International Conference on GIScience Short Paper Proceedings

Title

Building Consistent Multi-temporal Population Data at Fine Resolution through Spatially Refined Areal Interpolation

Permalink

https://escholarship.org/uc/item/89x537cr

Journal

International Conference on GIScience Short Paper Proceedings, 1(1)

Authors

Zoraghein, Hamidreza
Leyk, Stefan

Publication Date

2016

DOI

10.21433/B31189x537cr

Peer reviewed

Building Consistent Multi-temporal Population Data at Fine Resolution through Spatially Refined Areal Interpolation

H. Zoraghein[1], S. Leyk[1]

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

Abstract

Demographic data are aggregated over areal units to protect privacy and are often inconsistent over time. Areal interpolation methods are used to estimate population in one census year within the units of another year to construct temporally consistent small census units. This research enhances these methods by using three advanced spatial refinement approaches, tested in Mecklenburg County, North Carolina to estimate population in 2000 within census tracts from the 2010 census. The results demonstrate the effectiveness of spatial refinement in reducing estimation errors, systematically. The proposed methods can be used to analyze micro-scale spatio-temporal demographic processes with minimum estimation error.

1. Introduction

Spatial analysis on demographic data aggregated over incompatible boundaries represents a challenge, particularly when the data were collected over historical inconsistent units. To understand micro-scale spatio-temporal demographic processes, data need to be collected over temporally consistent fine-resolution census geographies such as census tracts. However, in reality, their boundaries change over time due to population fluctuations, especially in rapidly growing areas.

Areal interpolation transfers the variable of interest from source zones to target zones and is used in temporal demographic applications (Gregory 2002; Schroeder 2007). In such applications, source populations in one census year (enumerated in source zones) are estimated within enumeration units from the target census completed in another year (target zones).

If the underlying assumptions of areal interpolation methods are not met, accuracy can be very low. Therefore, recent studies have developed spatially refined interpolation techniques with the objective of decreasing population estimation errors (Buttenfield *et al.* 2015; Ruther *et al.* 2015).

Areal interpolation methods are based on population density and area calculations. It can be expected that if these methods are constrained to spatially refined inhabited sub-areas of source and target zones, area and population density estimates will be more precise and realistic. Commonly, the spatially refined sub-areas are delineated using ancillary variables presumably related to population distribution in a dasymetric mapping approach (e.g., Mennis 2003).

This research extends the previous efforts, leveraging three advanced spatial refinement strategies to estimate total population enumerated in census tracts in 2000 within census tract boundaries used in the 2010 census.

2. Study Area and Data

The study area spans Mecklenburg County, North Carolina. It includes both urban areas of Charlotte at its center and large rural areas at its margins and has a history of rapid population growth.

Primary datasets include census tracts and census blocks from the 2000 and 2010 decennial censuses. Residential parcels, NLCD 2001 and 2011 and TIGER/Line data for road networks are used as ancillary datasets for spatial refinement.

3. Methods

3.1 Unrefined Areal Interpolation

Target Density Weighting (TDW) as a versatile areal interpolation method is included in this research and assumes the ratios of population densities of atoms (intersections of source and target zones) to source zones remain the same over time (Schroeder 2007).

3.2 First Spatial Refinement

The first strategy applies TDW to only refined sub-areas of source and target zones delineated by residential parcels as the ancillary variable (Zoraghein *et al.* 2016). The built-year attribute that records when the main structure of a parcel was built is used to match parcels with the census year.

3.3 Second Spatial Refinement

In addition to the geometric footprints of residential parcels, the second refinement uses their housing type to cap or amplify population density within different residential zones. For example, the population density of parcels of type apartments is higher than parcels with single-family residences, and this inherent diversity is addressed in this strategy.

Expectation Maximization (EM) is an iterative statistical optimization technique (Dempster *et al.* 1977), used for the second strategy. All the residential parcels of the same type (e.g., condominium) form control zones, and population density for each zone is estimated through EM. Some control zones include parcels with high variability in area. Thus, assuming one population density value for these zones is unrealistic. Therefore, “Enhanced EM” (EEM) is applied to address this issue.

EEM first identifies the three control zones that represent the highest variability in parcel area measures and the three control zones with the highest number of parcels. Each of these control zones are divided into four homogeneous control sub-zones using a quantile classification scheme for parcel area. For example, instead of using only one single-family residential control zone, four sub-zones of that type are included in EEM. The remaining steps are the same as EM.

3.4. Third Spatial Refinement

The third strategy is not confined to only residential parcels. It leverages additional complementary ancillary variables such as NLCD developed classes (21, 22, and 23) and road network buffer zones (100m buffer distance). The NLCD class selection follows Ruther *et al.* (2015) for delineating refined areas.

This methodology refines initial residential parcels as follows: if a parcel contains instances of the developed NLCD classes, only those instances are used for spatial refinement. However, if no developed land exists, the intersection area of the parcel with road buffers is used to spatially refine the parcel.

This refinement specifically targets rural settings, where large residential parcels overestimate residential areas while NLCD underestimates developed land as a well-known limitation of such databases (e.g., Leyk *et al.* 2014).

3.5. Validation

To derive ground-truth population values for each target zone, block population values in 2000 are aggregated to the target zone boundaries. Accuracy metrics such as Mean Absolute Error (MAE), Median Absolute Error, Root Mean Square Error (RMSE) and 90% percentile of absolute error are calculated based on measured and estimated tract values, and compared across the methods.

4. Results

Table 1 summarizes the results of all three refinement levels.

Table 1. Accuracy metrics of unrefined and refined methods.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
TDW	330	138	531	931	Unrefined
Refined TDW	235	99	379	672	First
Modified Refined TDW	178	75	283	503	Third
EM	447	262	702	1352	Second
Modified EM	236	136	390	611	Third
EEM	192	101	334	498	Second
Modified EEM	152	66	274	382	Third

Figure 1 shows the error maps. Moreover, Table 2 includes the mean normalized absolute errors of the third refinement methods divided by the mean normalized absolute errors of either first or second refinement methods for both total and rural target tracts.

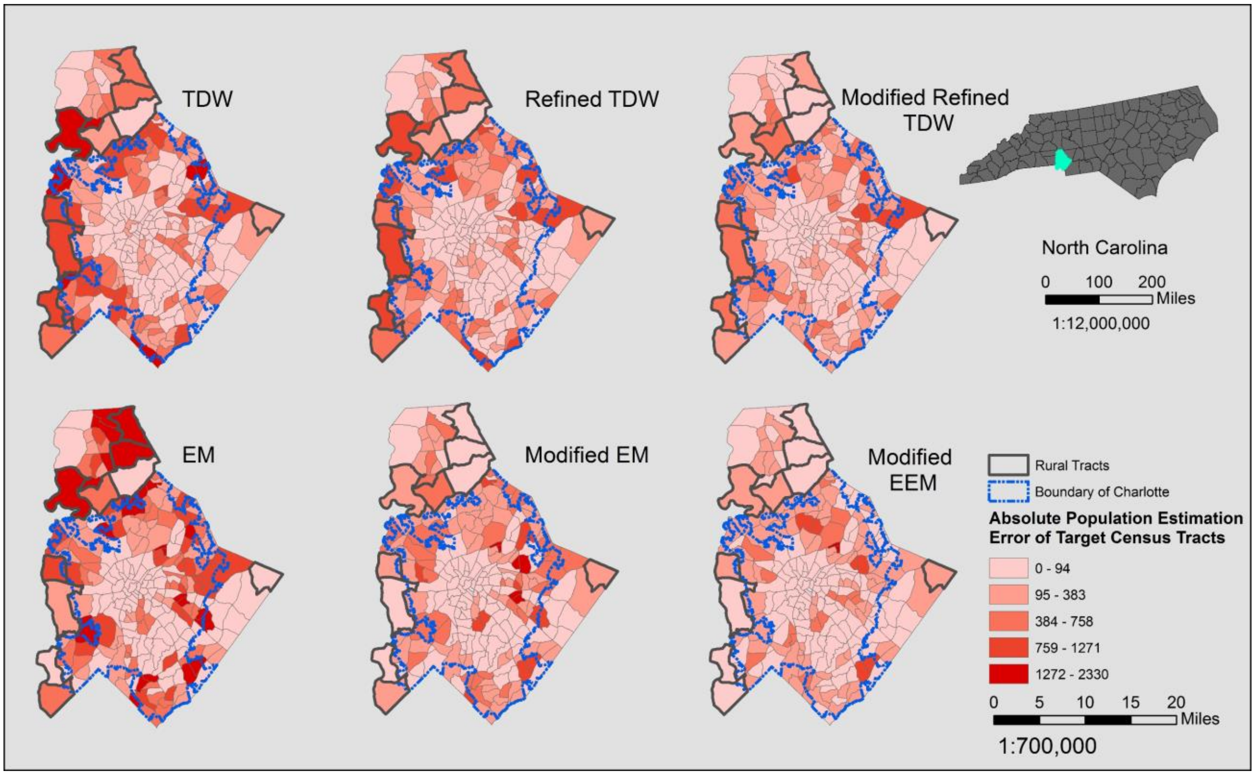

Figure 1. Absolute error maps of the methods.

Table 2. Comparison of the third refinement with first/second refinement.

Method	Total Tracts	Rural Tracts
ModRefAW/RefAW	0.82	0.28
ModRefTDW/RefTDW	1.13	0.53
ModRefPM/RefPM	0.87	0.29
ModEM/EM	0.88	0.22
ModEEM/EEM	1.25	0.38

As a coarse approximation for rural tracts, the number of rural households within each target tract is divided by its total count of households. Each tract with a proportion greater than 0.1 (10%) is considered rural.

5. Discussion and Future Research

Both Table 1 and Figure 1 demonstrate that spatial refinements reduce the error metrics, consistently. Refined TDW is more accurate than the unrefined method, and the third refinement is more accurate than the first. The pattern is similar in Areal Weighting (AW) and Pycnophylactic Modeling (PM) although not included in this paper. The third refinement outperforms EM and EEM as the second refinement methods. The most accurate method is Modified EEM.

As expected, the third refinement results in significant improvements for rural target tracts across all the methods even when it results in less accuracy for total target tracts (Table 2). A value lower than 1 indicates the mean normalized absolute error is lower for the third spatial refinement method than either the first or second spatial refinement approaches.

Future research will focus on data-driven optimization approaches for determining road buffer distance and expand the analyses to longer time periods and different study areas.

References

Buttenfield BP, Ruther M and Leyk S, 2015, Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. *Cartography and Geographic Information Science*, 42(5):449–459.

Dempster A, Laird N and Rubin D, 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Gregory IN, 2002, The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4):293–314.

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.

Mennis J, 2003, Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1): 31–42.

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

Zoraghein H, Leyk S, Ruther M and Buttenfield BP, 2016, Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, 58:19–28.