The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study

Fang Qiu[1]
University of Texas at Dallas, Program in Geospatial Information Sciences, 800 West Campbell Road, Richardson, Texas 75080

Caiyun Zhang
Florida Atlantic University, Department of Geosciences, 777 Glades Road, Boca Raton, Florida 33431

Yuhong Zhou
University of Texas at Dallas, Program in Geospatial Information Sciences, 800 West Campbell Road, Richardson, Texas 75080

Abstract: Areal interpolation has been an active research area, but its wide adoption in the general GIS community is limited due to the lack of implementation tools in commercial software. To bridge this gap, an areal interpolation extension is developed in ArcGIS with 4 popular algorithms, 10 raster and vector implementations, and 4 commonly used error measures. A comparative case study utilizing the extension shows that it provides general users with a user-friendly interface for performing areal interpolation without dealing with the complexities of the underlying algorithms and their implementation details.

INTRODUCTION

The advent of geographic information system (GIS) technology makes it possible to integrate various disparate data using a common spatial reference system. Further, through map overlay operations such as union and intersect, these spatial data can be combined or split into different sets of areal units. In almost all commercial GIS software systems supporting these operations, associated attributes are usually carried over without modification to the newly formed spatial objects. This is reasonable for categorical attributes (such as names and classes) and in some cases for numerical variables (such as ratios, percentages, and proportions) (Goodchild and Lam, 1980). However, for numerical attributes that are the result of aggregating individual observations (such as counts, area, and volume), it can be problematic to carry them over to the new objects (Flowerdew et al., 1991). For example, demographic information from the census is an attribute that is usually aggregated into enumeration units from individual survey records. A census unit with 100 residents may be split into two units

[1]Corresponding author; email: ffqiu@utdallas.edu

GIScience & Remote Sensing, 2012, **49**, No. 5, p. 644–663. http://dx.doi.org/10.2747/1548-1603.49.5.644

by a map overlay operation. If the attribute is simply carried over without modification to these two units, each will have 100 residents, resulting in an artificial inflation of the total population in the study area. To address this problem, an areal interpolation operation is needed. Areal interpolation is a set of methods that can estimate an aggregate attribute of one areal unit system (the newly formed polygons in this case) based on that of another, spatially incongruent, system in which the attribute data were collected (the original polygons in this case). The original units for which the attribute is known are often referred to as *source units* and those for which the attribute needs to be estimated are termed *target units* (Markoff and Shapiro, 1973).

In addition to areal units created from GIS operations such as overlay and buffering, areal interpolation is needed in many other applications, especially those involving the use of population data. Demographic information from national censuses is usually based on arbitrarily designated census units, such as census blocks, block groups, and census tracts in the U.S. case. To associate such information with other areal-based datasets such as market catchment areas, postal delivery zones, and police beats or areal units representing environmental phenomena, such as watersheds, soil types, and land use, areal interpolation needs to be performed because the boundaries of these areal units are seldom spatially congruent with those of the census.

Many areal interpolation algorithms have been proposed in the literature based on different assumptions regarding the underlying distribution of the population. The simplest algorithms are area-weighting, where population in the source units is spatially reallocated into target units using as a weight the area that each source unit contributes to the target area (Lam, 1983). This method is straightforward but its implicit assumption of a homogeneous internal population distribution in each source unit is rarely true in reality. The pycnophylactic method (Tobler, 1979) is another approach to areal interpolation which, like the area-weighting approach, relies only on the data being estimated. It disaggregates the population of source units at their centroids into a continuous raster surface, the cells of which are then used to estimate the target unit populations. It is an improvement over the area-weighting method because it is not assuming a homogeneous population distribution, and the continuous surface eliminates sharp transitions in population estimation across unit boundaries.

Because population distribution is often related to other socioeconomic phenomena, many so-called intelligent methods for areal interpolation were proposed to improve areal interpolation accuracy by using ancillary data to shed light on the underlying population distribution (e.g., Flowerdew, 1988; Flowerdew et al., 1991; Fisher and Langford, 1995; Xie, 1995; Eicher and Brewer, 2001; Mennis, 2003; Holt et al., 2004; Wu and Murray, 2005; Langford, 2007; Liu et al., 2008). Goodchild et al. (1993) were the first to introduce ancillary data into areal interpolation. The ancillary data are usually referred to as "control units" in general, with two-dimensional ancillary data (such as land use) called "control zones" and 1-D ancillary data (such as road networks) termed "control lines." Among intelligent methods, dasymetric approaches are arguably the most well-known solutions to the problems caused by the areal weighting's homogeneity assumption. The dasymetric method was originally designed to address the homogenous visualization issues of choropleth maps and was then adapted as an areal interpolation approach. Dasymetric methods allocate source unit populations to smaller control units that have different but internally consistent densities (Langford et. al., 1991; Langford, 2006) to achieve heterogeneous

population distribution. Various versions of dasymetric methods have been proposed, usually based on the size (e.g., area or length) of the control units used to disaggregate the source unit population and then to reaggregate to obtain the target unit population. According to Hawley and Moellering (2005), Fisher and Langford (1995) were the first to publish a dasymetric areal interpolation method using 2-D land use data as control zones. Their binary dasymetric method is based on the area of populated and unpopulated land units. A number of dasymetric approaches (Langford et al., 1991; Langford and Unwin, 1994; Yuan et al., 1997; Langford, 2006) utilize multiple residential classes instead of a simple binary division of populated and unpopulated land use. Because these approaches often rely on multivariate regression to achieve areal interpolation, they are known usually as multi-class regression dasymetric methods. Dasymetric methods can also be implemented using 1-D network information as ancillary data. Xie (1995) proposed three algorithms: the network length method, the network hierarchical weighting method, and the network house-bearing method. Many of these studies demonstrate that dasymetric methods can achieve higher accuracy compared with approaches that do not utilize ancillary data. Area-weighting, pycnophylactic, binary dasymetric, and multi-class regression dasymetric methods are now the most widely cited approaches in the areal interpolation literature. They have laid the foundation for further areal interpolation algorithm development and have often served as benchmarks for comparison with newly developed algorithms.

These newly developed algorithms include variants of dasymetric methods based on the size of water bodies, transportation structures, parks, urban lands, etc. (Deichmann, 1996; Turner and Openshaw, 2001; Hawley and Moellering, 2005), and predetermined densities of different land use types obtained through empirical sampling (Mennis and Hultgren, 2006). Statistically based methods have also been introduced, which establish a relationship between population and relevant (but not always directly related) socioeconomic variables (Flowerdew, 1988; Flowerdew and Green, 1991) or spectral values from digital satellite imagery (Harvey, 2002a, 2002b), rather than the size of the control units. Geostatistical approaches originally designed for point interpolation have also been extended into the field of areal interpolation (Kyriakidis, 2004; Kyriakidis and Yoo, 2005; Wu and Murray 2005). To produce an accurate areal interpolation, these algorithms often require extra adjustment procedures to achieve the pycnophylactic property (Tobler, 1979), or the volume-preserving requirement (Lam, 1983), which ensures that the original value of the source units is preserved in the transformation to the target units. This means that when the population count of a source zone is disaggregated and summed back into the source zone, the total should equal that of the original value.

To evaluate which areal interpolation algorithm is the most appropriate for an application, comparative studies have been conducted in the literature, with accuracy assessment based on a variety of error measures. Fisher and Langford (1995) used root mean squared error (RMSE) to quantify the error introduced by various areal interpolation methods. They concluded that area-weighting performed poorly compared with intelligent models that utilized ancillary data, and that better accuracy was achieved as the number of target zones declines. Sadahiro (1999) defined a source zone shape and size-based error index for areal interpolation and used it to examine the spatial distribution of the errors. He concluded that uniform error distributions led to better overall interpolation accuracy; the utility of ancillary information varied according to

its relevance to the variable of interest; and adding inappropriate ancillary information may actually reduce the accuracy of the interpolation. By using mean squared error (MSE), Sadahiro (2000) demonstrated that estimation accuracy was improved when source zones were relatively small compared with the target zones, and that the relative shapes of the source and target zones also affected accuracy. Gregory (2000, 2002) used adjusted RMSE (ARMSE) to examine the accuracy of several areal interpolation techniques suitable for use with historical data. He concluded that the effectiveness of the technique depended on the variable to be interpolated and the choice of target geography. Hawley and Moellering (2005) systematically compared four popular approaches: the area-weighting method, pycnophylactic method, binary dasymetric method using ancillary land use data, and the binary dasymetric method using ancillary road network data. Based on RMSE and ARMSE, they concluded that the dasymetric method using road networks achieved the best result for population data when transferring from the census tract level to the block group level.

Langford (2006) used mean absolute error (MAE) and RMSE as measures to investigate whether a regional regression between population and land cover outperformed global regression in the dasymetric method, and whether a 3-class dasymetric method improved upon the binary dasymetric model. Langford's results indicated that regional regression was superior and had the advantage of highlighting spatial nonstationarity. However, the benefits of a 3-class regression dasymetric model over a binary model were inconclusive. More recently, Gregory and Ell (2006) and Schroeder (2007) explored errors in areal interpolation for temporal census data. All of these studies suggest that the accuracy of areal interpolation depends on a combination of factors, including method used, the nature of the variable being interpolated, the nature of the ancillary data, and the shape and size of both the source and target zones. In general, these studies based their findings upon only one or two error measures, and different measures were used in each study. This makes comparison of results across different studies difficult.

Despite being an active research area in recent years, areal interpolation has not been widely embraced by the general GIS user. There is little evidence in the literature suggesting that these areal interpolation algorithms have stepped out of their research domain to assist the broader GIS community. This is likely due to the absence of areal interpolation tools within or outside of commercial GIS software packages. Implementing these algorithms can be complex and may challenge the average GIS user (Langford, 2007). Compared to studies on algorithm development and accuracy evaluation, the literature describing the software implementation of areal interpolation tools is relatively rare. Our search found only four pertinent works, with no publications devoted to this aspect. Flowerdew et al. (1991) implemented their proposed statistical method in an Arc/INFO workstation using AML and a Generalized Linear Interactive Modeling System (GLIMS). Bloom et al. (1996) provided an enhanced version of their method in the Mapinfo Desktop Mapping Package. Xie (1995) used Arc/INFO AML along with user-written C code to implement his road network–based dasymetric method. Fisher and Langford (1996) described their implementation of the dasymetric method with ancillary land cover data in IDRISI GIS with the realization procedures.

Beyond these four publications, authors typically describe their methodologies in mathematical terms and only briefly discuss the implementation of the methods.

A person wishing to adopt these authors' methods often must reinvent the wheel and redevelop the areal interpolation tools implemented by the authors. However, the complex ideas underlying the methodology, the demand for a clear understanding of GIS functionality, the associated technical details involved, and the need for at least rudimentary programming skills have greatly dampened the enthusiasm of potential users, and even sophisticated researchers, for embracing these algorithms. One solution to this situation is undoubtedly the creation of a simple and convenient areal interpolation package that includes the most popular algorithms and makes them readily available to the general user. Developing this as an extension within an existing commercial GIS system avoids the need to start everything from scratch and can reach more users who are already familiar with that system. To the best of our knowledge, the only extant areal interpolation toolbox was developed by Schneider et al. (2006), and as of this writing, the link to the website hosting the toolbox is no longer active. According to the conference paper reporting the toolbox, it was developed as a Toolbox in ArcGIS, and incorporated three methods: area-weighting, EM algorithm, and the dasymetric method based on 2-D ancillary data. No detailed implementation procedures for each method were provided, and the system did not include the widely used dasymetric approach using 1-D ancillary data. The applicability of this toolbox was also limited because it required a run-time version of Matlab software, which would need additional licensing expense for most in the general GIS community. Furthermore, the toolbox did not include any calculation of error measures, a much-needed function for researchers to assess the accuracy of their areal interpolation results.

In this paper, we develop a new areal interpolation extension in the popular ArcGIS software environment, using ArcObjects and VB.NET programming. Our extension includes the area-weighting algorithm and the pycnophylactic method because of their simplicity and the fact that they do not need ancillary data. We also incorporate the binary and 3-class regression dasymetric algorithms (both of which can take either 2-D control zones or 1-D control lines as ancillary information) due to their popularity in the literature and the associated high accuracy they tend to achieve. Both raster and vector implementations are provided for most of these algorithms. We did not implement the statistics- and geostatistics-based intelligent approaches to areal interpolation at this stage. Methodologically, the statistical-based approaches are similar to the 3-class regression dasymetric algorithm, but are often data dependent. Geostatistics-based approaches are quite complex computationally and are still a new and developing area with limited practical applications.

What is possibly the most unique aspect of our toolbox is that we also incorporate four commonly used error measurements, including mean absolute percent error (MAPE), value weighted MAPE (VWMAPE), root mean squared error (RMSE), and adjusted RMSE (ARMSE). These support accuracy analysis if reference data are available. Finally, the computation time used by each implementation can be reported to allow for the assessment of the efficiency of an algorithm.. The extension is available at http://www.utdallas.edu/~ffqiu/arealinterpolation/. We hope that this areal interpolation extension will encourage a wider adoption of the most popular and mature areal interpolation techniques among the broader GIS user community.

The detailed algorithms and definitions of the error measurements included in the extension are described in the Methodology section below. The Case Study section presents an example of how the areal interpolation extension was used in

analyzing population data for Collin County, Texas. We summarize our findings in the Conclusions section.

METHODOLOGY

Four commonly used algorithms—area-weighting, pycnophylactic, binary dasymetric, and 3-class regression dasymetric—were integrated into the extension. Software implementing all these methods was constructed using the vector, raster, and attribute analysis functions available via ArcObjects. ArcObjects is the developer toolkit for the commercial ArcGIS software system. It can create stand-alone Component Object Model (COM) executables using any of the COM-compliant program languages such as Visual Basic 6.0, C#, and Visual Basic.NET (VB.NET). Microsoft COM techniques allow developers to generate re-usable software components and to link them with other components to build larger applications. In this study, we used VB.NET to create an areal interpolation extension. The detailed description and associated equations of the methods have been widely published in the literature. Only a brief introduction to the methods and the specific procedures for each software implementation are provided below.

Area-Weighting Method

The area-weighting method is a straightforward algorithm for performing areal interpolation (Goodchild and Lam, 1980) and the most popular choice when ancillary information is not available. Under the assumption that population is evenly distributed in the sources zone, a constant population density of each source zone is first estimated. Then the size of each overlapping area between a target zone and a source zone is used as a weight to estimate the population for the target zones. Areal weighting can be implemented using raster and vector techniques (Fig. 1), where the size is the area for the vector implementation or the number of cells for the raster implementation. The area-weighting method is a volume-preserving algorithm, and therefore no adjustment procedure is needed to conserve the source zone population.

Figure 1 is a flow diagram for the raster- and vector-based procedures implementing the area-weighting method. The raster-based implementation is comprised of the following main steps: (1) converting the vector-source zone data to raster using the *feature-to-raster function*; (2) joining the VAT table of the raster source zone to the vector source zone using the *join function* to access its population; (3) computing the population density per cell in the attribute table of the source zone by dividing the population by the number of cells in each source zone using the *field calculator function* ; and (4) summing the value of each cell falling in each target zone to get the interpolated value using the *zonal statistics function* with the target zone being the zone layer. The vector implementation involves: (1) computing the area of each source zone by using the *calculate_geometry function*; (2) deriving the density for each source zone by computing the ratio of its population value and area using the *field calculation function*; (3) overlaying the source zone and target zone using the *intersection function*; (4) computing the area of the intersected zone using the *calculate geometry function*; (5) determining the new population value for each intersected zone by multiplying its area and the density of the source zone this intersection falls into using the

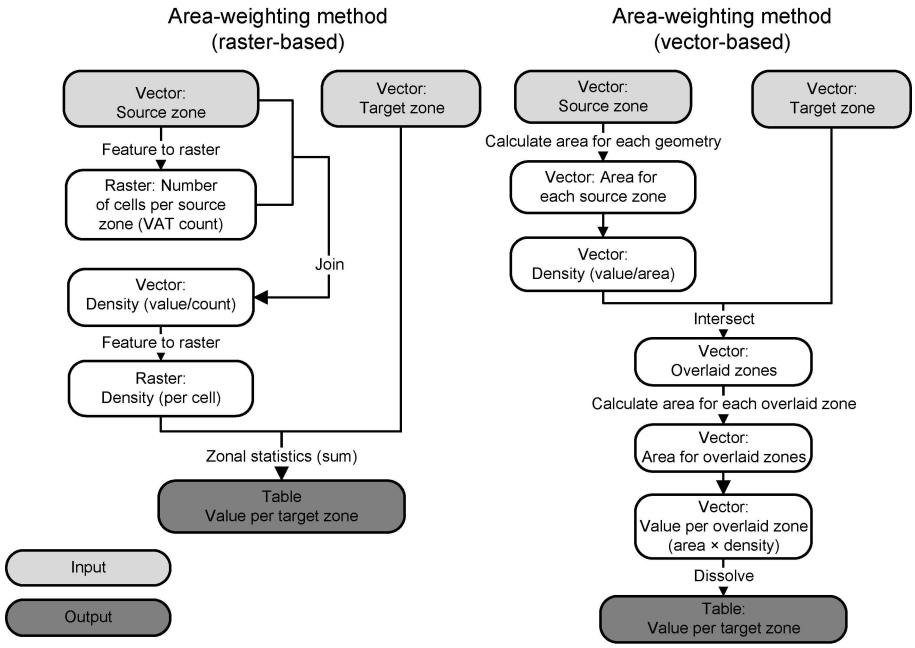

Fig. 1. Implementation steps for the area-weighting method.

field calculation function; and (6) obtaining the interpolated value of each target zone using the *dissolve function* for the intersected zone with target-ID as the key field and the new population value as the summary field.

Pycnophylactic Method

The pycnophylactic method only functions within a raster GIS environment due to the fact that it requires a raster smoothing process. The technique starts by computing a constant per-cell population density in each source zone as the area-weighting method. This per-cell population density is then smoothed by replacing it with an average density of its neighbors within a *n*-by-*n* moving window, with the size of the smoothing window *n* being customizable based on the resolution of the underlying data. To meet the pycnophylactic requirement of volume preservation, a procedure suggested by Mennis (2003) and Langford (2006) is employed to adjust the smoothed population density.

The raster-based procedures to implement the pycnophylactic method are shown in Figure 2. They entail: (1) obtaining the per-cell density using the same first three steps as the raster-based area-weighting method; (2) calculating a new density by replacing the old one with the mean of its neighbors using the *focal mean function*; (3) estimating the population for each source zone using the new per-cell density by the *zonal statistics function,* with the source zone being the zone layer; (4) adjusting the new density by multiplying each cell value with the ratio between the original population and the estimated total population of each source zone; and (5) summing

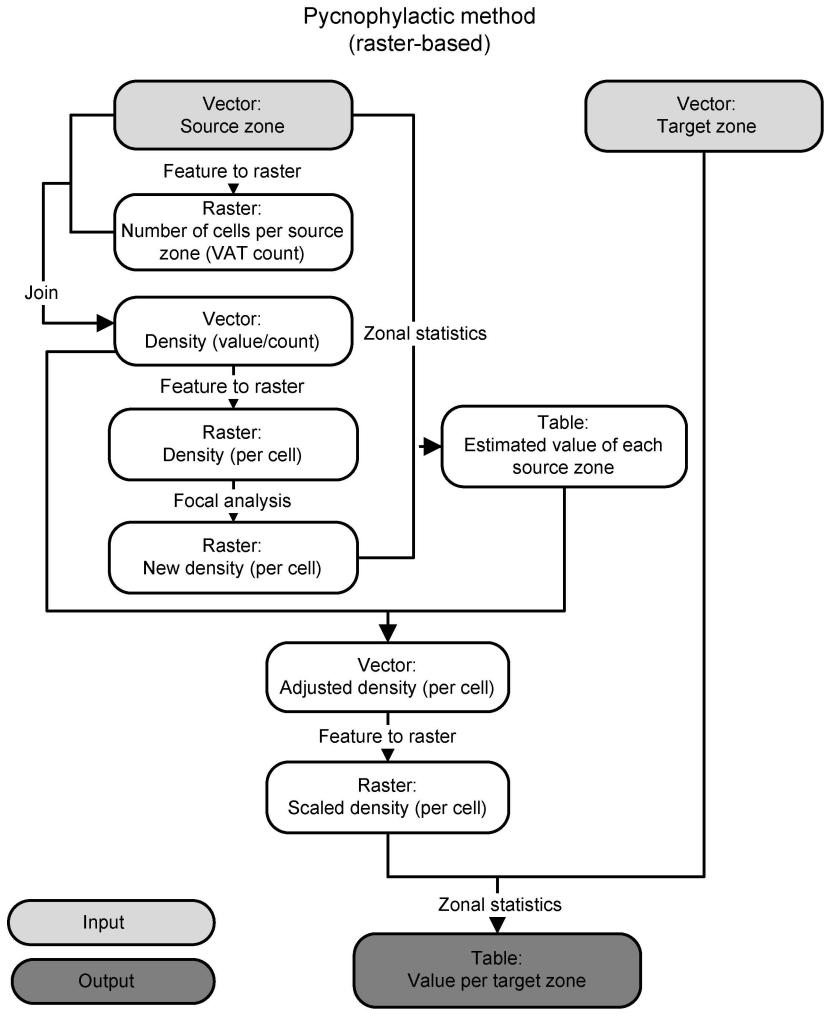

Fig. 2. Implementation steps for the pycnophylactic method.

the adjusted population density of each cell falling in each target zone to obtain the estimated population using the *zonal summary function*, with the target zone being the zone layer.

Binary Dasymetric Method

The binary dasymetric method is based on ancillary data (such as land use) that provides a binary divide between populated and unpopulated units. The use of this ancillary data as control units allows the population to be redistributed only to populated units and therefore does not assume an evenly distributed population in a source zone. However, it implicitly assumes that population density is constant in all the populated control units within a source zone, which is a weakness of the approach

since, as with the area-weighting approach, this does not often occur in reality. The population density in the control units is first calculated by dividing the population count of a source zone by the total size of all populated units falling within it, rather than by that of the source zone. Then the size of the overlapped unit between a target zone and a control unit is calculated and multiplied by the density that controls the unit's estimation of its population count. Finally, all control units falling into a target zone are summed to obtain its population.

The binary dasymetric method can be based on land use (such as residential and non-residential) or the street network. For the network-based method, the network length is used as ancillary data because of its simplicity; there is no requirement of detailed information regarding road classes and housing distributions along roads. Similar to the land use–based approach, the network length–based binary dasymetric methods assume that population is not evenly distributed across a source zone, but mainly located along the road network. Therefore, road segments are viewed as populated units in this case. The total length (in distance or number of cells) of the road segments (rather than the area) within a source zone is first computed and used to calculate population density per unit length in the source zone. The length of the network segments within the overlapping area between a source and a target zone is then obtained, which is multiplied by the density of their corresponding segment to obtain population counts distributed along these segments, which are summed to estimate the total population in each target zone.

The raster and vector implementation of the binary dasymetric method is illustrated in Figure 3. Note that the vector method is considerably more complex, although both techniques may produce similar interpolation results. The raster-based implementation involves: (1) dividing the ancillary raster data into populated (residential or road) and unpopulated (non-residential or non-road) units using the *reclassification function*; (2) obtaining the number of cells for the populated units within each source zone using the *zonal statistics function*; (3) calculating the per-cell density within each source zone by obtaining the ratio of the population for the source zone and the number of populated cells; (4) summing the per-cell density of the populated cell falling in each target zone to derive the interpolated value using the *zonal summary function,* with the target zone being the zone layer and density grid being the value layer. The vector-based implementation of the binary dasymetric method involves: (1) overlaying the 2-D or 1-D vector ancillary data with the source zone using the *intersect function*; (2) calculating the area or length of the overlaid units using the *calculate_geometry function*; (3) obtaining the total area or length of the overlaid units within each source zone using the *dissolve function*, with source-ID as the key field and area or length as the summary field; (4) calculating the density of the overlaid units by dividing the population of each source zone by the total area or length for that source zone; (5) overlaying the target zones and the overlaid units derived above using the *intersect function*; (6) calculating the area or length of the newly overlaid zones using the *calculate_geometry function*; (7) calculating a new population for each newly overlaid zone by multiplying the area or length with the density; and (8) obtaining the interpolated population of each target zone using the *dissolve function* for the new overlaid units, with target-ID as key field and new population as the summary field.

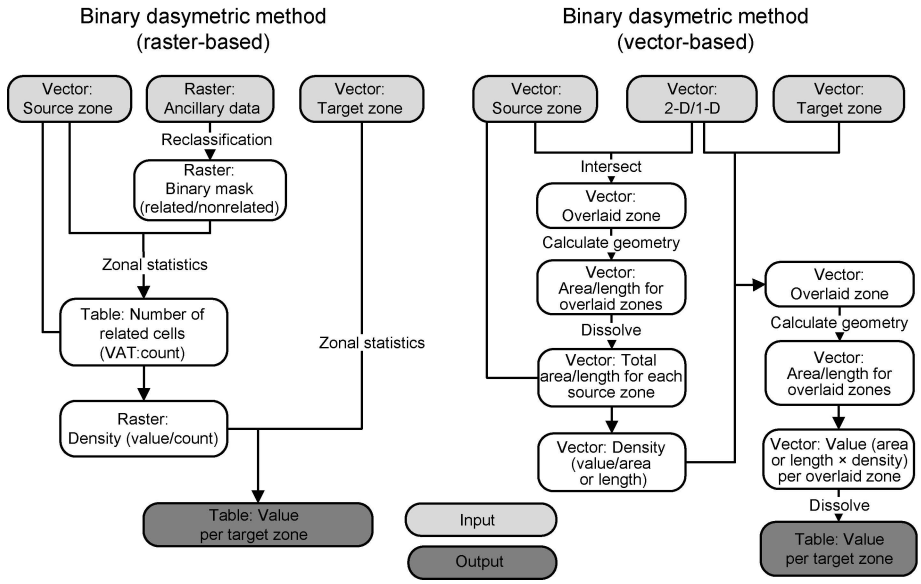

Fig. 3. Implementation steps for the binary dasymetric method.

3-Class Regression Dasymetric Method

The 3-class regression method is an extension of the binary dasymetric model and the statistical model involving ancillary data with more than two types. Rather than assuming population is evenly distributed in all residential units in a source, it allows, for example, single-family residential units to have a lower population density and multiple-family residential units a higher density. For this method, a regression model is needed to allocate population to low-density and high-density residential units, but not to the non-residential units (Langford et al., 1991; Langford and Unwin, 1994; Yuan et al., 1997). In the regression function, the population count of each source zone is the dependent variable and the total areas of each residential land use type are the independent variables. The regression parameters are in fact the population densities of their corresponding residential type. These population densities cannot be directly used for target zone population estimations if the source zone volume preserving property is to be achieved. To this end, the same rescaling operation (Mennis, 2003; Langford, 2006) that was employed in the pycnophylactic method is utilized to adjust the population density so that the population volume of the source zones can be preserved. The adjusted densities are then multiplied by the total area of their corresponding residential units; these outputs are then aggregated to estimate the population counts of the target zones.

There are two implementation strategies for the regression dasymetric method: the global regression and local regression approaches. Under global regression, data from all source zones are used to derive the density of each class. Under the local regression, a source zone and its neighbors are used to set up the regression function so that spatial non-stationarity can be modeled. The 3-class regression dasymetric

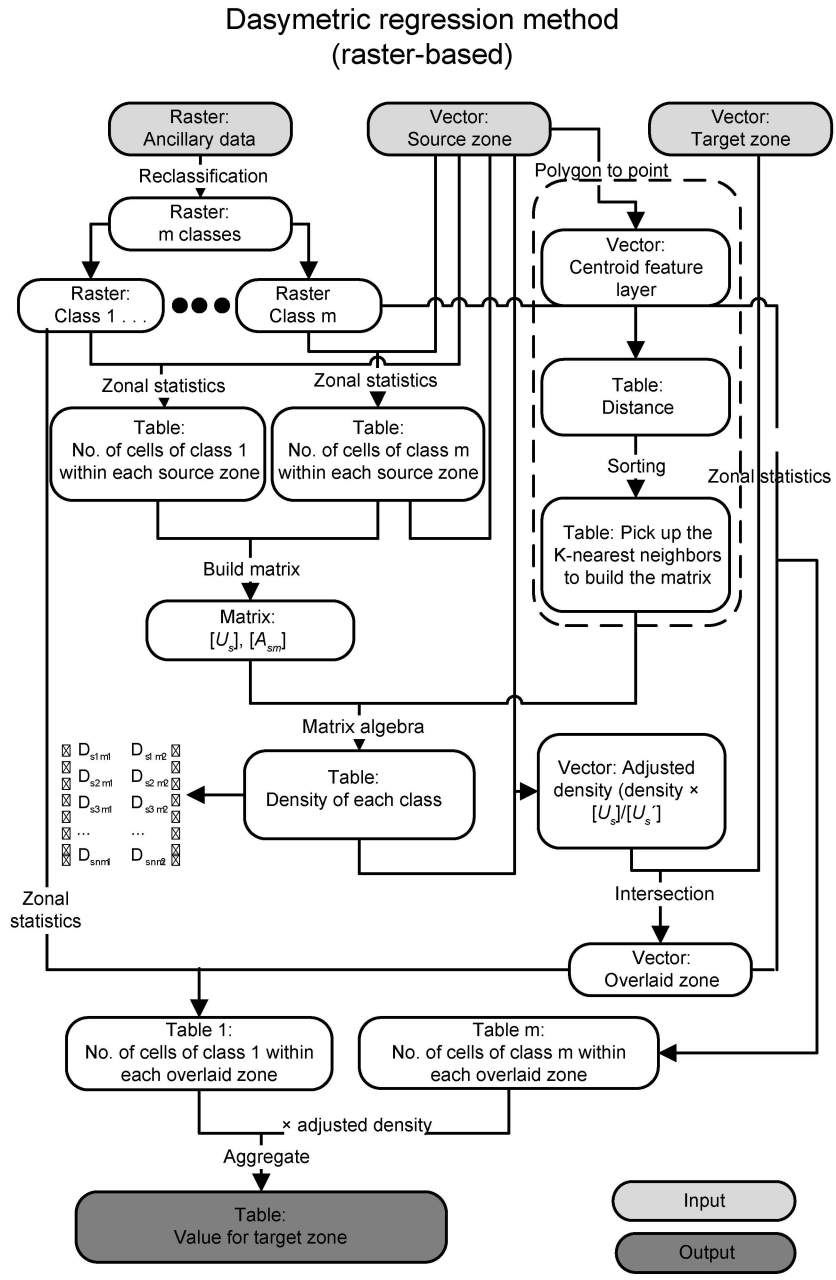

Fig. 4. Implementation steps for the regression dasymetric method.

method was implemented only in the raster environment due to its complex nature (Fig. 4).

The global regression dasymetric method is comprised of the following steps: (1) classifying the ancillary raster data into three classes (nonresidential, low-density residential, and high-density residential) using the *reclassification function*; (2) obtaining

the number of cells of each class within each source zone using the *zonal statistics function*, with source zone as the zonal layer; (3) building a source zone population matrix [U_s] and a residential area matrix [A_{sm}] using the number of cells of each residential class within each source zone; (4) calculating the per-cell population density per class [D_{sm}] based on regression matrix algebra; (5) estimating the population of each source zone by multiplying matrix [A_{sm}] with [D_{sm}]; (6) adjusting the per-cell density with the ratio of the original and estimated populations of the source zone; (7) overlaying the source zone with the target zone using the *intersect function*; (8) obtaining the number of cells of each land use class within each overlaid zone using the *zonal statistics function*; (9) calculating the population of each overlaid zone by multiplying the number of cells with the adjusted density of their corresponding residential class; and (10) summing the population of overlaid zones into the target zone using the *zonal statistics function* to estimate its population.

Note that the local regression method is similar to the global, but the construction of the matrix in step 3 depends on the neighbors of each source zone instead of all source zones. The neighbors can be selected based on either the distance between the centroids of the source zone (i.e., the *k* nearest neighbors) or the topological contiguity of the source zones (i.e., first-order neighbors being the immediate adjacent polygons, second-order neighbors being the immediate neighbors of the first-order neighbors, and so on). For the distance-based neighbor selection, the following steps are involved (highlighted by the dashed rectangle in Fig. 4): (1) converting the source zone polygons to point features using the *polygon to point function*; (2) calculating the distance between points using the *point distance function*; (3) picking up the *k*-nearest source zones by sorting the distance table for one source zone to build the matrices; (4) following steps 4–6 of the global regression dasymetric method to get the adjusted density for each residential class for one source zone; (5) repeating step 3 and step 4 until the density for each residential class within each source zone is obtained; and (6) following steps 7–10 of the global regression dasymetric method to get the final interpolated population for each target zone. For the topological contiguity–based neighbor selection, the algorithm is a little more complex and we coded this part mainly using the ArcObjects *IndexQuery function* to find each source zone's adjacent neighbors.

One of the key tasks in implementing the regression dasymetric method is to code the matrix algebra functions. Previous developers (Schneider et al., 2006) turned to commercial statistical packages such as Matlab through loose coupling techniques to accomplish this task. This obviously requires users of this earlier toolbox to have access to an external third-party software system. In this study, we implemented a Gauss-Jordan elimination algorithm to conduct the matrix inversion operation using VB.NET. This eliminates the need for third party software systems, reduces execution times, and improves overall software reliability. Our current system incorporates a maximum of three classes into the regression dasymetric method. If there is a need to use more than three classes for the control zones, the tool can be further expanded without altering its fundamental structure.

Error Measures

Four commonly used error measures were incorporated in the areal interpolation extension: the mean absolute percentage error (MAPE), value-weighted mean

absolute percentage error (VWMAPE), root mean square error (RMSE), and adjusted root mean square error (ARMSE). The detailed equations to calculate these measures can be found in Zhang and Qiu (2011).

Installation, User Interface, and Execution

To use the areal interpolation toolbox in ArcGIS, it needs to be registered as an extension within the ArcGIS system. Instructions (including the registering command) are provided along with the extension's executable files from the aforementioned website. Once installed, the areal interpolation extension toolbar has eight tools: Add Data, Area-Weighting Method, Pycnophylactic Method, Dasymetric Method, 3-Class Regression Dasymetric Method, Error Analysis, Help, and Copyright (Fig. 5). Clicking each tool will bring forth a corresponding dialog box.

The Add Data dialog box lets users bring in the source zones and target zones, as well as raster or vector ancillary data. After adding the data, areal interpolation can be conducted by selecting one of the four algorithms. For the Area-Weighting Method, users need to specify the source and target zones (by default they are the same as those set by the Add Data tool), the numeric population field to be interpolated, and a work space for the storage of intermediate and output files. If a raster-based approach is used, the cell size must also be set. After setting parameters, users then press the "Interpolate" button to initiate the execution and the final interpolated value will be automatically added to the attribute table of the target zone. For the Pycnophylactic Method, the user interface is similar to the Area-Weighting Method, except that users need to define the window size (usually a positive odd number), which by default is 3. For the Binary Dasymetric Method, users need to specify the ancillary data in addition to the required parameters, which are similar to those for the area-weighting method. For the 3-Class Regression Dasymetric Method, when the local regression preference is chosen, users need to define the number of neighbors (i.e., k) for raster data or lags (i.e., order of adjacency) for vector data. If the reference population of the target zones is available, the areal interpolation errors can be reported and stored as an ASCII file by using the Error Analysis dialog box. All the raster-based implementation requires to be defined by users is the cell size, an important parameter that impacts the interpolation accuracy and efficiency.

CASE STUDY

To demonstrate the utility of the developed Areal Interpolation Extension for ArcGIS, a case study in Collin County, Texas was conducted. Located on the north edge of the Dallas/Fort Worth Metroplex, Collin County is one of most rapidly growing suburban areas in the country, with a total population increase of more than 50% from 2000 to 2010. For areal interpolation, the 85 census tracts in the county were used as source zones, while 32 ZIP code tabulation areas (ZCTA) were used as target zones (Fig. 6).

To perform dasymetric-based areal interpolations, ancillary data are needed. For the binary dasymetric method, land use data from the North Central Texas Council of Govenments was employed with the land use types grouped into two classes, populated and unpopulated areas. For the 3-class regression dasymetric model, the same

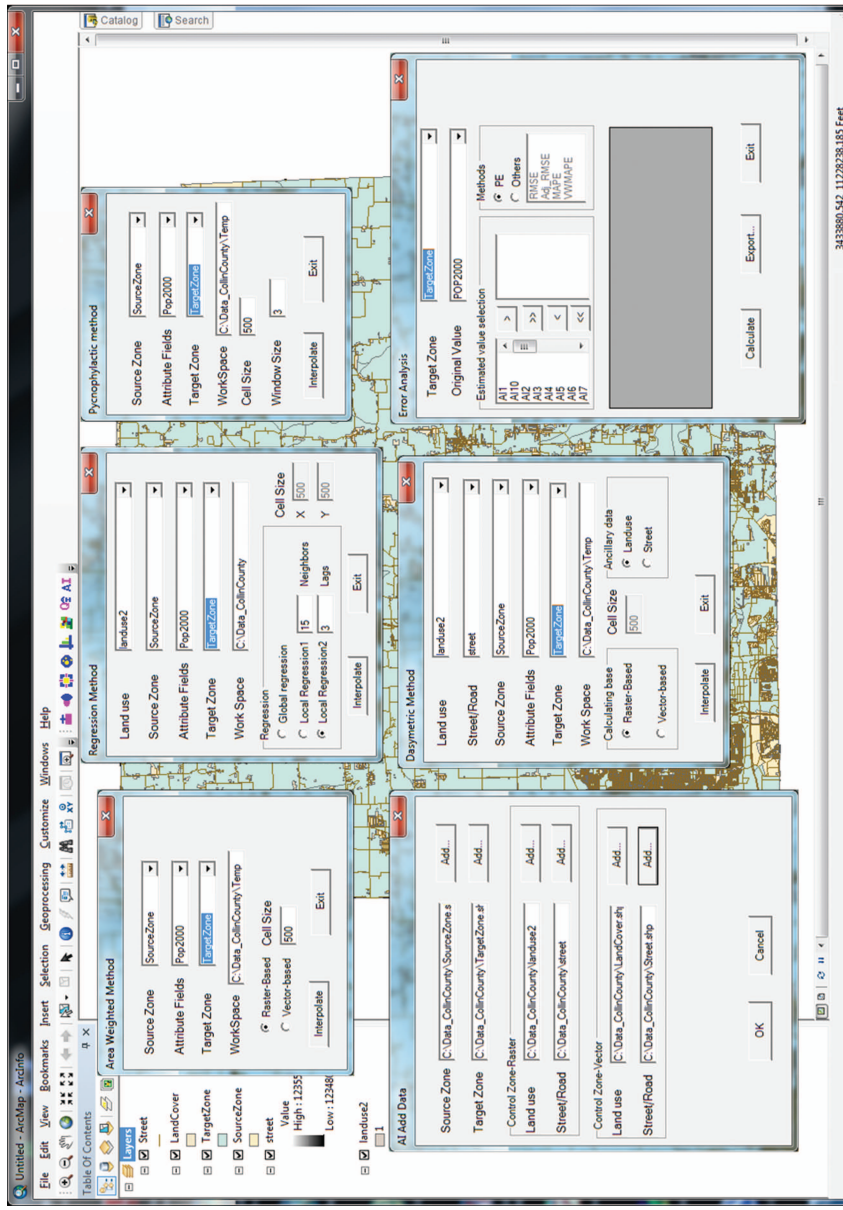

Fig. 5. Interfaces of the areal interpolation toolbox in ArcMap.

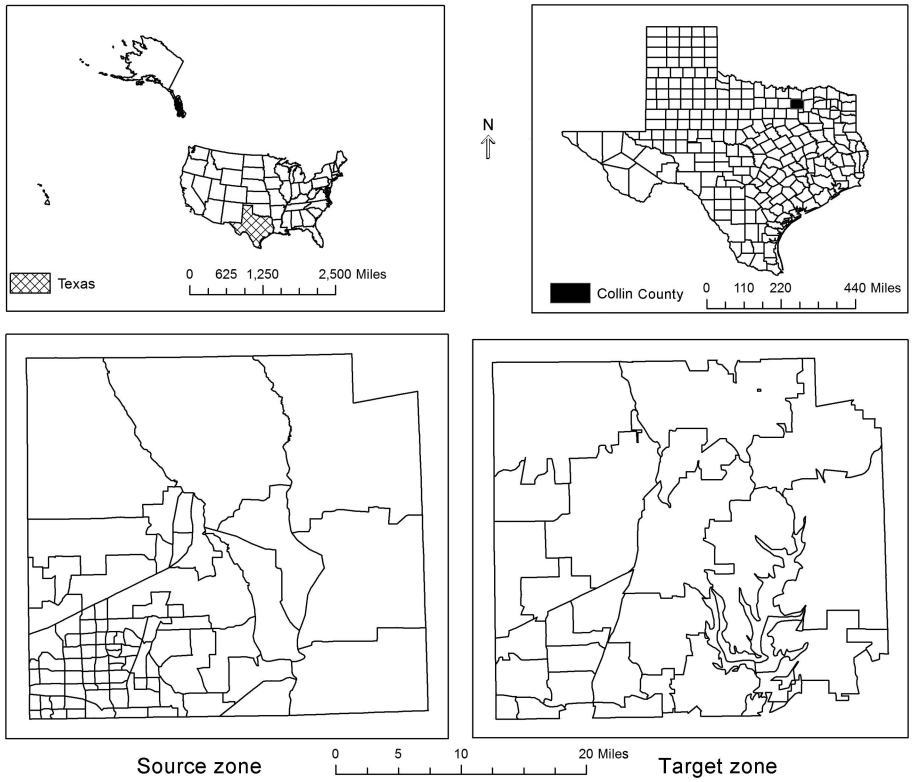

Fig. 6. Test area, Collin County, Texas, and related data.

data were regrouped into nonresidential, low-density residential, and high-density residential classes. For the network length–based binary dasymetric approach, 31,951 road segments extracted from TIGER/Line network data were used as ancillary data.

Each of the four areal interpolation methods incorporated in the Areal Interpolation Extension were used to transform the population data from census tracks to ZIP code regions, and each transformation was evaluated using each of the toolbox's four accuracy assessment measures. The resulting interpolation errors and computation time of each algorithmic implementation are shown in Table 1 and Figure 7 for the purposes of comparison.

From Table 1 and Figure 7, we note that:

1. The network length binary dasymetric method with the vector implementation generated the most accurate results, followed by its raster counterpart. This agrees with the results from Xie (1995), Reibel and Bufalino (2005), and Hawley and Moellering (2005), all of whom indicated that the road network algorithm was in general the most accurate technique for population interpolation.

2. The binary dasymetric method with land use as ancillary data produced better results than the 3-class dasymetric approach, which is a surprise because a

Table 1. Total Errors of Each Implementation

Algorithm[a]	MAPE (pct.)	VWMAPE (pct.)	RMSE	ARMSE
AWTR	13.50	7.20	2760	0.18
AWTV	13.50	7.10	2731	0.18
PYCR	13.80	7.50	2718	0.18
BD2R	9.80	7.00	2415	0.14
BD1R	10.40	5.70	1813	0.15
BD2V	11.10	7.40	2201	0.16
BD1V	8.80	5.90	1767	0.12
GRDR	11.90	10.00	2945	0.15
LDRD	10.30	7.70	2494	0.14
LDRT	11.60	9.60	2828	0.15

[a] AWTR = area-weighting method (raster-based); AWTV = area-weighting method (vector-based); PYCR = pycnophylactic method (raster-based); BD2R = binary dasymetric method using land use data (raster-based);
BD1R = binary dasymetric method using street data (raster-based); BD2V = binary dasymetric method using land use data (vector-based); BD1V = binary dasymetric method using street data (vector-based); GRDR = global regression dasymetric method using 3-class land use data (raster-based); LDRD = local regression dasymetric method using 3-class land use data and selecting 16 nearest neighbors based on the distance between each source zone's centroid; LDRT = local regression dasymetric method using 3-class land use data and selecting neighbors based on third-order topological relationship (i.e., first-, second-, and third-order neighbors were used).

more detailed division of the ancillary data is used in the 3-class dasymetric approach, which should have produced superior results. Langford (2006) also found that a 3-class regression dasymetric model did not always prove advantageous over a binary model. Pending further studies, this must be attributed to the specific characteristics of the data sets investigated rather than to inherent features of the methods.

3. Langford (2006) also found that using local regression could possibly improve the performance of the 3-class regression dasymetric approach by better incorporating spatial non-stationarity. A similar result was also obtained in our case study, where local regressions achieved a better areal interpolation than the global regression. This is likely due to the existence of spatial non-stationarity in the relationship between population and land use classes in this study area.

4. The algorithms using ancillary data outperformed those without ancillary data, but were less computationally efficient in general. Similar accuracies were obtained from each of the three interpolation implementations that did not use ancillary data.

5. The raster-based binary dasymetric methods with the default cell size (500 feet) consumed less computation time than their vector counterparts, an

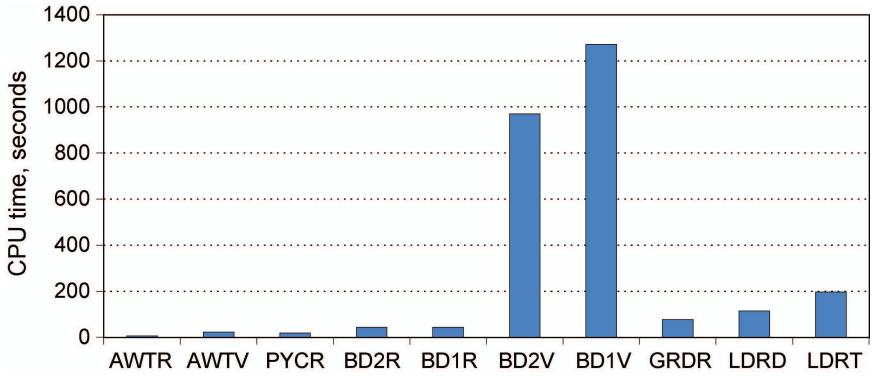

Fig. 7. Computational cost of each implementation.

inclusive statement that is likely cell size dependent. Because the vector data used in this case study consisted of a great number of highly detailed polygons and/or lines, and vector-based implementations rely on overlay operations that involve complex topological reconstruction, the computational expense is apparently very high. Raster implementations, on the other hand, may be computationally less expensive and more stable than their vector counterparts, but with comparable accuracy if an appropriate cell size is chosen. When a large volume of vector data is involved in an areal interpolation application, users are encouraged to consider the tradeoff between efficiency and accuracy before determining which algorithm and implementation to use.

CONCLUSIONS

To bridge the gap between the abundance of areal interpolation methods in the literature and the scarcity of their application in the real world, a set of tools implementing several established areal interpolation methods was developed in the form of an extension to the most commonly used ArcGIS software in this study. Detailed implementation procedures in both vector and raster GIS environments were presented with a case study of residential population interpolation in a suburbanized area conducted to demonstrate usage of the tools. It also provides a systematic comparison of the various algorithms and implementations incorporated into the extension based on the several error measures available. To the best of our knowledge, this paper is the first comprehensive treatment of areal interpolation software development with detailed algorithm descriptions and implementation procedures. The areal interpolation extension presented in this paper has the following advantages:

1. It integrates four commonly used algorithms (area-weighting, pycnophylactic, binary dasymetric, and 3-class regression dasymetric methods) and a total of 10 implementation techniques into a single, easy-to-use system. The multiple implementations provide general users with more alternatives to accommodate their individual preferences.

2. The 3-class regression dasymetric algorithm is implemented without using any external statistical packages. Compared to previous systems that relied on external resources, this feature makes the extension appealing to users who do not want to purchase a license for expensive statistical software for just areal interpolation applications.

3. The 3-class local regression dasymetric methods have two different neighbor selection schemes, the source zone's topological contiguity and the distance-based *k*-nearest neighbors, an attractive feature to deal with potential spatial non-stationarity in the data.

4. The extension incorporates four commonly used error measures (MAPE, VWMAPE, RMSE, and ARMSE), making it possible for users to evaluate interpolation performance and conduct cross-comparison of different algorithms and implementations.

5. A friendly graphical user interface is provided for general users to perform areal interpolation by a few simple button clicks, without having to deal with the complexities of the underlying algorithms and their implementation details. It has the potential to encourage more GIS users to adopt areal interpolation in real-world applications.

6. Finally, for researchers who are interested in developing their own areal interpolation algorithms, this extension (along with the established algorithms, the error measures, and the testing data included) provides them with a set of benchmarks to compare against. We also hope that the implementation details described in this paper will be helpful for those who would like to program their own areal interpolation tools.

REFERENCES

Bloom, M. L., Pedler P. J., and G. E. Wragg, 1996, "Implementation of Enhanced Areal Interpolation Using Mapinfo," *Computers & Geoscience*, 22:459–466.

Deichmann, U, 1996, *A Review of Spatial Population Database Design and Modelling*, Santa Barbara, CA: National Center for Geographic Information and Analysis, Technical Report 96-3.

Eicher, C. and C. A. Brewer, 2001, "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation," *Cartography and Geographic Information Science*, 28:125–138.

Fisher, P. F. and M. Langford, 1995, "Modelling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation. *Environment and Planning A*, 27:11–24.

Fisher, P. F. and M. Langford, 1996, "Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping," *Professional Geographer*, 48:299–309.

Flowerdew, R., 1988, *Statistical Methods for Areal Interpolation: Predicting Count Data from a Binary Variable,* Newcastle upon Tyne, UK: Northern Regional Research Laboratory, Research Report, No. 15.

Flowerdew, R., Green, M., and E. Kehris, 1991, "Using Areal Interpolation Methods in Geographic Information Systems," *Papers in Regional Science*, 70:303–315.

Goodchild, M. F., Anselin, L., and U. Deichmann, 1993, "A Framework for the Areal Interpolation of Socioeconomic Data," *Environment and Planning A*, 25:383–397.

Goodchild, M. F. and N. S.-N. Lam, 1980, "Areal 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 Long-term Change in England and Wales," in *Proceedings of the Fifth International Conference on Geocomputation*, Greenwich, UK, August 23–25.

Gregory, I. N., 2002, "The Accuracy of Areal Interpolation Techniques: Standardizing 19th and 20th Century Census Data to Allow Long-Term Comparisons," *Computers, Environment, and Urban Systems*, 26:293–314.

Gregory, I. N. and P. S. Ell, 2006, "Error-Sensitive Historical GIS: Identifying Areal Interpolation Errors in Time-Series Data," *International Journal of Geographical Information Scie*nce, 20:135–152.

Hawley, K. and H. Moellering, 2005, "A Comparative Analysis of Areal Interpolation Methods," *Cartography and Geographic Information Science*, 32:411–423.

Harvey, J. T., 2002a, "Estimating Census District Populations from Satellite Imagery: Some Approaches and Limitations," *International Journal of Remote Sensing*, 2:2071–2095.

Harvey, J. T., 2002b, "Population Estimation Models Based on Individual TM Pixels," *Photogrammetric Engineering and Remote Sensing*, 68:1181–1192.

Holt, J. B., Lo,C. P., and T. W. Hodler, 2004, "Dasymetric Estimation of Population Density and Areal Interpolation of Census Data," *Cartography and Geographic Information Science*, 31:103–121.

Kyriakidis, P. C., 2004, "A Geostatistical Framework for Area-to-Point Spatial Interpolation," *Geographical Analysis*, 36:259–289.

Kyriakidis, P. C. and E.-H. Yoo, 2005, "Geostatistical Prediction and Simulation of Point Values from Areal Data," *Geographical Analysis*, 37:124–151.

Lam, N. S.-N., 1983, "Spatial Interpolation Methods: A Review," *The American Cartographer*, 10:129–149.

Langford, M., 2006, "Obtaining Population Estimates in Non-census Reporting Zones: An Evaluation of the 3-Class Dasymetric Method," *Computers, Environment and Urban Systems*, 30:161–180.

Langford, M., 2007, "Rapid Facilitation of Dasymetric-Based Population Interpolation by Means of Raster Pixel Maps," *Computers Environment and Urban Systems*, 31:19–32.

Langford, M., Maguire, D. J., and D. J. Unwin, 1991, "The Areal Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Framework," in *Handling Geographical Information*, Masser, I. and M. Blakemore (Eds.), London, UK: Longman, 55–77.

Langford, M. and D. J. Unwin, 1994, "Generating and Mapping Population Density Surfaces within a Geographical Information System," *The Cartographic Journal*, 31:21–26.

Liu, X. H., Kyriakidis, P. C., and M. F. Goodchild, 2008, "Population-Density Estimation Using Regression and Area-to-Point Residual Kriging," *International Journal of Geographical Information Science*, 22:431–447.

Markoff, J. and G. Shapiro, 1973, "The Linkage of Data Describing Overlapping Geographical Units," *Historical Methods Newsletter*, 7:34–46.

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

Mennis, J. and T. Hultgren, 2006, "Intelligent Dasymetric Mapping and its Application to Areal Interpolation," *Cartography and Geographic Information Science*, 33:179–194.

Reibel, M. and M. E. Bufalino, 2005, "Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems," *Environment and Planning A*, 37:127–139.

Sadahiro, Y., 1999, "Accuracy of Areal Interpolation: A Comparison of Alternative Methods," *Journal of Geographical Systems*, 1:323–346.

Sadahiro, Y., 2000, "Accuracy of Count Data Estimated by the Point in Polygon Method," *Geographical Analysis*, 32:64–89.

Schneider, P., Kyriakidis, P. C., and M. F. Goodchild, 2006, "Improving Spatial Support Interoperability in GIS Using Geostatistics: An Areal Interpolation Toolbox," in *Proceedings, GIScience 2006*, Münster, Germany, September 20–23.

Schroeder, J. P., 2007, "Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data," *Geographical Analysis*, 39:311–335.

Turner, A. and S. Openshaw, 2001, "Dissaggregative Spatial Interpolation," paper presented at GISRUK 2001, Glamorgan, UK, April 18–20.

Tobler, W., 1979, "Smooth Pycnophylactic Interpolation for Geographical Regions," *Journal of the American Statistical Association*, 74:519– 530.

Wu, C. and A. T. Murray, 2005, "A Cokriging Method for Estimating Population Density in Urban Areas," *Computers, Environment, and Urban Systems*, 29:558–579.

Xie, Y., 1995, "The Overlaid Network Algorithms for Areal Interpolation Problem," *Computers, Environment, and Urban Systems*, 19:287–306.

Yuan, Y., Smith, R. M., and W. F. Limp, 1997, "Remodeling Census Population with Spatial Information from Landsat TM Imagery," *Computers, Environment, and Urban Systems*, 21:245–258.

Zhang, C. and F. Qiu., 2011, "A Point-Based Intelligent Approach to Areal Interpolation," *Professional Geographer*, 63:262–276.