Computers, Environment and Urban Systems 54 (2015) 23–31

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A local polycategorical approach to areal interpolation

Jie Lin [a,*], Robert G. Cromley [b,1]

[a] School of Earth Sciences, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang 310027, PR China
[b] Department of Geography, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 2 December 2014
Received in revised form 28 May 2015
Accepted 29 May 2015
Available online 14 June 2015

Keywords:
Areal interpolation
Local statistical
Geographically weighted regression
Quantile regression

ABSTRACT

Areal interpolation is a technique used to transfer attribute information from source zones with known values to target zones with unknown values. This paper presents and describes a new polycategorical method that integrates positive aspects of both geographically weighted regression (GWR)-based and quantile regression (QR)-based interpolators for solving areal interpolation problems. Two different types of neighborhoods for selecting observations used to estimate ancillary control densities are presented: one that is spatially based and one that is statistically based. The new polycategorical methods are evaluated against a number of existing methods – areal weighting, pycnophylactic, binary dasymetric, intelligent dasymetric mapping, and GWR using test data from the 2010 census population, the National Land Cover Database 2006 (NLCD2006) and the Topologically Integrated Geographic Encoding and Reference (TIGER) line graph files. The evaluations include several overall error measurement indices as well as maps of the spatial distribution of the error associated with selected methods. Results suggest that with appropriate land cover categories and neighborhoods, the new polycategorical methods provide comparable results to local regression models but with much less computation complexity.

1. Introduction

Often, spatial data are aggregated into areal units for further analysis, even though the data may be collected at the individual level. This aggregation occurs for several reasons (Openshaw & Taylor, 1981): (1) spatial data concerning personal information are restricted by privacy and confidentiality; (2) data in the aggregated form are convenient and require less volume for storage, and also have a computational advantage over data in a disaggregated form; and (3) geography has a long tradition of studying data at the regional level. However, different applications, variations in geographic scale, and the distinct nature of a phenomenon's distribution have worked against the unification of areal units into a single standardized system (Visvalingam, 1991). It is also well known that results of spatial analyses are sensitive to the choice of zoning system associated with the aggregation, which is known as the modifiable areal unit problem (MAUP) (Openshaw & Taylor, 1981).

Related to these issues is the change of support problem in which data are collected for one measurement support but must be transferred to a different support system before analysis is performed. This change of support requires values to be estimated at locations different from those at which the data have been observed (Gelfand, Zhu, & Carlin, 2001). For data collected and reported in areal units, it means that values must be estimated for an alternative zoning system. Areal interpolation (AI) is a procedure for transferring attribute values from one partition of geographic space (a set of source regions) to a different partition (a set of target regions) (Goodchild & Lam, 1980; Lam, 1983). The development of Geographic Information Systems (GIS) has also increased the necessity for AI as a GIS analysis frequently generates new layers of areal units for which non-spatial attribute information must be estimated.

A number of different AI methods have been developed over time to improve the efficiency and accuracy of the transfer procedures. Local models, in which the relationship between an attribute variable and ancillary information is estimated using a selected subset of observations from the full set, are now fairly common in AI procedures as they address the heterogeneity problem with respect to any relationship. By close scrutiny of two statistical interpolators that emphasize local variation, a geographically weighted regression (GWR)-based (Lin, Cromley, & Zhang, 2011) and a quantile regression (QR)-based interpolator (Cromley, Hanink, & Bentley, 2012), this study proposes a new local polycategorical AI procedure that integrates the positive aspects of both aforementioned regressions. QR is a regression with varying parameter estimates like GWR, but these regression models differ in two respects: (1) QR minimizes the sum of absolute deviations whereas GWR minimizes the sum of squared deviations, and (2) QR estimates are a function of a position in the statistical distribution (the quantile

[*] Corresponding author. Tel.: +86 571 87952453.
E-mail addresses: jielin@zju.edu.cn (J. Lin), robert.cromley@uconn.edu (R.G. Cromley).
[1] Tel.: +1 8604862059.

http://dx.doi.org/10.1016/j.compenvurbsys.2015.05.007
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

level) rather than a position in geographic space. The term 'local' used in this paper mainly stresses locally-varying model parameter estimates, which is consistent with Fotheringham's (1997) discussion, but not the selection of a subset of observations used in model calculation.

The rest of this paper is organized as follows: related work is provided in Section 2. Section 3 explains the proposed polycategorical method. The data and comparative methods are presented in Section 4. Section 5 presents and compares the results of different AI methods in conjunction with various control data. Finally, conclusions and future work are discussed in Section 6.

2. Background

AI methods are used to estimate attribute data that are associated with a density surface. The accuracy of the AI method then is a function of how accurately the underlying density surface can be approximated. The assumptions made by the AI method regarding the nature of the density surface within the source and target spatial units have a major impact on model performance. AI methods are sometimes dichotomized as being either *simple* or *intelligent*. Simple AI methods, transfer data from source zones to target zones without using any ancillary data whereas intelligent areal interpolation methods use some form of ancillary data that provide insight to the underlying density surface in order to improve estimation accuracy (Langford, 2006; Langford, 2007; Langford, Maguire, & Unwin, 1991). Ancillary data then are used to infer the internal structure of the density surface within each source zone. For estimating population, two readily available ancillary datasets that are commonly used are remotely sensed land cover data and road networks. In the US land cover data are easily accessible from the National Land Cover Database (NLCD) and road networks can be obtained from the U.S. Census Bureau's 2010 TIGER/Line files. Another ancillary data used more recently in the areal interpolation of population is parcel data. Maantay, Maroko, and Herrmann (2007) developed the so-called cadastral-based expert dasymetric system to redistribute population in urban area, while Tapp (2010) used county address points and parcels to estimate population in rural and transitional areas. Although the assumption in these methods is very straightforward, cadastral data are not as readily available as road networks or land cover data. Moreover, due to a rapid expansion in the use of Web 2.0 applications, numerous forms of volunteered geographic information (VGI) are being produced by general public. These open access geographic data together with other traditional ancillary data can also be used for the areal interpolation of population. Bakillah, Liang, Mobasheri, Jokar Arsanjani, and Zipf (2014) proposed a framework using OpenStreetMap points-of-interest and pre-classified land use land cover categories to infer population at a building level. Lin and Cromley (2015) evaluated geo-located nighttime tweets collected from Twitter.com both as single control data and as an enhancement to other control data for areal interpolation of population as well as different age-specific population groups.

2.1. Areal weighting

Areal weighting, the easiest AI method, assumes that the density surface is uniform within each source polygon (the overall surface is a 3D prism). It calculates the geometric overlay of the source and target zones, and values are estimated for each intersection zone by proportionally weighting the data counts by the area of each intersection polygon with respect to the area of the source polygon that contains it (Goodchild & Lam, 1980). It is the most widely used method among all other methods due to its intuitively simple theory, low data and computation requirement, and it can be easily included in GIS software (Xie, 1995). However, areal weighting usually produces poorer estimates when compared against results of other methods because of its simplistic representation of the density surface and lack of any ancillary data.

2.2. Pycnophylactic interpolation

Tobler's pycnophylactic method (1979), another simple AI method, creates a smooth density surface of raster units from an initial 3D prism surface associated with the source zones. Its original purpose was to construct isopleth maps. Tobler noted that the method could also be used as an interpolation technique for transferring data from one support to another by aggregating the raster values of the surface by each target zone that contains them. The regular grid has also been modeled as an irregular triangular network (TIN) surface in Rase's (2001) pycnophylactic interpolation procedure. Using this method as an AI interpolator in GIS is more difficult because it requires a pycnophylactic surface interpolator, and vector-to-raster and raster-to vector operations as intermediary steps.

2.3. Binary dasymetric method

A direct extension of areal weighting that uses ancillary data and is easy to implement in a GIS is the binary dasymetric method (Fisher & Langford, 1996). For estimating population, the ancillary information is classified into areas containing population and areas that do not. For land cover data, this only requires that land cover categories are reclassified into either a populated or an unpopulated category. For road networks, the length of the road network within each source could replace the area of the source zone if one assumes that the population is located along the road itself (Xie, 1995), or a buffer zone can be created around the road network that would contain the population (Mrozinski & Cromley, 1999). The areas within each source zone that do not contain population are "erased" from the total area of the source zone so that only areas that contain population are retained. The density surface is viewed as a dichotomous prism within each source zone. For binary dasymetric AI, areal weighting is then used to estimate target zone populations from the source zones based on populated areas. Thus binary dasymetric AI can also be termed an areal weighting of populated areas.

2.4. Dasymetric mapping-based polycategorical methods

A finer granularity of the density surface can be estimated by polycategorical AI methods. For example, instead of grouping land cover categories into populated and unpopulated, these categories can be grouped into different levels of population densities. Polycategorical methods either follow the principles of dasymetric mapping (Wright, 1936) or some statistical method. Eicher and Brewer (2001) implemented two polycategorical approaches based on Wright's dasymetric method but they used a totally subjective scheme to estimate the population densities within the different categories. The limiting variable method includes more land cover groups and produces significantly better results than other methods, while the three-class method produces only slightly better results than the binary dasymetric approach.

In contrast, Mennis and Hultgren (2006) developed the intelligent dasymetric mapping (IDM) method to redistribute population between different land cover types in a more objective manner. The method first overlays the layer of source zones against the layer of ancillary zones. Next, the density of each ancillary class is estimated by first associating source zones with a specific ancillary class with respect to one of three criteria: (1) the source zone is completely contained within the ancillary class, (2) the centroid of the source zone lies within the ancillary class, or (3) at least a specific percentage of the source zone is within the ancillary class. The density of a given ancillary class then is the ratio of sum of the total population of all source zones associated with a class divided by the total area of all such source zones. In this version, the density associated with any ancillary class is constant over the study area. To model spatial variation in the relationship between density and an ancillary class, Mennis and Hultgren (2006) incorporated an additional data set of region zones and estimated density for each ancillary class separately for each individual region.

2.5. Statistically-based polycategorical methods

Most statistical approaches to polycategorical AI estimate the density value of each ancillary class within some region is based on some form of regression analysis. Langford et al. (1991) proposed several regression models to describe the population as some function of the pixel counts for each land use class. Two logical flaws were pointed out by the authors after examining the results of the ordinary least square (OLS) models: (1) even if there is no residential land use present, population still can exist when the estimated intercept in the regression is not zero, and (2) a negative population estimation is possible if the density estimate for a particular ancillary class is negative. In order to address these logical dilemmas, they simplified their regression model by reducing the independent variables and forcing the intercept to zero at the cost of the interpolation accuracy.

Flowerdew and Green (1989) noted that Poisson regression is theoretically preferable for modeling counts, as negative population estimation could be avoided by using Poisson regression. In order to use existing target zone information to make areal interpolation more effective, Flowerdew and Green (1991) improved their Poisson model by synthesizing the Expectation and Maximum-likelihood (EM) algorithm, which was originally developed by Dempster, Laird, and Rubin (1977) to solve problems of missing data. Flowerdew and Green (1994) also extended the EM method to deal with continuous variables having a normal distribution.

In these regression approaches there is usually a residual associated with each observation; a non-zero residual means that the volume-preserving property is not maintained in the estimated density surface. In this case, the initial estimated surface is scaled by multiplying it by the ratio of each source zone's observed population to that source zone's estimated population. This ensures that the pycnophylactic property exists for all source zones before the final step that reassigns the adjusted densities to the target zones. Such a scaling step is used by Reibel and Agrawal (2007) in conjunction with regression-based AI methods and the National Land Cover Dataset (NLCD) to interpolate population in eastern Los Angeles County, California.

However, the early regression models generally underperformed when compared to areal-based dasymetric models (Fisher & Langford, 1995). Population density varied more between different places than between different ancillary classes. This is a problem of the global nature of regression analysis. In order to make the relationship between population and ancillary class in regression model not only vary across ancillary classes but also over geographic space, Yuan, Smith, and Limp (1997) developed a regional model that regressed population against land cover types in each county of their study area. This approach is similar in nature to Mennis and Hultgren's (2006) use of region zones to increase the level of spatial variation within an ancillary class. The model fitness in three of the four counties outperformed the globe regression model based on *R*-square values. However, Langford (2006) argues that the counties used by Yuan et al. are, after all, administrative areas, which have no basis for determining the underlying distribution of population. This is a problem of the *a priori* delineation of regions. Furthermore, the locally-varying model parameters between counties indicate this variation in the relationship between population density and land cover type is likely to be continuous, because variation is likely within regions if it exists between regions (Fotheringham, Brunsdon, & Charlton, 2002).

In order to model the continuous spatial variation in the relationship between density and ancillary classes, Lo (2008) used GWR (Fotheringham et al., 2002) to estimate population in the city of Atlanta, Georgia. In his empirical test, the local GWR model outperformed a global OLS model based on error measurements derived from residuals in source tracts. Later, Lin et al. (2011) implemented GWR in an AI framework. GWR AI outperformed binary dasymetric and global regression interpolators in their test trials. They also pointed out that choices of bandwidths and scales of target zones have a significant impact on the performances of GWR-based AI models, while the locations of estimation points did not matter as much as the former two factors.

Cromley et al. (2012) have also used quantile regression (Koenker & Bassett, 1978) as the basis for estimating the density surface for ancillary classes. By associating every observation (source zone) with a quantile level, density values for each ancillary class within the observation can be estimated. Also, the residual associated with the observation is zero which means that QR-based areal interpolation model is inherently volume-preserving. The major drawback to the implementation of a QR interpolator is the time and computation necessary to associate observations with specific quantile levels. Whereas GWR estimates are directly associated with a given location or observation, more time is needed to make the association in QR with observations than to perform the quantile regression itself. However, aspects of the QR approach are used in the next section to develop local polycategorical interpolators for which estimates can be easily associated with observations.

3. Local polycategorical areal interpolation

When using geographically weighted regression and quantile regression as the basis for areal interpolation, a series of regression models are run to determine the density value for each ancillary category within each source zone. In these models, the relationship between population and the density estimate for each ancillary class is expressed as the following linear function:

$$Y_i = \sum_j \beta_{ij} X_{ij} + \varepsilon_i \tag{1}$$

where

Y_i = population for the *i*th source zone;

β_{ij} = the density estimate for the *j*th ancillary category for the *i*th source zone;

X_{ij} = the area (or number of pixels) of the *j*th ancillary category for the *i*th source zone;

ε_i = residual for the *i*th source zone.

In GWR, each observation is weighted by a kernel density function having a given bandwidth that is centered on the source zone for which the density values are estimated. For QR, the residuals in each regression are weighted by the parameters ρ and $(1-\rho)$ in which ρ is the quantile level that ranges between zero and one; ρ is the weight associated with a negative residual and $(1-\rho)$ is associated with a positive residual. The linear equation has no intercept term so that no population will exist in the absence of an ancillary category.

If ε_i equals zero and all β_{ij} are positive, then the method is volume preserving and does not need a scaling step. These conditions are easy to ensure in QR but not in GWR and so a scaling step must be used in the latter instance. On the other hand it is much easier in GWR to associate local density estimates with the source zone *i* on which the weights are centered. Instead of the association being known *a priori* as in GWR, the association between density estimates and the source zone *i* is determined *a posteriori* for QR. The coefficients of a quantile-regression are assigned *manually* as density values for each ancillary category of the source zone for which the regression line is a perfect fit (see Cromley, Hanink, and Lin (2013) for the details of this procedure). This is a barrier for implementing the QR approach in a software system unless an *a priori* association between local density estimates and source zone can be established.

The following local polycategorical methods integrate the positive aspects of both forms of regression for ease of implementation. In regression analysis, density coefficients are found by solving a system of linear equations. In OLS regression and GWR, this is the system of normal equations. In QR, it is a system of linear equations in a mathematical programming model. To solve for n unknowns at least n equations are necessary. If n ancillary categories are included in the areal interpolation procedure, then at least n source zones are needed to build a system of linear equations. Let one of these source zones be the unit for which the density coefficients are being estimated and let the remaining source zones be units in the "neighborhood" of that source zone. Two different kinds of "neighborhoods" are used here to choose the remaining units: one based on geographic proximity and one based on statistical proximity (see Fig. 1). The former mirrors the logic of GWR-based interpolation and the latter mirrors the logic of QR-based interpolation. For geographic proximity, the m source zones nearest in distance to the center source zone are selected (where $m \geqslant n - 1$). For statistical proximity, the source zones are first sorted by their overall population density and the m source zones closest in rank order to the center source zone are selected; this simulates the notion of a ρ quantile level because quantiles are rank ordered.

A second model difference is based on the volume preserving property. As discussed, the QR model ensures the volume preserving property by structuring the model such that all coefficients are non-negative and there is a zero residual for a source zone associated with one quantile-regression. For the GWR model, the volume preserving property is ensured by a scaling step that is applied after the initial population estimates are calculated. Following the QR approach, the density coefficients for the ancillary classes associated with the center source zone is estimated by the following linear program:

$$\text{Minimize } M\,\lambda_c^- + M\,\lambda_c^+ + \sum_i (\lambda_i^- + \lambda_i^+) \tag{2}$$

$$\sum_j^n \beta_j X_{cj} + \lambda_c^- + \lambda_c^+ = P_c \tag{3}$$

$$\sum_j^n \beta_j X_{ij} + \lambda_i^- + \lambda_i^+ = P_i \quad \text{for all } i \in N_c \tag{4}$$

$$\beta_j, \lambda_c^-, \lambda_c^+, \lambda_i^-, \lambda_i^+ \geq 0 \tag{5}$$

where

M is a very large positive number;

λ_c^- is a deviational variable representing the amount of underestimation for the center source zone;

$\lambda_c{}^+$ is a deviational variable representing the amount of overestimation for the center observation;

β_j is the estimated density value for the *jth* ancillary class;

X_{cj} is the pixel count of *jth* ancillary class in the center source zone or the area of that ancillary class;

P_c is the population count for the center source zone;

X_{ij} is the pixel count of *jth* ancillary class in the *ith* neighboring source zone or the area of that ancillary class;

λ^- is the deviational variable representing the amount of underestimation for the *ith* observation by the linear equation;

λ_i^+ is the deviational variable representing the amount of overestimation for the *ith* neighboring observation;

P_i is the population count for the *ith* neighboring observation;

n is the number of ancillary classes;

N_c is the set of source zones in the center source zone's neighborhood.

The objective function minimizes the sum of the deviational variables so that the best fit of the density coefficients can be found with respect to the source zone information. The M parameter in the objective function ensures that the deviation variables associated with the center source zone will be equal to zero, and consequently the population prediction for the center source zone will be exact. This condition plus the fact that the density coefficients are all non-negative ensures that the volume-preserving property will be maintained. The proposed local polycategorical models based on geographic and statistical proximity filtering windows are tested against one another and previously developed AI models in the following sections.

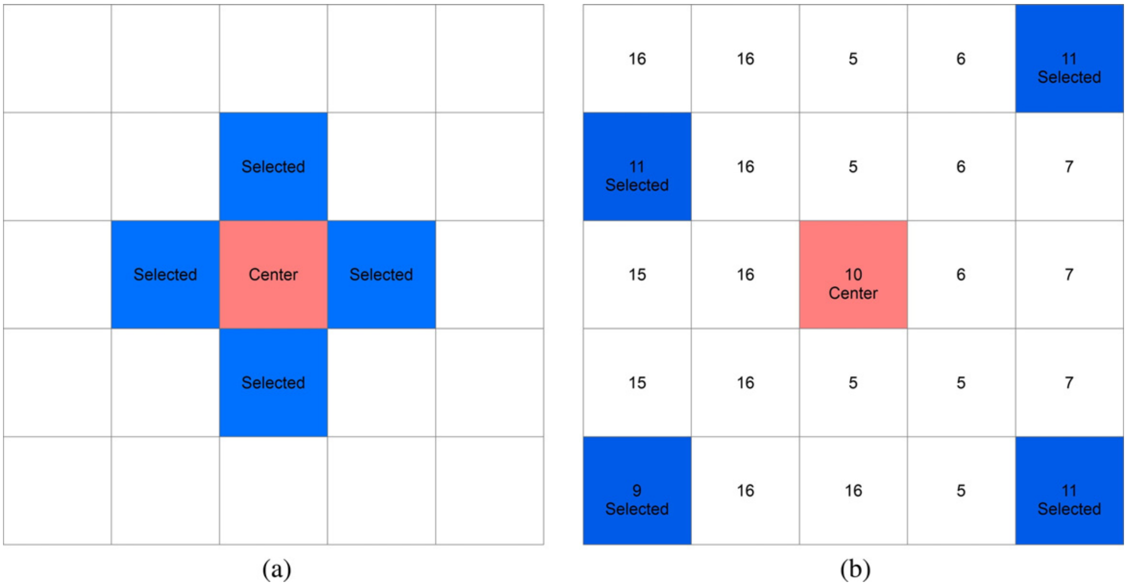

Fig. 1. Geographic (a) and statistical (b) proximities based neighborhoods for proposed polycategorical areal interpolation.

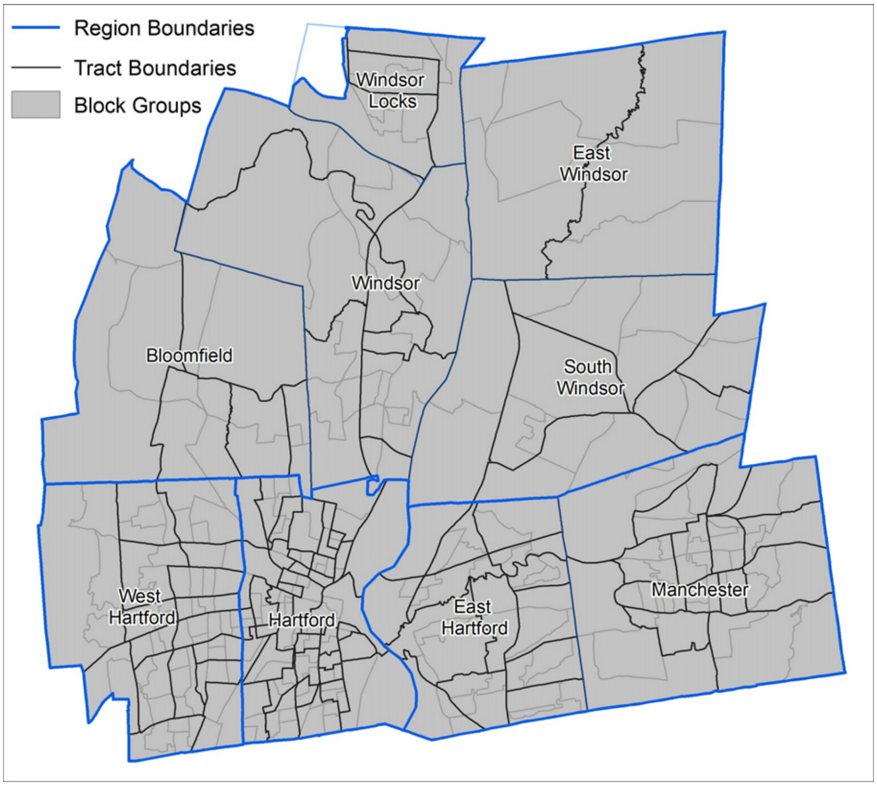

Fig. 2. The nine-town study area overlaid with the source and target zones.

4. Data and research design

In this research, the study area consists of nine towns in Hartford County, Connecticut. In 2010 census, the study area had a total population of 396,435. The study area is characterized by various types of land uses, including residential areas with different population densities, commercial and industrial, and open space. The census tracts and census block groups are used as source zones and target zones respectively. Shapefiles for these units were downloaded from the U.S. Census Bureau's 2010 Topologically Integrated Geographic Encoding and Reference (TIGER/Line) Shapefiles Main Page (https://www.census.gov/cgi-bin/geo/shapefiles2010/main). Fig. 2 shows the study area overlaid with the target and source zones. The population data were extracted from 2010 Census Summary File 1 (SF1) Table P1, and were downloaded from U.S. Census Bureau's web data retrieve interface American FactFinder (http://factfinder2.census.gov). The geography boundary files and the demographic data were then joined together in ArcGIS 10.2.

The pre-classified land use/land cover (LULC) dataset used as the ancillary data in this study is the National Land Cover Database 2006 (NLCD2006), which is publicly available from the Multi-Resolution Land Characteristics Consortium (MRLC) (http://www.mrlc.gov/), derived by Fry et al. (2011) based primarily on the unsupervised classification of Landsat 7 Enhanced Thematic Mapper+(ETM+) satellite imagery of 2006 with a 30 × 30 m spatial resolution. The initial pre-classified NLCD2006 dataset was reclassified based on the similarities of LULC types and possible population densities, and then four categories were selected for the areal interpolation analysis: (1) developed, low intensity (DL); (2) developed, medium intensity (DM); (3) developed, high intensity (DH); and, (4) deciduous, evergreen, and mixed forest (FOR). Fig. 3a shows the distribution of the four land cover classes. These four categories were chosen as populated areas because they have the highest correlation with population distribution in descending order from DL to FOR.

A second ancillary dataset was compiled based on road networks, which were downloaded from U.S. Census Bureau's 2010 TIGER/Line Shapefiles Main Page (https://www.census.gov/cgi-bin/geo/shapefiles2010/main). A 100 foot buffer was constructed around on each side of secondary roads (MTFCC S1200), local neighborhood roads, rural roads, and city streets (MTFCC S1400). Certain road features such as primary highways, highway ramps, and parking lot roads in the original data were not used. The road buffer was used to create a binary spatial control – the area inside the buffer is considered populated and the area outside the buffer is considered unpopulated.

There are six types of areal interpolation models included in this study: areal weighting (AW), pycnophylactic (PYCNO), binary dasymetric (BDAS), intelligent dasymetric mapping (IDM), geographically weighted regression (GWR), and the proposed polycategorical method (PC) described in the previous section. AW and PYCNO have the lowest data requirements as they do not use any ancillary data. BDAS also is not polycategorical but is included because of its low computational requirements. IDM calculated population density for each ancillary class based on the source zones within which that class appears most frequently by regions, in which the ancillary class density also varies by a predefined region – in this case four regions (see Fig. 2 for the delineation of these regions). The regions were based on differing levels of population density. The GWR interpolator, serves as a representative of local regression interpolators, used a Gaussian kernel density function and an adaptive bandwidth that minimized the cross-validation score. AW, BDAS, and IDM were performed in ArcGIS 10.2. PYCNO was implemented by a Python script using the Arcpy module under the ArcGIS 10.2 environment (this script is available from the corresponding author upon request). GWR was executed by the spgwr package (Bivand & Yu, 2014) under the R environment, the estimated coefficients were then imported into ArcGIS 10.2 for the remaining interpolation steps.

The local polycategorical model had two sub-types based on defining the center source zone neighborhood as either based on distance

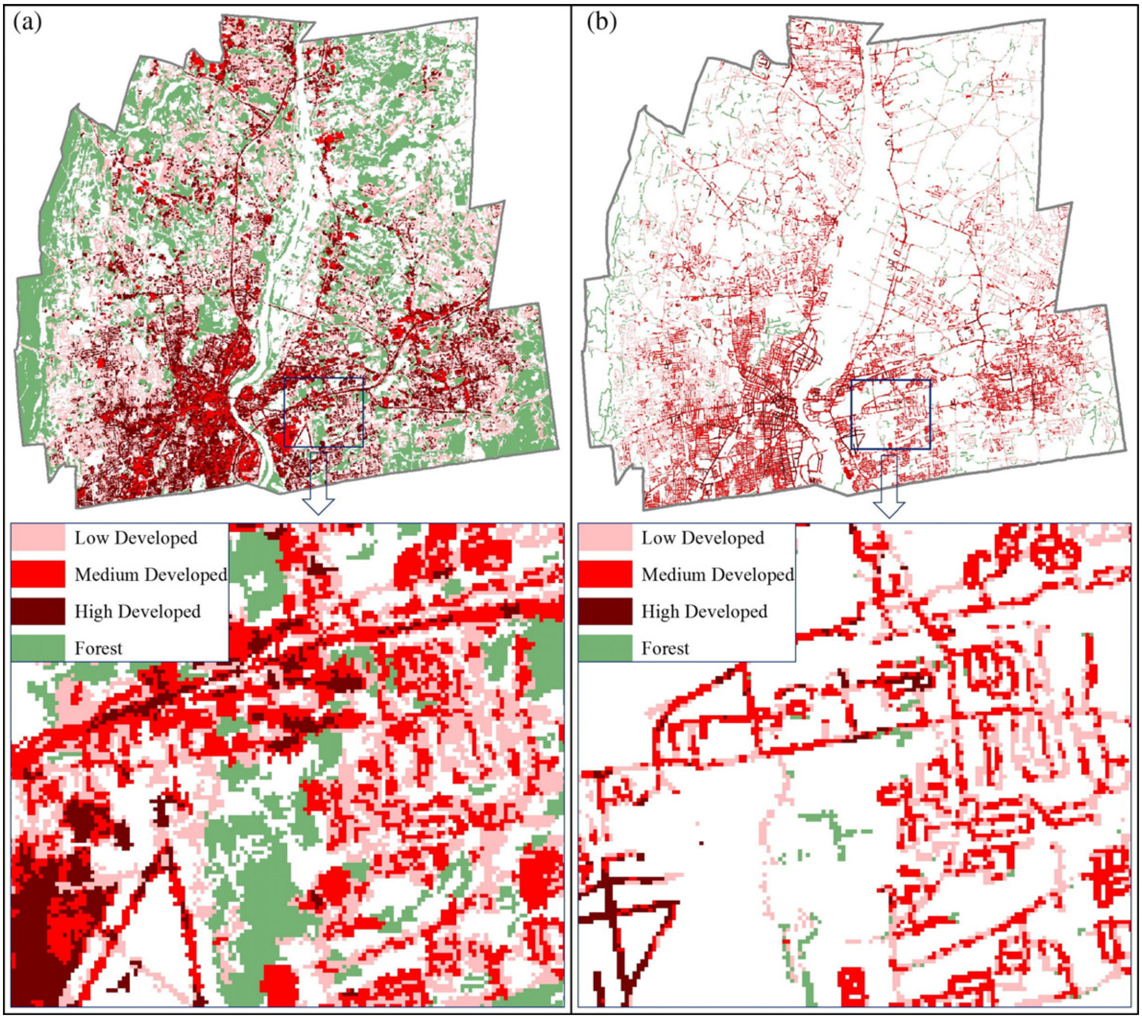

Fig. 3. Four reclassified land cover categories used in intelligent areal interpolation models: (a) whole; and (b) reduced by the 100-foot road buffer.

(PCD) or similar position in the statistical distribution of population density (PCS). These two forms of neighborhoods are selected because a similar relationship between population density and each land cover category can be expected for places nearby or in the same part of the data distribution. Each local polycategorical method was based on a neighborhood of nine observations. Nine is an arbitrary number but it is greater than the maximum number of ancillary categories and a little higher than the average number of first order geographic neighbors. A Fortran program was developed to solve the linear programing problem at the linear system based stage (this program is available from the corresponding author upon request); the calculated coefficients were then imported into ArcGIS 10.2 for the remaining interpolation steps in the same manner as the GWR coefficients. Given the differences in software used to perform the different AI methods, a comparative empirical computational study was not performed. Because the number of equations was based on *m* neighbors (nine in this study) rather than the total number of observations, *n*, fewer computations are required for the PCD and PCS interpolators than for the GWR interpolator.

Each areal interpolation method, except AW and PYCNO, had three scenarios based on using an increasing number of ancillary land cover classes. Each scenario is named as the combination of an areal interpolation model and the number of land cover classes. The extension “2” included the DL and DM classes, “3” added the DH class to the previous ones, and “4” added the FOR class to the others. For the binary dasymetric method, the extension number means how many classes were lumped together as the single control area. For example, BDAS2 means that DL and DM were aggregated together. In a second set of scenarios, the land cover ancillary classes were reduced in size by the 100-foot road buffer; in these scenarios it is assumed that populated areas only occur in the specified land cover classes within 100 feet of a designated road. Fig. 3b shows the distribution the reduced four land cover categories that are used to drive the second set of AI procedures.

The overall accuracy of each interpolation method is evaluated using the root mean square error (RMSE), the adjusted root mean square error (adj-RMSE) (Gregory, 2000), the mean absolute error (MAE), and the adjusted absolute error (adj-MAE). The adj-MAE scales absolute error by the observed population in each target zone in the same manner as adj-RMSE does (Lin, Cromley, Civco, Hanink, & Zhang, 2013). RMSE is based on the squared difference between the estimated and actual population counts whereas MAE is based on the absolute difference between the estimated and actual population counts. Thus the former is more sensitive to individual extreme errors. The adjusted measures account for differences in the size of the actual counts.

5. Results and discussion

Table 1 shows the values of the four overall error scores for all methods using two, three, and four land cover categories as the control layer. Overall, increasing the number of control categories did not improve the results. Models using just the DL and DM categories had the best results for every method except the GWR. In general, the two methods that did not use ancillary data had worse error measures than those methods that used ancillary data. For these two, AW had higher error values than PYCNO, which is consistent of the argument that methods that allow no variation in population density within the source region should not perform as well as those that permit some variation.

For results derived from the NLCD only, the GWR method with three categories had the lowest score for each of the four error measures. BDAS2 was the easiest to implement and second best among all AI models investigated here, although the results of IDM2 and PCS2 were in the same range of values. Statistical neighbors rather than geographic neighbors seem to have more in common with respect to control category density values, given that local polycategorical models based on statistical neighborhoods (PCS) outperformed their distance-based neighborhood counterparts (PCD). The greatest change in error results within the same method occurred for the IDM model. It performed reasonably well for two categories but increased significantly in error for three and four categories. This due to the fact that at the regional level, there was no source zone for which a particular land cover was the modal category, and the densities for these land cover categories had to be imputed.

Using the road buffer control in conjunction with land cover always resulted in lower error scores for the same AI method. This decrease was often greater than differences between methods only using land cover data. For example, the MAE value for the BDAS2 model decreased from 262.11 to 236.24 if only land cover classes were used with the GWR3 model, but this error value decreased to 224.50 if the same method, BDAS2, now used both land cover and road buffer controls. Among the local polycategorical methods, statistical neighborhoods still outperformed geographical neighborhoods. PCS2 was the best among all local polycategorical models and second best among all AI models over all error measures. GWR3 again was the best overall. BDAS2 and IDM2 were slightly worse than PCS2, but were still better than PCD2. It is also worth noting that because the local polycategorical methods use only neighbor observations to estimate the density for each land cover category of center observation, they are usually more sensitive to the accuracy of the ancillary information, which is reflected by the large variance of error values for each polycategorical method with respect to different land cover scenarios.

In addition to this global analysis of error, a local evaluation of AI results is done by mapping the spatial distribution of total absolute error for each source zone to see how well selected models internally redistribute the population of the source zones. The spatial distribution of areal interpolation errors for GWR3 (the method with the overall lowest error measures) when using only land cover data as the control layer (Fig. 4a) are compared against the results for the same method using both land cover and the road buffer (Fig. 4b). Next, the error maps for PCS2 (the method with the second lowest error measures for land cover and the road buffer) are given in Fig. 4c and d. As expected, a large decline in absolute error, especially for those tracts in the rural periphery area, occurs after the road buffer was added both for GWR3 and PCS2 (however, some tracts did increase in error). Because the decline was more for GWR3 within whole study area, Fig. 4b has far fewer observations in the extreme error class than Fig. 4d. Another pattern worth noting in Fig. 4 is that PCS2 outperforms GWR3 in the rural periphery area while the performances reverse in denser urban areas. A possible explanation to this pattern can be that the population density of each land cover category is more similar in the same part of the data distribution for a rural area, while it is more similar for nearby places for an urban area.

6. Conclusions

This paper has proposed and evaluated linear system-based polycategorical AI methods solved by using subsets of geographical and statistical neighbors. It contributes to a growing literature about modeling spatial non-stationarity in the relationship between population density and land cover without using a subjective scheme to estimate density (Eicher & Brewer, 2001) or a prior regional delineation to fit model locally (Mennis & Hultgren, 2006; Yuan et al., 1997). Remotely sensed land cover is commonly used as ancillary data to estimate population distribution; however, caution is needed when selecting the most likely residential categories because adding more land cover classes offers little benefit, or indeed makes the interpolation worse, which has also been found in earlier studies by Langford (2006) and Cromley et al. (2012). Furthermore, the original remotely sensed data overlaid against other data correlated with population, such as a road buffer, to refine residential areas can lead to a significant improvement in AI performance. A case study with the nine-town area of

Table 1
Overall errors of areal interpolation methods.

	NLCD only				NLCD and road buffer			
	RMSE	Adj-RMSE	MAE	Adj-MAE	RMSE	Adj-RMSE	MAE	Adj-MAE
AW[a]	576.89	0.584	390.60	0.360				
PYCNO[a]	558.47	0.536	379.23	0.338				
BDAS2	**438.36**	**0.441**	**262.11**	**0.235**	**385.04**	**0.377**	**224.50**	**0.202**
BDAS3	458.98	0.463	274.12	0.247	393.99	0.395	226.86	0.207
BDAS4	547.18	0.550	355.86	0.323	416.61	0.418	252.77	0.230
IDM2	**440.17**	**0.444**	**264.99**	**0.238**	**379.69**	**0.374**	**222.32**	**0.200**
IDM3	617.07	0.590	395.70	0.352	470.04	0.459	281.91	0.257
IDM4	541.65	0.545	343.82	0.313	456.49	0.449	273.00	0.248
GWR2	459.28	0.458	278.97	0.250	383.61	0.373	232.00	0.210
GWR3	**370.67**	**0.382**	**236.24**	**0.214**	**358.76**	**0.355**	**209.18**	**0.188**
GWR4	393.85	0.399	250.52	0.226	388.56	0.374	228.78	0.203
PCD2	**448.27**	**0.461**	**275.80**	**0.251**	**399.26**	**0.380**	**244.54**	**0.220**
PCD3	505.52	0.513	312.77	0.288	432.97	0.415	272.70	0.247
PCD4	549.24	0.515	351.71	0.317	494.24	0.448	315.68	0.278
PCS2	**441.94**	**0.441**	**264.86**	**0.236**	**378.04**	**0.367**	**222.42**	**0.200**
PCS3	463.22	0.463	282.12	0.253	408.81	0.407	242.37	0.219
PCS4	560.89	0.569	360.48	0.330	427.32	0.435	261.29	0.239

Note: NLCD = national land cover database; RMSE = root mean square error; Adj-RMSE = adjusted root mean square error; MAE = mean absolute error; Adj-MAE = adjusted root mean square error; AW = areal weighting; PYCNO = pycnophylactic; BDAS = binary dasymetric; IDM = intelligent dasymetric mapping; GWR = geographically weighted regression; PCD = polycategorical distance; PCS = polycategorical statistic.
Bold values indicate the smallest error values for each areal interpolation model with different ancillary land cover classes.

[a] Neither AW nor PYCNO used any ancillary data.

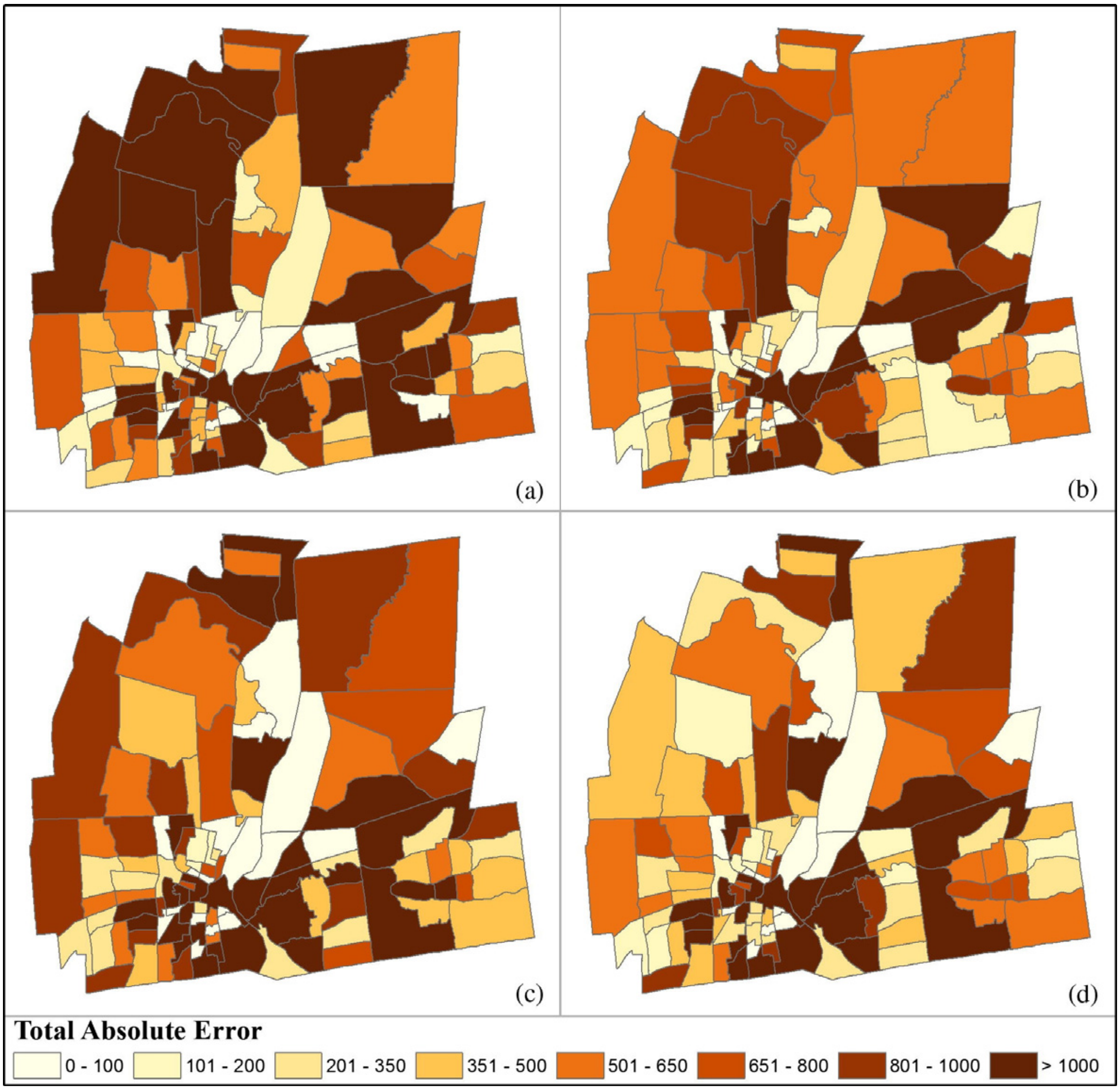

Fig. 4. The spatial distribution of total absolute error for each source tract: (a) results of GWR3 with NLCD only; (b) results of GWR3 with NLCD and road buffer; (c) results of PCS2 with NLCD only; and (d) results of PCS2 with NLCD and road buffer.

Connecticut, US, indicates small margins of difference in the results of the proposed polycategorical methods, BDAS, and IDM when low and medium developed categories are used as ancillary data. Theoretically, more complex relationships between population and land cover is allowed in the proposed polycategorical methods than that in BDAS and IDM. Thus future case studies should focus on areas with more complicated patterns of density distribution to verify if the findings of small margins could be reversed in another situation. The notion of statistical neighborhoods was also introduced as an alternative to geographic distance-based neighborhoods in developing a local interpolator. The logic of statistical neighborhoods is that control categories in source zones having similar population densities should also have similar relationships between ancillary categories and population density.

Obviously, estimate accuracy of any method that uses a kernel function is sensitive to the bandwidth values. The subjective bandwidth of nine neighborhoods used in this study, especially for those based on statistical similarity, was competitive in its performance when compared with other test methods. Another finding with the small number of neighbor observations is the increased sensitivity to the accuracy of ancillary information. Future research will investigate the tradeoff of model complexity and performance by testing different number of neighbors in solving the system of linear equations, and at the same time determining if small marginal improvements can be justified given the extra computation involved. Furthermore, the assumption that the number of neighborhoods selected based on same criteria is equally good for all observations across the study area may oversimplify the complexity of population distribution. Instead, models that allow locally varied numbers of neighborhoods and selection schemes in estimating density coefficients for each observation may be more reasonable and thus worth exploring in future research.

Areal interpolation is a widely used procedure for estimating unknown attribute values for a set of target zones from known attribute values for a set of source zones. The development of GIS software packages has both created a greater need for areal interpolation as well as provided a platform for making areal interpolation available to a broader user group. Furthermore, several free and open source implementation of AI methods (or methods that could be adopted for AI purposes) are available as R packages, such as the aigis package by Bryant and Westerling (2012), the pycno package by Brunsdon (2014), and the rtop package by Skoien (2014). All these three packages were developed based on simple areal interpolation or geostatistic methods and do not incorporate any control data. In recent years, however, a variety of different intelligent methods have been proposed as areal interpolators with better performance than simple interpolators but most are not used by the general public because general public has limited accessibility to these methods. Thus integration of the proposed local polycategorical AI method into existing software packages is a focus of future efforts.

References

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.

Bivand, R., & Yu, D. (2014). *Package 'spgwr'*. <http://cran.r-project.org/web/packages/spgwr/index.html> Accessed March 2015.

Brunsdon, C. (2014). *Package 'pycno'*. <http://cran.r-project.org/web/packages/pycno/index.html> Accessed March 2015.

Bryant, B., & Westerling, A. (2012). *Package 'aigis'*. <https://github.com/cran/AIGIS> Accessed March 2015.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2012). A quantile regression approach to areal interpolation. *Annals of the Association of American Geographers*, *102*(4), 763–777.

Cromley, R. G., Hanink, D. M., & Lin, J. (2013). Developing choropleth maps of parameter results for quantile regression. *Cartographica*, *48*, 177–188.

Dempster, A., Laird, N., & 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.

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

Fisher, P., & Langford, M. (1995). Modeling the errors in areal interpolation between zonal systems by monte carlo simulation. *Environment and Planning A*, *27*, 211–224.

Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. *The Professional Geographer*, *48*(3), 299–309.

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

Flowerdew, R., & Green, M. (1989). Statistical methods for inference between incompatible zonal systems. In M. Goodchild, & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 239–247). London: Taylor and Francis.

Flowerdew, R., & Green, M. (1991). Data integration: Statistical methods for transferring data between zonal systems. In I. Masser, & M. Blakemore (Eds.), *Handling geographical information* (pp. 38–54). London: Longman.

Fotheringham, A. S. (1997). Trends in quantitative methods I: Stressing the local. *Progress in Human Geography*, *21*(1), 88–96.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: Wiley.

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., et al. (2011). Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *77*(9), 858–864.

Gelfand, A., Zhu, L., & Carlin, B. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, *2*(1), 31–45.

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

Gregory, I. N. (2000). An evaluation of the accuracy of the areal interpolation of data for the analysis of longterm change in England and wales. In *Proceedings of the 5th international conference on geocomputation*. Kent: University of Greenwich, August 23–25.

Koenker, R., & Bassett, G. (1978). Regression quantile. *Econometrica*, *46*(1), 33–50.

Lam, N. S. -N. (1983). Spatial interpolation methods: A review. *Cartography and Geographic Information Science*, *10*(2), 129–150.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, *30*(2), 161–180.

Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems*, *31*(1), 19–32.

Langford, M., Maguire, D., & Unwin, G. (1991). The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In I. Masser, & M. Blakemore (Eds.), *Handling geographic information: Methodology and potential applications* (pp. 55–77). London: Longman.

Lin, J., & Cromley, R. G. (2015). Evaluating geo-located twitter data as a control layer for areal interpolation of population. *Applied Geography*, *58*, 41–47.

Lin, J., Cromley, R. G., Civco, D. L., Hanink, D. M., & Zhang, C. (2013). Evaluating the use of publicly available remotely sensed land cover data for areal interpolation. *GIScience & Remote Sensing*, *50*(2), 212–230.

Lin, J., Cromley, R. G., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS*, *17*(1), 1–14.

Lo, C. P. (2008). Population estimation using geographically weighted regression. *GIScience & Remote Sensing*, *45*(2), 131–148.

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.

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

Mrozinski, R. D., & Cromley, R. G. (1999). Singly- and doubly-constrained methods of areal interpolation for vector-based GIS. *Transactions in GIS*, *3*(3), 285–301.

Openshaw, S., & Taylor, P. J. (1981). The modifiable areal unit problem. In N. Wrigley, & R. Bennett (Eds.), *Quantitative geography: A British view* (pp. 60–69). London: Routledge and Kegan Paul.

Rase, W. -D. (2001). Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, *3*(2), 199–213.

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

Skoien, J. O. (2014). *Package 'rtop'*. <http://cran.r-project.org/web/packages/rtop/index.html> Accessed March 2015.

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

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

Visvalingam, M. (1991). Areal units and the linking of data. In L. Worrall (Ed.), *Spatial analysis and spatial policy using geographic information systems* (pp. 12–37). London: Belhaven Press.

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

Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, *19*(4), 287–306.

Yuan, Y., Smith, R. M., & Limp, W. F. (1997). Remodeling census population with spatial information from Landsat TM imagery. *Computers, Environment and Urban Systems*, *21*(3), 245–258.