Deriving Spatially Refined Consistent Small Area Estimates over Time Using Cadastral Data

H. Zoraghein[1,*], S. Leyk[1], M. Ruther[2], B. P. Buttenfield[1]

[1]Department of Geography, University of Colorado, Boulder, CO, 80309, USA
Email:{ hamidreza.zoraghein , stefan.leyk , babs }@colorado.edu

[2] Department of Urban and Public Affairs, University of Louisville, Louisville, KY, 40208, USA
Email: matthew.ruther@louisville.edu

[*]Corresponding author; Telephone: (+1)7208782151

Abstract

Areal interpolation is the process of transferring data collected over source zones to target zones. One of the applications of areal interpolation is to construct temporally consistent areal units for comparing socioeconomic data over time. Dasymetric modeling is the process of employing ancillary data to spatially refine the distribution of socioeconomic variables. In this paper, three areal interpolation methods, namely areal weighting (AW), target density weighting (TDW) and pycnophylactic (PM), with and without spatial refinement, are used to interpolate census tract populations in 2000 (source zones) to census tract boundaries in 2010 (target zones). The spatial refinement is performed using residential parcels. The study area is Hennepin County, Minnesota. Accuracy assessment is based on the interpolated population of each target zone and its benchmark population resulting from aggregating the population values of census blocks within it. According to accuracy comparisons, spatial refinement has potentials to improve areal interpolation results. However, the improvement level depends on the employed areal interpolation method.

Keywords: Areal Interpolation, Dasymetric Modeling, Small Area Estimates, Cadastral Data, Census Tracts.

1. Introduction

Areal interpolation techniques such as Areal Weighting (AW) (Goodchild and Lam 1980), pycnophylactic modeling (PM) (Tobler 1979), and Target Density Weighting (TDW) (Schroeder 2007) have been used for constructing temporally consistent enumeration units in cases where these units are incompatible. Dasymetric modeling is a special type of areal interpolation that employs ancillary data to spatially refine the distribution of the variable of interest (Wright 1936, Mennis 2009), allowing the development of advanced small area estimation techniques (Kim and Yao 2010, Leyk et al. 2013). Dasymetric refinement prior to areal interpolation for temporal analysis can improve the precision and accuracy of population change estimates. Holt et al. (2004) show that refining census units using developed land cover classes at one point in time improves areal weighting results.

Existing national land cover data is limited in rural areas because its coarse resolution tends to underestimate residential areas (Leyk et al. 2014). In this study, parcel data are tested to depict populated places more reliably. Population estimates from the U.S.

Decennial Census for 2000 and 2010 are dasymetrically refined using residential parcels prior to areal interpolation for temporal analysis. Each of three areal interpolation methods are run with and without refinement to interpolate census tract populations in 2000 to census tract boundaries for 2010, and estimation errors are assessed.

2. Study Area, Data and Preprocessing Steps

Hennepin County, Minnesota (fig. 1) contains dense urban places in the east and sparsely populated rural portions in the west, making it an ideal area to evaluate methods performance under different conditions.

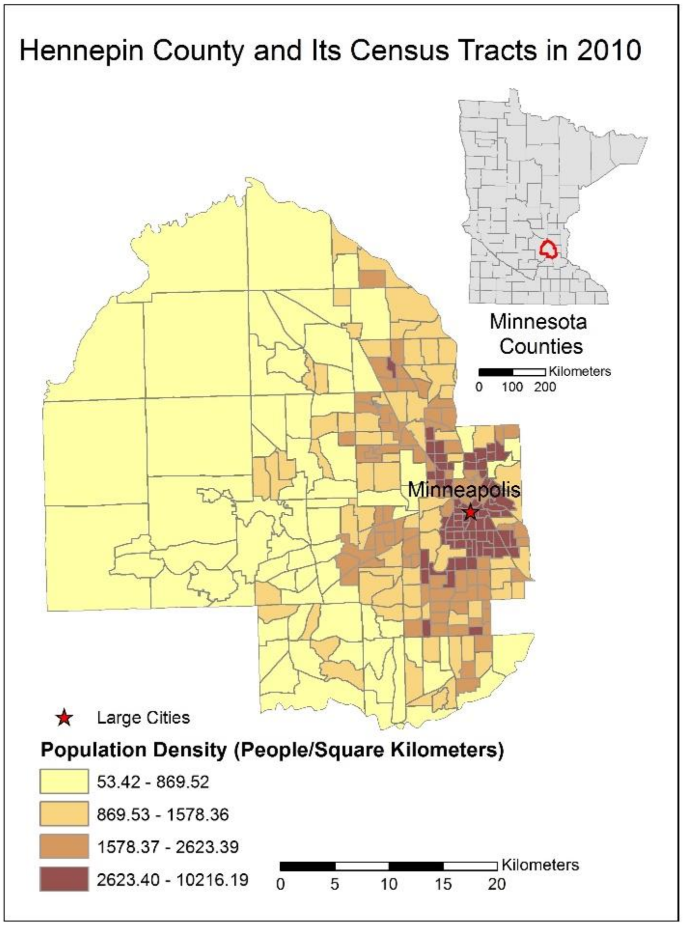

Figure 1. The study area.

Changes in census tract boundaries between 2000 and 2010 were divided into unchanged tracts (less than 900 m^2), tracts that changed slightly (more than 900 m^2 but less than 45,000 m^2), and those that changed significantly (more than 45,000 m^2). These thresholds allow for differentiation between changes caused by improved precision in the 2010 TIGER files and ‘real’ changes in boundaries at two different levels. Of the 298 study area tracts in 2010, 129 were categorized as slightly changed from 2000 and 52 as significantly changed.

All parcel types relating to residential use were selected. The most important attribute in this parcel dataset was the year built, which was used to create distributions of residential parcels with buildings built before 2000 and buildings built before 2010. To validate, Census block data from 2000 were aggregated to tracts and compared with the interpolated tract values to calculate error measures.

3. Methods

3.1. Areal Weighting

AW estimates source population in target zone boundaries based on the overlapping area between source and target zones (i.e., atoms). The underlying assumption is that population is uniformly distributed within a source zone (equation 1):

$$pop_{st} = (\frac{A_{st}}{A_s}) \times pop_s \tag{1}$$

A_{st} is the atom area, A_s is the source zone area, pop_s is the source zone population, and pop_{st} is the atom population. The population of target zone t is calculated by summing the population counts of all the atoms within it.

In refined AW, population is assumed to be homogenously distributed within the residential portion of a source tract. Accordingly, all area elements in equation 1 are refined into residential areas. Fig. 2 illustrates how spatial refinement modifies areas.

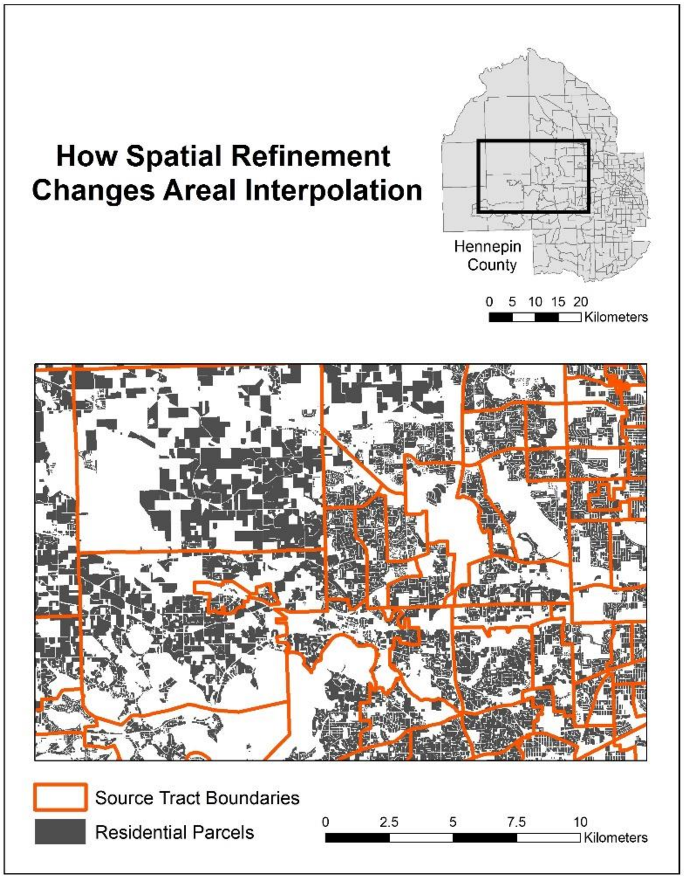

Figure 2. Residential parcel areas are less than tract areas, and this modifies areal interpolation results.

3.2. Target Density Weighting

TDW is based on two assumptions: First, within a source zone, the spatial distribution of the variable of interest Y among atoms is proportionally the same as the distribution of an ancillary variable Z. Second, the density of Z in any atom equals the density of Z in the corresponding target zone (equation 2).

$$\frac{z_{st}}{A_{st}} = \frac{z_t}{A_t} \tag{2}$$

z_{st} and z_t are values of Z for atom st and target zone t, respectively. In this study, Z is the population distribution in 2010, and the variable of interest Y is the population distribution in 2000 to calculate:

$$y_t = \sum_s y_{st} = \sum_s \frac{\frac{A_{st}}{A_t} \times z_t}{\sum_\tau \frac{A_{s\tau}}{A_\tau} \times z_\tau} \times y_s \tag{3}$$

In equation 3, y_t, y_{st}, and y_s are the variable of interest for target zone t, atom st, and source zone s, respectively. τ is a target zone index, iterating through all target tracts intersecting source zone s.

In refined TDW, only residential portions of source and target zones are used. All area elements in the above equations are replaced by residential areas.

3.3. Pycnophylactic Modeling

PM assumes a smooth volume-preserving density function avoiding abrupt changes across adjacent zones. First, the population of source zones is assigned to a grid of cells and adjusted in an iterative process. Once stabilized, cell values are aggregated to target zones. In contrast, refined PM assigns population to each residential cell of a source zone and then creates a smooth surface based on the residential cells within the local neighborhood.

PM and refined PM are run on a grid with 30 meter resolution using a circular neighborhood with a radius of 25 cells (750 meters). The surfaces stabilize after 25 iterations.

4. Results

Tables 1 and 2 show the absolute errors for each interpolation method for all slightly changed tracts, and all significantly changed tracts, respectively. In AW, refinement results in higher MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) measures. However, median absolute errors of refined AW are lower than for unrefined AW, indicating that refined AW produces more accurate estimates for target zones below the median.

TDW shows lower errors than AW, indicating that the assumptions underlying TDW are more realistic than the AW assumption. Refined TDW produces the most accurate estimates overall. Moreover, while refined AW shows inconsistent refinement performance across error estimates, refined TDW shows a consistent refinement effect. No significant changes in error estimates are observed for refined pycnophylactic modeling when compared to the unrefined version.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error
AW	138	36	321	496
Refined AW	148	7	402	568
TDW	101	27	230	236
Refined TDW	70	6	177	180
PM	145	33	326	520
Refined PM	140	28	311	533

Table 1. Absolute error measures of all methods for slightly changed target tracts.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error
AW	298	97	501	712
Refined AW	332	22	629	1014
TDW	207	92	357	537
Refined TDW	138	22	270	539
PM	312	113	508	780
Refined PM	303	114	485	801

Table 2. Absolute error measures of all methods for significantly changed target tracts.

The maps in fig. 3 show the spatial distribution of absolute errors of the methods within the study area. For refined maps, the errors pertain to residential areas, and are visualized over tracts for a better comparison. The maps show that refined AW reduces smaller but not larger errors in comparison to AW. Refined TDW generally shows fewer tracts with medium to high errors than TDW. PM results are mixed and difficult to interpret.

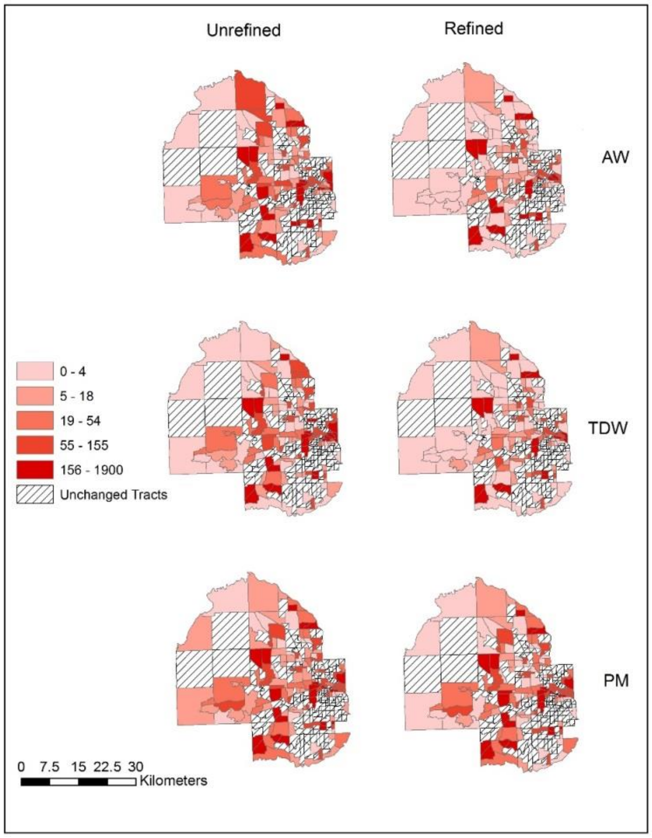

Figure 3. The absolute error (absolute differences between estimated and measured population) maps of the methods at the 2010 tract level.

5. Discussion and Future Research

Refined TDW provides consistent improvement, mainly because the two underlying assumptions of this method result in more robust results for this short period of time. The PM smooth density function may have become less reliable after refinement, but there is potential to improve this method through parameter adjustment and processing. Future research will include residential unit types (e.g., multi-unit or single family) and additional ancillary data to further improve temporal analysis of spatially consistent demographic small area estimates.

6. Acknowledgements

Funded by the U.S. National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, Project BCS-0961598 awarded to University of Colorado Boulder.

7. References

Goodchild M and Lam NSN, 1980, Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, 1:297–312.

Holt JB, Lo CP and Hodler TW, 2004, Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartography and Geographic Information Science*, 31(2):103–121.

Kim H and Yao X, 2010, Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, 31(21):5657–5671.

Leyk S, Nagle NN and Buttenfield BP, 2013, Maximum entropy dasymetric modeling for demographic small area estimation. *Geographical Analysis*, 45(3):285–306.

Leyk S, Ruther M, Buttenfield BP, Nagle NN and Stum AK, 2014, Modeling residential developed land in rural areas: A size-restricted approach using parcel data. *Applied Geography*, 47:33–45.

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

Tobler WR, 1979, Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530.

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