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

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

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

Anna Dmowska, Tomasz F. Stepinski *

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

ARTICLE INFO

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

Keywords:
Dasymetric modeling
Gridded population data
Census
NLCD
Land use

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

2. Input data

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

2.1. Census data

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

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