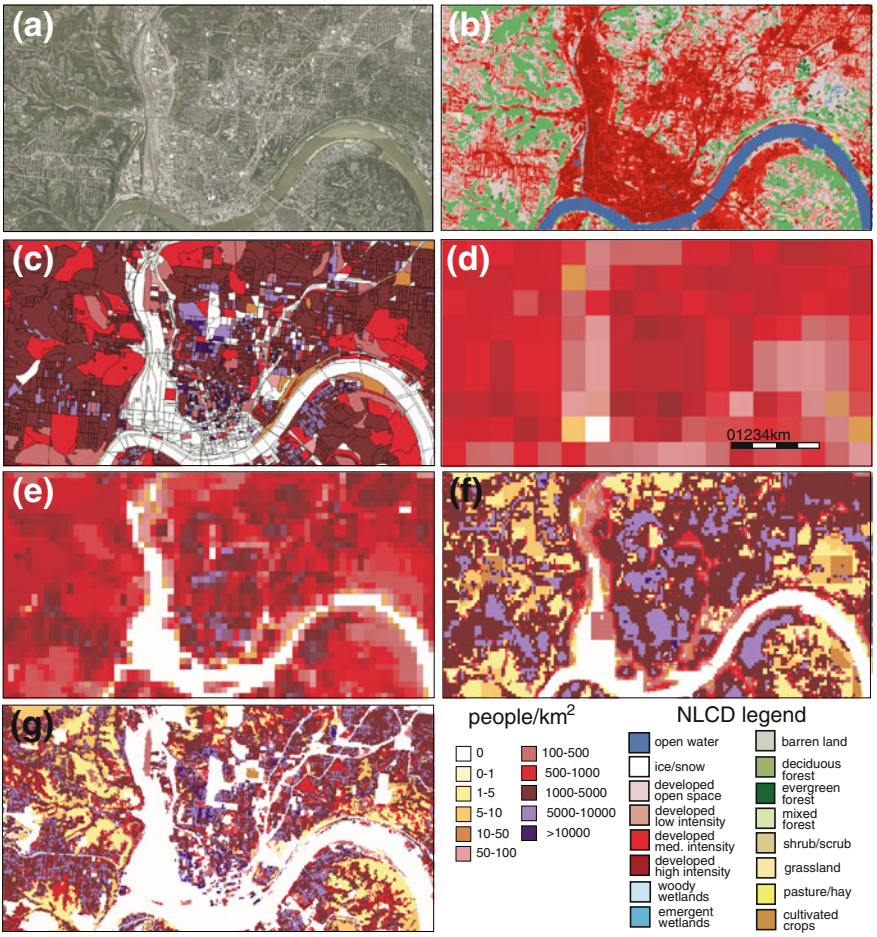

Fig. 1 A comparison of population grids for the Cincinnati (Ohio) site. **a** Satellite image (Google Maps), **b** land cover map (NLCD 2011), **c** census block-based map of population density, **d** SEDAC 1 km grid, **e** SEDAC 250 m grid, **f** Gen-1 90 m grid, **g** Gen-2 30 m grid

Figure 2 compares different population density grids in a rural site located around the village of Somerset (Ohio). For reference, Fig. 2a shows a satellite image of the site, and Fig. 2b shows a land cover map (NLCD 2011) of the site. The site consists mostly of agricultural land. Apart from the village, the population is concentrated in farmhouses located predominantly along roads. Because the site is sparsely populated, census blocks are relatively large. Four population density maps are compared: census block-based (Fig. 2c), SEDAC 1 km grid (Fig. 2d), Gen-1 90 m grid (Fig. 2e), and Gen-2 30 m grid (Fig. 2f). All grids have the same legend (see Fig. 1 for the legend). Note that, unlike the Cincinnati site, the SEDAC 250 m grid is not available for this area.

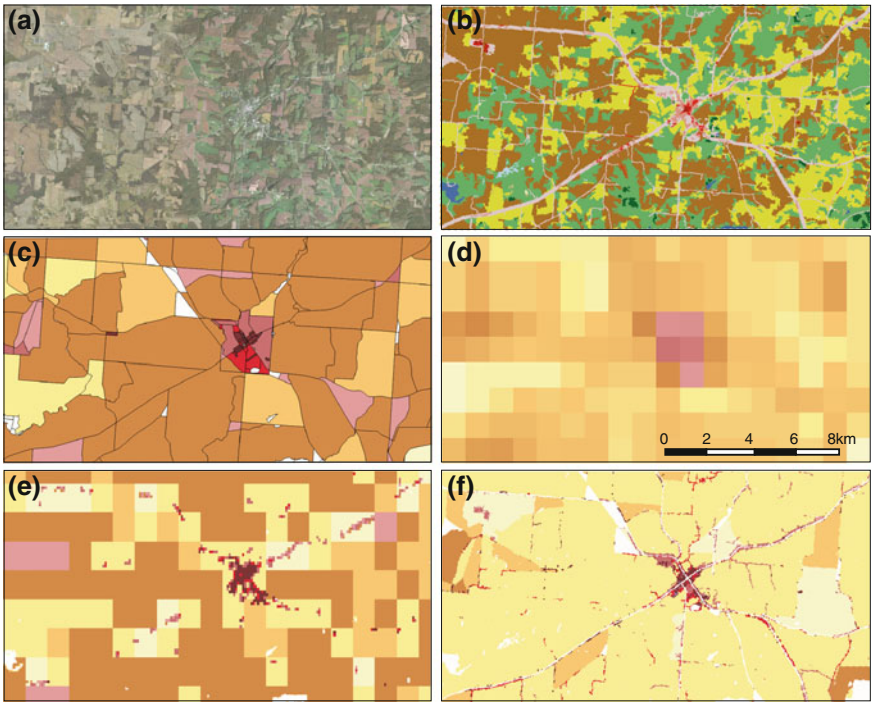

Fig. 2 A comparison of population grids for the Somerset (Ohio) site. **a** Satellite image (Google Maps), **b** land cover map (NLCD 2011), **c** census blocks-based map of population density, **d** SEDAC 1 km grid, **e** Gen-1 90 m grid, **f** Gen-2 30 m grid. For legends, see Fig. 1

The block-based map (Fig. 2c) does not correctly reflect an actual distribution of population in this site. The color assigned to the blocks to depict their values of population density reflects an average density over each entire block, whereas most of the block area is uninhabited or very sparsely populated because it is covered by crops and pastures. The SEDAC 1 km grid (Fig. 2d) offers a fair approximation to an overall distribution of population, but without any details for the village of Somerset (population 1,418). The Gen-1 grid (Fig. 2e) was able to resolve the village, but its population density over farmland is too high (although this value is still low, given that the area is rural). The Gen-2 grid (Fig. 2f) recognizes individual farm houses, and thus lowers the population density of farmland as people are assigned to very small regions at the locations of farm houses. Additional information from the land use data delineates uninhabited land such as the state forest (the light yellow area at the right edge of the region).

Figure 3 shows maps of racial diversity for an urban site covering Cincinnati, Ohio (Fig. 3a and b) and a rural site around Fresno, California (Fig. 3c and d). Our racial diversity maps (Fig. 3a and c) are 90 m grids produced using the Gen-1 disaggregation method on 2000 census data. The total population is segmented with respect

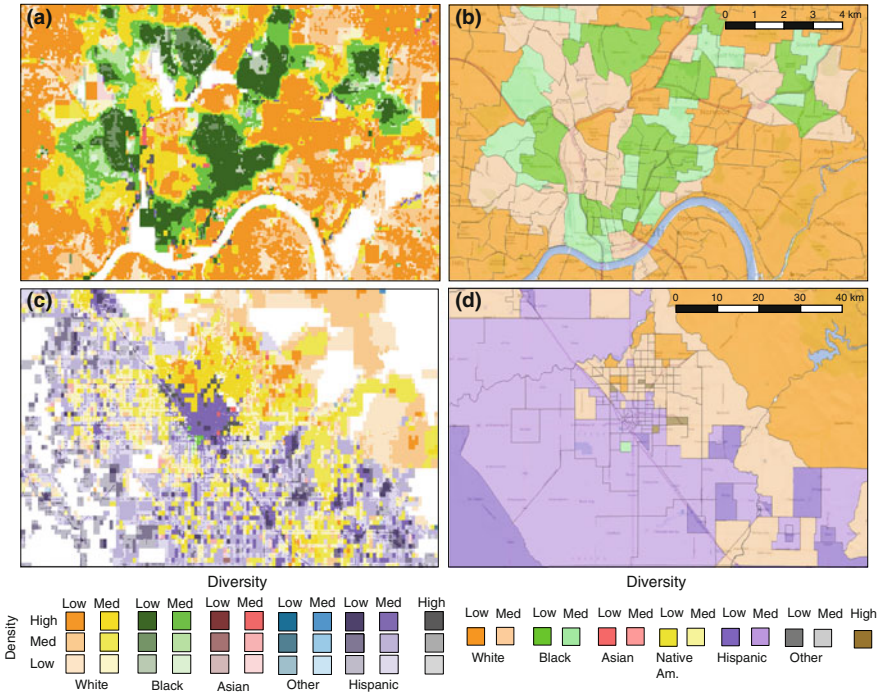

Fig. 3 A comparison of racial diversity maps for sites in urban and rural settings. **a** Racial diversity, 90 m grid for the Cincinnati (Ohio) area, **b** Mixed Metro census tracts-based map for the Cincinnati area, **c** Racial diversity, 90 m grid for the Fresno (California) area, **d** Mixed Metro census tracts-based map for the Fresno area

to race/ethnicity into the following groups: whites, blacks, Asians, Hispanics, and others. Each group is disaggregated separately in a way to preserves the value of total population in each 90 m cell. Using values of population densities for all race groups, we classify grid cells into 33 categories, taking into account diversity level (low, medium or high), dominant race, and population density (low, medium, and high). Uninhabited areas are grouped into the separate 34th category. Details of this classification are given in Dmowska and Stepinski (2014). The two diversity grids shown in Fig. 3 are parts of the U.S.-wide racial diversity grid that can be explored in the SocScape web application. For comparison, we also show (Fig. 3b and d) racial diversity maps of the same sites that are available from http://mixedmetro.com/. The Mixed Metro maps are not grids; instead, they are based on census tracts. They are produced using a classification of population into 13 diversity/race categories. Their classification is similar to ours, but does not include population density, and uses division into six instead of five racial groups. The legends of the two classifications have been constructed to correspond to each other as much as possible.

Comparing the two maps in an urban site (Fig. 3a and b), we observe that they roughly correspond to each other in delineating white-dominated and black-dominated areas. The grid-based map has a better resolution, and also distinguishes between inhabited and uninhabited areas. Therefore, it is more useful than the Mixed Metro map for assessing racial diversity of a neighborhood at the street scale. The same conclusion holds when comparing the two maps in the rural site (Fig. 3c and d). Because census tracts in more sparsely populated areas are larger, the resolution of the Mixed Metro map is worse than in the urban setting. The Mixed Metro map also can be misleading due to its lack of delineation of uninhabited areas and its lack of information about population density. A good example of this problem is observed in the upper right-hand corner of the map (Fig. 3d), which indicates a large area with population dominated by whites. However, our map (Fig. 3c) correctly depicts this region as uninhabited or very sparsely populated. The Mixed Metro map also fails to convey that Hispanic-dominated areas are urban clusters that are not as wide-spread as the Mixed Metro map suggests.

4 Conclusions

Our project to develop high-resolution population and demographic grids for the entire United States has resulted in 90 m population density grids for the years 1990 and 2000, and a 30 m grid for 2010. It has also resulted in 90 m racial diversity grids for 1990 and 2000. These grids are available for exploration and downloading from the SocScape web application at http://sil.uc.edu.

The Gen-2 method, used to calculate a 2010 edition of the population grid, cannot be fully applied to 1990 and 2000 data because the land use data (Theobald 2014) pertain only to 2010. However, the Gen-2 method can recalculate population grids for 1990 and 2000, but with land cover as the only ancillary data. The Gen-2 method is a significant improvement over our older Gen-1 method (Dmowska and Stepinski 2014) because it only uses original census data, which gives us a full control over the quality of its results. It is difficult-to-impossible to perform a formal assessment of our grids, accuracy. Such an assessment requires comparison with sub-block resolution data, such as, for example, parcel data. This is not feasible for the entire United States but can be performed with very small regions for which parcel data have been utilized to calculate population density. Using results available in the literature, we conclude that our 30 m population grid agrees well with parcel-derived population density maps for a small area in Alachua County, Florida (Jia et al. 2014).

Grids of demographic variables other than population density can be calculated using weights established by the population model. Examples of such variables (available at the census block level) are race, age, and income. No ancillary data specific to race, age, or income exist that would allow directly disaggregating these variables within a block; but we can disaggregate them according to the population model. By narrowing the locations where people live within a block, we increase the spatial resolution of these variables, although we would not be able to account for

variation of, for example, racial diversity within a block. Nevertheless, as illustrated by Fig. 3, using grids of demographic variables (like racial diversity) instead of maps based on census areal units results in a much better depiction of the actual spatial distribution of these variables.

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

References

Anderson JR, Hardy EE, Roach JT, Witmer RE (1976) A land use and land cover classification system for use with remote sensor data. Tech rep, Geological Survey Professional Paper 964

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

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

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. Appl Geogr 24(2):97–117

Dmowska A, Stepinski TF (2014) High resolution dasymetric model of US demographics with application to spatial distribution of racial diversity. Appl Geogr 53:417–426

Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA (2000) LandScan: a global population database for estimating populations at risk. Photogram Eng Remote Sens 66(7):849–857

Dong P, Sathya R, Nepali A (2010) Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data. Int J Remote Sens 31(2):5571–5586

Eicher CL, Brewer CA (2001) Dasymetric mapping and areal interpolation: implementation and evaluation. Cartogr Geogr Inf Sci 28:125–138

Fry JA, Coan MJ, Homer CG, Meyer DK, Wickham JF (2009) Completion of the National Land Cover Database (NLCD) 1992–2001 land cover change retrofit product. Tech rep, U.S. Geological Survey Open-File Report 2008–1379

Gallego FJ (2010) A population density grid of the European Union. Popul Environ 31(6):460–473

Gallego FJ, Batista F, Rocha C, Mubareka S (2011) Disaggregating population density of the European Union with CORINE land cover. Int J Geogr Inf Sci 25(12):2051–2069

Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ (2013) High resolution population distribution maps for Southeast Asia in 2010 and 2015. PLoS One 8(2):e55,882

Goodchild M, Anselin L, Deichmann U (1993) A framework for the areal interpolation of socioeconomic data. Environ Plann A 25:383–397

Jia P, Qiu Y, Gaughan AE (2014) A fine-scale spatial population distribution on the high-resolution gridded population surface and application in Alachua County, Florida. Appl Geogr 50:99–107

Langford M, Unwin D (1994) Generating and mapping population density surfaces within a geographical information system. Cartogr J 31(1):21–26

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 7(2):e31,743

Mennis J (2003) Generating surface models of population using dasymetric mapping. Prof Geogr 55(1):31–42

Mennis J, Hultgren T (2006) Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci 33(3):179–194

Qiu F, Cromley R (2013) Areal interpolation and dasymetric modeling. Geogr Anal 45(3):213–215

Reibel M, Bufalino ME (2005) Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. Environ Plann A 37(1):127–139

Tatem AJ, Noor AM, vonHagen C, DiGregorio A, Hay SI (2007) High resolution population maps for low income nations: combining land cover and census in East Africa. PLoS One 2(12):e1298
Theobald DM (2014) Development and applications of a comprehensive land use classification and map for the US. PloS One 9(4):e94,628
Weber N, Christophersen T (2002) The influence of non-governmental organisations on the creation of Natura 2000 during the European Policy process. For Policy Econ 4(1):1–12
Wright J (1936) A method of mapping densities of population: with Cape Cod as an example. Geogr Rev 26(1):103–110