High resolution population grid for the entire United States

A. Dmowska, T. F. Stepinski

Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA
Telephone: +1 513 .556.3583
Fax: +1 513.556.3370
Email: dmowskaa@ucmail.uc.edu, stepintz@uc.edu

Abstract

We have developed a 30 m resolution population grid for the entire United States on the basis of the 2010 Census block-level data. Dasymetric model was used to disaggregate population in each of ~ 11 millions census blocks into their constituent 30m cells. The model uses land cover (NLCD 2011) and the new U.S.–wide land use dataset as auxiliary variables. Both auxiliary datasets have 30 m resolution; the land use dataset permits distinctions between populated and unpopulated impervious areas. Visual comparison of population density maps based on our model with the population maps based on previous U.S.–wide population density resources indicates much better utility of the new grids. Our population grid is accessible (for exploration and download) through SocScape – an internet-based application available at http://sil.uc.edu.

Keywords: population grid; dasymetric modeling; big data.

1. Introduction

Quick and convenient access to high resolution population and demographic data over the entire U.S. would be an important input for a large-scale quantitative socio-economic analysis and a valuable resource for supporting planning and management decisions on the scale of the entire country. Unfortunately, an inherent structure of the U.S. Census data does not support such access. The solution is to construct census grids which support efficient algorithmic analysis of the data. For the U.S. census grids at 1 km resolution (250 m at major metropolitan areas) have been developed by the Socioeconomic Data and Application Center (SEDAC) (http://sedac.ciesin.columbia.edu/) using simple areal weighting interpolation from census blocks. However, for many applications the resolution of SEDAC grids is insufficient. Also, SEDAC grids are only available for 1990 and 2000 censuses, so they cannot be used for mapping population on the basis of 2010 census. The Oak Ridge National Laboratory is developing (Bhaduri et al., 2007) the LandScan USA - 90 m grid obtained by disaggregating census blocks using dysametric modeling. 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. We used dasymetric modeling utilizing National Land Cover Dataset (NLCD) to sharpen SEDAC grids to the resolution of 90 m (Dmowska and Stepinski, 2014). Resultant grids are conveniently accessible through SocScape – a GeoWeb application available at http://sil.uc.edu. Although SocScape grids offer significant improvement over SEDAC grids, their accuracy are restricted by the quality of SEDAC grids and by using land cover as the only auxiliary data. Here we report on our second generation of high resolution demographic U.S.-wide grids. The two major improvements over the previous methodology are: (a) dasymetric modeling directly from census blocks (no SEDAC grids used as an intermediate step), (b) using new, 30 m resolution U.S. land use map (Theobald, 2014) as the second auxiliary data in addition to the NLCD. The new approach requires much bigger computational effort but results in an improved quality of the grids. This paper focuses on population grid, but grids of other demographic variables can be obtained in the same fashion.

2. Data and Methods

2.1 Census data and its rasterization

The source of the base data is the 2010 U.S. Census block-level data that consists of two components, shapefiles (TIGER/Line Files) with geographical boundaries of ~ 11 millions blocks and summary text files that lists population data for each block. For each state separately we joined the boundaries shapefile and data from the summary file to form a vector file. The vector files were rasterized to a 30 m resolution grids. The 30 m resolution was selected because it is the resolution of both auxiliary datasets and having all data on the same grid expedites the computation. The cells in the initial grid have homogeneous values across each block – a result of census aggregation of the data. All computations were performed using Python scripts written for GRASS GIS 7.0 software which is especially well adapted to work with large volume of data.

2.2 Auxiliary data

The purpose of dasymetric modelling is to disaggregate blocks population counts using auxiliary data information so the values of the cells within a block are heterogeneous increasing the spatial resolution of population information. We use land cover (NLCD) as an auxiliary data. However, land cover classes are established based on spectral information and cannot recognize between populated and unpopulated buildings. Therefore we also use a land use map (also 30 m resolution) that makes such distinction. Land use map is used to determine uninhabited areas regardless of their land cover class.

2.3 Dasymetric modeling

We reclassify NLCD into 6 classes: developed open space area, developed low intensity, developed medium intensity, developed high intensity, vegetation, and uninhabited (from the land use map plus water and barren NLCD classes). The population in each (rasterized) block is redistributed to its cells using block-specific weights assigned to the cells having different classes. The weights are assigned based on two factors, relative density of population for each class and the area of each block occupied by each class (Mennis, 2003). Following Mennis and Hultgren (2006) representative population density for each class is established using a set of blocks (selected from the entire U.S.) having relatively homogenous land cover (90% for developed classes and 95% for vegetation classes). Once the weights are calculated the homogeneous values in the block's cells are multiplied by the weights yielding the sought-after disaggregation.

3. Results

The U.S.-wide 30 m resolution population grid is too large to be distributed via SocScape, but it is available upon request directly from the authors. The re-sampled 90 m resolution version of the grid will be available through SocScape. This downscaled version of the grid would be still more accurate than the current SocScape grids because of additional information in the land use data and because the modelling is done directly on the original block-level data. To demonstrate a difference between our present and previous grids we have chosen two sample locations, one in an urban setting (Cincinnati, OH) and another in the rural setting (Somerset, OH).

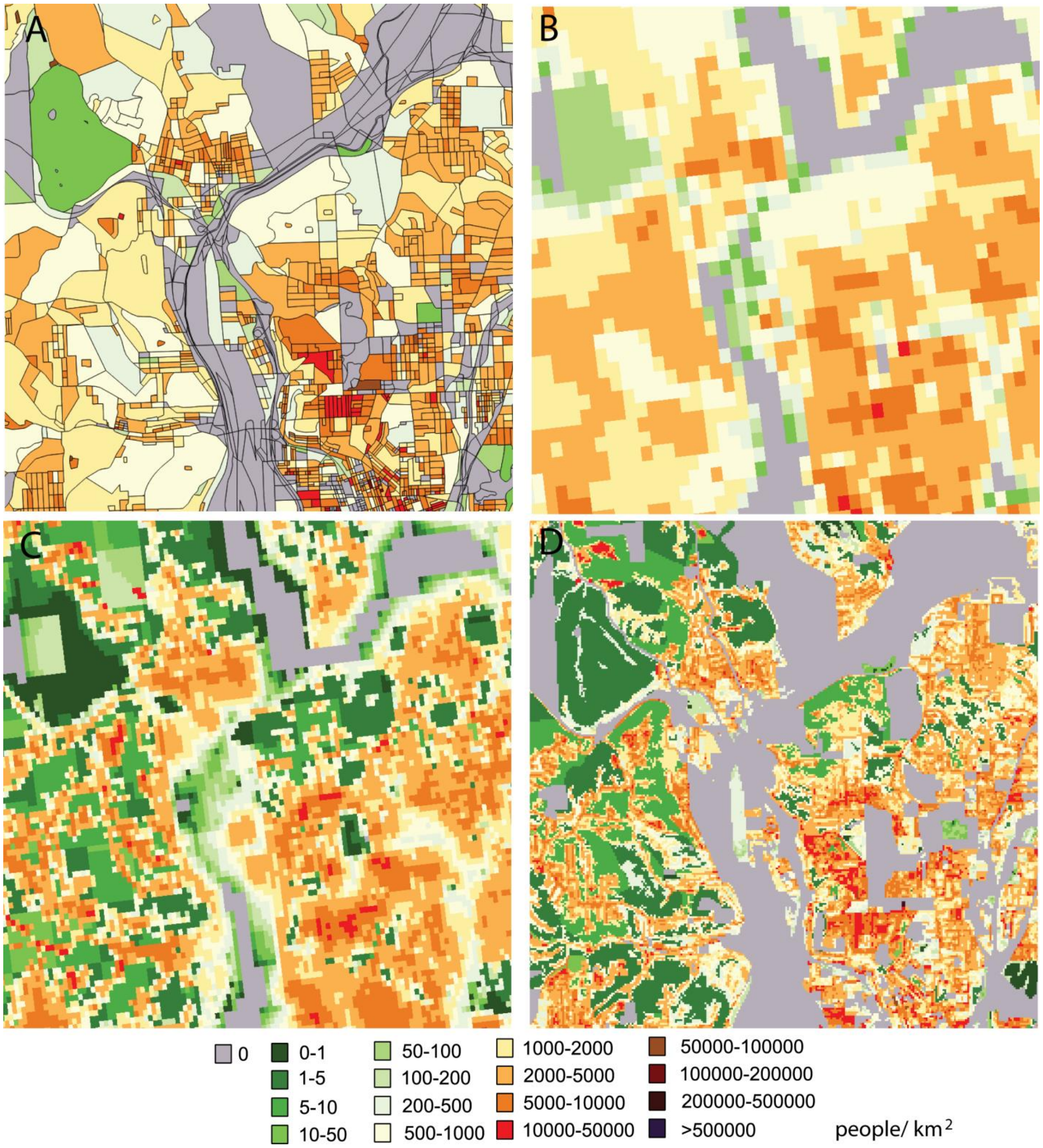

Figure 1. Population density maps of the area in Cincinnati, OH based on: (A) census blocks, (B) SEDAC 250 m grid, (C) our previous 90 m grid, (D) our current 30 m grid.

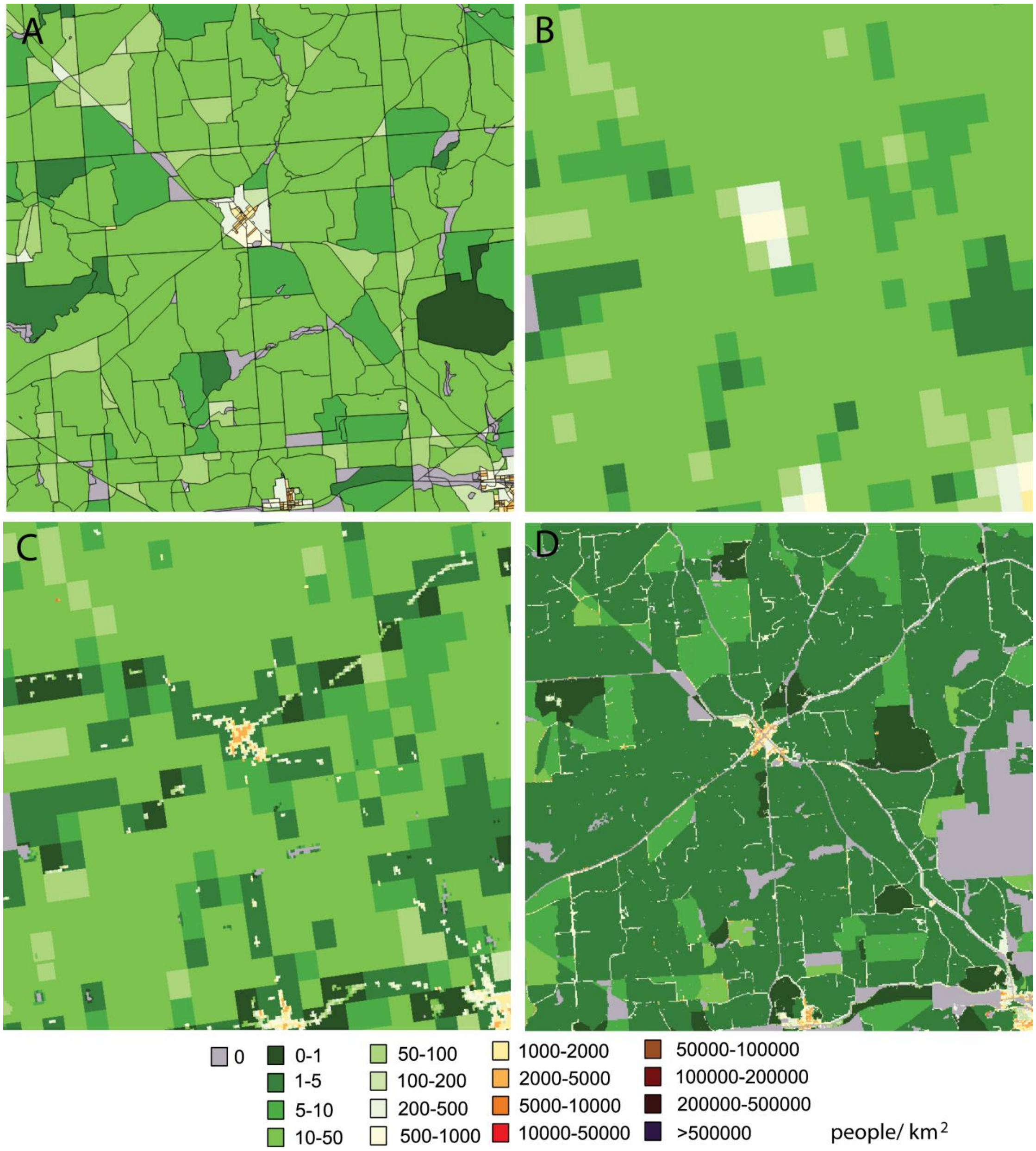

Figure 2. Population density maps of the Somerset, OH area based on: (A) census blocks, (B) SEDAC 250 m grid, (C) our previous 90 m grid, (D) our current 30 m grid.

Fig.1 compares maps of population density in a portion of the Cincinnati area using census blocks (A), SEDAC 250 m grid (B), our previous 90 m grid (C), and our new 30 m grid (D). As the blocks in the urban areas are small, the SEDAC grid does not offer spatial resolution improvement over the block-based map, although it is still more convenient to use. The 90 m grid offers overall resolution improvement over the block-based map but suffers from the lack of information on uninhabited areas with spectral signatures of buildings. The new 30 m grid is clearly superior to all other maps.

Fig.2 compares maps of population density in the Somerset area using census blocks (A), SEDAC 1 km grid (B), our previous 90 m grid (C), and our new 30 m grid (D). SEDAC grid offers a fair approximation to an overall distribution of population but without any details in the village of Somerset (population 1,418). The 90 m grid resolves the village but shows too high (although still low) population density over the farmland. The 30 m grid recognizes individual farm houses and thus lowers the population density of farmland (darker green) as people are assigned to very small regions at the locations of farm houses. Additional information from the land use data delineates uninhabited such as the state forest (grey area at the right edge of the region).

4. Conclusions

We have made further progress toward our goal to develop high resolution demographic grids covering the entire conterminous United States. Although disaggregation methods (including dysametric modeling) are well established and straightforward to apply, their application to "big data" has been hindered by the need to handle big datasets, the need for efficient computational methods, and by the limited availability of high resolution auxiliary data. We were able to improve on our previous grids because 30 m land use data over the entire U.S. become available (Theobald, 2014) and because we have developed efficient software to disaggregate population counts directly from the census blocks instead of from a coarse SEDAC grid as in our previous method.

Formal assessment of accuracy of 30 m grid is only possible by comparison to data with sub-block resolution, such as, for example, parcel data. Such assessment is not feasible for the entire U.S., but can be performed for small regions for which parcel data has been utilized for population density. Our map agrees well with parcels-derived population density map for Alachua County, Florida (Jia et al., 2014).

Grids of demographic variables other than population count 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. There are no auxiliary data specific to race, age, or income that would allow disaggregating directly these variables within a block, but we can disaggregate them according to the population model. By narrowing down the locations where people live within a block we increase spatial resolution of these variables although we would not be able to account for variation of, say, racial diversity within a block.

5. Acknowledgements

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

6. References

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

Dmowska A, Stepinski TF, 2014, High resolution dasymetric model of US demographics with application to spatial distribution of racial diversity. *Applied Geography* 53: 417-426.

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. *Applied Geography* 50: 99-107.

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

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

Theobald DM, 2014, Development and Applications of a Comprehensive Land Use Classification and Map for the US. *PloS One* 9(4): e94628.