DASYMETRIC MODELLING OF POPULATION DISTRIBUTION – LARGE DATA APPROACH

ANNA DMOWSKA

Department of Geoinformation, Institute of Geoecology and Geoinformation, Adam Mickiewicz University in Poznań, Poland

Manuscript received: January 4, 2019
Revised version: February 11, 2019

DMOWSKA A., 2019. Dasymetric modelling of population distribution – large data approach. *Quaestiones Geographicae* 38(1), Bogucki Wydawnictwo Naukowe, Poznań, pp. 15–27. 3 figs, 1 table.

ABSTRACT: Existing resources of population data, provided by national censuses in the form of areal aggregates, have usually insufficient resolution for many practical applications. Dasymetric modelling has been a standard technique to disaggregate census aggregates into finer grids. Although dasymetric modelling of population distribution is well-established, most literature focuses on proposing new variants of the technique, while only few are devoted to developing broad-scale population grids that could be used for real-life applications. This paper reviews literature on construction of broad-scale population grids using dasymetric modelling. It also describes an R implementation of fully automated framework to calculate such grids from aggregated data provided by national censuses. The presented implementation has been used to produce high resolution, multi-year comparable, U.S.-wide population datasets that are the part of the SocScape (Social Landscape) project.

KEY WORDS: population grids, dasymetric modelling, R

Corresponding author: Anna Dmowska, dmowska@amu.edu.pl

Introduction

An access to high resolution data on population distribution is needed for a wide range of applications related to urban and transport planning (Benn 1995, Murray et al. 1998, Pattnaik et al. 1998), resources management (Gleick 1996), disaster/relief mitigation (Bhaduri et al. 2002), assessment of human pressure on the environment (Weber, Christophersen 2002) and quantifying environmental impact on population (Vinkx, Visee 2008). Reliable information on population distribution is also essential for characterizing population at risk from natural hazards (Dobson et al. 2000, Chen et al. 2004, Tralli et al. 2005, Thieken et al. 2006, McGranahan et al. 2007, Maantay, Maroko 2009, Mondal, Tatem 2012, Tatem et al. 2012, Berke et al. 2015, Tenerelli et al. 2015, Calka et al. 2017) and for public health applications such as disease burden estimation and epidemic modelling (Hay et al. 2005, Tatem et al. 2008, 2011, 2012).

The quality of population data varies from one country to another, especially between low-income and high-income countries. Low-income countries are often lacking population data or data has a poor quality (Tatem et al. 2007). High-income countries usually have resources to collect data for each household, but such data is only released in the form of areal aggregates to protect

sciendo

doi: 10.2478/quageo-2019-0008
ISSN 0137-477X, eISSN 2081-6383

privacy (Langford 2013, Bakillah et al. 2014). The size of aggregation units differ between the countries. However, even the smallest aggregated units have often insufficient spatial resolution for many practical applications. Aggregated data also has many limitations (Schroeder 2007, Dmowska et al. 2017), including:

1. spatial resolution depends on the choice of census units, and it is spatially varying (low in rural areas, higher in urban areas),
2. mapped population is distributed uniformly within each census unit, even if majority of the area is uninhabited (covered by parks, forest, water, etc.), and
3. spatial extents of census units change with time, which makes difficult to conduct year-to-year comparisons.

Furthermore, such data is usually available in the form of attribute tables, with the option to join them to vector files containing boundaries of aggregation units in order to perform GIS-based analysis. This makes the data difficult to work with, especially for the large areas.

Most of aforementioned limitations of aggregated data can be overcome by gridded (raster) data. The advantage of using gridded data include:

1. spatial resolution, defined by the size of the cell, is high and spatially constant over the whole area;
2. the extend of cells does not change between years making year-to-year comparison easy to perform, and
3. the uninhabited areas can be properly identifies and properly mapped via dasymetric modelling thus making population maps more accurate.

Several methods for disaggregating census data into grid cells (or smaller areal units) have been introduced over the years. Such methods can be divided into two groups: areal weighting (Goodchild, Lam 1980, Flowerdew, Green 1992, Goodchild et al. 1993) and dasymetric modelling (Wright 1936, Langford, Unwin 1994, Eicher, Brewer 2001, Mennis 2003). Areal weighting is a type of an areal interpolation used to transform geographic data from one set of boundaries to another. Areal weighting assigns to each grid cell population value based on its percentage area of the host areal units (Mennis 2003). Dasymetric modelling uses ancillary information of higher spatial resolution to help refine location of population during the process of disaggregating spatial data to finer units (Mennis 2003).

Dasymetric modelling takes advantage of a correlation (a model) between population density and values of ancillary variable; the stronger the correlation (the better the model) the more accurate is the resulting population grid. Dasymetric modelling is well established in the literature (for a review see Petrov 2012). It has been defined and developed in 1911 by Benjamin (Veniamin) Petrovich Semenov-Tyan-Shansky (Bielecka 2005, Petrov 2012) and popularized by Wright (1936). After 2000, an interest in dasymetric mapping had significantly increased due to the progress in the GIS and remote sensing technologies (Mennis 2009, Petrov 2012). Published papers described development of new approaches to dasymetric modelling based on different ancillary data and a variety of techniques to establish relation between population density and values of ancillary variable. These papers focus on the *theory* and do not provide actual datasets which are the results of proposed techniques.

Among the ancillary data used to disaggregate population, the most popular is the land cover data (Wright 1936, Mennis 2003, Bielecka 2005, Gallego et al. 2011, Linard et al. 2011, Dmowska, Stepinski 2014, 2017a, b, Dmowska et al. 2017). Land cover data is provided in the form of a categorical grid, with different categories indicating types of land cover. Broad-scale land cover datasets are obtained by classifying large mosaics of remotely sensed multispectral images. They have spatial resolution higher than the resolution of census aggregated units. One problem with land cover datasets is that they are based on surface spectral properties leading to possible confusion between populated and unpopulated objects (for example buildings) having same spectral signatures. This problem can be minimized by adding land use data as an additional ancillary variable (Dmowska, Stepinski 2017a).

The other source of ancillary data is high resolution satellite images (Lu et al. 2010, Ural et al. 2011, Lung et al. 2013), LIDAR data (Lu et al. 2010), tax parcel data (Maantay et al. 2007, Tapp 2010, Jia et al. 2014, Jia, Gaughan 2016), street density (Reibel, Bufalino 2005), density of point of interests (Bakillah et al. 2014), light emission data (Briggs et al. 2007, Sridharan, Qiu 2013) and

address datasets (Zandbergen 2011). Recently, social media data also are used (Patel et al. 2017). Such datasets can be used individually or in combination to construct a dasymetric model.

Many papers concentrate on establishing relation between population and ancillary data in dasymetric modelling. These approaches had changed over the years from using predetermined weights (binary approach or limiting variable estimation (Eicher, Brewer 2001), through using empirical sampling (Mennis 2003, Mennis, Hultgren 2006), to employ statistical techniques such as regression analysis (Flowerdew, Green 1992, Briggs et al. 2007) or random forest (Stevens et al. 2015). An overview of developed methods can be found among others in Wu et al. (2005) and Maantay et al. (2007).

Despite the increasing body of the literature describing various techniques for dasymetric modelling, there is still lack of high resolution population grids. There are only few products which provide high resolution population grids on global or continental scale (Table 1).

LandScan and Gridded Population of the World, Version 4 (GPWv4) provide population grids at global scale at a resolution of 30 arc-seconds (approximately 1 km at the equator). LandScan is developed by the Oak Ridge National Laboratory using the best available census data for particular regions. It is a product of dasymetric modelling based on land cover, roads, slope, urban areas, village locations, and high resolution imagery analysis as ancillary data and sub-national level census counts for each country as population data. LandScan population grid is a combination of locally adoptive models that are tailored to account the differences in spatial data availability, quality, scale, and accuracy of data for each individual country and region (ORNL 2019).

Gridded Population of the World, Version 4 (SEDAC 2019) provides gridded population estimates with a resolution of 30 arc-seconds (approximately 1 km at the equator) for the years 2000, 2005, 2010, 2015, and 2020. The census data, collected around 2010 (between 2005 and 2014) are extrapolated to a series of output years. GPWv4 is the result of uniform areal weighting approach (Doxsey-Whitfield et al. 2015).

The WorldPop project (2019) provides population grids at a resolution of 1km at the continental scale and 100 m/cell for most individual countries in Africa, Asia, as well as in South and Central Americas. It was initiated in 2013 by combining the AfriPop, AsiaPop and AmeriPop population mapping projects (Gaughan et al. 2013, Tatem et al. 2013). Population grids are the result of dasymetric modelling, performed for each country separately based on census data (or official population estimates) at the finest level of aggregation available for each country and using remotely-sensed and geospatial datasets (e.g. settlement locations, settlement extents, land cover, roads, building maps, health facility locations, satellite night lights, vegetation, topography, refugee camps) as ancillary data. Dasymetric modelling follows a procedure described by Stevens et al. (2015).

The WorldPop project (2019) also provides data for mapping births and pregnancies (Tatem et al. 2014), age and sex structure (Alegana et al. 2015) and population dynamics based on cell phone data (Deville et al. 2014). The recent initiative (WorldPop Archives) aims at providing

Table 1. Characteristics of broad-scale population grids.

Project	Region	Resolution	Timestamp	Availability
WorldPop	South America, Central America, Africa, Asia	100 m (country), 1 km (continent)	2010–2020 with 5 year interval	http://www.worldpop.org.uk
LandScan	world-wide	1 km	2000–2017 with 1 year interval	https://landscan.ornl.gov
GPWv4	world-wide	1 km	2000, 2010, 2015, 2020	http://sedac.ciesin.columbia.edu
E.U. pop grid	European Union countries	100 m	2000	http://www.eea.europa.eu
Australian pop.grid	Australia	1 km	2011	http://www.abs.gov.au
SEDAC-USA	United States	1 km (USA), 250 m (MSA)	1990, 2000, 2010	http://sedac.ciesin.columbia.edu
SocScape	United States	30 m	1990, 2000, 2010	http://sil.uc.edu

uniform, resampled and co-registered spatial data layers at two different resolutions (3 and 30 arc-second) ready-to-use for modelling and mapping population distribution (Lloyd et al. 2018).

A 100 m/cell population grid (Gallego 2010, Gallego et al. 2011) has been developed for the European Union (EU) countries; it is available from the European Environment Agency data warehouse (EEA 2019). This dataset is a result of dasymetric modelling calculated using population data from the 2000/2001 round of censuses aggregated to nearly 115,000 areal units and 100 m/cell raster version of CORINE Land Cover 2000 as an ancillary data.

Batista e Silva et al. (2013) reported on producing 2006 population estimates at 100 × 100 meter cells, for the territory of the EU27 (except Greece) and Andorra, Norway, Iceland, San Marino, Monaco, Lichtenstein, the Vatican City. Authors tested several different approaches to establish relation between population data and ancillary variables as well several different ancillary datasets to check whether using more detailed ancillary data in the dasymetric mapping leads to improved accuracy. The final map uses population data aggregated to 100,925 local administrative units (LAU2) downloaded from EUROSTAT with a refined version of CORINE Land Cover 2006 and information on the soil sealing degree. The final map is only made available in the PDF format as a supplementary material to paper Batista e Silva et al. (2013).

In the North America WorldPop project (WorldPop 2019) provides data for Mexico and a few separate projects provide data for the United States. Until recently, the only available data for the entire U.S. were the population grids developed by SEDAC as a result of aerial interpolation of U.S. census data. These grids are available of 30 arc-seconds (approximately 1 km at the equator) resolution for the U.S. for 1990, 2000 and 2010 year and at a 7.5 arc-second (approximately 250 m) resolution for major metropolitan areas (MSA) for 1990 and 2000 year. Although there are prepared for 1990, 2000, 2010, they cannot be used for direct comparison studies due to different format (2000 year dataset use integer counts, whereas 1990 and 2010 real number counts which makes those datasets non-comparable). Also ~1km resolution is not sufficient for many practical applications.

Since 2014 the other resource of population grids are provided by SocScape project (Dmowska, Stepinski 2017a, Dmowska et al. 2017). SocScape project makes available two types of products for the conterminous U.S.:

1. 30m high resolution grids of the entire population and for race/ethnicity subpopulations in 1990, 2000, 2010,
2. racial diversity grids.

High resolution grids are the product of dasymetric modelling performed on block-level census data (the smallest level of aggregation in the U.S.) and 30 m National Land Cover Datasets (NLCD). Racial diversity grids show spatial character of racial diversity across the U.S. in the form of three-dimensional classification of grid cells based on population density, dominant race and diversity level expressed by standardized entropy (Dmowska et al. 2017). SocScape data are the only available on the public domain broad-scale population grids comparable between years that can be used for quantitatively assessment of population changes.

Although dasymetric modelling is a well-known technique and straightforward to apply, its application for producing broad-scale, high resolution grids presents several challenges. The main challenge is the availability of the high resolution ancillary data, which must be available for the entire area of interests in uniform fashion and quality and must be comparable between different years if population grids are intended to be use for change analysis. Producing broad-scale high resolution maps require development of an efficient, fully automated algorithm to work with large datasets, so calculations can be performed within a reasonable time.

This paper is not focused on the development of new techniques and testing different types of ancillary data for the dasymetric modelling, but on developing fully automated computational framework and applying it to provide actual, broad-scale, multi-year comparable population grids that can be an input to the wide range of applications. This paper consists of three parts:

1. an extensive review of the literature on constructing broad-scale population grids,
2. description of the development of R-based implementation of computational framework,
3. showing examples of resultant population grids.

In addition, Section 2 briefly describes data and presents short overview of methodology used to produce high-resolution, multi-year compatible population grids for the entire U.S, which are the part of the SocScape project. Final conclusions are drawn in Section 4.

Data and methods

To produce population grids, which are temporarily comparable, requires the following condition on the data:

1. usage of contemporarily collected population and ancillary data to construct a grid for a given epoch (for example, 2000 census data should be coupled with circa 2000 land cover data), and
2. ancillary data should have the same meaning over all epochs (for example, land cover data at different years should have the same categories).

Data in the SocScape project fulfils those conditions, thus making the resultant population grids comparable between different years.

Population data

The source of the population data in the SocScape project is the 1990, 2000, and 2010 decennial U.S. Censuses data aggregated at the block level. The block level is the smallest aggregated units of the U.S. Census. This data consists of two components: shapefiles (TIGER/Line Files), indicating blocks geographical boundaries, and summary text files which lists population data for each block. Data has been downloaded from National Historical Geographic Information System (NHGIS) (MPC 2019). NHGIS project distributes population and shapefiles with additional key identifier making easier joining boundaries with an attribute tabular data. Tabular data are available as a one file for the entire U.S., whereas shapefiles are provided at the state level. Size of shapefiles containing block boundaries and their population counts vary from 34 MB for District of Columbia to 4037 MB for the state of California. The overall size is 39 GB. Number of blocks in 1990, 2000 and 2010 is ~7.15 million (1990), ~8.2 million (2000), and ~11.15 million (2010).

Ancillary data

SocScape project uses land cover datasets as ancillary data. This choice is dictated by the fact that land cover is the only ancillary data for which a single dataset, the National Land Cover Dataset or NLCD, covers the entire conterminous U.S. (or CONUS) at the same spatial resolution (30 m per cell) and the same quality.

NLCD datasets are available for 1992, 2001, 2006 and 2011. However, NLCD1992 has a legend which is incompatible with later editions. For comparison between 1992 and 2001 NLCD 1992/2001 Retrofit Land Cover Change Product should be used. It is a product based on Anderson Level 1 classification and consists of 8 *unchanged* categories (open water, urban, barren, forest, grass/shrub, wetlands, ice snow) and 55 categories (the combination of those 8) indicating changes between 1992 and 2001. Based on those categories 1992/2001 Retrofit Land Cover Change Product can be divided into two separate maps (for 1992 and for 2001) representing 8 categories of land cover types. NLCD2001 and NLCD2011 consist of 16 classes of land cover categories (including 4 categories of developed areas; see Fig. 3 for legend). Ancillary data are based on the 1992 land cover maps derived from 1992/2001 Retrofit Land Cover Change Product and 2001 and 2011 edition of NLCD (for example see panels A–C in Fig. 3). Land cover data from 1992, 2001 and 2011 match closely population data from 1990, 2000 and 2010 U.S. Decennial Censuses. To transform all three NLCD maps to a common legend land cover maps are reclassified to just three categories: urban (represents 4 developed categories in NLCD and urban category in 1992 land cover map), vegetation (represents forest and agriculture categories), and uninhabited (represents water, ice/snow, barren land categories). Example of ancillary data is shown in Fig. 3. These reclassified maps are used as ancillary data for a dasymetric model.

Methods

The overall dasymetric model follows the methodology introduce by Dmowska and Stepinski (2017a) to obtain 2010 population grid. The only difference is in using 3-class land cover data instead of the combination of land cover/

land use classes as ancillary data. According to this methodology the population in each block is redistributed to its cells using block-specific weights assigned to the cells having different ancillary classes. The weights are assigned based on the relative density of population for each ancillary class and the area of each block occupied by each class (Mennis 2003). The population in each cell is calculated by multiplying the number of people in the block by a weight assigned to the cell based on its ancillary class.

The important step in dasymetric modelling is the establishment of the relationship between ancillary and population data. A presented model uses the set of characteristic (or representative) values of population densities in each ancillary class (Mennis, Hultgren 2006). Representative population density for each class is established using a set of blocks (selected from the entire conterminous U.S.) having relatively homogeneous land cover (90% for urban class and 95% for vegetation class). The representative density for particular ancillary class is calculated by dividing the sum of population living in the selected blocks by the overall area of these blocks. Representative densities are required to establish the relative density of population used to calculate block-specific weights. Relative density of population for each ancillary class is calculated by dividing the representative density for this class and the sum of representative densities for all ancillary classes.

Computation and results

R implementation of dasymetric modelling

The major challenge to calculating 30m dasymetric model of population density for the entire conterminous U.S. is the size of input and output data. Population grids provided by SocScape project are the result of disaggregating ~11 millions of census blocks into over 8 billion (8,651,157,015) of grid cells. The choice of output resolution (30m) is dictated by the resolution of ancillary data as it is most convenient to disaggregate census data to the resolution of the ancillary data.

Traditionally, dasymetric modelling was computed in a GIS environment, such as ESRI ArcGIS (Sleeter, Gould 2007), QGIS (Mileu, Margarida 2018), GRASS GIS (Dmowska, Stepinski 2014). However, for a broad-scale model such approach is computationally inefficient. Processing such amount of data requires fully automated and flexible computational environment. Dasymeteric model used in SocScape is implemented in R (R Core Team 2018). R is a comprehensive computational environment that includes libraries to work with different types of data: tabular data, geospatial data (libraries *sp, sf, raster*). It also provides tools for binding to external data sources such as GRASS GIS (library *rgrass7*), GDAL (library *rgdal*) or standard relational databases (libraries *DBI, RSQLite*). R provides some advantage over GIS software and other programming languages. The main advantage is that it allows building efficient, flexible and fully automated computational environment to work with large dataset without advanced programming skills.

The key factor of calculation for the broad-scale model is to manage data storage requirements and to control the time of computation (see Fig. 1).

In order to handle a large dataset in R, geospatial data is first divided into separate counties using region concept in GRASS GIS. In GRASS GIS, region settings determine the spatial extent and resolution of the grid. Geospatial data is next

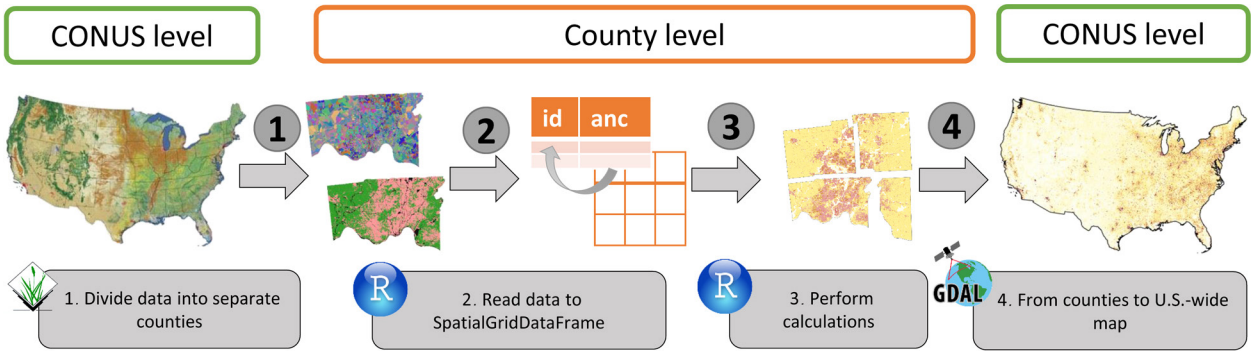

Fig. 1. Process of handling large datasets in R.

read into R and stored in *SpatialGridDataFrame* object. *SpatialGridDataFrame* is one of the spatial objects provided by *sp* library to work with spatial data in R (Pebesma, Bivand 2005). It integrates bounding box, the Coordinate Reference System and grid topology with attribute data stored in data.frame (tabular format in R). Working with *SpatialGridDataFrame* object allows integration of spatial content (census boundary, ancillary data) with Census population data into a single relational model, performing calculation at tabular level (data.frame), and, in the last step, propagating data into grids cells. The dasymetric modelling is performed in R for each county separately. In the last step dasymetric maps for individual counties are joined into a map for the entire conterminous United States.

Figure 2 shows in details the R-based implementation of dasymetric modelling in SocScape project. Calculation process has been divided into 5 steps:

1. pre-processing of Census and geospatial data,
2. the establishment of the relationship between ancillary and population data,
3. performing dasymetric modelling,
4. propagate dasymetric model for geospatial grids,
5. post processing: prepare hi-res population maps for the entire U.S.

The whole procedure is implemented in R. In addition to R, GRASS GIS 7.0, SQLite and GDAL library has been used in the pre-processing and post processing steps. The computational framework consists of several scripts used for reading U.S. Census text file into SQLite database, reading geospatial data from GRASS GIS to *SpatialGridDataFrame* object in R, performing dasymetric modelling, exporting population grids into GeoTiff and joining GeoTiffs into U.S.-wide map.

The first step of calculation procedure is the pre-processing of population and ancillary data. In this step U.S.-wide census block level data are imported from text file to SQLite database using R tools designed to work directly with a database (library *DBI, RSQLite*). SQLite is a public-domain, single-user relational database management system which stores the entire database as a single cross-platform file and implements a subset of the SQL 92 standard, including the core table creation, updating, insertion, and selection operations. *RSQLite* package embeds the SQLite database engine in R, providing a DBI-compliant interface (Mller et al. 2018).

RSQLite provides functions to read data from R to SQLite, write data directly from database table into data.frame in R, performing SQL queries. This functionality will be used to extract population data for the particular county and read them from a database directly into R data.frame object.

Geospatial data is pre-processed using GRASS GIS software (step not shown in Fig. 2.), before it is imported to *SpatialGridDataFrame* object in R. Block level census boundaries are available as state level shapefiles. Those shapefiles are imported to GRASS GIS, rasterized to match NLCD grid topology and divided into separate counties. Land cover data, used here as ancillary datasets, are stored as U.S.-wide files. Pre-processing of ancillary data includes extracting land cover data for 1992 from NLCD 1992/2001 Retrofit Land Cover Change Product, reclassifying NLCD maps into 3 classes (uninhabited, urban, vegetation) and dividing data into separate counties using region concept in GRASS GIS. Rasterized census block's boundaries and 3-class ancillary data are imported to *SpatialGridDataFrame* object in R using *rgrass7* package (Bivand 2017). Package *rgrass7* provides interpreted interface between GRASS geographical information system, version 7 and R (Bivand 2017). The interface uses classes defined in the *sp* package to hold spatial data (Pebesma, Bivand 2005). This package allows reading raster data directly from GRASS GIS to *SpatialGridDataFrame* (or SGDF) object in R. The SPGD object for each county with two layers (census boundaries and ancillary data) is stored as rds file.

In the next step, data for a particular county is read to R to perform dasymetric modelling. Population data is extracted from SQLite database and read directly to data.frame object in R. Geospatial data is restored from rds files containing SGDF object with census boundaries and ancillary data. Before performing dasymetric modelling ancillary data are upgraded based on population data by assigning uninhabited class to blocks with population equal to 0.

Next, area of each ancillary class in each block is calculated and stored in data.frame. Notice, that at this point two data.frames are available – one of them containing block id and population

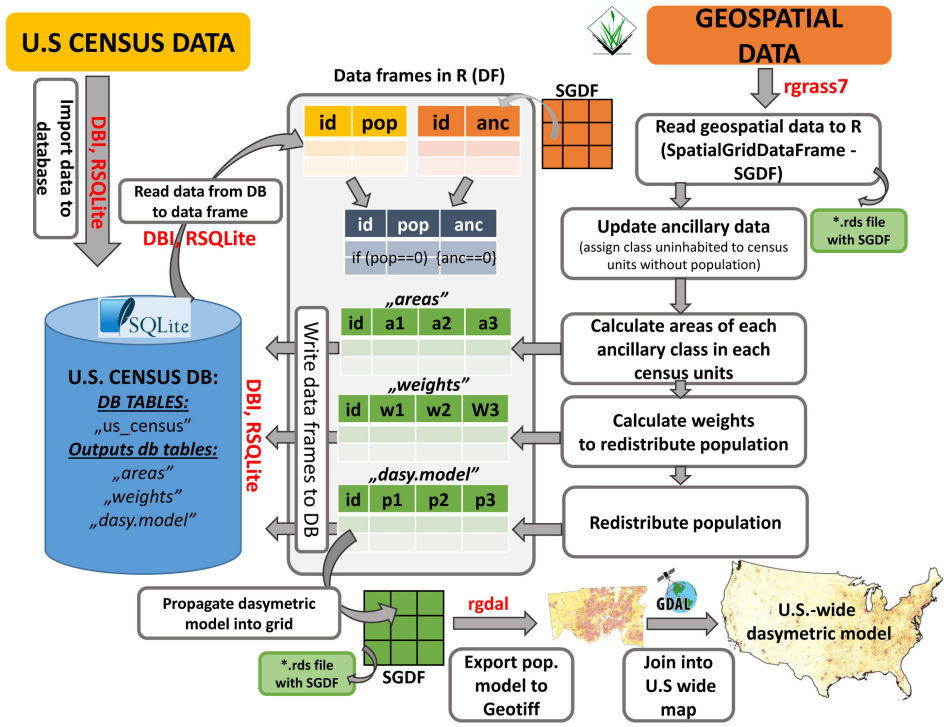

Fig. 2. Framework of calculation dasymetric modelling in R.

data and the second containing block id and the area of each ancillary class in this block. Using these two types of information stored in data. frames the weights to redistribute population among different ancillary classes are calculated for each block. Then population in each block is multiplied by weights. It results in creating the data.frame with the number of people assign to each type of ancillary class for each block. Areas, weights and the result of dasymetric procedure are stored in data.frames and are written to SQLite database to be used for further analysis.

The result of dasymetric modelling is also propagated into grid and stored as additional layer (together with ancillary information and census block boundaries) in *SpatialGridDataFrame* object. In the last step dasymetric population grid for each county is exported from R to Geotiff using *rgdal* library. Package *rgdal* provides bindings to the *Geospatial Data Abstraction Library* (GDAL) (>= 1.11.4) and access to projection/transformation operations from the PROJ.4 library (Bivand et al. 2018). Finally, GDAL library is used to create U.S. conterminous population grid based on counties Geotiff. First, Virtual Dataset (VRT) that is a mosaic of the counties Geotiffs is built using *gdalbuildvrt* program provides by GDAL. Next VRT object is converted to U.S.-wide Geotiff using *gdal_translate* program provided by GDAL. The result is U.S.-wide population grid at 30m resolution.

The calculation of a dasymetric model for a single county (containing 10, 000 blocks) takes 14 seconds. In comparison, using dasymetric modelling toolbox (Sleeter, Gould 2007) in ArcGIS software calculations takes 600 seconds. The whole procedure from the pre-processing steps to obtaining the final dasymetric map for the entire conterminous U.S. takes 55 h using a PC computer with Intel 3.4 GHz, 4-cores processor and 16 GB of memory running the Linux system. The most time consuming step is data pre-processing (37 h). Determining a relation between population and ancillary data and performing dasymetric model takes 6 h, and creating one map from counties' dasymetric models takes 12 h.

Examples of U.S.-wide population grids

Described implementation of dasymetric model has been used to produce high resolution, multi-year comparable, U.S.-wide population grids which are the part of the SocScape (Social Landscape) project. This project provides an open access to high resolution (30 m) population, subpopulation (separate race/ethnicity group) and

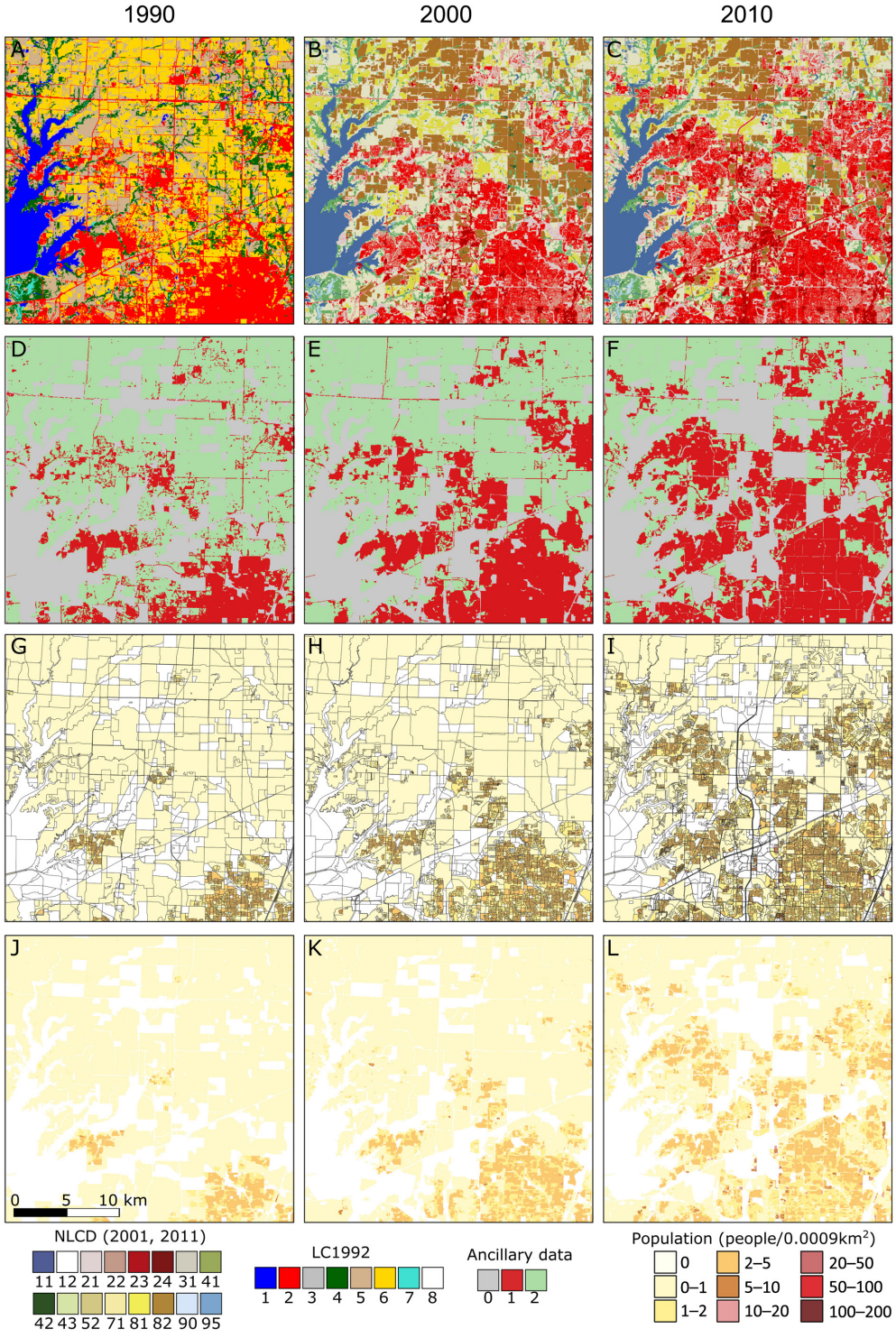

Fig. 3. The city of Frisco, TX located in Collin and Denton county.
(A) Land cover map for 1992 based on 1992/2001 Retrofit Land Cover Change Product. (B) NLCD 2001. (C) NLCD 2011. (D–F) Land cover data reclassified into 3 categories. (G–I) Block level population data. (J–L) Population density shown in 30 m resolution grids.
NLCD 2001, 2011: 11 – open water; 12 – ice/snow; 21 – developed, open space; 22 – developed, low intensity; 23 – developed, medium intensity; 24 – developed, high intensity; 31 – barren land; 41 – deciduous forest, 42 – evergreen forest, 43 – mixed forest, 52 – shrub/scrub; 71 – grassland; 81 – pasture/hay; 82 – cultivated crops, 90 – woody/wetlands, 95 – emergent wetlands.
LC1992: 1 – open water; 2 – urban; 3 – barren; 4 – forest; 5 – grass/shrub; 6 – wetlands; 7 – wetlands; 8 – ice/snow
Ancillary data: 0 – uninhabited; 1 – urban areas; 2 – nonurban areas.

racial diversity grids for the entire conterminous United States for 1990, 2000, 2010 (Dmowska, Stepinski 2017b, Dmowska et al. 2017). SocScape project consists of two parts – GeoWeb application designed to explore U.S-wide population and racial diversity grids and SocScape data website, which provides data for each county and for 363 MSA as a zip archive.

Figure 3 shows an example of population grids for the area centred on the city of Frisco, TX located in the Collin and Denton county. Frisco, TX is a part of the Dallas-Fort Worth metropolitan area and it is considered as the fastest-growing city in the United States from 2000–2009 with the population of 116,989 people at the 2010 census. Figure 3 is divided into 12 panels arranged into 3 columns (corresponding to 1990, 2000, 2010 year respectively) and 4 rows (corresponding to different types of data). The panels A–I show population and ancillary data used as an input to dasymetric modelling. Panels A–C present land cover data with the original legends whereas panels D–F present ancillary data reclassified into three categories, which are fully comparable between years. Panels G–I show census block level population data. The population grids are presented in the Panel J–L.

In this example, census block and population grids show the main features of population distribution in a similar way. The main limitation of block level data is that they cannot be used to quantitatively assess changes in population distribution. The boundary of aggregated units changed between 1990 and 2010 year. The urbanization process, which is seen in the land cover maps, caused an increase in the number of blocks in the presented area from 2100 in 1990 year to 10360 in 2010 year. On the other hand, population grids can be directly used to assess changes in population distribution, as they are produced based on multi-year comparable ancillary data.

Conclusion

This paper reviewed literature on production of broad-scale population grids and reported on the R implementation of an automated framework to perform dasymetric modelling to produce such grids. Described implementation of dasymetric model has been used to produce high resolution, multi-year comparable, U.S.-wide population grids for 1990, 2000, 2010 year.

Main advantages of using R to perform dasymetric calculation are:

1. no advanced programming skills are required,
2. less processing steps are required than using GIS software,
3. no intermediate layers are produced,
4. increased flexibility and automation, and
5. easily expandable to variables other than total population.

The framework has been implemented to work with U.S. Decennial Census population data available for 1990, 2000, 2010 years. However, it can be easily modified to work with other source of data and for other levels of aggregation (i.e. census tracts, block groups). The practical advantage of presented framework has been already illustrated by computing high resolution demographic grids for race/ethnicity subpopulation using weights established by population model. Preparing U.S.-wide demographic grids, using already established weights, take 13 h and it does not required any pre-processing steps. The weights established by population model are stored in SQLite database. The other demographic grids can be calculated by importing U.S.-wide block level data to SQLite database and by multiplying its counts by weights. Also the presented framework can be used to preparing high resolution population grids for 2020, when U.S Decennial Census data become available.

Presented framework can be also easily expandable to calculate other types of maps which use as an input the results of dasymetric modelling. Examples of such maps are racial diversity maps (Dmowska, Stepinski 2017b, Dmowska et al. 2017) and racial dots maps (Dmowska, Stepinski 2019).

Acknowledgments

SocScape project has been developed in the Space Informatics Lab at the University of Cincinnati and it is available on http://sil.uc.edu. All R scripts are available from http://dmowska.home.amu.edu.pl. Author would like to thank both reviewers for their helpful and insightful comments on the paper.

References

Alegana V.A., Atkinson P.M., Pezzulo, C., Sorichetta A., Weiss D., Bird T., Erbach-Schoenberg E., Tatem A.J., 2015. Fine resolution mapping of population age-structures for health and development applications. *Journal of the Royal Society Interface* 12(105): 20150073.

Bakillah M., Liang S., Mobasheri A., Jokar Arsanjani J., Zipf A., 2014. Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science* 28(9): 1940–1963.

Batista e Silva F., Gallego J., Lavalle C., 2013. A high-resolution population grid map for Europe, *Journal of Maps* 9(1): 16–28.

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

Berke P., Newman G., Lee J., Combs T., Kolosna C., Salvesen D., 2015. Evaluation of networks of plans and vulnerability to hazards and climate change: A resilience scorecard. *Journal of the American Planning Association* 81: 287–302.

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

Bielecka E., 2005. A dasymetric population density map of Poland. In *Proceedings of the 22nd International Cartographic Conference:* 9–15.

Bivand R., 2017. *rgrass7: Interface Between GRASS 7 Geographical Information System and R. R package version 0.1-10*. Online: https://CRAN.R-project.org/package=rgrass7 (accessed 10 February 2019).

Bivand R., Keitt T., Rowlingson B., 2018. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.3-4*. Online: URL https://CRAN.R-project.org/package=rgdal (accessed 10 February 2019).

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

Calka B., Nowak Da Costa J., Bielecka E., 2017. Fine scale population density data and its application in risk assessment. *Geomatics, Natural Hazards and Risk* 8(2): 1440–1455,

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

Deville P., Linard C., Martin S., Gilbert M., Stevens F.R., Gaughan A.E., Blondel V.D., Tatem A. J., 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111(45): 15888–15893.

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

Dmowska A., Stepinski T.F., 2017a. A high resolution population grid for the conterminous United States: The 2010 edition. *Computers, Environment and Urban Systems* 61: 13–23.

Dmowska A., Stepinski T.F., 2017b. Mapping changes of racial composition in the United States: 1990–2010. In: *Population Association of America 2017, Chicago, April 26–29, 2017*. Online: https://paa.confex.com/paa/2017/meetingapp.cgi/Paper/10564 (accessed 10 February 2019).

Dmowska A., Stepinski T.F., 2019. Mapping racial diversity using grid-based racial dot maps and racial diversity maps. Accepted In: *Population Association of America 2019, Austin, April 10–13, 2019*. Online: http://paa2019.populationassociation.org/abstracts/191266 (accessed 10 February 2019).

Dmowska A., Stepinski T.F., Netzel P., 2017. Comprehensive framework for visualizing and analyzing spatio-temporal dynamics of racial diversity in the entire United States. *PLoS ONE* 12(3): e0174993.

Dobson J.E., Bright E.A., Coleman P.R., Worley B.A., 2000. LandScan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing* 66 (7): 849–857.

Doxsey-Whitfield E., MacManus K., Adamo S.B., Pistolesi L., Squires J., Borkovska O., Baptista S.R., 2015. Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography* 1 (3): 226–234.

EEA [European Environment Agency], 2019. Data and maps. Online: www.eea.europa.eu/data-and-maps (accessed February 11, 2019).

Eicher C.L., Brewer C.A., 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28: 125–138.

Flowerdew R., Green M., 1992. Developments in areal interpolation methods and GIS. *Annals of Regional Science* 26: 67–78.

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

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

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

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

Goodchild M., Anselin L., Deichmann 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.

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

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

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

Langford M., 2013. An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis* 45(3): 324–344.

Langford M., Unwin D.J., 1994. Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal* 31(1): 21–6.

Linard C., Gilbert M., Tatem A.J., 2011. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* 76(5): 525–538.

Lloyd C. T., Sorichetta A., Tatem A.J., 2017. High resolution global gridded data for use in population studies. *Scientific data 4:* 170001.

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

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

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

Maantay J.A., Maroko A.R., Herrmann C., 2007. Mapping population distribution in the urban environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science* 34 (2): 77–102.

McGranahan G., Balk D., Anderson B., 2007. The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization* 19(1): 17–37.

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

Mennis J., 2009. Dasymetric mapping for estimating population in small areas. *Geography Compass* 3(2): 727–745.

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

Mileu N., Queirós M., 2018. Development of a QGIS plugin to dasymetric mapping. *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings* 18: 9. Online: scholarworks.umass.edu/foss4g/vol18/iss1/9 (accessed 10 February 2019).

Mller K., Wickham H., James D. A., Falcon S., 2018. *RSQLite: 'SQLite' Interface for R. R package version 2.1.1*. Online: https://CRAN.R-project.org/package=RSQLite (accessed 10 February 2019).

Mondal P., Tatem A.J., 2012. Uncertainties in measuring populations potentially impacted by sea level rise and coastal flooding. *PLoS ONE* 7(10): 1–7.

MPC [Minnesota Population Center], 2019. National Historical Geographic Information System. Online: www.nhgis.org (accessed February 11, 2019).

Murray A.T., Davis R., Stimson R. J., Ferreira. L., 1998. Public transportation access. Transportation Research Part D: *Transport and Environment* 3(5): 319–328.

ONRL [Oak Ridge National Laboratory], 2019. LandScan™. Online: landscan.ornl.gov (accessed February 11, 2019).

Patel N.N., Stevens F.R., Huang Z., Gaughan A.E., Elyazar I., Tatem A.J., 2017. Improving large area population mapping using geotweet densities. *Transactions in GIS* 21(2): 317–331.

Pattnaik S.B., Mohan S., Tom V.M., 1998. Urban bus transit route network design using genetic algorithm. *Journal of transportation engineering* 124(4): 368–375.

Pebesma E.J., Bivand R.S., 2005. *Classes and methods for spatial data in R*. R News 5 (2), 9–13. Online: https://CRAN.R-project.org/doc/Rnews/ (accessed 10 February 2019)

Petrov A., 2012. One hundred years of dasymetric mapping: back to the origin. *Cartographic Journal* 49(3): 256–264.

R Core Team, 2018. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Online https://www.R-project.org/ (accessed 10 February 2019).

Reibel M., Bufalino M.E., 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 37(1): 127–139.

Schroeder J.P., 2007. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39(3): 311–335.

SEDAC [Socioeconomic Data and Applications Center], 2019. Gridded Population of the World (GPW), v4. Online: sedac.ciesin.columbia.edu/data/collection/gpw-v4 (accessed February 11, 2019).

Sleeter R., Gould M., 2007. *Geographic Information System software to remodel population data using dasymetric mapping methods*.

Sridharan H., Qiu F., 2013. A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis* 45(3): 238–258.

Stevens F.R., Gaughan A.E., Linard C., Tatem A.J., 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE* 10(2): 1–22.

Tapp A.F., 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science* 37 (3): 215–228.

Tatem A.J., Adamo S., Bharti N., Burgert C.R., Castro M., Dorelien A., Fink G., Linard C., John M., Montana L., Montgomery M.R., Nelson A., Noor A.M., Pindolia D., Yetman G., Balk D., 2012. Mapping populations at risk: improving spatial demographic data for infectious disease modelling and metric derivation. *Population Health Metrics* 10(1): 8.

Tatem A.J., Campbell J., Guerra-Arias M., De Bernis L., Moran A., Matthews Z., 2014. Mapping for maternal and newborn health: the distributions of women of childbearing age, pregnancies and births. *International Journal of Health Geographic* 13(1): 2.

Tatem A.J., Campiz N., Gething P.W., Snow R. W., Linard C., 2011. The effects of spatial population dataset choice on estimates of population at risk of disease. *Population Health Metrics* 9(1): 4.

Tatem A.J., Gaughan A.E., Stevens F.R., Patel N.N., Jia P., Pandey A., Linard C., 2013. Quantifying the effects of using detailed spatial demographic data on health metrics: a systematic analysis for the AfriPop, AsiaPop, and AmeriPop projects. *The Lancet* 381: S142.

Tatem A.J., Guerra C.A., Kabaria C.W., Noor A.M., Hay S.I., 2008. Human population, urban settlement patterns and their impact on plasmodium falciparum malaria endemicity. *Malaria Journal* 7(1): 218.

Tatem A.J., Noor A.M., von Hagen C., Di Gregorio 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): 1–8.

Tenerelli P., Gallego J.F., Ehrlich D., 2015. Population density modelling in support of disaster risk assessment. *International Journal of Disaster Risk Reduction* 13: 334–341.

The WorldPop project, 2019. *The WorldPop project*. Online: http://www.worldpop.org.uk/ (accessed 11 February 2019).

Thieken A.H., Müller M., Kleist L., Seifert I., Borst D., Werner U., 2006. Regionalisation of asset values for risk analyses. *Natural Hazards and Earth System Science* 6(2): 167–178.

Tralli D.M., Blom R.G., Zlotnicki V., Donnellan A., Evans D., 2005. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of Photogrammetry and Remote Sensing* 59(4): 185–198.

Ural S., Hussain E., Shan J., 2011. Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation* 13(6): 841–852.

Vinkx K., Visee T., 2008. Usefulness of population files for estimation of noise hindrance effects. 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.

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

Wu S., Qiu X., Wang L., 2005 Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience & Remote Sensing* 42(1): 80–96.

Zandbergen P.A., 2011. Dasymetric mapping using high resolution address point datasets. *Transactions in GIS* 15: 5–27.