Spatio-temporal Small Area Analysis for Improved Population Estimation Based on Advanced Dasymetric Refinement

Hamidreza Zoraghein, Stefan Leyk, Barbara Buttenfield and Matt Ruther

ABSTRACT: Demographic datasets are aggregated over areas to protect privacy. To study micro-scale demographic processes, those datasets have to be collected over temporally consistent small areas. However, the availability of such data is limited. That is, the demographic data is either aggregated over large geographical areas (i.e., counties), or collected over small population-derived census units that are temporally inconsistent (i.e., census tracts).

Areal interpolation methods transfer the variable of interest from source zones to target zones. The methods can be used in temporal demographic applications to create temporally consistent population estimates over small areas by transferring population values from the areal units of one census year (i.e., source zones) to the units of another census year (i.e., target zones).

In this research, spatial refinement is incorporated into areal interpolation methods to enhance their population interpolation accuracy. Moreover, one method called Enhanced Expectation Maximization (EEM) is introduced. Areal interpolation methods -- with and without spatial refinement – are used to estimate total population values from census tracts in 1990 to census tract boundaries in 2010 in Mecklenburg County, North Carolina. Based on validation results, EEM is the most accurate method to create temporally consistent population estimates for the 1990-2010 period in the study area.

KEYWORDS: Areal Interpolation, Spatial Refinement, Expectation Maximization, Population, Accuracy

Introduction

Data enumerated over areal units is common in applications such as demographic analyses or health related studies where the protection of privacy is a priority. The spatial or temporal incompatibility of reporting units determined by different sources (agencies, authorities) is a common issue in such applications. For example, demographic data collected over school districts cannot be analyzed in conjunction with those from block groups because the boundaries do not coincide. This issue is aggravated when demographic data collected for different points in time are to be used to characterize fine resolution processes relevant to demographic, economic or health-related changes. In such applications, the analyst requires that the data for the different points in time enumerated within temporally compatible (i.e., identical) fine-resolution units. Unfortunately, data with such characteristics is very limited, in particular for finer resolution units.

U.S. Census data are published at different aggregation levels. Those data that are released for smaller units (e.g., statistical geographies such as census tracts) are often not temporally consistent because their boundaries are sensitive to population changes. In

contrast, coarser resolution units (e.g., legal geographies such as counties) are usually temporally consistent over time but do not allow studying fine scale population processes.

The resulting inconsistency of small census geographies over time impedes the effectiveness of studying fine resolution temporal changes of demographic attributes. Studies to date relied on highly aggregated data, created minimum comparable areas to maintain consistency (e.g., Barufi et al., 2012) or used areal interpolation (e.g., Gregory, 2002). The use of highly aggregated data or minimum comparable areas compromise the granularity of the data, whereas areally interpolated estimates can be error-prone if the underlying assumptions of the utilized areal interpolation methods are not fulfilled.

Areal interpolation transfers the variable of interest from source zones to target zones and represents the default solution often implemented in temporal applications (e.g., Gregory, 2002; Schroeder, 2007; Logan et al., 2014). In such applications, source populations in one census year (enumerated within source zones) are estimated within enumeration boundaries from the target census completed in a different year (target zones). Therefore to preserve granularity, the finest resolution enumerated areas in one census year can be set as the target boundaries, and the demographic attributes of other census years are transferred (i.e., re-allocated) to these target zones.

However, areal interpolation methods rely on assumptions, and if these assumptions are not met, the errors of the resulting demographic estimates can be very high. Therefore, recent studies have begun to explore various advanced approaches to make areal interpolation methods for temporal analysis more robust (e.g., Schroeder and Van Riper, 2013; Logan et al., 2014; Buttenfield et al., 2015; Ruther et al., 2015; Zoraghein et al., 2016). As demonstrated in some of these studies, the incorporation of spatial refinement through dasymetric modeling is a strategy that has led to more accurate population estimates in those research efforts.

Dasymetric modeling is a spatial analytical method that incorporates ancillary data into areal interpolation approaches correlated to the population outcome. It depicts quantitative areal data using boundaries that divide the mapped area into zones of relative homogeneity in population density with the purpose of more accurately portraying the underlying statistical surface (Eicher and Brewer, 2001). Dasymetric modeling relies on two types of ancillary data: limiting and related variables. The former constrains the study area to inhabitable regions and exhibits a binary relationship with the population distribution, while the latter can have more complex relationships with the population distribution to cap or amplify density estimates at a finer spatial resolution (Leyk et al., 2013). Researchers have used various ancillary data – including land cover, road networks and address points – to model the population distribution reliably (e.g., Reibel and Bufalino, 2005; Mennis and Hultgren, 2006; Reibel and Agrawal, 2007; Tapp, 2010).

Areal interpolation methods use population density and area calculations as input. It can be hypothesized that if these methods are applied to spatially refined sub-areas of source and target zones that are actually inhabited (as compared to unrefined enumeration units that can contain uninhabited land), the area and population density calculations will be

more precise and realistic. Thus, areal interpolation coupled with spatial (dasymetric) refinement will most likely lead to more accurate population estimates for target zones. The spatially refined sub-areas are delineated by ancillary variables used for dasymetric modeling.

In this research, *two different spatial refinement strategies* are tested to interpolate population values from census tracts in 1990 (source zones) to the census tract boundaries in 2010 (target zones), with the objective of constructing population estimates over consistent units from 1990 to 2010 in Mecklenburg County, North Carolina. In previous research efforts, the National Land Cover Database (NLCD) has been used successfully as an ancillary variable for spatial refinement (e.g., Reibel and Agrawal, 2007; Buttenfield et al., 2015; Ruther et al., 2015). In this research, residential parcels, which have been found to be promising in such analyses (Zoraghein et al., 2016), are used as the ancillary variable to refine census units.

The *first* spatial refinement strategy identifies only those sub-areas of source and target zones that are delineated as populated based on the geometric footprints of residential parcels, which are thus used as the limiting ancillary variable.

The *second* spatial refinement employs a novel combination of limiting and related ancillary variables for more effective dasymetric refinement prior to temporal interpolation. Different housing types of residential parcels have different population density values. For example, single-family parcels are less populated than condos and this diversity in population density should be incorporated into the spatial refinement step. Thus, the housing type of residential parcels is employed as a related ancillary variable in this research in order to create more accurate depictions of the population distribution.

Background

Areal Interpolation

Areal interpolation can be applied to apportion population estimates from enumeration units for one time period into units created for another time period to achieve temporally consistent enumeration units (Schroeder, 2007; Schroeder and Van Riper, 2013). Several areal interpolation methods have been developed to date, including Areal Weighting (AW) (Goodchild and Lam, 1980; Lam, 1983), Target Count Weighting (TCW) (a term introduced by Schroeder (2007) after the method presented by Howenstine (1993) and Mugglin and Carlin (1998)), Pycnophylactic Modeling (PM) (Tobler, 1979), and Target Density Weighting (TDW) (Schroeder, 2007). These methods are described briefly below.

AW

The AW method estimates the variable of interest in target zone boundaries based on the overlapping area between source and target zones (i.e., intersections or “atoms”). An underlying assumption is that the variable of interest is uniformly distributed within a source zone:

$$y_{st} = (\frac{Area_{st}}{Area_s}) \times y_s$$

where $Area_{st}$ is the area of the atom created by the overlap between source zone s and target zone t, $Area_s$ is the source zone area, y_s is the variable of interest for the source zone and y_{st} is the variable of interest for the atom. The variable of interest for target zone t is then simply derived by aggregating the calculated values of all the atoms within it.

TDW

Schroeder (2007) introduced TDW as an areal interpolation method appropriate for temporal analysis of census data. TDW makes two assumptions. First, within a source zone, the spatial distribution of the variable of interest *Y* among atoms is assumed to be proportionally the same as the distribution of an ancillary variable *Z* (Schroeder, 2007). The second assumption states that the density of *Z* in any atom equals the density of *Z* in the corresponding target zone:

$$\frac{z_{st}}{Area_{st}} = \frac{z_t}{Area_t}$$

where z_{st} and z_t indicate the ancillary variable *Z* for atom *st* and target tract *t*, respectively; and $Area_{st}$ and $Area_t$ are the corresponding areas. The variable of interest for target zone t (y_t) is calculated as follows:

$$y_t = \sum_s y_{st} = \sum_s \frac{(\frac{Area_{st}}{Area_t}) \times z_t}{\sum_\tau (\frac{Area_{s\tau}}{Area_\tau}) \times z_\tau} \times y_s$$

where y_{st} is the variable of interest for atom *st*, and y_s is the variable of interest for source zone *s*. The term τ is a target zone index, independent of *t*, defined for each target zone intersecting source zone *s*. As Equation 3 suggests, y_{st} is calculated based on the proportional distribution of the ancillary variable Z among atoms, and y_t is determined by aggregating all y_{st} values intersecting the target tract.

PM

The PM method assumes the existence of a smooth density function and incorporates the densities of adjacent zones. The density function must be pycnophylactic, i.e., volume-preserving: it must reproduce the original value of a source zone if applied to it (Tobler, 1979).

To interpolate population estimates from source zones to target zones, first a grid is superimposed on the study area. Then, population density per cell is calculated, and cell values per source zone are aggregated and compared to the original value to maintain the pycnophylactic property. This process is iterated until a stopping criteria is fulfilled. Finally, the last cell values are aggregated to target zone boundaries.

Study Area and Data

The study area is Mecklenburg County, North Carolina, which includes both urban areas of Charlotte at the center of the county and large rural areas at its margins. The county exhibited rapid population growth over time (Figure 1).

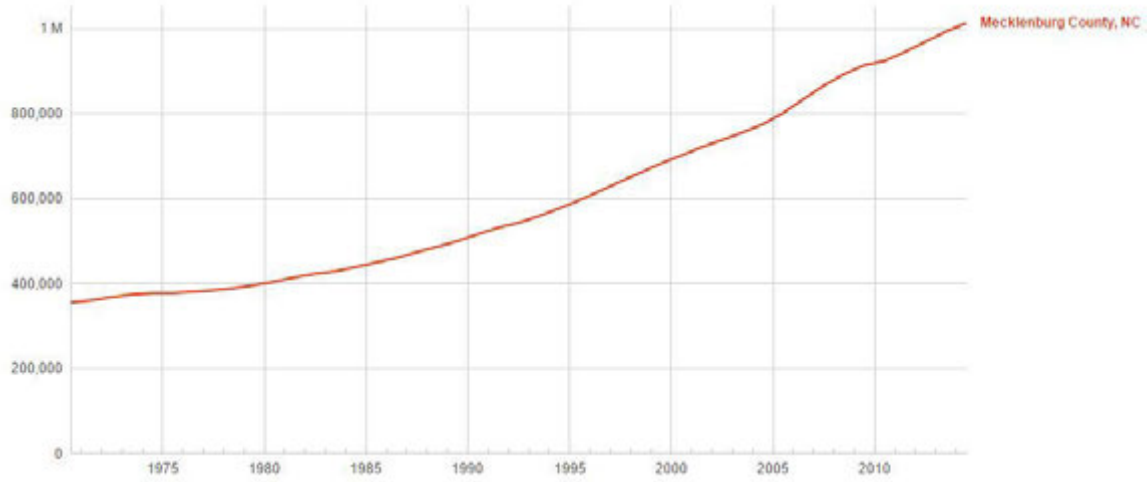

Figure 1: Temporal changes of population in Mecklenburg County (data from U.S. Census)

The required data for this research are divided into the primary and ancillary datasets. The primary datasets include census data. For this research, census tracts (i.e., source and target zones) and census blocks (for validation) of Mecklenburg County from the decennial censuses in 1990 and 2010 were used. The census boundaries and population values for 2010 were downloaded from the TIGER products (https://www.census.gov/cgi-bin/geo/shapefiles2010/main) and American FactFinder (http://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml) data portals as parts of the Census website, respectively. The historic census boundaries and population data for 1990 were accessed through National Historical Geographic Information System (NHGIS) website (https://www.nhgis.org). The ancillary dataset for the study area includes residential parcels and was accessed through the County website.

Area values are one of the main elements in all the methods. Therefore, it is important to have the projection systems of all the datasets transformed to a projection system that preserves area. Thus, USA Contiguous Albers Equal Area Conic was used as the projected coordinate system in the analyses.

Method

Spatially Refine Enumeration Units prior to Areal Interpolation

The first spatial refinement uses residential parcels as a *limiting* ancillary variable. The parcel type attribute is used to determine residential vs. non-residential parcels. The built-year attribute – which records the built year of the main structure within each parcel – is used to delineate residential parcels built before 1990 and those built before 2010. This allows the ancillary data to better align with the two censuses that provide the source and target zones.

The spatial refinement in AW modifies its underlying assumption as follows: population is homogenously distributed within the residential area of a source zone, and no population is assigned to non-residential parts.

Refined TDW uses residential areas within both source and target zones. Thus, all TDW assumptions related to population distribution within source and target zones as well as atoms are applied to the residential land within these zones. Accordingly, all zone areas in TDW equations (i.e., $Area_{st}$ and $Area_t$) are replaced by and refined to residential areas.

Kim and Yao (2010) proposed the spatially refined PM for non-temporal small area estimation. The method creates smooth surfaces dependent on the neighborhood of each cell, and is defined over refined areas, thus allowing more precise depiction of populated areas and neighborhood relations. It uses the same iterative process as the unrefined PM. However, instead of dividing the population of each source zone by the number of all cells within it, the method divides the zone population by the number of all cells comprising the residential part within it. After the pycnophylactic iterative process reaches a stable surface, all cells that have an assigned population count in each target zone are aggregated to compute target zone estimates.

Enhance Refinement by the Incorporation of Housing Characteristics

While the first spatial refinement strategy uses only the geometric footprints of residential parcels as a limiting ancillary variable, the association between population and ancillary variables is not necessarily a binary relationship. Ancillary data can also reduce or amplify the likelihood of the presence of population and the resulting estimates of population density. This type of association is addressed by the related ancillary variable approach in dasymetric refinement. Expectation Maximization (EM) uses an iterative process to optimize population density weights for different conditions defined by the ancillary data, thereby offering an appropriate framework for implementing the second spatial refinement.

The EM algorithm provides a robust framework for model fitting and maximum likelihood estimation in settings of incomplete data. Its name, coined by Dempster et al. (1977) refers to the two steps that comprise each iteration of the algorithm. First, the expectation (E) step "completes" the data by computing the conditional expectation for missing data, given a set of observed data and estimated model parameters. The maximization (M) step then fits the model, estimating model parameters by maximum likelihood given the "complete" data from the E step. A feedback loop between E and M steps is established and repeated until convergence (Schroeder and Van Riper, 2013).

Flowerdew and Green (1994) demonstrated how the EM algorithm can be applied in areal interpolation applications. In this research, EM is used to calculate the population density weight for each control zone. Here, a control zone is defined by all residential parcels that have the same housing type. This is justified by the fact that different housing characteristics can be related to varying population densities, and this variation should be reflected and incorporated.

In the E step, the algorithm estimates the values of $\widehat{y_{sc}}$, i.e., the population counts for the intersections between source zone *s* and control zone *c*.

$$\widehat{y_{sc}} = y_s \left(\frac{\widehat{\lambda_c} A_{sc}}{\sum_k \widehat{\lambda_k} A_{sk}} \right)$$

where y_s is the population count of source zone s, $\widehat{\lambda_c}$ is the estimated density of control zone *c*, A_{sc} is the area of the region of intersection between *s* and *c*, and *k* is a second control zone index, independent of *c* to reflect all control zones intersecting s. The first E step is essentially similar to AW and assumes equal weights for all housing types. Then, the M step re-estimates all λ_c values using the equation below.

$$\widehat{\lambda_c} = \frac{\sum_s \widehat{y_{sc}}}{A_c}$$

The estimates of $\widehat{\lambda_c}$ from the M step are used to estimate $\widehat{y_{sc}}$ in the next E step. This iterative process is carried out until a convergence criterion is fulfilled. Here, the criterion is met when the maximum absolute difference between the density values (i.e., λ_c values) from the last two runs is lower than 0.001. At that point, the final $\widehat{y_{sc}}$ values are used to calculate the population count for target zone *t*.

EM assumes that the population density is constant within one control zone. However, this assumption can become problematic. If the residential parcels of the same type that form a control zone are diverse in area, the assumption of constant population density for the whole control zone is often not realistic. This research introduces Enhanced EM (EEM) as its primary focus to address this issue. EEM first identifies the three control zones that represent the highest variability in the areas of their underlying parcels and the three control zones that have the highest number of parcels. It then categorizes each of the selected control zones (housing types) to four new, more homogeneous control sub-zones based on area quantiles. For example, instead of using only one multi-family type residential control zone in the algorithm, four sub-classes of that type are included in EEM based on the area quartiles. By doing so for all selected housing types, the number of control zones in the study area increases from 12 to 30. The remaining steps of EEM are the same as in EM, described above.

Validation

The validation of the estimated tract-level results is done using census block statistics. After transferring population estimates from source zones to target zones, each 2010 census tract is linked with an estimated population count in 1990. The next step obtains a ground-truth population count for a target zone in 1990. To validate these 1990 estimates

in target zones, population counts of census blocks in 1990 are aggregated to the target zone boundaries. Therefore, for each method two values for each 2010 census tract are derived: the estimated population count based on the utilized interpolation method and the measured population based on census blocks for validation. Different error measures are calculated such as Mean Absolute Error (MAE), median absolute error, Root Mean Square Error (RMSE) and 90% percentile of absolute error. These error measures can be compared across methods and expand understanding of the estimation error distribution in different aspects, leading to a more comprehensive comparative analysis of 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 error can be used to describe the upper end of the error distribution and placement of extreme absolute error values.

Results

Table 1 summarizes four absolute error measures for population estimates in 1990 within 2010 target zone boundaries. Refined methods are applied only to developed areas of source and target zones delineated by residential parcels. In Table 1, the first spatial refinement methods are those preceded by "Refined". EM and Enhanced EM (EEM) represent the results derived for the second spatial refinement.

Figure 2 shows the absolute error maps of the methods. PM maps are not included because the accuracy level of PM ranges between those of AW and TDW. Therefore, AW represents a better contrast with the more accurate methods and is chosen as the benchmark method to show the improvement effect.

According to Table 1, the effect of the first spatial refinement is consistent for all the methods, i.e., all four error measures are lower in the refined implementations of AW, TDW and PM. This confirms that the application of areal interpolation methods to only residential sub-areas of source and target zones leads to more accurate population estimates for target zones. Refined TDW results in the most accurate estimates among the methods that use the first spatial refinement. Figure 2 also depicts the superiority of Refined TDW over unrefined and other first spatial refinement methods.

According to both Table 1 and Figure 2, EEM – as the suggested method of this research – is the most accurate method, indicating its great potential for temporal areal interpolation of population estimates. This demonstrates the effectiveness of utilizing the related ancillary data in the form of housing types of residential parcels in decreasing the absolute error measures of population estimation. EEM reduces the four measures by 64%, 68%, 60% and 64%, respectively, relative to AW as the benchmark method and by 20%, 24%, 14% and 6%, respectively, relative to Refined TDW as the second best performing method.

Moreover, Figure 2 depicts that the effect of spatial refinement in absolute error reduction is not constrained to only either urban target tracts or rural tracts. Rather, it shows that the

improvement effect can be observed for both the target tracts within the boundary of Charlotte and those outside.

Table 1: Absolute error measures of different areal interpolation methods for 1990 (source zones) to 2010 (target zones).

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error
AW	546	346	832	1477
Refined AW	326	175	530	916
TDW	387	255	575	829
Refined TDW	247	146	382	558
PM	526	263	824	1395
Refined PM	332	168	559	899
EM	432	207	712	1051
EEM	197	111	329	524

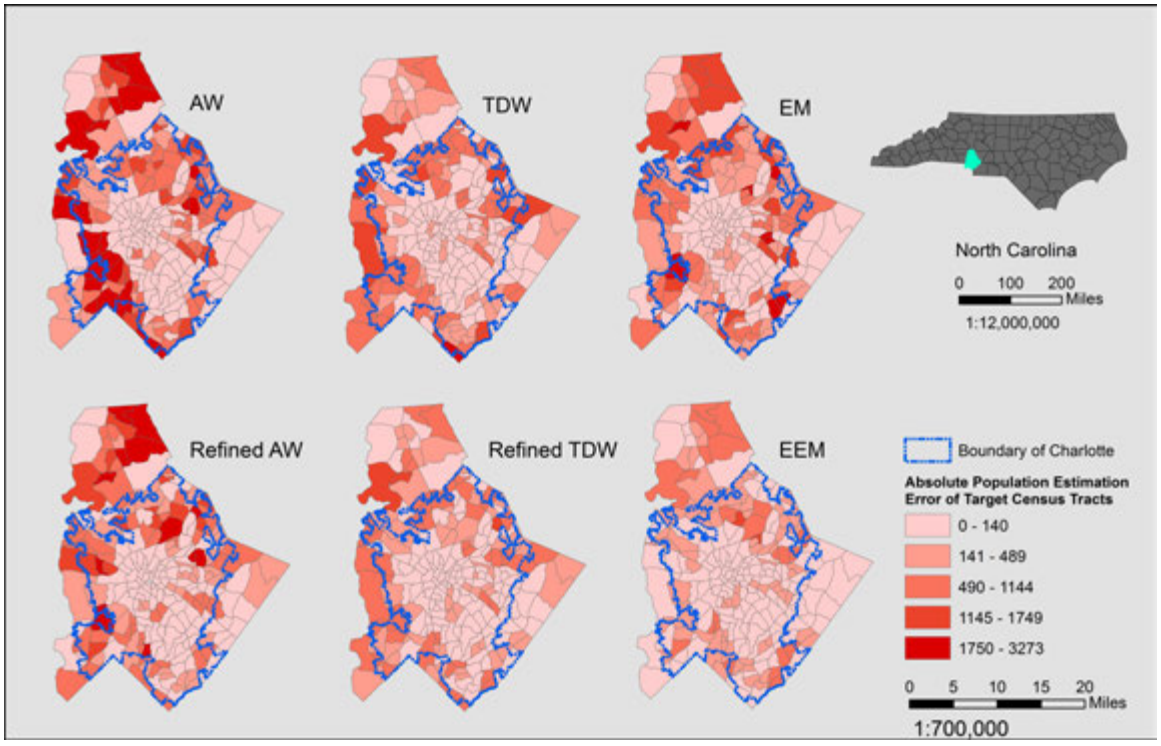

Figure 2: Absolute error maps of the areal interpolation methods

Discussion

Among the methods using the first spatial refinement strategy, Refined AW and Refined PM use refined areas from only 1990, whereas Refined TDW uses refined areas from both 1990 and 2010 by design. Therefore, Refined TDW seems to leverage the changes in developed areas over time in an effective way, and consequently demonstrates a more pronounced improvement effect of the spatial refinement.

The baseline EM method inherently uses ancillary data and employs the different categories as related ancillary information. However, the error measures of EM are rather high, labeling it the least accurate method after AW and PM. One main reason for this observation could be the high area variation in the underlying residential parcels forming one control zone. EEM applies the EM algorithm on a set of smaller, presumably more homogeneous set of control zones. The remarkable improvement effect is observed from Table 1 and in comparing the maps in Figure 2.

Future research will improve EEM and make it more robust by developing an objective strategy for choosing the control zones that should be further categorized, expand the analyses to more study areas, demographic attributes and time periods, and make a detailed comparison between the results of spatial refinement for rural and urban target zones.

References

Barufi, A. M. Haddad, E. and Paez, A. (2012) Infant mortality in Brazil, 1980-2000: A spatial panel data analysis. *BMC Public Health*, 12, 1, pp. 181–195.

Buttenfield, B. P. Ruther, M. and Leyk, S. (2015) Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. *Cartography and Geographic Information Science*, 42, 5, pp. 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, pp. 1–38.

Eicher, C. L. and Brewer, C. A. (2001) Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28, 2, pp. 125–138.

Flowerdew, R. and Green, M. (1994) Areal interpolation and types of data. In S. Fotheringham and P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 121–145).

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

Gregory, I. N. (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, pp. 293–314.

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

Lam, N. S. N. (1983) Spatial interpolation methods: a review. *The American Cartographer*, 10, 2, pp. 129–150.

Leyk, S. Buttenfield, B. P. Nagle, N. N. and Stum, A. K. (2013) Establishing relationships between parcel data and land cover for demographic small area estimation. *Cartography and Geographic Information Science*, 40, 4, pp. 305–315.

Logan, J. R. Xu, Z. and Stults, B. J. (2014) Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database. *The Professional Geographer*, 66, 3, pp. 412–420.

Mennis, J. and Hultgren, T. (2006) Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science*, 33, 3, pp. 179–194.

Mugglin, A. S. and Carlin, B. P. (1998) Hierarchical Modeling in Geographic Information Systems: Population Interpolation over Incompatible Zones. *Journal of Agricultural, Biological, and Environmental Statistics*, 3, 2, pp. 111–130.

Reibel, M. and Agrawal, A. (2007) Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, 26, 5-6, pp. 619–633.

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

Ruther, M. Leyk, S. and 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, pp. 158–178.

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

Schroeder, J. P. and 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, pp. 216–237.

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

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

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

Hamidreza Zoraghein, Department of Geography, University of Colorado at Boulder, Boulder, CO 80309

Stefan Leyk, Department of Geography, University of Colorado at Boulder, Boulder, CO 80309

Barbara Buttenfield, Department of Geography, University of Colorado at Boulder, Boulder, CO 80309

Matt Ruther, Department of Urban and Public Affairs, University of Louisville, Louisville, KY 40208