International Cartographic Association
Association Cartographique Internationale

Estimating changes in urban land and urban population using refined areal interpolation techniques

Hamidreza Zoraghein,[a] Stefan Leyk[a]

[a] *Department of Geography, University of Colorado, Boulder, Colorado, USA; hamidreza.zoraghein@colorado.edu, stefan.leyk@colorado.edu*

Abstract: The analysis of changes in urban land and population is important because the majority of future population growth will take place in urban areas. U.S. Census historically classifies urban land using population density and various land-use criteria. This study analyzes the reliability of census-defined urban lands for delineating the spatial distribution of urban population and estimating its changes over time. To overcome the problem of incompatible enumeration units between censuses, regular areal interpolation methods including Areal Weighting (AW) and Tar-get Density Weighting (TDW), with and without spatial refinement, are implemented. The goal in this study is to estimate urban population in Massachusetts in 1990 and 2000 (source zones), within tract boundaries of the 2010 census (target zones), respectively, to create a consistent time series of comparable urban population estimates from 1990 to 2010. Spatial refinement is done using ancillary variables such as census-defined urban areas, the National Land Cover Database (NLCD) and the Global Human Settlement Layer (GHSL) as well as different combi-nations of them. The study results suggest that census-defined urban areas alone are not necessarily the most meaningful delineation of urban land. Instead, it appears that alternative combinations of the above-mentioned ancillary variables can better depict the spatial distribution of urban land, and thus make it possible to reduce the estimation error in transferring the urban population from source zones to target zones when running spatially-refined temporal areal interpolation.

Keywords: Urban population, Urban land, Areal interpolation, Spatial refinement, Census

1. Introduction

The analysis of changes in urban land and urban population is important because a large proportion of human population resides in urbanized or peri-urban areas, and this proportion is continuously increasing. Knowledge of such trends has important implications in interdisciplinary contexts including climate change and energy consumption, risk assessment and crisis management as well as land-use and urban planning, to name a few. However, such trends are difficult to measure using existing, temporally inconsistent population data. Therefore, this research employs areal interpolation methods coupled with spatial refinement to analyze urban land and urban population in different census years, from 1990 to 2010, within consistent fine-resolution census units such as census tracts.

Historically, the U.S. Census Bureau has defined urban areas for each census year based on criteria related to population density and land-use. However, these criteria have changed over time, and consequently the urban lands in 1990 or 2000 underlie different definitions than those in 2010 (U.S. Census Bureau 2011). The main objective of this study is to assess how the urban areas defined in 1990, 2000 and 2010 actually reflect the spatial distribution of urban population and how this spatial depiction can be improved using other ancillary variables for spatial refinement.

Areal interpolation coupled with spatial refinement has been demonstrated as an effective approach to reduce estimation errors in temporally interpolating population enumerated in a set of source zones (source census year) to target zones defined by the boundaries of the target census year (e.g. Ruther et al. 2015; Zoraghein et al. 2016). In this study, this approach is tested for estimating urban population of census tracts in 1990 and 2000 (i.e., the source zones) within census tract boundaries in 2010 (i.e., the target zones) using different ancillary variables for spatial refinement to create a temporally consistent time series of urban population distributions at the tract level. The analysis is carried out for the whole state of Massachusetts. The validation results are evaluated to determine which ancillary variables represent urban land most reliably.

Figure 1 shows the census-defined urban areas of Massachusetts in 1990, 2000 and 2010. Massachusetts is a highly urbanized state; according to the U.S. Census, its urban proportion of the total population has changed from 84.3% to 92% during 1990 to 2010. Figure 1 also depicts a growing pattern in the urban areas of the state (from around 5093 km2 in 1990 to around 8045 km2 in 2010). This study explores if these areas represent the distribution of urban population reliably, and proposes other variables to delineate areas where the urban population lives.

Proceedings of the International Cartographic Association, 1, 2017. This contribution underwent single-blind peer review based on submitted abstracts | https://doi.org/10.5194/ica-proc-1-130-2017 | © Authors 2017. CC BY 4.0 License.

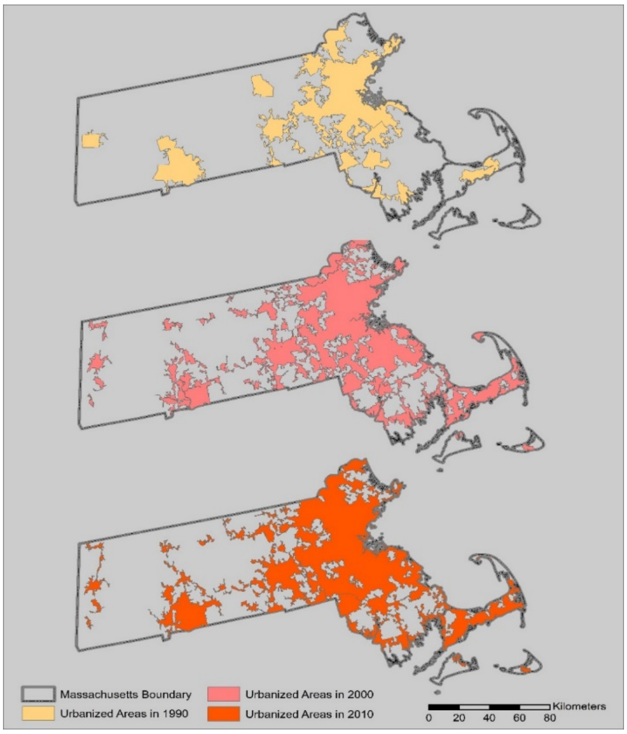

Fig. 1. The state of Massachusetts and its census-defined urban areas in 1990, 2000 and 2010.

2. Data

The boundaries of census tracts in 1990, 2000 and 2010 along with their urban population counts found in the summary files are the focus in this study. Census blocks represent the smallest enumeration units published by the Census. Blocks are labeled either urban or rural in all the three census years. Therefore, urban-labeled blocks in 1990 and 2000 as well as their population values are used here as reference data to evaluate the estimated urban population counts at the tract level. The tract-level and block-level boundaries and population values for 1990 were retrieved from the National Historical Geographic Information System (NHGIS) (Minnesota Population Center 2016) whereas they were extracted from the Census website for 2000 and 2010.

Three ancillary variables associated with the distribution of urban population are used in this study. They include census-defined urban areas in 1990, 2000 and 2010, the National Land Cover Database (NLCD) in 1992, 2001 and 2011 and the Global Human Settlement Layer (GHSL). NLCD is a Landsat based national land cover dataset at 30m resolution. Its primary objective is to provide nationally complete, current, consistent, and public domain in-formation on the nation's land cover. The dataset presents different land cover types in different classes (Homer et al. 2007). The main focus in this study is on developed land cover classes that could be related to human settlement (i.e., classes 21, 22 and 23 in 1992 and classes 21, 22, 23 and 24 in 2000 and 2010). The GHSL represents global spatial information about the human presence on the planet over time. In this study, the Landsat based fine resolution (38m) version of GHSL is used. It contains built-up land from before 1975 to 2014 (Pesaresi et al. 2016).

3. Methodology

Two areal interpolation methods, namely Areal Weighting (AW) (Goodchild and Lam 1980) and Target Density Weighting (TDW) (Schroeder 2007) are implemented to estimate the urban population in Massachusetts in 1990 and 2000 within target tract boundaries used for the census survey in 2010. All methods described are run for two time periods, 1990 to 2010 and 2000 to 2010, respectively. The methods are briefly described below, but the reader can refer to previous works (e.g., Zoraghein et al. 2016) for more detailed explanations and mathematical formulae. Importantly, in this study the spatially refined temporal interpolation framework is applied to urban population using ancillary variables that are known to be associated with urban lands and thus delineate areas where urban population is expected to reside.

AW is the most basic areal interpolation method and assumes the population density is constant within source zones. The method estimates source population in target zone boundaries based on the overlapping area between source and target zones (i.e., intersections or "atoms"). The population of each target zone is then simply calculated by summing up the population counts of all the atoms within it.

Spatially refining source zones prior to areal interpolation is supported by different ancillary variables and modifies the underlying assumption as follows: population is homogenously distributed within the developed land of a source zone, and no population is assigned to non-developed parts. This assumption is expected to be more realistic and generally results in more precise reapportionment of population counts.

Schroeder (2007) introduced TDW as an areal interpolation method appropriate for temporal analysis of census data. TDW is based on the assumption that the spatial distribution of population density in the source year among atoms and with regard to the encompassing source zones remains proportionally the same over time. For example, if population density is distributed in a 2:1 ratio between two atoms in 2010, it is assumed that this ratio was the same in 2000.

Based on previous studies, TDW often outperforms AW (Schroeder 2007; Schroeder and Van Riper 2013), suggesting that it is more reasonable to assume that the ratio of population density of atoms to their encompassing source zones remains constant than to assume that population is homogeneously distributed within source zones.

Refined TDW uses only developed/built-up areas within both source and target zones. This refinement implies that the underlying assumption of unrefined TDW be modified. In a first step, source and target zones are spatially refined using the areas labeled by the ancillary variable. Then TDW is applied to these refined areas under the assumption that the ratio of refined population densities of atoms to refined population densities of source zones re-mains the same temporally.

While refined AW uses developed areas of only the source year, refined TDW incorporates those areas of both the source and target years.

Spatial refinement is done using census-defined urban areas, GHSL built-up area, various combinations of NLCD developed classes as well as different intersections of those three datasets. Consequently, for each refinement scenario a new set of population estimates is created.

Once urban population estimates are derived for target tract boundaries, the interpolation results are validated using urban blocks of the source year (1990 and 2000, respectively). For example, for validating the estimated urban population values in 1990 within target tract boundaries in 2010, urban blocks in 1990 are used and aggregated to the target boundaries to constitute ground-truth values.

4. Results

Four absolute error metrics are used to evaluate different implementations of the two methods. Those metrics are Mean Absolute Error (MAE), median absolute error, 90th percentile of absolute errors, and Root Mean Square Error (RMSE). MAE is calculated by averaging the absolute differences between estimated and block-aggregated values of target tracts. Median absolute error is determined by taking the value of the 50th percentile of absolute errors. RMSE is calculated based on absolute differences between estimated and block-aggregated values of target tracts by taking the square root of the mean of squared differences.

These error measures help to characterize the error distributions, leading to more comprehensive comparative analysis on the performance of the established methods. That is, the MAE and RMSE measures demonstrate the overall representative behavior of the estimation error and are sensitive to outliers while the median absolute error and 90% percentile of absolute errors can be used to describe the upper end of the error distribution and placement of extreme absolute error values.

Tables 1 and 2 show the results of a selected set of implementations used in this study for the periods 1990-2010 and 2000-2010, respectively. Along with regular AW and TDW, refined AW and refined TDW using different ancillary variables are also included in Tables 1 and 2. Refinement through urban areas means using census-defined urban areas as the ancillary variable. The label “Most Reliable” indicates the refinement scenario using some combi-nation of NLCD, GHSL and census-defined urban areas that resulted in lowest estimation errors when compared to the urban block populations.

Method	MAE	Median Absolute Error	RMSE	90% of Absolute Error
AW	353	58	835	1146
Refined AW (Urban Areas)	486	63	1168	1574
Refined AW (Most Reliable)	147	39	309	471
TDW	166	62	337	437
Refined TDW (Urban Areas)	354	68	994	653
Refined TDW (Most Reliable)	142	53	278	407

Table 1. Absolute error measures of different areal interpolation methods for 1990 to 2010

Method	MAE	Median Absolute Error	RMSE	90% of Absolute Error
AW	322	11	1393	1006
Refined AW (Urban Areas)	181	6	589	532
Refined AW (Most Reliable)	138	5	445	393
TDW	60	10	152	164
Refined TDW (Urban Areas)	76	8	237	170
Refined TDW (Most Reliable)	51	6	145	143

Table 2. Absolute error measures of different areal interpolation methods for 2000 to 2010

The most reliable solution for spatially refining AW for estimating urban population in 1990 within target tract boundaries of the 2010 census is the intersection of NLCD developed classes 21 and 22 and the GHSL built-up are-as whereas TDW performs most reliably when tracts are refined using the intersection of NLCD developed classes 21, 22 and 23 and GHSL in the source year and the intersection of NLCD developed classes 22, 23 and 24 and GHSL in the target year. For the 2000-2010 time period, AW is refined most effectively using the intersection of NLCD developed classes 22, 23 and 24, GHSL and the census-defined areas in the source year. The most reliable solution for spatially refined TDW for 2000-2010, however, can be found when employing NLCD developed classes 22, 23 and 24 for the both source and target years. Figure 2 shows for each year the derived delineations of “re-vised” urban land used for spatial refinement in the most reliable scenarios for TDW shown in Tables 1 and 2. These revised urban land depictions are the optimal spatial distributions for transferring urban population statistics in 1990 and 2000 to tract boundaries of the 2010 census, respectively.

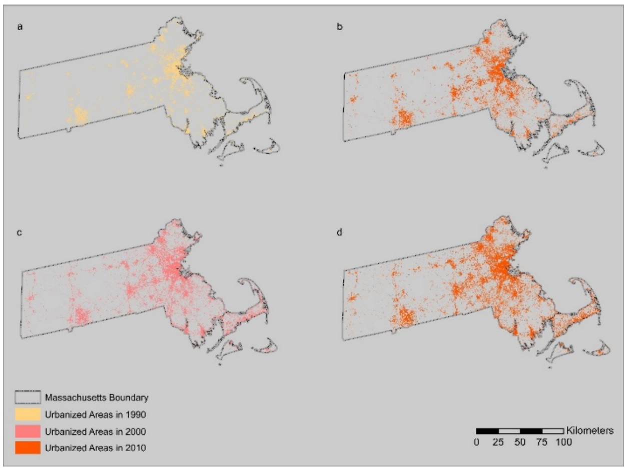

Fig. 2. The most reliable solutions of spatial refinement for TDW in 1990 (a) and 2010 (b) for the 1990-2010 time period and in 2000 (c) and 2010 (d) for the 2000-2010 time period.

5. Discussion and Conclusions

According to Table 1, refined AW and refined TDW using census-defined urban areas increase the absolute errors as compared to regular, unrefined implementations of the two methods. In the case of refined AW, this could mean that urban areas in 1990 do not explain the distribution of urban population effectively. For refined TDW this increase in error could indicate that the census-defined urban areas in 1990 and 2010 do not adequately describe the changes of urban footprints between the two years.

However, the use of above-described combinations of ancillary variables for spatial refinement results in considerable improvement of both refined AW and refined TDW. These observations imply that the spatial distributions composed of the optimal combination of ancillary variables can be seen as a more representative delineation of the urban settings in 1990 and seem to more reliably reflect changes in urban lands between 1990 and 2010.

Table 2 allows similar interpretations for the time between 2000 and 2010 except that refined AW using census-defined urban areas results in higher accuracy levels than AW, meaning that those areas are appropriate ancillary data for spatial refinement of urban population estimates. However, other ancillary variables in combination appear to represent more reliable urban footprints in 2000 and reflect more reliable changes in urban land between 2000 and 2010.

It is acknowledged that the U.S. Census classifies urban areas using many criteria and aims to improve the classification process to make it more consistent and reflective of urban criteria. This study represents an initial step to-ward evaluating the existing urban areas and possibly improving their classification using other nationally and globally available ancillary datasets that can be used to delineate areas of urban population. The tremendous potential of improvement especially for modeling changes of urban land and urban population from 1990 to 2010 were observed in Massachusetts using the exogenous ancillary variables. The analysis will be repeated for different states and possibly at the national level to assess the consistency of the improvement results. Moreover, the resulting modified representations of urban land need to be analyzed in conjunction with other social and physical pro-cesses such as migration, land-use change, energy consumption and crisis management to see how the modeling of these processes can benefit from the new establishments of urban land.

6. Acknowledgements

Research reported in this publication was, in part, supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD066613 and through a development grant received from its Population Center (CUPC) at the Institute of Behavioral Science. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was also funded, in part, by the US National Science Foundation award #1416860 to the City University of New York, the Population Council, the National Center for Atmospheric Research and the University of Colorado at Boulder. Finally, Innovative Seed Grant funding from the University of Colorado are acknowledged.

7. References

Goodchild, M., & Lam, N. S. N. (1980). Areal interpolation: a variant of the traditional spatial problem. Geo-Processing, 1, 297–312.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., et al. (2007). Completion of the 2001 national land cover database for the conterminous United States. Photogrammetric Engineering and Remote Sensing, 73(4), 337–341.

Minnesota Population Center. (2016). National Historical Geographic Information System: Version 11.0 [Database]. University of Minnesota. http://doi.org/10.18128/D050.V11.0

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., et al. (2016). Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. JRC Technical Report; European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen: Ispra, Italy.

Ruther, M., Leyk, S., Buttenfield, B. P. (2015). Comparing the Effects of an NLCD-derived Dasymetric Refinement on Estimation Accuracies for Multiple Areal Interpolation Methods. GIScience & Remote Sensing, 52(2), 158–178.

Schroeder, J. P. (2007). Target density weighting interpolation and uncertainty evaluation for temporal analysis of census data. Geographical Analysis, 39(3), 311–335. http://doi.org/10.1111/j.1538-4632.2007.00706.x

Schroeder, J. P., & Van Riper, D. C. (2013). Because Muncie's densities are not Manhattan's: Using geographical weighting in the EM algorithm for areal interpolation. Geographical Analysis, 45(3), 216–237. http://doi.org/10.1111/gean.12014

U.S. Census Bureau. (2011). Differences Between the Census 2000 and 2010 Census Urban Area Criteria. http://www2.census.gov/geo/pdfs/reference/ua/2000_2010uadif.pdf. Accessed 26 February 2017

Zoraghein, H., Leyk, S., Ruther, M., Buttenfield, B. P. (2016). Exploiting temporal information in parcel data to refine small area population estimates. Computers, Environment and Urban Systems, 58, 19–28. http://doi.org/10.1016/j.compenvurbsys.2016.03.004