High resolution dasymetric model of U.S demographics with application to spatial distribution of racial diversity

Anna Dmowska[a,b], Tomasz F. Stepinski[a,*]

[a] *Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, USA, OH 45221-0131, USA*
[b] *Geoecology and Geoinformation Institute, Adam Mickiewicz University, Dziegielowa 27, 60-680 Poznan, Poland*

Abstract

Population and demographic data at high spatial resolution is a valuable resource for supporting planning and management decisions as well as an important input to socio-economic academic studies. Dasymetric modeling has been a standard technique to disaggregate census-aggregated units into raster-based data of higher spatial resolution. Although utility of dasymetric mapping has been demonstrated on local and regional scales, few high resolution large-scale models exist due to their high computational cost. In particular, no publicly available high resolution dasymetric model of population distribution over the entire United States is presently available. In this paper we introduce a $3''$ ($\sim$ 90 m) resolution dasymetric model of demographics over the entire conterminous United States. The major innovation is to disaggregate already existing $30''$ ($\sim$ 1 km) and $7.5''$ ($\sim$ 250 m) SEDAC (Socioeconomic Data and Applications Center) Census 2000 grids instead of the original census block-level data. National Land Cover Dataset (NLCD) 2001 is used as ancillary information. This allows for rapid development of a U.S.-wide model for distribution of population and sixteen other demographic variables. The new model is demonstrated to markedly improve spatial accuracy of SEDAC model. To underscore importance of high spatial resolution demographic information other than total population count we demonstrate how maps of several population characteristics can be fused into a "product" map that illustrates complex social issues. Specifically, we introduce a "diversity" categorical map that informs (at nominal $3''$ resolution) about spatial distribution of racial diversity, dominant race, and population density simultaneously. Diversity map is compared to a similar map based on census tracts. High resolution raster map allows study of race-diversity phenomenon on smaller scale, and, outside of major metropolitan areas, revels existence of patterns that cannot be deduced from a tract-based map. The new high resolution population and diversity maps can be explored online using our GeoWeb application DataEye available at http://sil.uc.edu/. Both datasets can be also downloaded from the same website.

Keywords:
Population density, gridded demographic data, dasymetric modeling, racial diversity, census

1. Introduction

High resolution (hi-res) data on population distribution is needed to address many important issues including assessing human pressure on environment (Weber and Christophersen, 2002), quantifying environmental impact on population (Vinkx and Visee, 2008), and characterizing populations at risks from natural hazards (Dobson et al., 2000; Chen et al., 2004; Tralli et al., 2005; Thieken et al., 2006). Bhaduri et al. (2002) list many more potential applications for hi-res population distribution data. In addition, if such data is in a raster format and covers large spatial extent (continental or global scale) it could be used for comparative analysis of its local, constituent patterns. Recently, tools for query and retrieval of local spatial patterns have been developed (Jasiewicz and Stepinski, 2013; Stepinski et al., 2014) in the context of high resolution land cover data; generalization of such tools to population/census rasters is straightforward. Such tools provide means for rapid, real time exploration of vast population distribution datasets. They could be particularly valuable if the population data is enhanced by demographic information (sex, age, race) and socio-economic

*Corresponding author. *Email address:* stepintz@uc.edu
Email addresses: dmowskaa@ucmail.uc.edu (Anna Dmowska), stepintz@uc.edu (Tomasz F. Stepinski)

information (income, education, etc.).

For the purpose of this paper we arbitrarily define the hi-res datasets as those having spatial resolution of at least 100 m. Many local scale hi-res population/census data have been developed, but the availability of hi-res, large scale datasets is still limited. Population, demographic, and socio-economic data are collected through censuses at the household level (an ultimate high resolution) but released in an aggregated form because of confidentiality and privacy concerns. For example, the smallest aggregation areal unit released by the U.S. Census Bureau is a census block; there are over 8 millions such polygonal units in the U.S. The arbitrary nature of areal unit partitioning (the size and population of census blocks varies greatly) limits the usefulness of unit-based dataset for spatial analysis. Additional limitation stems from changes to the boundaries of aggregation units from one census to another making change analysis difficult. A standard approach to obtaining spatial analysis-ready population/census dataset is to transform unit-based data into raster-based data having cell size smaller than a majority of areal units. Such procedure is referred to as disaggregation, spatial decomposition (Bhaduri et al., 2007) or downscaling (Gallego, 2010).

Numerous methods for achieving units-to-raster transformation has been developed; they falls into two basic categories: areal weighting and dasymetric modeling. In areal weighting method (Goodchild and Lam, 1980; Flowerdew and Green, 1992; Goodchild, 1993) a regular grid is intersected with units polygons and each grid cell is assigned a value based on the proportion of the polygon contained in each cell. Dasymetric modeling (Wright, 1936; Langford and Unwin, 1994; Eicher and Brewer, 2001; Mennis, 2003) refers to a process of disaggregating unit-based data to a finer grid-based data using ancillary data to refine population distribution. Land use/land cover (LULC) maps are most often used (Tian et al., 2005; Monmonier and Schnell, 1984) as ancillary data because of high correlation between LULC category and population density. Dasymetric modeling is not restricted to using a single ancillary data source, using multiple sources have been proposed (Langford and Unwin, 1994) and applied in practice (Bhaduri et al., 2007).

Disaggregation methods are well established and straightforward to apply. However, their application to production of global or continental scale hi-res population datasets are hindered by the need to handle big datasets and the limited availability of hi-res LULC maps covering required spatial scale. For the countries in the European Union (EU) the hi-res, 100 m/cell population grid has been developed (Gallego, 2010; Gallego et al., 2011) using population data from the 2000/2001 round of censuses aggregated to nearly 115,000 areal units called "communes" and performing dasymetric mapping using 100 m/cell raster version of CORINE Land Cover 2000 as an ancillary data. This dataset is available from the European Environment Agency data warehouse (http://dataservice.eea.europa.eu/). The WorldPop project (http://www.worldpop.org.uk/) provides 100 m/cell population grids for most countries in Africa Asia, as well as South and Central Americas. WorldPop uses population census data, official population size estimates and corresponding administrative unit boundaries at the highest level available for each country. For Africa census data is disaggregated (Tatem et al., 2007; Linard et al., 2012) using land cover data from GlobeCover (http://due.esrin.esa.int/globcover/). For Asia and Americas census data is disaggregated (Gaughan et al., 2013) using land cover data from MDA GeoCover.

Finally, for the United States the Oak Ridge National Laboratory is developing (Bhaduri et al., 2007) the LandScan USA - 90 m/cell population distribution dataset that uses the U.S. Census Bureau block-level population data and the National Land Cover Dataset (NLCD) 30 m/cell land cover auxiliary data as its primary components. Alone among other products, LandScan USA provides both nighttime as well as daytime population distributions. It uses a multi-dimensional dasymetric modeling approach utilizing variety of data in addition to LULC data to disaggregate the census blocks. However, LandScan USA is not currently available, nor is it expected to be in the public domain once it become available thus limiting its utility to the scientific community.

In this paper we document our development of 90 m/cell demographic grids for the conterminous United States. Instead of using aggregated areal units as an input for our modeling we use already existing grids of coarser resolution. Specifically, we use U.S. Census Grids - a ~1 km/cell dataset developed by the Socioeconomic Data and Application Center (SEDAC) (http://sedac.ciesin.columbia.edu/). SEDAC grids cover not only population counts but also demographic data (age, race, sex) and socio-economic data (income, education, etc.). They are a product of simple areal weighting interpolation from census blocks, no auxiliary data has been used for their creation. We sharpen selected SEDAC grids from ~1 km/cell to ~90 m/cell using land cover (NLCD) auxiliary data. Thus, we transform coarser raster to finer raster instead of transforming census polygons to a raster. This allows for fast

development of grids with markedly higher spatial resolution than the original SEDAC model. Models can be calculated for 1990 and 2000 censuses for which SEDAC data presently exist. Here we focus on sharpening selected SEDAC data pertaining to 2000 census using 2001 NLCD. Our model is best utilized for socio-economic research where spatial accuracy combined with moderate population count accuracy is sufficient to bring new insights. To demonstrate its utility we compute a "diversity" categorical map that informs (at nominal $3''$ resolution) about spatial distribution of racial diversity, dominant race, and population density simultaneously.

2. Data and its pre-processing

The input to our dasymetric model is provided by the SEDAC 2000 U.S. Census Grids. SEDAC grids are created by taking population and housing counts at the block level and performing areal weighting interpolation to a latitude-longitude quadrilateral grid. SEDAC uses TIGER/Line files for the census block boundaries and the U.S. Census Bureau summary file 1 (SF1) and summary file 3 (SF3) tables for the demographic and socioeconomic characteristics of each census block.

SEDAC provides two sets of grids at different spatial resolutions. First, the $30''$ (approximately 1 km) resolution set covers the entire conterminous U.S.; we refer to this set as SEDAC-US data. Second, the $7.5''$ (approximately 250 m) resolution set covers top 50 metropolitan statistical areas (MSA) with population over 1 million; we refer to this set as SEDAC-MSA data. Both, SEDAC-US and SEDAC-MSA sets consist of 40 different grids each characterizing different demographic or socioeconomic characteristic. Some of these grids are calculated from the SF1 while other from the SF3. The SF3 data is released at the census block group level and needs to be proportionately allocated to census blocks using the distribution of the underlying SF1 population before performing areal weighting interpolation to a latitude-longitude quadrilateral grid.

We have selected 17 demographic/socioeconomic characteristics available as SEDAC grids to perform dasymetric modeling and to obtain hi-res grids; our selection is listed in Table 1 together with indication whether the original data comes from SF1 or SF3. Note that definitions of characteristics in "Income" group differ from agglomerative definitions of original census data. The last column in Table 1 shows difference (expressed in %) between total conterminous U.S. population (within a group) according to SEDAC-US and our dasymetric model (see discussion in section 6). We use $30''$ SEDAC grids as a base to be sharpened by dasymetric modeling and we use the 30 m resolution National Land Cover Dataset 2001 (NLCD 2001) and $7.5''$ SEDAC-MSA data as ancillary data to refine the base grids to $3''$ (approximately 90 m) resolution. The NLCD 2001 is given in the Albers USGS projection and is reprojected to a latitude-longitude quadrilateral grid with resolution of $1''$ in order to be used as ancillary data in dasymetric modeling.

2.1. Pre-processing of SEDAC data

We have encountered several issues with the SEDAC grids and needed to pre-process them in order to avoid inconsistencies within our model. First, the SEDAC-MSA data does not distinguish between "no data" and data with value equal to zero; both are given the value of 0. This causes problems especially in cases when two MSAs overlay each other (which is the case for some MSAs, especially those located in the eastern seaboard). We pre-process the SEDAC-MSA data in order to disambiguate between cells having "no data" values and those having 0 values. We assume that MSA cells surrounded by cells with non-zero values are having values of 0, other cells, those located on the periphery of the city and not surrounded by cells with non-zero values are assigned "no-data" assignation.

The second issue is that summing over all subclasses of SEDAC values within a group does not result in the SEDAC total population value within a given grid cell. For example summing up cell values of all seven age categories (as shown in Table 1) should result in the value of total population in this cell but it does not. This problem can be traced to the fact that SEDAC provides only integer values for population counts even so actual values must have been non-integer as they are the results of areal weighting interpolation. Rounding the number up introduces the observed issue which is most noticeable in the regions characterized by low population density. For demographic/socioeconomic characteristics listed in Table 1 (except total population) we pre-process the values of SEDAC grids using the following formula:

$$R_i^{\text{new}} = R_i^{\text{old}} \frac{Pop}{\sum_i R_i^{\text{old}}} \tag{1}$$

where R_i^{new} is the re-calculated cell value of i-th characteristic in a group, R_i^{old} is the original cell value of i-th characteristic in a group, Pop is the cell value of total population, and summation is over all subclasses in the group. Thus, for example, for the "age" group in Table 1, the summation is over seven age subclasses. This

Table 1: Selected demographic/socioeconomic characteristic

Groups	Selected characteristics	Abb.	Summary File	Difference [%]
Total population	Population	pop	SF1	0.001
Race and Ethnicity	Non-Hispanic White	nhw	SF1	0.026
	Non-Hispanic Black	nhb	SF1	0.056
	Asian alone	as	SF1	0.007
	American Indian and Alaska Native alone	am	SF1	0.137
	Native Hawaiian and other Pacific Islander alone	pi	SF1	0.027
	Hispanic	hi	SF1	0.002
Age	Population under age 1	a1	SF1	0.166
	Population ages 1 to 4	a2	SF1	0.221
	Population ages 5 to 17	a3	SF1	0.245
	Population ages 18 to 24	a4	SF1	0.188
	Population ages 25 to 64	a5	SF1	0.217
	Population ages 65 to 79	a6	SF1	0.273
	Population age 80 and older	a7	SF1	0.204
Income	Population living below 50% of the poverty level	sevp	SF3	0.001
	Population living between 50% and 100% of the poverty level	pov	SF3	0.001
	Population living between 100% and 200% of the poverty level	lowi	SF3	0.001

transformation assures consistency of the data. Note that it makes the values in SEDAC grids non-integer.

3. Methodology

As mentioned above we use two ancillary data sources to refine the SEDAC-US grids. The NLCD 2001 data is the primary ancillary source, and it is the only data source in the regions where higher resolution SEDAC-MSA is not available. NLCD 2001 has an overall accuracy of over 83%, the regional accuracy changes slightly (Wickham et al., 2010) reflecting diversity of dominant land cover classes. The accuracy of urban classes is higher than average. The SEDAC-MSA data is used as a secondary source of refinement and is applied together with the NLCD 2001 in regions where it is available.

First consider only the SEDAC-US total population grid (the first entry in Table 1). Each $30''$ grid cell gives a total population count. This cell contains 900 smaller, $1''$ cells for which NLCD 2001 labels are assigned. NLCD legend for the conterminous U.S. has 16 different land cover/land use categories (http://www.mrlc.gov/nlcd2001.php), and each SEDAC-US cell can be considered as a pattern or mosaic of these categories. The idea of dasymetric modeling is to redistribute the total population count within a cell into its constituent sub-cells using NLCD information. We use two different types of NLCD information for dasymetric modeling. The first information type is an average population density for each NLCD category which provides an empirical association between land cover/land use category and population density. The second information type is a description of the category in the NLCD legend which provides semantic guide for association between land cover/land use category and population density. Average densities may not be always sufficient because they are calculated from data sources (SEDAC-US and NLCD) with vastly different spatial scales. For example, a category "developed, open space" has an average density of 162 people/km^2 which seems at odds with its semantic description as "areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes." This description, together with comparison of the NLCD 2001 map with satellite image mosaic, points out to the calculated average density being an overestimation resulting from the fact that, in the urban settings and on the spatial scale of $30''$, this category is intermingled with other urban categories having high population density.

Based on an empirical average densities and our judgment resulting from NLCD legend we have reclassified the 16 NLCD categories into 6 classes for which we have assigned relative population densities (d) as follows: (1) "developed, high intensity" ($d_1 = 14.0$), (2) "developed, medium intensity" ($d_2 = 13.0$), (3) "developed, low intensity" ($d_3 = 6.0$), (4) "developed, open space" ($d_4 = 0.01$), (5) "water" ($d_5 = 0$), (6) "other" ($d_6 = 0.01$). Only the relative values of those densities

are important, not their absolute magnitudes.

In principle we could refine the SEDAC-US total population grid all the way down to 1″ which is the resolution of the ancillary NLCD data; this is a standard dasymetric modeling procedure. However, we don't believe there is a justification for having demographic data at such fine resolution. Instead, we refine the SEDAC-US data to 3″ (∼ 90 m) which is a sufficiently high resolution for such data. Thus one SEDAC-US grid cell contains 10×10=100 target cells. In turn, each target (3″) cell contains 3×3=9 NLCD cells but is assigned only a single population count value.

Let (i, j), $i = 1, \ldots, 10$, $j = 1, \ldots, 10$ be an index (x and y coordinates) of a target cell in a given SEDAC-US cell. A weight $W^{(i,j)}_{\mathrm{NLCD}}$ is a fraction which, when multiplied by a population count for the entire SEDAC-US cell, yields a population count in the (i, j) target cell. $W^{(i,j)}_{\mathrm{NLCD}}$ is given by the following formula:

$$W^{(i,j)}_{\mathrm{NLCD}} = \frac{A^{3''}_1(i,j)\, d_1 + \cdots + A^{3''}_6(i,j)\, d_6}{A^{30''}_1\, d_1 + \cdots + A^{30''}_6\, d_6} \qquad (2)$$

where $A^{30''}_k$ is an area of the k-th reclassified NLCD category within a SEDAC-US cell and $A^{3''}_k(i, j)$ is an area of the k-th reclassified NLCD category within a target cell (i, j). Note that the denominator in eqn. 2 is the sum (over the entire range of indices i and j) of numerator terms for all target cells. Thus, the sum of all weights is equal to 1. The weights redistribute population count in the SEDAC-US cell into its constituent target cells; target cells with large weights "attract" people whereas target cells with small weights "repulse" people. If a SEDAC-US cell contains only a single NLCD category there is no redistribution; each target cell is assigned an equal share of population count.

From a procedural point of view it is convenient to consider SEDAC-MSA as an ancillary data to be used to sharpen the SEDAC-US data. We first resample the SEDAC-MSA from 7.5″ to 3″ resolution using linear interpolation so each SEDAC-US grid cell contains 100 target cells for which SEDAC-MSA data provides population counts and other characteristics. The sharpening weights due to SEDAC-MSA are given by the following formula:

$$W^{(i,j)}_{\mathrm{MSA}} = \frac{V(i,j)}{\sum_{i,j} V(i,j)} \qquad (3)$$

where $V(i, j)$ is the SEDAC-MSA value associated with a target cell (i, j) and the summation in the denominator is over the entire range of indices i and j. Multiplying $W^{(i,j)}_{\mathrm{MSA}}$ by a population count for the entire SEDAC-US cell, yields a population count in the (i, j) target cell. Note that this count may be somewhat different from the $V(i, j)$ because SEDAC data lacks consistency between 7.5″ and 30″ resolution editions.

In places where SEDAC-MSA is available we need to combine sharpening information from NLCD and SEDAC-MSA. The problem is analogous to that encountered in the field of risk analysis when combining multiple expert judgments (Clemen and Winkler, 1999). Here, we have two "expert judgments", one provided by the NLCD and another by the SEDAC-MSA. In the absence of any likelihood function associated with the experts' information we use multiplicative averaging (called a logarithmic opinion pool in risk analysis). The final weights are given by the following formula:

$$W^{(i,j)}_{\mathrm{TOT}} = \begin{cases} \frac{W^{(i,j)}_{\mathrm{NLCD}} \times W^{(i,j)}_{\mathrm{MSA}}}{\sum_{i,j} W^{(i,j)}_{\mathrm{NLCD}} \times W^{(i,j)}_{\mathrm{MSA}}}, & \text{in MSA} \\ W^{(i,j)}_{\mathrm{NLCD}}, & \text{elsewhere} \end{cases} \qquad (4)$$

For modeling grids other than total population (the last 16 entries in Table 1) we also use eqn. 4 except that the weights $W^{(i,j)}_{\mathrm{MSA}}$ are calculated using values $V(i, j)$ indicating a given characteristic rather than a total population count. Thus, within the MSAs our model redistributes each demographic/socioeconomic characteristic independently, but outside the MSA, due to lack of data, all characteristics are redistributed using the same redistribution weights based on the land cover/land use categories.

4. Computation and results

The major challenge to calculating 3″ (∼90 m) dasymetric models of population density and other demographic/socioeconomic characteristics for the entire conterminous U.S. is the sizes of the input and output files. Table 2 summarizes sizes and properties of input and output files. SEDAC-MSA data is not listed because it consists of 50 smaller, separate files. The sizes and properties of 3″ models for demographic/socioeconomic characteristics other than total population are the same as those for the population model listed in Table 2. The diversity model is a product computed from race and ethnicity characteristics and is discussed in section 5. All computations were performed using Python script written for GRASS 7.0 software (Neteler and Mitasova, 2007).

The computation consists of several steps that follows the methodology described in the previous section. The

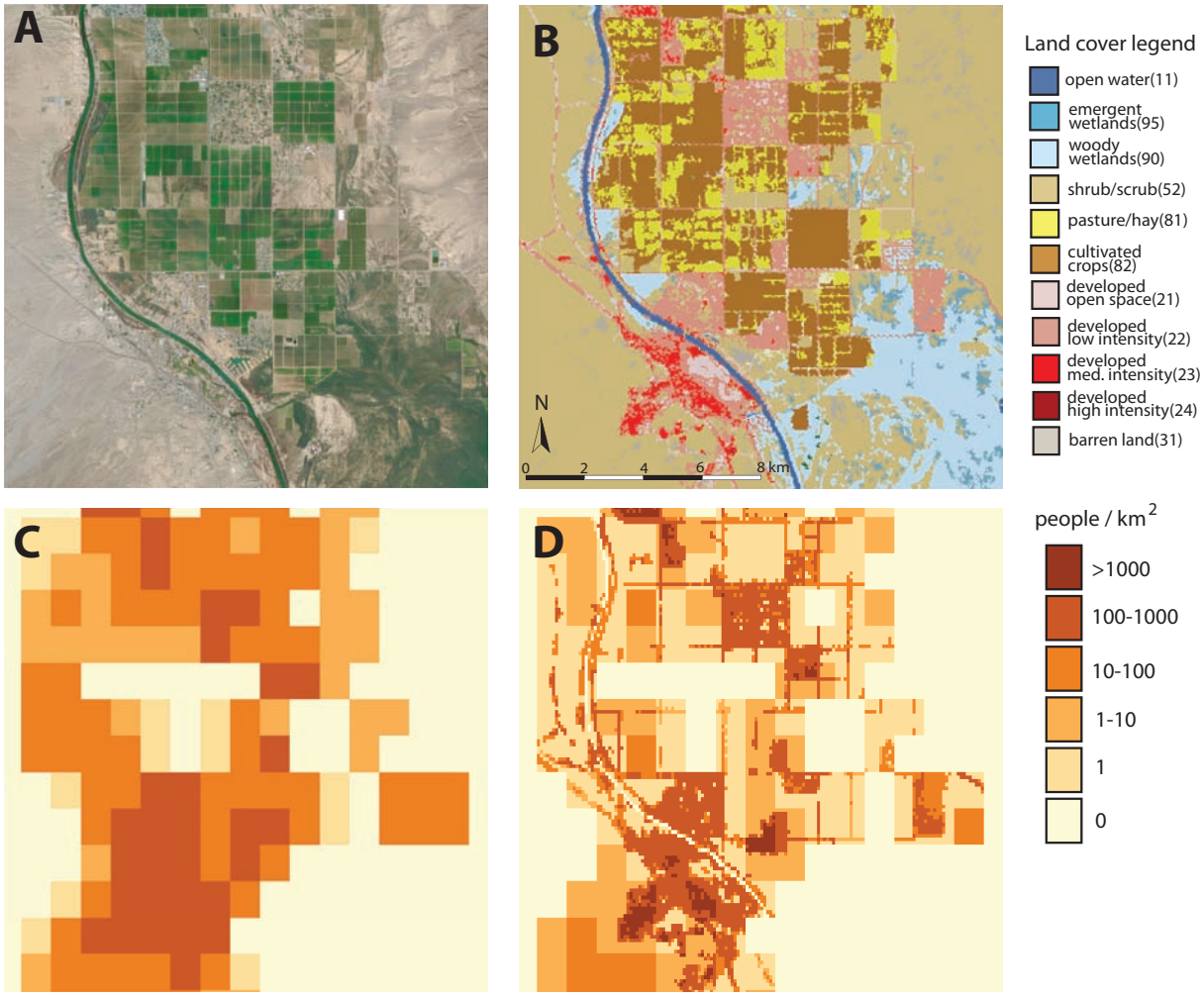

Figure 1: The city of Needles located on the banks of the Colorado River at the California-Arizona-Nevada border.(A) Satellite image (Google Maps). (B) Land cover/land use map (NLCD 2001). (C) Population density as shown by 30″ (~1 km) SEDAC-US grid. (D) Population density as shown by our 3″ (~90 m) dasymetric model. NLCD 2001 legend is shown with names of land cover categories and their numerical codes. Maps are shown in the Google (Mercator) projection.

Table 2: Sizes of input and output datasets

Data sets	Resolution	Rows	Columns	# of cells	NoData cells	Type	Size	Zipped size
				Input				
NLCD 2001	1"	104400	244800	25,557,120,000	58.85%	8-bit int	19,500 MB	1,100 MB
SEDAC-US	30"	3480	8160	28,396,800	58.85%	16-bit	5.9 MB	5.5 MB
				Output				
population model	3"	34800	81600	2,839,680,000	58.85%	32-bit int	11,400 MB	383 MB
diversity model	3"	34800	81600	2,839,680,000	58.85%	8-bit int	3,460 MB	76 MB

most time-consuming step of computation is the calculation of $W_{\mathrm{NLCD}}^{(i,j)}$ (Eqn. 2). To achieve this step we have divided the area of the entire conterminous U.S. into 48 tiles each having size of 6°in the north-south direction and 10° in the east-west direction. The weights and the final values for our dasymetric model were calculated in each tile separately and the results were combined into a single map. This part of our calculations took $\sim$ 12 hours using a computer with Intel 3.4GHz, 4-cores processor and 16GB of memory running the Linux system.

The resultant 3″ models of total population and diversity are available for download from http://sil.uc.edu/ as GeoTiff rasters in latitude/longitude projection. A value of a cell in the population dataset is a number of people per cell. Notice that this number may not be an integer as we keep model results without rounding them off to integers. However, for distribution purposes all cell values are multiplied by 10^6 and saved as 32-bit integers. A value of a cell in the diversity dataset is one of 33 diversity labels (see section 5). In addition, these two models can be explored in their full extent and full spatial resolution using our interactive GeoWeb application DataEye available from the same website. The DataEye versions of the models are 90 m resolution and are shown in Albers USGS projection in order to conform to other resources available within this tool. The DataEye version of the population model has been reclassified to just ten categories and should only be used for exploration purposes. For access to the models of the remaining demographic/socioeconomic characteristics listed in Table 1 please contact the corresponding author.

Our models offer significant improvements in spatial accuracy over the original SEDAC grids. In order to fully appreciate these improvements the new models need to explored using the GeoWeb tool (see above). Here we present two examples, one from a region where only SEDAC-US data is available and another from a region where also SEDAC-MSA data is available.

Fig. 1 shows a city of Needles located in the Mojave Desert on the western banks of the Colorado River in San Bernardino County, California. A satellite image (panel A) and land cover map (panel B) reveal desert landscape with irrigated agricultural activity; they show a ground truth for spatial distribution of population. Populated areas are concentrated in the city center (lower-left part of the site on the western bank of the Colorado River), along banks of the river, along the highways (Interstate I-40 crossing the city and a road parallel to the river and heading north). Additional populated areas are interwove with agricultural land located in the central and north-central parts of the site. The SEDAC-US 30″ grid (recalculated from counts to population density) is shown in panel C and our 3″ model is shown in panel D. Comparison of these two population maps to each other and to the satellite image and land cover map clearly shows advantages of our model over SEDAC in spatial accuracy. Our model correctly reflects major features of the population distribution in this site. On the other hand, the SEDAC grid is too coarse to capture most features of the population distribution. In particular, the presence of the river and the roads are not captured by the SEDAC grid.

Fig. 2 shows an area centered on downtown Cincinnati, OH. This is a site for which both SEDAC-US and SEDAC-MSA grids are available. Satellite image (panel A) and land cover map (panel B) reveal presence of 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. The SEDAC-US grid (panel C) with the resolution of 30″ captures only most basics features of population distribution - enhanced population density in the northwest and the northeast parts of the site with somewhat decreased density in the central and southern parts of the site. The presence of the river and of the industrial-transportation corridor cannot be deduced from this map. The SEDAC-MSA grid (panel D) with the resolution of 7.5″ shows some additional details with the river and the industrial-transportation corridor now reflected in the population distribution However,

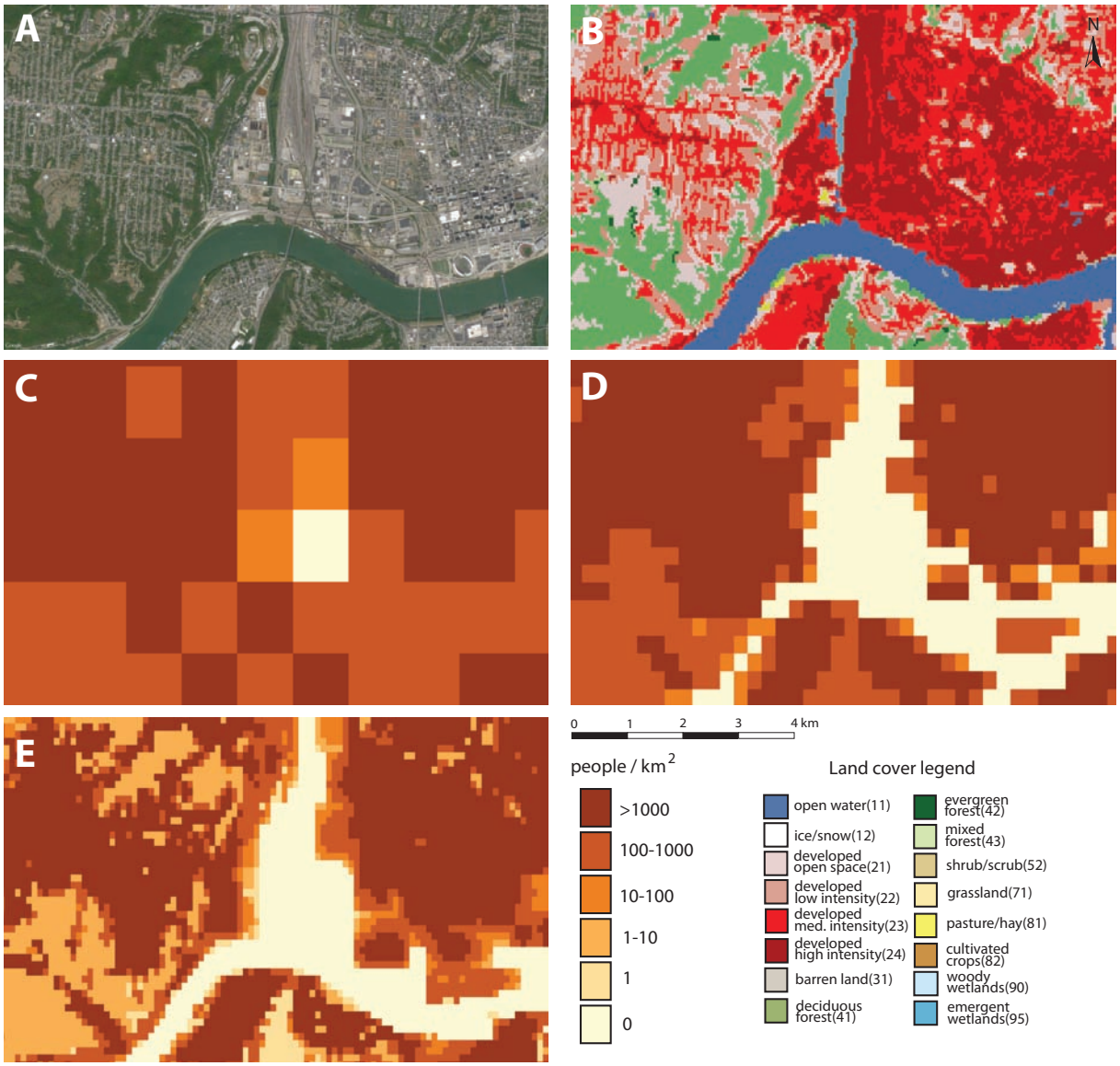

Figure 2: Downtown Cincinnati, Ohio. (A) Satellite image (Google Maps). (B) Land cover/land use map (NLCD 2001). (C) Population density as shown by 30″ (~1 km) SEDAC-US grid. (D) Population density as shown by 7.5″ (~250 m) SEDAC-MSA grid. (E) Population density as shown by our 3″ (~90 m) dasymetric model. NLCD 2001 legend is shown with names of land cover categories and their numerical codes. Maps are shown in the Google (Mercator) projection causing square cells to appear slightly elongated.

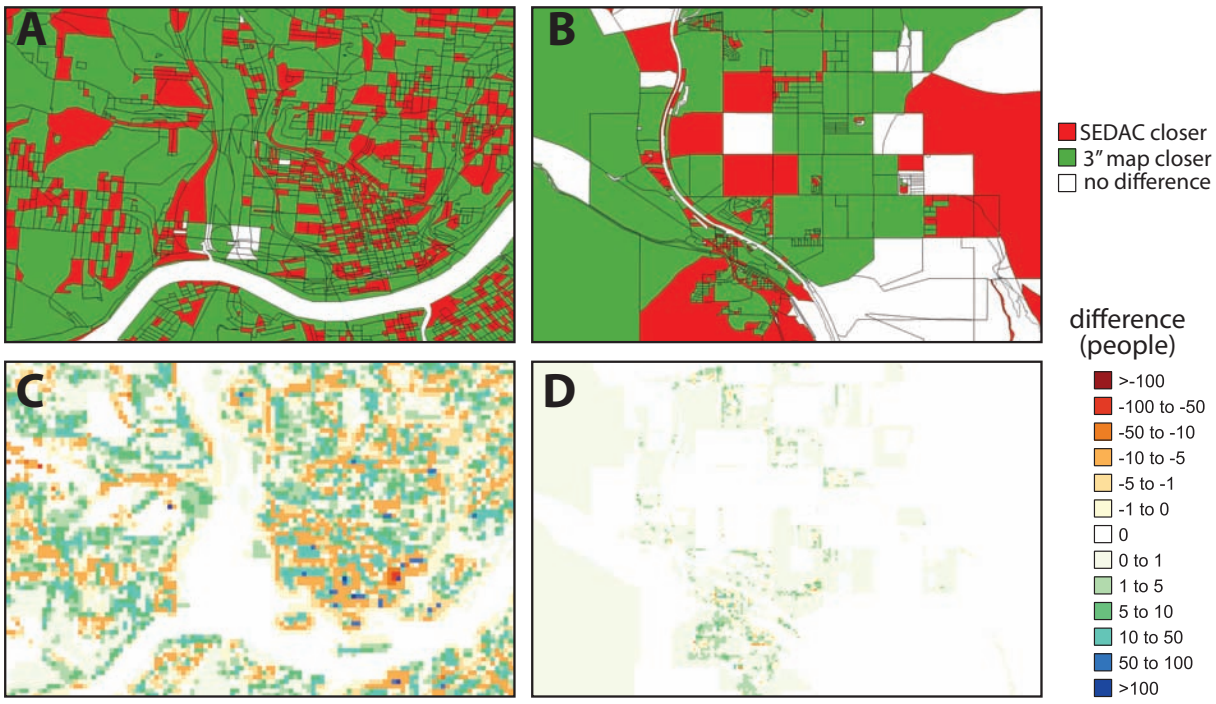

Figure 3: (A-B) Integration of SEDAC and our model over individual census blocks; color indicates which model integrates closer to the census block-level population count; (A) downtown Cincinnati, Ohio, (B) Needles, California. (C-D) Difference in population counts between the standard dasymetic model and our model; (C) downtown Cincinnati, Ohio, (D) Needles, California.

populated areas still suffer from poor resolution with parks and other green spaces not distinguished from built-up areas where population concentrates. Our dasymetric model grid (panel E) with the resolution of ~90 m offers superior details in comparison with SEDAC-MSA. The river and the industrial-transportation corridor are well delineated and parks and green spaces are clearly distinguished from built-up areas. DataEye tool should be used for more in-depth exploration of population distribution within this site.

As expected, our model offers only moderate improvement over SEDAC in an accuracy of population counts when integrated back to the block-level. To make a comparison between the two models we integrate the values of each model cells over extent of a census blocks and compare the results to population counts as given in the block-level census data. The results of this comparison are shown for the Cincinnati area (Fig. 3A) and for the Needles area (Fig. 3B). The red color indicates blocks for which integration of SEDAC values yields population closer to a block-level census count, the green color indicates blocks for which integration of our model values yields population closer to a block-level count, and the white color indicates census blocks for which there is no significant difference between the two models. Statistics over all census blocks in the state of Ohio shows that our model get closer to block-level values for 47% of the blocks, the SEDAC model gets closer for 21% of the blocks, and there is no difference for 32% of the blocks.

We also compare an accuracy of population counts between our model and the standard dasymetric model calculated by disaggregating census blocks using the NLCD as ancillary information. For the purpose of such comparison we have calculated the standard dasymetric model for the two test areas (Cincinnati and Needles) and compared the difference in population counts between the two models. Fig. 3C shows the results for the Cincinnati area and Fig. 3D shows the results for the Needles area. A displayed variable is a difference between standard dasymetric model and our model. The negative values are in the places where our model overcounts the population and the positive values are in places where it undercounts the population with respect to the standard model. Differences are most pronounced in densely populated areas, where census blocks are small.

5. Application to distribution of racial diversity

To demonstrate how our hi-res grids can be utilized as a resource for social sciences we present a specific example relating to the issue of spatial distribution of racial diversity within the U.S. population. This is not meant to be a throughout investigation of this issue but rather an illustration of how our grids can be combined into a single "product" – a high resolution map of diversity categories – that serves as an input to further research.

The U.S. society is becoming more racially diverse. Mapping spatial distribution of race and diversity and understanding its correlation with other factors, such as, for example, access to institutional and economic resources, exposure to crime, pollution, and health disparities is needed to inform social policy decision makers. Consequently, a body of work exists on mapping race and diversity (Holloway et al., 2012; Farrell and Lee, 2011; Reardon and Firebaugh, 2002), although it focuses exclusively on major metropolitan areas. Two issues had arisen: (1) How to measure diversity quantitatively? (2) How to select a spatial scale over which diversity is assessed? Using our hi-res demographic grids can help to address both of these issues, but it is the issue #2 that we focus on here.

The existing body of work maps diversity at the level of census tract. A usual rationale (Iceland and Weinberg, 2002) is that census tracts, which typically have between 2,500 and 8,000 people, are defined with local input and to represent neighborhoods; they typically do not change much from census to census, except to subdivide. However, mapping diversity on census tract level is unsatisfactory (Lee et al., 2008; Kramer et al., 2010) for at least two reasons. First, tracts have large variation in spatial sizes, so the spatial resolution of U.S.-wide map changes between metropolitan areas, where tracts are smaller, and the rest of the country where they are larger. However, racial diversity may occur in areas of low population density on scales even smaller than those observed in large cities reflecting an overall smaller sizes of these settlements. Second, using census tracts presumes uniform conditions throughout them which is not always true despite an effort to make tracts represent neighborhoods. As a result tract-based maps may show uniform level of diversity over areas where in reality a level of diversity is varying. They may also show variation (on the boundaries between tracts) where none exist. Using our hi-res grid solves these problems. The grid has fine enough resolution to render issues related to a scale of areal unit irrelevant. Diversity can be calculated and mapped at the maximum resolution of the grid and integrated to other units if desired.

We have used 3″ grids from the "Race and Ethnicity" group in Table 1 to classify the grid cells into categories pertaining to levels of diversity and domi-

nant races. Population of a cell indexed by (i, j) is represented by a normalized histogram $\{p_1^{(i,j)}, \ldots, p_K^{(i,j)}\}$ where $p_k^{(i,j)}$, $k = 1, \ldots, K$ is the proportion of population belonging to racial/ethnic group k in the cell (i, j). We consider $K = 5$ race/ethnicity categories: non-Hispanic white, non-Hispanic black, Asian, other (American Indian and Alaska Native alone and Native Hawaiian and other Pacific Islander alone) and Hispanic origin without regard of race. Our goal is to derive a 3″ product grid that allows simultaneous comprehension of racial composition, diversity, and population density.

Many different approaches to measure diversity has been proposed (Reardon and Firebaugh, 2002; Farrell and Lee, 2011; Holloway et al., 2012). We follow an approach of Holloway et al. (2012) with several important differences: (1) We use 3″ lon-lat grid rather than census tracts thus eliminating problems related to spatial scale (see discussion above). (2) We define five racial/ethnic categories instead of six as we decided to combine American Indians with Alaska Native and Native Hawaiian into a single category which, for the conterminous U.S. is dominated by American Indians. (3) We use three-dimensional classification based on diversity, race, and population density instead of two-dimensional classification (diversity and race) used by Holloway et al. (2012). Following Holloway et al. (2012) we categorize racial diversity on the basis of standardized informational entropy (Shannon, 1948) with modifications made to assure agreement between obtained categories and customary notions of group dominance (Farrell and Lee, 2011). The standardized entropy of population histogram in a cell (i, j) is given by:

$$E^{(i,j)} = -\frac{1}{E_{\max}} \sum_{k=1}^{K} p_k^{(i,j)} \ln(p_k^{(i,j)}) \qquad (5)$$

where $E_{\max} = -\ln(1/K)$ is the maximum value of entropy for histogram with K categories. Its presence standardizes entropy values to the range between 0 (if histogram has only a single bin indicating no diversity) and 1 (if all histogram's bins are equal indicating maximum diversity).

All cells are divided into three diversity classes:

- Cell (i, j) belongs to the *low diversity* class if its histogram fulfills two conditions: (1) $E^{(i,j)} < 0.41$, and (2) $\max_{1 \le k \le K} p_k^{(i,j)} > 0.8$ (dominant race constitute more than 80% of cell's population).

- Cell (i, j) belongs to the *high diversity* class if its histogram fulfills three conditions: (1) $E^{(i,j)} > 0.79$, (2) $\max_{1 \le k \le K} p_k^{(i,j)} < 0.5$ (dominant race constitute less than 50% of cell's population), and (3) a sum of two most dominant races constitute less than 80% of cell's population.

- Cell (i, j) belongs to the *moderate diversity* class if it does not belong to neither high nor low diversity classes.

All cells are also divided into three population density classes as follows:

- *Low density* class if population density assigned to the cell is less than 3 people/km^2. 75% of all cells belong to this class

- *Medium density* class if population density assigned to the cell is 3-30 people/km^2. 20% of all cells belong to this class.

- *High density* class if population density assigned to the cell equal to or greater than 30 people/km^2. 5% of all cells belong to this class.

Combining three diversity categories with three density categories results in nine combined categories. Six (those associated with low and moderate diversity) of those nine categories are further sub-divided with respect to five possible dominant races. By definition, the high diversity category does not have a dominant race and does not need further division. The result is a diversity–race–density classification of population cells into 33 categories (see Fig. 4C).

The resultant 3″ diversity map is available for download from http://sil.uc.edu/ as a GeoTiff raster in the latitude/longitude projection. In addition, this map can be explored using our interactive GeoWeb application DataEye available from the same website. Small portions of diversity map can be downloaded directly from DataEye in GeoTiff format using built-in "download-what-you-see" utility.

Our diversity map could be compared with diversity maps and data available from http://mixedmetro.us/ and based on Holloway et al. (2012) methodology. There are differences between the two resources stemming from their resolutions, formats, legends, coverages, and method of accessibility. Our map is a 3″ ($\sim$ 90 m) raster showing 33 categories resulting from diversity–race–density classification. The map pertains to Census 2000 data. The entire conterminous U.S. can be explored online and small portions of the map can be downloaded directly from the exploration tool. MixedMetro resources are census tracts-based shapefiles showing 9

categories resulting from diversity-race classification. Data pertaining to Census 1990, 2000, and 2010 is available for all 50 states. Maps can be explored online on the state-by-state basis. Separately, diversity data can be downloaded on the state-by-state basis. The biggest advantage of our diversity map is it high spatial resolution. Here we present two examples of comparison between our diversity map and diversity maps downloaded from MixedMetro.

The first example is the city of San Francisco, California – a high density urban setting where census tracts are small and tracts-based map has the highest spatial resolution. Fig. 4A shows our map and Fig. 4B shows a corresponding map from MixedMetro. Boundaries of census tracts are shown on both maps for reference. Despite different legends (see panels C and D) the maps can easily be compared. The comparison reveals that, as expected, the higher resolution map provides details not captured by the tracts-based map. First, some tracts include uninhabited areas within their boundaries. Prominent examples are the Presidio of San Francisco (norther tip of the map), Golden State Park (elongated rectangle in the northwestern part of the map) and the Olympic Club (in the southwestern part of the map). Averaging population over the entire tract give false impression of diversity distribution. Our map does not assign any diversity category to uninhabited areas and shows them in white. Second, some tracts show clear divisions within their boundaries with respect to race and diversity, a distinction which is lost in averaging over the entire tract. Finally, boundaries between some differently-labeled tracts cut through regions of homogeneous diversity.

The second example shows maps (Fig. 5) of Bullhead City which is located at the Arizona-Nevada border and has population of $\sim 39,000$. This area is in general sparsely populated but includes clusters of housing developments built to accommodate population growth due to economic opportunities in casinos and ancillary services across the border in Nevada. Census tracts here are mostly large and averaging racial diversity over them misses important information. Our map (Fig. 5A) shows the population agglomerated along the state border (the Colorado River) and exhibiting rich mosaic of diversity categories. Even if differences related to population density are disregarded the city is still a mosaic of white-dominated and latino-dominated neighborhoods with different degrees of diversity. There are also some small spots where Asians and American Indians (other) dominate. This richness is lost in the tracts-based map (Fig. 5B) which indicates that most of the area is dominated by one category – "low diversity whites", with the remaining area categorized as "medium diversity whites" and a single tract classified as "low diversity latino". Tracts-based map does not reflect the fact that most of the site's area is uninhabited. Large spatial extents of tracts leads to loss of important details – the demographic character of the city deduced exclusively from the map on Fig. 4B would not correspond to reality.

6. Discussion and future directions

In this paper we reported on our effort to improve the U.S.-wide SEDAC census grids by performing dasymetric modeling using the NLCD. The results are uniform $3''$ grids of population and sixteen other demographic variables which are freely available for download and can be explored online in their entirety through a GeoWeb application.

A decision to improve upon SEDAC rather than to calculate the standard dasymetric model was dictated by the computational complexity of disaggregating $\sim$8 millions census block polygons to a grid. Disaggregating coarser grid to finer grid is a much less complex computational task. There are some disadvantages of disaggregating SEDAC grids instead of census blocks. First, although our model has high spatial accuracy, its population count accuracy is not as good as that of the standard dasymetric model, especially in regions where census blocks are small. This is because population is disaggregated over an area which is large as compared to small blocks, so smaller scale information contained in sub-partitions by such blocks is lost. Second, inconsistencies of SEDAC data needed assumptions to be resolved. Finally, SEDAC does not provide at present grids for 2010, although it is expected that they will become available eventually. Our calculations themselves are very accurate. As can be seen from the last column in Table 1 the difference between the values resulting from integration of all $\sim$15 billion $3''$ cells and the values resulting from integration of $\sim$14 million SEDAC $30''$ cells differ by small fraction of 1% indicating high accuracy of our model. Moreover, difference listed in Table 1 are not attributed to inaccuracy of our model, but, instead, can be attributed to inconsistencies between $30''$ and $7.5''$ versions of SEDAC maps.

Overall, our $3''$ grids offer a significant upgrade to the SEDAC grids. They are well suited for U.S.-wide socio-economic research, like in our "diversity" map example. However, they should not be substituted for standard dasymetric model for local or regional area where calculation of standard model is computationally viable. Future work will concentrate on calculating $3''$ grids for

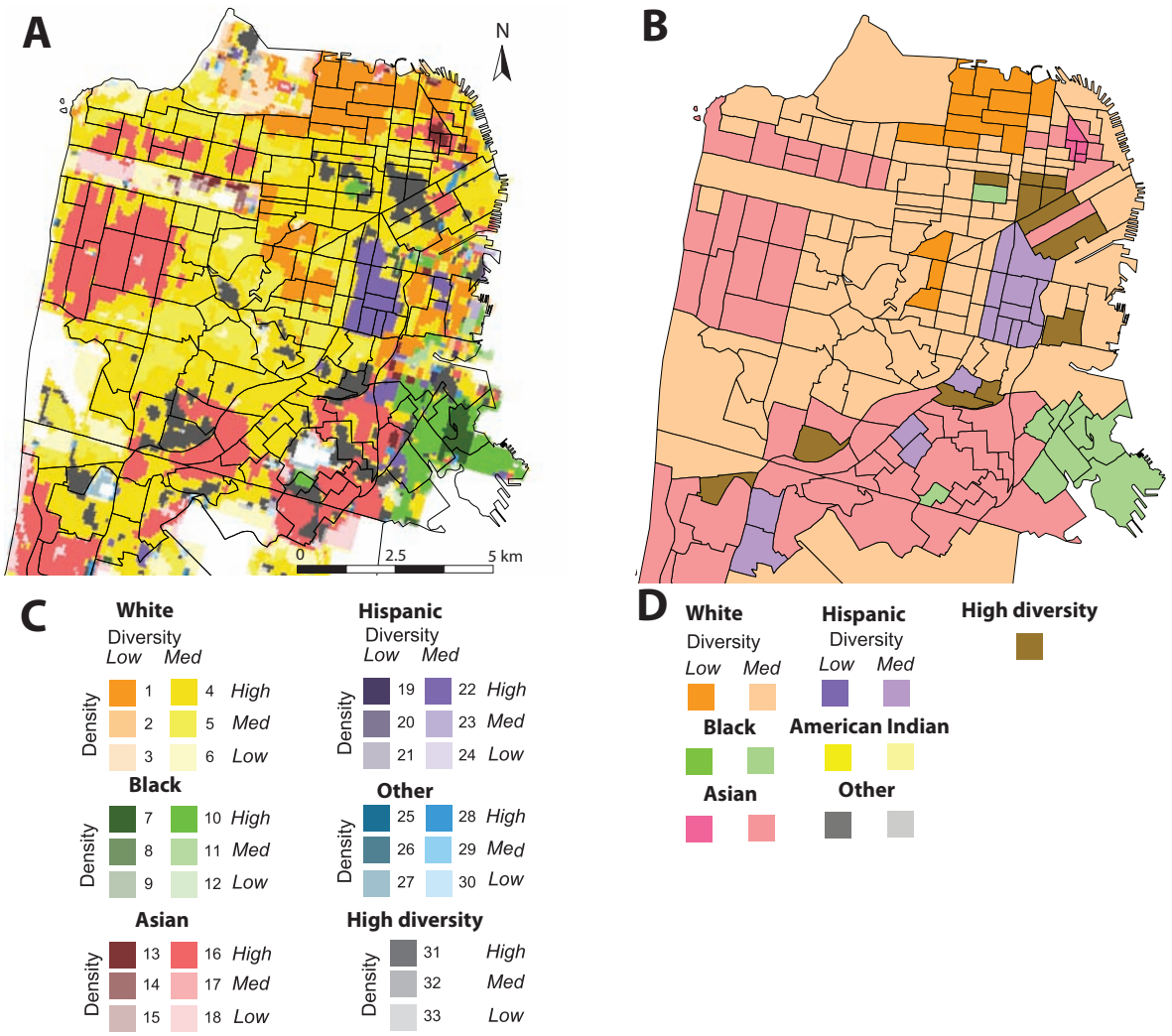

Figure 4: San Francisco, California. (A) Racial diversity map derived from our 3″ (~90 m) dasymetric model; census tracts boundaries are shown for reference. (B) Racial diversity map based on census tracts Holloway et al. (2012). (C) Legend to our diversity map. (D) Legend to Holloway et al. (2012) diversity map. Maps are shown in the projection Albers USGS projection.

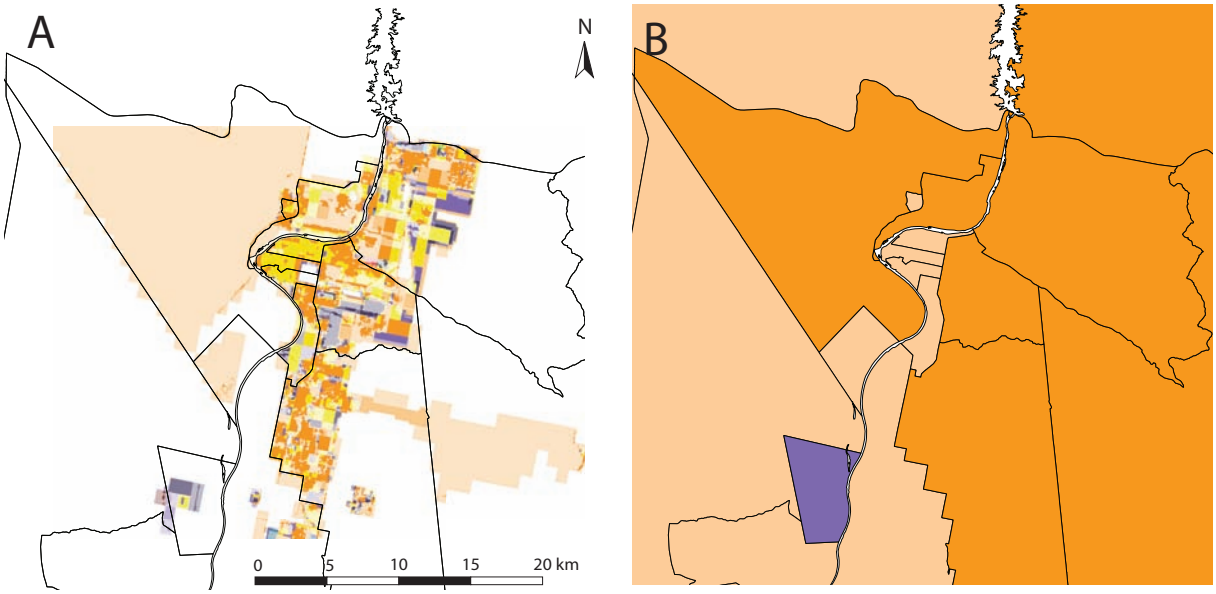

Figure 5: Bullhead City located on the Arizona-Nevada boarder. (A) Racial diversity map derived from our 3″ (~90 m) dasymetric model; census tracts boundaries are shown for reference. (B) Racial diversity map based on census tracts Holloway et al. (2012). See Fig. 4 for diversity categories legends. Maps are shown in the projection Albers USGS projection.

1990 census data (with NLCD 1992 serving as ancillary data) and extending the range of calculated characteristics. When SEDAC will release their grids based on 2010 Census we will calculate 3″ grids for 2010 data using NLCD 2011 edition. We can also consider using recently published (Theobald, 2014) U.S.-wide 30 m resolution land use dataset instead of NLCD as ancillary information.

Ability to combine multiple hi-res grids into a single hi-res "product" grid makes possible spatial studies of complex social phenomena. We use the name "product" for categorical grid that combines information from several demographic characteristics as collected by the census. In this paper we presented one such product which we called "diversity". In fact this product shows more than just a degree of racial diversity, it also encodes a dominant race (if any) and density of the population. Because it is designed to illustrate three different demographic characteristics the diversity product has 33 categories. With this many categories it is difficult-to-impossible to select a legend with colors that have logical meaning and, at the same time, they are completely dissimilar to each other. The colors in our legend (Fig. 3C) appear quite distinct, but some of them may be confused with each other on an actual map. This is why DataEye – our GeoWeb application for exploration of the grids – has an utility that shows a specific category of a pixel when a user right-click on it. This makes exploring diversity product very easy.

Our diversity product offers a significant upgrade to the MixedMetro dataset which is based on census tracts. The higher resolution provides the most important improvement. This increased resolution comes from two separate sources, first SEDAC grids use census blocks instead of census tracts data, second, our dasymetric model provided further refinement. The difference in resolution is quite dramatic everywhere, but most dramatic outside large cities (see Fig. 5). Thus, using our diversity product it is now possible to study racial segregation not only in large metropolitan areas but also in smaller cities as well as rural areas. In addition to higher spacial resolution, our diversity product provides information on population density that is absent from MixedMetro dataset.

Other interesting products can be constructed from grids listed in Table 1. Variables from the "Age" group can be combined into an "age diversity" product and variables from the "Income" group can be combined into an "income diversity" product. In addition, variables in the "Race and Ethnicity" group could be combined with variables from either "Age" group or "Income" group or both to produce even more complex products. When constructing these composite products one have to be careful to assure that a population in a

single cell is large enough to support meaningful classification. On the other hand, small cell size makes possible studying local spatial patterns of demographics. Existing literature focuses on binary patterns in the context of residential segregation. Using indices developed to quantify five dimensions of segregation (Massey and Denton, 1988) a quantitative description of binary segregation style in a metropolitan area can be obtained. Using hi-res grids a more general analysis, one that involves multiple races and is able to assess an overall measure of similarity between segregation patterns, can be conducted using a methodology (Jasiewicz and Stepinski, 2013; Stepinski et al., 2014) originally developed for comparison of land cover patterns. Using this methodology would allow to search the U.S. for locations where segregation/diversity patterns are similar to a given template/example.

Acknowledgments. The authors wish to thank J. Jasiewicz and R. Szczepanek for helpful comments and discussions and P. Netzel for integrating population and diversity maps into the DataEye app. This work was supported by the University of Cincinnati Space Exploration Institute.

References

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, M. L., 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. Applied Geography 24(2), 97–117.

Clemen, R. T., Winkler, R. L., 1999. Combining probability distributions from experts in risk analysis. Risk Analysis 19(2), 187–203.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., Worley, B. A., 2000. LandScan: a global population database for estimating populations at risk. Photogrammetric engineering and remote sensing 66, no. 7 (2000): 849-857. 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.

Farrell, C. R., Lee, B. A., 2011. Racial diversity and change in metropolitan neighborhoods. Social Science Research 40(4), 1108–1123.

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, no. 6 (2010): 460-473. 31(6), 460–473.

Gallego, F. J., 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(12), 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.

Goodchild, M., A.-L. . D. U., 1993. A framework for the areal interpolation of socioeconomic data. Environment and Planning, A 25, 383–397.

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

Holloway, S. R., Wright, R., Ellis, M., 2012. The Racially Fragmented City? Neighborhood Racial Segregation and Diversity Jointly Considered. The Professional Geographer 64, 63–82.

Iceland, J., Weinberg, D. H., 2002. Racial and ethnic residential segregation in the United States 1980-2000. Tech. rep., Bureau of Census.

Jasiewicz, J., Stepinski, T. F., 2013. Example-Based Retrieval of Alike Land-Cover Scenes From NLCD2006 Database. Geoscience and Remote Sensing Letters 10, 155–159.

Kramer, M. R., Cooper, H. L., Drews-Botsch, C. D., Waller, L. A., Hogue, C. R., 2010. Do measures matter? Comparing surface-density-derived and census-tract-derived measures of racial residential segregation. .International Journal of Health Geographics 9, 29.

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

Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., O'Sullivan, D., 2008. Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. American Sociological Review 73(5), 766–791.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., Tatem, A. J., 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 7(2), e31743.

Massey, D. S., Denton, N. A., 1988. The dimensions of residential segregation. Social forces 67(2), 281–315.

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

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

Neteler, M., Mitasova, H., 2007. Open source GIS: a GRASS GIS approach, 3rd Edition. Springer, New York.

Reardon, S. F., Firebaugh, G., 2002. Measures of multigroup segregation. Sociological methodology 32(1), 33–67.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Stepinski, T. F., Netzel, P., Jasiewicz, J., 2014. LandEx - A GeoWeb tool for query and retrieval of spatial patterns in land cover datasets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(1), 257–266.

Tatem, A. J., Noor, A. M., vonHagen, C., DiGregorio, A., Hay, S. I., 2007. High resolution population maps for low income nations: combining land cover and census in East Africa. PLoS One 2(12), e1298.

Theobald, D. M., 2014. Development and Applications of a Comprehensive Land Use Classification and Map for the US. PloS One 9, no. 4 (2014): e94628. 9(4), e94628.

Thieken, A., Mueller, M., Kleist, L., Seifert, I., Borst, D., Werner, U., 2006. Regionali-sation of asset values for risk analyses. Natural Hazards and Earth System Sciences, 6, 6, 167–178.

Tian, Y., Yue, T., Zhu, L., Clinton, N., 2005. Modeling population density using land cover data. Ecological modelling 189, no. 1 (2005) 189(1), 72–88.

Tralli, D. M., Blom, R. G., Zlotnicki, V., Donnellan, A., L.Evans, D., 2005. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. ISPRS Journal of Pho-

togrammetry and Remote Sensing 59(4), 185–198.

Vinkx, K., Visee, T., 2008. Usefulness of population files for estimation of noise hindrance effects. In: ICAO Committee on Aviation Environmental Protection. CAEP/8 Modelling and Database Task Force (MODTF). 4th Meeting. Sunnyvale, USA. pp. 20–22.

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.

Wickham, J. D., Stehman, S. V., Fry, J. A., Smith, J. H., Homer., C. G., 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sensing of Environment 14(6), 1286–1296.

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