# A Quantile Regression Approach to Areal Interpolation

Robert G. Cromley, Dean M. Hanink, and George C. Bentley

*Department of Geography, University of Connecticut*

Areal interpolation has been developed to provide attribute estimates whenever data compilation or an analysis requires a change in the measurement support. Over time numerous approaches have been proposed to solve the problem of areal interpolation. Quantile regression is used in this study as the basis of the areal interpolator because it provides estimates conditioned on local parameters rather than global ones. An empirical case study is provided using a data set in northern New England. Land cover data, provided by the National Oceanic Atmospheric Administration, derived from remotely sensed images for 2001 captured by the LANDSAT Thematic Mapper at a resolution of 30 × 30 meters, are used for the ancillary variables for the regression model. The utility of quantile regression as an areal interpolation method is evaluated against simple averages, areal weighting, dasymetric interpolation, and ordinary least squares and spatial regression methods. For the empirical data set used in the study, results show that quantile regression was a better interpolator for the given data set but that binary dasymetric interpolation was a close second. These results were only for one data set and further evaluation is necessary before more general conclusions can be made. *Key Words: areal interpolation, dasymetric mapping, linear programming, quantile regression.*

面插值技术被发展出来提供在需要改变测量支持时数据汇编或分析的属性估计。随着时间的推移，许多方法已被提出来解决面插值的问题。本研究用分位数回归作为面插值器的基础，因为它提供了以局部参数而不是整体参数为条件的估计。本研究使用新英格兰北部的一组数据，提供一个实证案例研究。由国家海洋大气管理局提供的，从由陆卫专题成像仪获取的，30×30 米分辨率的，2001 年的遥感图像所提取的土地覆盖数据，被用作回归模型的辅助变量使用。把分位数回归用作一种面插值方法进行评估，并与简单平均，加权面，分区密度插值，以及普通最小二乘法和空间回归方法相比较。对于在研究中使用的实证数据集，结果表明分位数回归对给定的数据集是一个更好的的插值器，但二进制分区密度插值紧随其后。这些结果只基于一个数据集，在作出更一般的结论之前，需要进一步的评估。关键词：面插值，分区密度映射，线性编程，分位数回归。

La interpolación areal ha sido desarrollada para proveer estimativos de atribución cuandoquiera que la compilación de datos o un análisis requiera un cambio en el apoyo de la medición. Con el tiempo se han propuesto numerosos enfoques para resolver el problema de la interpolación areal. La regresión de cuantiles se utiliza en este estudio como la base de la interpolación areal porque esta proporciona estimativos más condicionados en parámetros locales que en los globales. Se presenta un estudio empírico de caso en el que se utilizó un conjunto de datos referidos a la parte norte de Nueva Inglaterra. Se usaron datos de cobertura del suelo suministrados por la Administración Nacional Oceánica Atmosférica – que habían sido derivados en 2001 de imágenes de percepción remota captadas por el Mapeador Temático LANDSAT a una resolución de 30 × 30 en metros – para las variables auxiliares del modelo de regresión. La utilidad de la regresión de cuantiles como método de interpolación areal es evaluada frente a promedios simples, peso por área, interpolación dasimétrica, mínimos cuadrados ordinarios y métodos de regresión espacial. Para el conjunto de datos empíricos usados en el estudio, los resultados muestran que la regresión de cuantiles fue un mejor interpolador para el conjunto de datos dado, aunque la interpolación binaria dasimétrica estuvo muy cerca en el segundo lugar. Estos resultados se refieren a un solo conjunto de datos, por lo que se hace necesaria una evaluación adicional antes de que se puedan formular conclusiones más generales. *Palabras clave: interpolación areal, mapeo dasimétrico, programación linear, regresión de cuantiles.*

In spatial analysis, data often are collected using one measurement support, but the analysis to be performed might use a different support. The widespread use of geographic information systems (GIS) has increased the need to change support because these systems integrate data from different sources into a common database. The change of support problem is then concerned with inferences or estimates of values at or for locations different from locations at which values have been observed (Gelfand, Zhu, and Carlin 2001).

*Annals of the Association of American Geographers*, 102(4) 2012, pp. 763–777 
Initial submission, August 2010; revised submissions, December 2010 and March 2011; final acceptance, April 2011
Published by Taylor & Francis, LLC.

Spatial interpolation procedures are used to solve the change of support problem, but alternate forms of spatial interpolation are needed because there are fundamental differences in the types of spatial data involved. Geographic data have been classified as either being field-based or object-based data (Worboys and Duckham 2004). An attribute of interest might vary over space as a continuous field for which measurements can be made anywhere. The field-based model views geographic data as collections of spatial distributions. The alternative view is that space is populated with objects (entities). In the object-based model, space is a void except where objects are located and measurement occurs only at these locations. To protect privacy and for other reasons, individual objects are often also aggregated into areal units that act as collector zones. The values for these collector zones are expressed either as spatially extensive counts such as total population or as spatially intensive rates or averages such as population density (Goodchild and Lam 1980).

*Point* interpolation is normally used to change support for field data. These interpolators include inverse weighted distance, ordinary kriging, spline functions, and trend surface analysis (Lam 1983). *Areal* interpolation, on the other hand, is used to change support for entity data that have been aggregated into areal units. Areal interpolation generally refers to procedures for transferring values from one partition of space to a different one (Goodchild and Lam 1980; Lam 1983). The observed geography is referred to as the *source layer* and the inferred geography is the *target layer*. Subsequent spatial analyses are performed using the target zones. Whereas point interpolation is used to estimate values for a wide range of field attributes such as elevation, temperature, and pollutants, areal interpolation has focused mainly on population estimates. The purpose of this study is to investigate the use of quantile regression as an areal interpolator of population.

## Background

A basic characteristic of areal interpolation that distinguishes it from point interpolation is its volume-preserving property (Lam 1983). A seminal contribution to areal interpolation methods was developed by Tobler (1979) as an outgrowth of a procedure to create isopleth maps of population density of the United States from polygon data. Pycnophylatic (volume-preserving) interpolation was used to transform a polygonal prism surface to a smooth, rasterized, continuous surface. Because the volume of the isopleth surface had the same volume as the polygonal prism, Tobler noted that the raster units could also be reaggregated into target zones different from the original source polygons. In this transfer of units, no population count would be lost or gained. Because points are the geometric dual of areas (White 1979), point interpolation procedures such as kriging, least-squares fitting with splines, or distance-weighted least squares can be used for areal interpolation (Lam 1983; Xie 1995). The conservation of total surface volume must then be maintained, however, by applying a scaling step in which the initial target unit estimates are multiplied by the ratio of observed source unit values to inferred source units values based on the target estimates.

Over time, a plethora of procedures directly intended for areal interpolation have been developed and a number of ways for classifying these procedures exist. Individual methods might fall into different categories within various classifications. One classification is based on how the need to estimate unknown values for areal units arises from four different geometric situations associated with creating and manipulating areal units of different map layers (Mrozinski and Cromley 1999). The first need is for the "missing data" problem that occurs when one or more unknown values exist for units within the same data layer but do not involve a change of support. This problem can be solved using cartographic methods based on spatial proximity (Tobler and Kennedy 1985) and statistical methods using other attribute information to generate maximum likelihood estimates (Griffith, Bennett, and Haining 1989).

The second need is for solving the "alternative geography" problem mentioned earlier. This involves the transfer of unit data values for a known source geography of areal units to the units of an incongruent target geography at the same scale. A special form of the alternative geography problem is when the target geography is hierarchically nested within the source geography so that the intersection of the two geographies is congruent to the target geography. This normally occurs when the target geography represents a change in scale. The third and fourth situations requiring areal interpolation again result from polygon overlay operations but the target layer is now the intersection layer rather than one of the two polygon input layers. The difference between the third and the fourth situation is that values are known for only one input layer in the third situation, whereas values are known for both input layers in

the fourth situation. The former requires that the volume of only one input layer be preserved and the latter requires that volume must be preserved with respect to both input layers (Mrozinski and Cromley 1999).

Another classification is based on the underlying data model—vector versus raster. Usually the source layer uses the vector data model, but the target layer can use different data models including raster, vector, or triangular irregular networks (TIN) data models. The original vector-based method is areal weighting, in which values are estimated as averages proportionally weighted by the area of each intersection polygon to the area of the source polygon that contains it (Goodchild and Lam 1980). Areal weighting's major weakness is the assumption that the attribute being interpolated is uniformly distributed within each source zone. Even though it performs poorly in most evaluations (Fisher and Langford 1995; Langford 2006), it is frequently used because of its low data requirements.

Even if the final target layer uses a vector data model, vector-to-raster and raster-to-vector operations can be used as intermediary steps and the interpolation itself is calculated over a grid. Point interpolation and Tobler's pycnophylatic method are examples of this approach. Kyriakidis (2004) has formulated areal interpolation as a geostatistical area-to-point kriging problem and Yoo, Kyriakidis, and Tobler (2010) have shown how geostatistical methods are by their design pycnophylatic. Ensuring the volume-preserving property requires solving the problem with quadratic programming methods, however (Yoo, Kyriakidis, and Tobler 2010). Tobler's pycnophylactic procedure using regular grids also has been extended to a surface representation based on a TIN model by Rase (2001). Other raster-based approaches use the centroid of a source region as the location of average density within the region and estimate the population density of other cells based on an inverse distance weighting (see Bracken and Martin 1989; Martin 1989; Bracken 1991; Martin and Bracken 1991). Each raster cell receives some population from all centroids within its distance window similar to a singly constrained spatial interaction model. Although the purpose is often to generate local surfaces of population density, the cells can be reaggregated into new target zones. When doing so, a major concern is the size of the raster cell in relation to the size of the target unit (Martin 1996).

Another factor that increases the use of raster models is the use of ancillary data for improving estimation accuracy, especially remotely sensed satellite imagery that is raster based. *Simple* areal interpolation methods, such as Tobler's pycnophylatic method, areal weighting, and ordinary kriging, 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 correlates to the attribute being estimated (Langford, Maguire, and Unwin 1991; Okabe and Sadahiro 1997; Langford 2006). Intelligent interpolators are based on the principles of dasymetric mapping (Wright 1936) in which additional control variables are used to identify zones having different population densities. Within intelligent methods, there is usually a difference between those that are logical extensions of Wright's (1936) original dasymetric mapping (see Langford, Maguire, and Unwin 1991; Fisher and Langford 1996; Eicher and Brewer 2001; Mennis 2003, 2009; Holt, Lo, and Hodler 2004; Langford 2006; Mennis and Hultgren 2006; Tapp 2010) versus those that are based on a statistical technique (see Flowerdew and Green 1989, 1991, 1994; Green 1990; Flowerdew, Green, and Kehris 1991; Goodchild, Anselin, and Deichmann 1993; Bloom, Pedler, and Wragg 1996; Mugglin and Carlin 1998; Mugglin et al. 1999; Mugglin, Carlin, and Gelfand 2000; Gelfand, Zhu, and Carlin 2001; Reibel and Agrawal 2007; Merwin, Cromley, and Civco 2009). Different correlates such as remotely sensed land cover categories (Fisher and Langford 1996; the most commonly used one), road length and road category (Xie 1995), and road buffer areas (Mrozinski and Cromley 1999) have been used as control variables. Recently, cadastral units have been used in urban areas as a control variable (Maantay, Maroko, and Herrmann 2007) and Tapp (2010) has even used parcel boundaries to estimate population in rural areas in counties that have digitized their cadastral data.

The simplest dasymetric interpolation method is the binary dasymetric method (Fisher and Langford 1996) in which there are only populated and unpopulated areas within each source and target zone. Rather than using the total area in calculating a population density for each source zone, the area of the populated portion is used to calculate the population density. Initially polycategorical dasymetric methods either assigned a fixed proportion of the total population to each land category (see Eicher and Brewer 2001) or selected a sample of target zones that are completely homogeneous with respect to a specified land category (see Mennis 2003). Mennis and Hultgren (2006) relaxed these conditions to permit the analyst first to set a threshold level and then to select those source zones for a category whose proportion is equal to or exceeds that threshold to be used in calculating the density level of that category.

In statistical methods, a functional relationship is established between the ancillary data and the values being estimated, usually using some form of regression. Besides the global nature of ordinary least squares (OLS) regression, Langford, Maguire, and Unwin (1991) noted several problems with standard ordinary least squares regression when used as the basis for areal interpolation: (1) regression models normally have an intercept, which means that even when the values for all independent variables are zero, population can still exist; (2) because regression coefficients are not restricted to be nonnegative, the estimated value for certain units could be negative; and (3) the regression coefficients are global over the entire study region. The first issue can easily be overcome by forcing the regression line through the origin, an option that is part of most regression software. Several approaches also have been proposed to overcome the negative coefficient and negative estimates. The presence of spatial autocorrelation in OLS errors can affect the values of the coefficients themselves, even with respect to their sign (Pace and Gilly 1997), because the problem is effectively the result of a missing explanatory variable in the regression equation. Spatial error models (SEMs) directly account for spatial autocorrelation in errors in their specification (Anselin 2003) and so are often used as an alternative global regression model in the analysis of spatial data such as population densities. SEMs, though, can still have negative coefficents. Instead, Flowerdew and Green (1989) suggested using Poisson regression, which is theoretically preferable for modeling counts and negative population estimates are precluded. Moxey and Allanson (1994) avoided negative regression coefficients by using inequality restricted least squares. This optimization technique requires quadratic programming to ensure that all coefficients are nonnegative. Yuan, Smith, and Limp (1997) proposed adding a scalar to each estimated population value equal to the lowest negative estimate. The initial estimates are then scaled to ensure the pycnophylatic property for all source zones.

The last issue for intelligent areal interpolation models is related to the problem of heterogeneity in which the relationship between a set of independent variables and the outcome (i.e., population) differs over subsets of the data. The global nature of most regression models limits their effectiveness with respect to heterogeneity without some extensions. Binary dasymetric areal interpolation solves the heterogeneity problem by restricting the subset to an individual observation. Mennis and Hultgren (2006) handled heterogeneity for polycategorical dasymetric models by allowing the density values of ancillary classes to be set within different region zones. Similarly, Yuan, Smith, and Limp (1997) used regional regressions in which the observations were partitioned by administrative units such as counties, but Langford (2006) noted that this partitioning is arbitrary with respect to the underlying spatial distribution.

The global models thus have limited accuracy; Fisher and Langford (1995) reported that statistical methods perform less satisfactorily than dasymetric areal interpolation. Any statistical method can easily be made local by applying a scaling step to enforce the volume-preserving property on the source zones, which Langford (2006) called a hybrid model, but this does not directly address heterogeneity in the relationships, especially when a variety of different additives can be used in the scaling step. The next section describes an alternative regression model as the basis of areal interpolation in which changing relationships among variables can be identified among subsets of observations.

## A Quantile Regression Model

The three issues associated with regression models outlined earlier do not exist with quantile regression. Quantile regression (QR) is a form of weighted regression, as is geographically weighted regression (GWR; Fotheringham, Brunsdon, and Charlton 1992) but the weights are not spatially determined. QR is useful because the homoscedasticity assumption of OLS regression often fails, and focusing only on central tendency cannot capture other trends in the distribution of the dependent variable (Hao and Naiman 2007). QR is based on minimizing the sum of absolute deviations about the regression line rather than minimizing the sum of the squared deviations as in OLS regression; as such, it is an extension of conditional median regression. Conditional median regression is more appropriate in situations when dependent variable distributions are skewed (Koenker 2005; Hao and Naiman 2007), as is frequently the case when analyzing population density. QR was developed to investigate relationships beyond the conditional median regression line, and its main advantage over OLS regression is its ability to model data with heterogeneous conditional distributions (Koenker and Bassett 1978). In areal interpolation, each source observation corresponds to a place, and places in the same part of the data distribution can have similar relationships between the dependent and independent variables. For example, in one tail of the distribution places could be mainly urban and thus have a

different relationship between population density and land cover than for observations in other portions of the distribution.

The QR model can be expressed as:

$$y_i = \alpha^{\rho} + \sum_j \beta_j{}^{\rho} x_{ij} + \varepsilon_i{}^{\rho} \quad (1)$$

where $y_i$ is the dependent variable for the ith obsevation, $\alpha^{\rho}$ is the intercept term for the quantile at $\rho$, $\beta_j{}^{\rho}$ is the regression coefficient for the jth independent variable for quantile at $\rho$, $x_{ij}$ is the ith observation for the jth independent variable, $\varepsilon_i{}^{\rho}$ is the error term for the ith observation for the quantile at $\rho$, and $0 < \rho < 1$ indicates the proportion of observations having values below the quantile at $\rho$ (Hao and Naiman 2007). If $\rho = 0.5$, the model would correspond to a conditional median regression. By analyzing the full range of values for $\rho$, one can investigate the effects of covariates on the complete distribution. As such, each covariate coefficient can vary over the range of $\rho$ values.

The quantile regression model is formulated as a linear program and can be solved using the simplex method (see Hadley [1964] or any linear programming text for a discussion of the simplex method). Solving a linear program using the simplex method is also easier than using a quadratic program to solve the inequality restricted least squares regression. Quantile regressions can have an intercept and can permit coefficients with negative values, but these features add more variables to the linear programming formulation. A quantile regression model applied to the problem of areal interpolation is formulated as:

$$\text{Minimize } \sum_i \rho\, \lambda_i{}^{-} + (1-\rho)\lambda_i{}^{+} \quad (2)$$

$$\sum_j \beta_j X_{ij} + \lambda_i{}^{-} + \lambda_i{}^{+} = P_i \text{ for all i} \quad (3)$$

$$\beta_j, \lambda_i{}^{-}, \lambda_i{}^{+} \geq 0; \quad (4)$$

where $\rho$ is the quantile parameter (or weight) that ranges from zero to one, $\beta_j$ is the estimated population density for the jth land cover class, $X_{ij}$ is the amount of the jth land cover class associated with the ith observation, $P_i$ is the population count for the ith observation, $\lambda_i{}^{-}$ is a deviational variable representing the amount of underestimation for the ith observation by the regression model, and, $\lambda_i{}^{+}$ is a deviational variable representing the amount of overestimation for the ith observation by the regression model.

In this model there is no variable representing an intercept term, only the population density value, $\beta_j$, associated with each land cover class is estimated. The objective function minimizes the sum of the residuals about the regression line. In the model, there is one constraint of the form represented by Equation 3 for every observation. In linear programming, a variable can have a nonnegative value only if it is part of the basis (such a variable is called *basic*). A basis is a square matrix having a rank equal to the number of independent constraints. At optimality, therefore, there can be only as many basic variables as there are observations. Also at optimality, at most only one deviational variable associated with an observation will be in the basis because no observation can be simultaneously over- and underestimated. If neither deviational variable associated with an observation at optimality is basic, then the regression line is a perfect fit for that observation. This means that there will be as many observations with a perfect fit as there are basic $\beta_j$ values because the total number of basic variables must equal the number of observations. Every observation will have a perfect fit associated with some value of $\rho$, although if there is more than one basic $\beta_j$ value, some observations will have a perfect fit for more than one value of $\rho$.

Although the weights are not spatial, they can be interpreted as being place based when using quantile regression as an areal interpolator. The quantile regression line that is a perfect fit for an observation or place is chosen as the model for determining the population density values associated with each land cover for that observation or place. By doing so, quantile regression becomes a pycnophylatic interpolator (the total volume of population is preserved and a scaling step is not needed) because the pixel count for land cover category j at the more aggregate scale is the sum of the pixel counts for the same category at the finer scale. Multiplying the finer scale values by the same $\beta_j$ value and then summing those numbers will result in the same population as multiplying the aggregate pixel count by $\beta_j$. If the population values for each land cover category are summed, this number will equal the original population count (the right-hand side value in Equation 3) because the deviational variables both equal zero.

Figure 1 displays a scatterplot of eleven data points for a model with dependent variable Y representing population and one independent variable X. For areal interpolation, X corresponds to an ancillary variable such as land cover. For quantile regressions with no intercept term, the line QR1 corresponds to a model in which $\rho = .09$ and 91 percent of the observations lie above the regression line and the line passes through point A. Similarly, the line QR2 passing through point

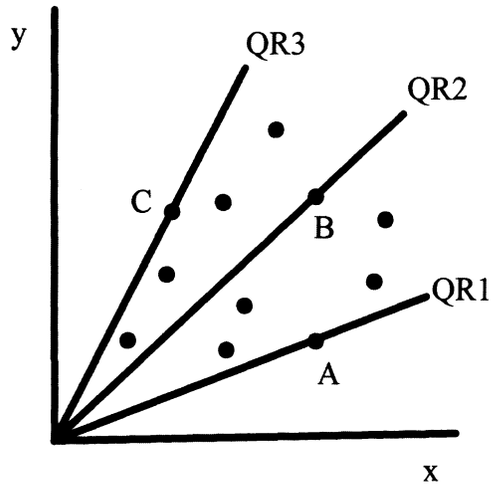


**Figure 1.** An example bivariate data plot with different quantile regression lines.

B corresponds to the median regression model ($\rho = 0.5$) in which half of the points lie above the regression line and half below it. Finally, the line QR3 passing through point C corresponds to the quantile regression in which $\rho = 0.91$ and 91 percent of the observations lie below the regression line. By choosing the appropriate value of $\rho$, a regression line directly passing through each point can be determined. The slope of each line is the population density per unit of X, and that density value is associated with a certain data value. It can be seen from this example that a quantile regression with no intercept and one independent variable would produce the exact same results as the binary dasymetric areal interpolator. The dasymetric areal interpolator then is just a special case of the more general quantile regression model.

In addition, it can be easily shown that any estimator based on a regression with no intercept and a single independent variable with a scaling step will produce results identical to the binary dasymetric areal interpolator. Let *gc* be the global coefficient, $a_i$ be the total area of interest within the ith unit, $P_i$ be the actual population of the ith unit, and $a_{ij}$ be the area of interest within the jth areal subunit within the ith areal unit. The binary dasymetric estimate for the population of the jth subunit is:

$$BDP_{ij} = (a_{ij}/a_i)\, P_i. \tag{5}$$

The initial regression estimate for the population of the jth subunit is:

$$RP_{ij} = gc\ a_{ij}. \tag{6}$$

The scaling factor (SF) to ensure that the actual population for the ith unit is preserved is:

$$SF = P_i \Big/ \left(\sum_j gc\ a_{ij}\right) \text{ or } P_i \Big/ \left(gc \sum_j a_{ij}\right). \tag{7}$$

The denominator of the scaling is the initial estimate of the population of the ith unit. The scaling step then multiplies the initial population estimate by the scale factor so that:

$$SRP_{ij} = gc\ a_{ij} \left(P_i \Big/ \left(gc \sum_j a_{ij}\right)\right). \tag{8}$$

The global coefficient *gc* cancels out and the sum of j subunits equals $a_i$ so that Equation 8 is equivalent to Equation 5. Because the binary dasymetric areal interpolator is much easier to calculate than either a global regression with a scaling step or local quantile regressions, the following empirical analysis does not use any regression models with a single independent variable—the binary dasymetric areal interpolator is their substitute.

## Study Area and Data

In this research, the study area consists of Cumberland County and York County in southern Maine and Rockingham County and Stafford County in neighboring New Hampshire (see Figure 2). In 2000, the four-county study region had a population of 841,946. These counties were chosen because they contain a variety of population densities from dense urban centers such as Portland, Maine, to sparsely populated rural New England woodlands. Given the diversity of settings, one would not necessarily expect a global model to capture accurately the relationship between population density and land cover.

In the first analysis, population counts for the 179 census tracts are used to estimate the population of the block groups nested within the tracts (of the original 182 tracts, 3 with zero population were removed in processing). This examines the problem of using areal interpolation for the issue of changing scales. For 8 tracts of the 179, only one block group was found in the tract. In these situations, the tract was arbitrarily subdivided into two, three, or four subunits; population data at the block level were aggregated to determine the population for each new subunit. The original 598 block groups were expanded to 612 block groups (Figure 3) as a result of this process. In the second analysis, population


**Figure 2.** The four-county study area.

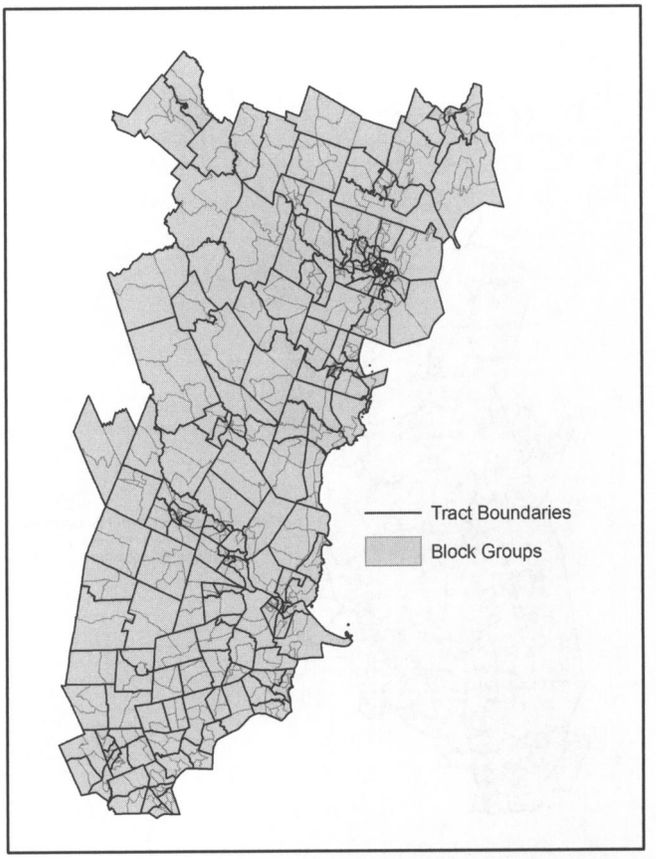


**Figure 3.** The final set of tract boundaries and block groups.

of the census tracts is used to estimate the population of an alternative geography of 179 polygons at the same scale (Figure 4). This examines the problem of using areal interpolation for the second issue of changing the partition at the same scale. The alternative geography of 179 polygons was compiled by dissolving the newly created layer of 612 block groups into this new geography. The actual population of these new polygons is the sum of the population of the block groups that formed each unit. Census 2000 data were used in this study to compile the population and tract and block group boundaries. The population count of persons and the geographic boundary files were downloaded from the U.S. Census Bureau Web site.

The preclassified land cover data used in this study were downloaded from the National Oceanic and Atmospheric Administration's Coastal Change Analysis Program (C-CAP) Web site. This remotely sensed image was captured by the LANDSAT Thematic Mapper in 2001 at a resolution of 30 × 30 meters pixels. The accuracy assessment for the data is 85 percent. The original classification had twenty-one land cover categories present in the study area, but this was reduced to just fourteen categories (Table 1) by eliminating water, open shore, and estuarine-related categories. The data were then imported into ArcGIS 9.3 (ESRI 2009).

**Table 1.** Distribution of land cover categories

| Land cover type | Number of pixels |
|---|---|
| Developed, High Intensity (DEVHINT) | 56,407 |
| Developed, Medium Intensity (DEVMINT) | 219,809 |
| Developed, Low Intensity (DEVLINT) | 424,495 |
| Developed, Open Space (DEVOPEN) | 146,163 |
| Cultivated Crops (CULTCRP) | 114,767 |
| Pasture/Hay (PASTHAY) | 643,133 |
| Grassland/Herbaceous (GRASS) | 59,998 |
| Deciduous Forest (DECFOR) | 1,364,546 |
| Evergreen Forest (EVERFOR) | 1,529,078 |
| Mixed Forest (MIXFOR) | 2,461,982 |
| Scrub/Shrub (SCRSHB) | 395,144 |
| Forested Wetland (FORWET) | 526,437 |
| Shrub Wetland (SHBWET) | 156,371 |
| Barren Land (BARREN) | 55,530 |

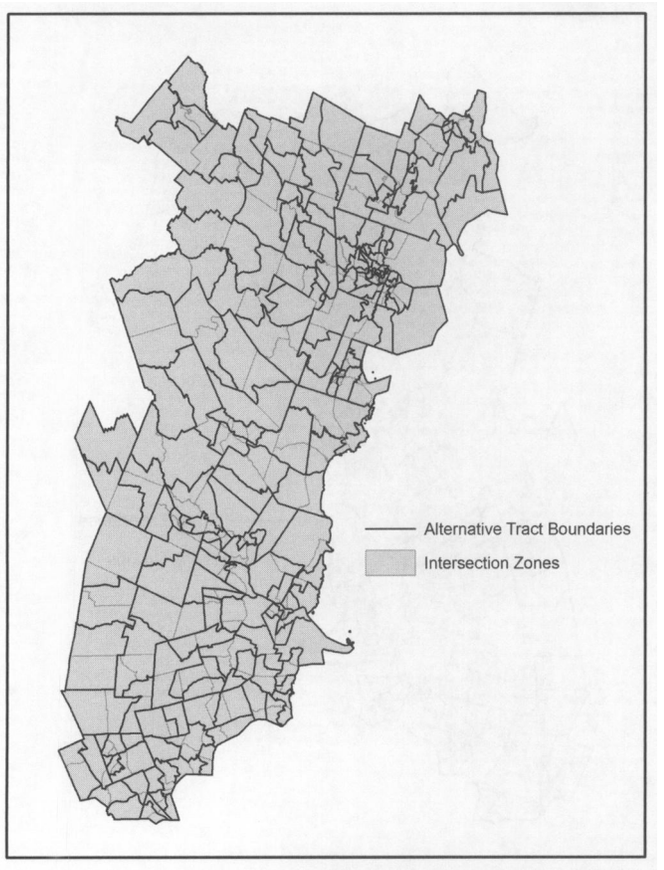


**Figure 4.** The final set of alternate tract boundaries and intersection zones.

Using the study area block group boundary layer, the land cover data were clipped to match the study area. The "Tabulate Area" tool from the Zonal toolbox was then used to calculate the area of the fourteen land cover types. Finally, the area value was divided by 900 square meters to convert the area value into the number of pixels.

## Test Design

As mentioned in the previous section, two analyses are performed—one estimating block group populations from known tract-level populations (changing scales) and the second estimating populations for an alternative tract geography (changing zones). For each analysis, three land cover scenarios are used. In scenario 1 all fourteen land cover categories are used as ancillary data. In scenario 2 the three land cover categories most associated with population were used as ancillary data. At the tract level, among all land cover categories listed in Table 1, the two having the highest Pearson's correlation coefficient with population were DEVMINT (0.65) and DEVLINT (0.52); DEVHINT's correlation coefficient (0.16) was much lower but it was still included because this category includes heavily built-up urban centers. Finally, in scenario 3 these three land cover categories are added together to form a single ancillary variable.

For each analysis, results from quantile regression-based areal interpolation are compared against results from areal weighting, binary dasymetric interpolation, global OLS regression, and global SEM regression (Table 2). Areal weighting is included because it is the most widely used method and other studies use it as the benchmark (Langford 2006). In this study, two OLS and two SEM regressions having different sets of independent variables without an intercept term are used to calculate initial population density coefficients for different land cover categories. These coefficients are then used to estimate populations for block groups and alternative geography polygons. In OLS14 and SEM14, tract population is regressed against the pixel counts of the fourteen land cover types using OLS and SEM regression, respectively (scenario 1). In OLS3 and SEM3, tract population is regressed against the pixel counts of DEVHINT, DEVMINT, and DEVLINT using OLS and SEM regression respectively (scenario 2). A final scaling step is applied to each of these four global regressions to convert them into local estimates. The two quantile regressions with no intercept terms corresponding to scenarios 1 and 2, respectively, are QR14 and QR3. The binary dasymetric interpolator that uses the total pixel count of (DEVHINT + DEVMINT + DEVLINT) as the populated area of interest is the only model associated with scenario 3 (as noted previously, single variable regression models with scaling produce equivalent results).

The accuracy of each interpolation method is evaluated using the root mean square (RMS) error and the

**Table 2.** Global regression coefficients (population density) for selected land cover categories

| Model | DEVHINT | DEVMINT | DEVLINT |
|---|---|---|---|
| OLS14 | –1.4746 | 1.9975 | 1.1678 |
| OLS3 | –1.5184 | 1.7904 | 1.0041 |
| SEM14 | –1.2396 | 1.1588 | 0.7529 |
| SEM3 | –1.5258 | 1.3604 | 0.4437 |

*Note:* DEVHINT = Developed, High Intensity; DEVLINT = Developed, Low Intensity; DEVMINT = Developed, Medium Intensity; OLS = ordinary least squares; SEM = spatial error model.

mean absolute error (MAE). The RMS error is calculated as the square root of the average of the squared difference between the estimated population value and the actual value. The MAE is calculated as the average of the absolute deviation between the estimated and actual values. These summary error measures allow an overall evaluation of the comparative performance of the areal interpolation methods.

## Results

In the first analysis, the differences between the global regression models and the local quantile models are readily apparent. Table 2 presents the regression coefficients for the three main land cover categories (DEVHINT, DEVMINT, and DEVLINT) for the regression models OLS14, OLS3, SEM14, and SEM3. The value for DEVHINT is negative for each of the global regression models. For the quantile model Q14, however, the coefficient for DEVHINT is high (about 11) for quantile values near zero and declines steadily until it reaches zero at the 37th quantile (Figure 5A). Spatially, the quantile values near zero are associated with source zones located in urban areas so the density coefficient for DEVHINT should be higher in these areas; the quantile values near one are associated with rural areas and should have a lower density coefficient for the same land cover category. The coefficient for DEVHINT in model QR3 had a similar pattern. Near a quantile value of zero, the coefficient value was close to 17 and declined until it reached zero at the 36th quantile (Figure 6A). Similarly, the highest coefficient value for DEVMINT among the global regressions was about two, but this coefficient ranged in QR14 from a high of 11 to zero at the 95th quantile (Figure 5B) and in QR3 it ranged from a high of almost 10 to zero at the 96th quantile (Figure 6B). For the developed low-intensity, rural residential land cover DEVLINT the differences were not as dramatic. Its coefficient in the OLS regression was just over one and somewhat under one in the SEM regressions. Likewise for QR14, its coefficient ranged from 0.6 to 1.2 between the 2nd and 97th quantiles (Figure 5C). For QR3, however, the coefficient steadily decreased from a high of 2.3 near a quantile value of zero to about 0.4 at a quantile value of one (Figure 6C). These results demonstrate the hetereogeneity in parameter estimates for these two important land cover categories. Overall, regression coefficients in the quantile regressions reflect the pattern of urban and rural differences in the density coefficients for the same land cover categories.

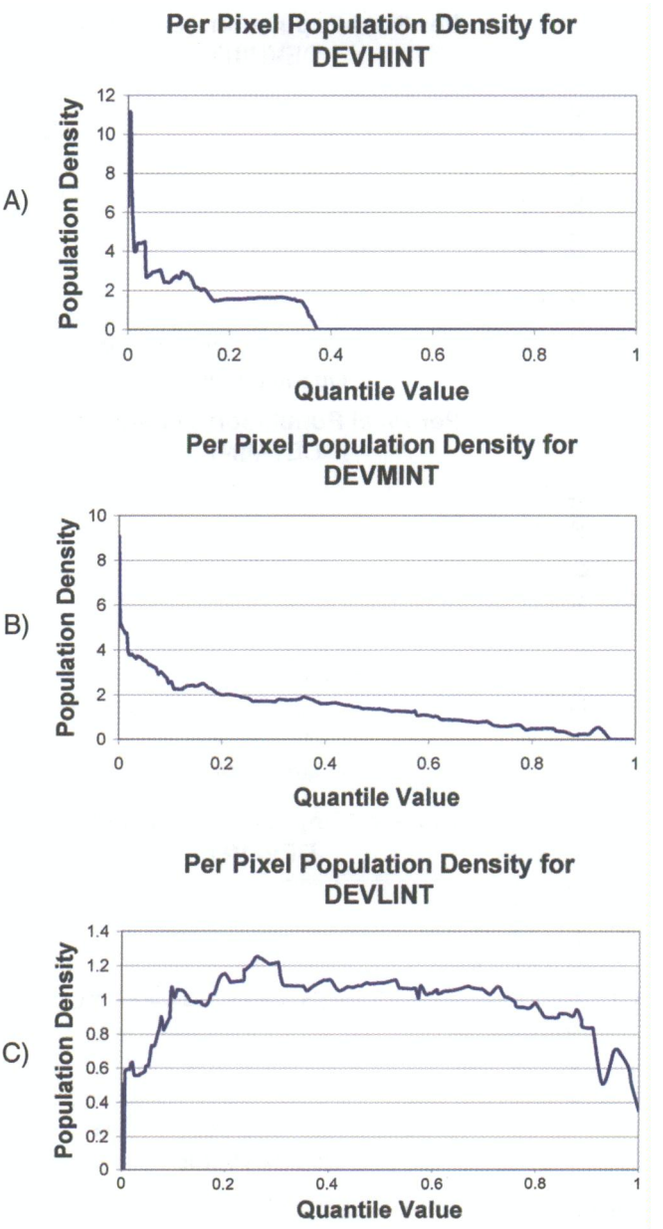


**Figure 5.** The distribution of quantile regression coefficients for the Q14 model: (A) Population density values for DEVHINT (Developed, High Intensity); (B) Population density values for DEVMINT (Developed, Medium Intensity); (C) Population density values for DEVLINT (Developed, Low Intensity). (Color figure available online.)

The differences in the estimated population density values for the different land cover types produced different population estimates at the block group level. Table 3 provides the accuracy measures for the estimations done at the block group level. Overall the QR14 model, which has the greatest number of land use categories, produced the best results, even after the scaling

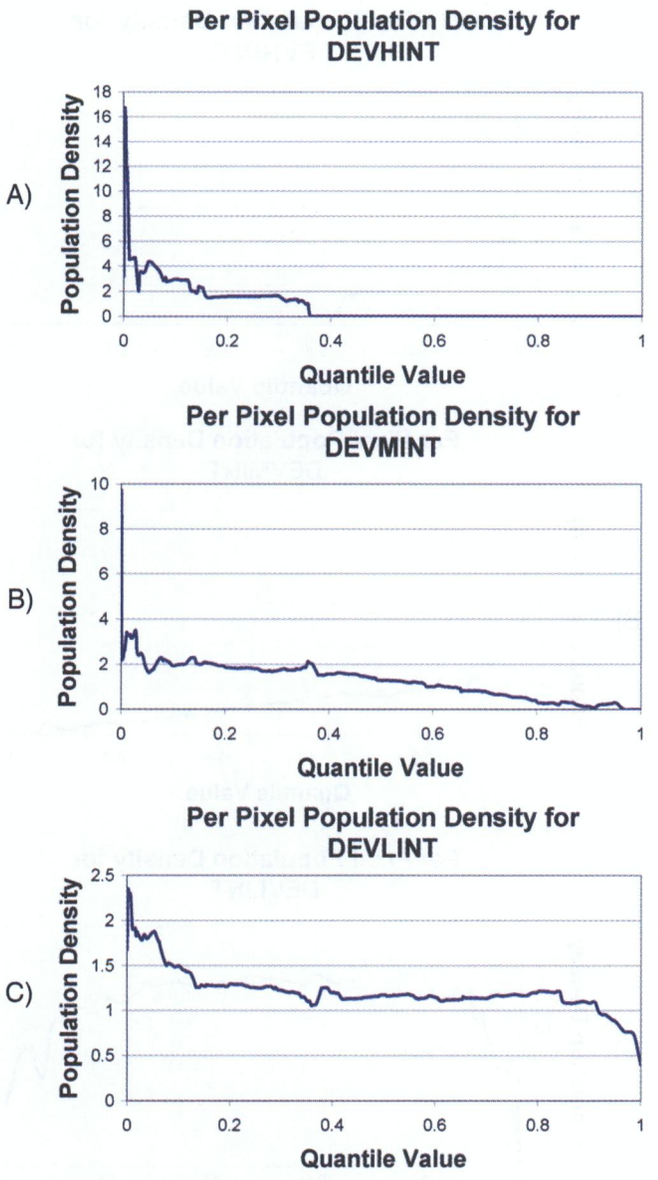


**Figure 6.** The distribution of quantile regression coefficients for the Q3 model: (A) Population density values for DEVHINT (Developed, High Intensity); (B) Population density values for DEVMINT (Developed, Medium Intensity); (C) Population density values for DEVLINT (Developed, Low Intensity). (Color figure available online.)

step was applied to the OLS and SEM models. QR14 was also marginally better than QR3 and the binary dasymetric interpolator. Figure 7 displays the spatial distribution of the absolute errors for each block group using a natural breaks classification so that each class would be relatively homogeneous with respect to MAE values. Overall the pattern of error is random, with both urban and rural areas having block groups that are fairly accurate and inaccurate.

**Table 3.** Results of different areal interpolation models for estimating block group populations

| | Estimates before scaling | | Estimates after scaling | |
|---|---|---|---|---|
| Interpolator | RMS value | MAE value | RMS value | MAE value |
| Areal weighting | 750 | 544 | 750 | 544 |
| Dasymetric | 457 | 345 | 457 | 345 |
| OLS14 | 665 | 493 | 471 | 354 |
| OLS3 | 678 | 510 | 470 | 357 |
| SEM14 | 944 | 776 | 613 | 449 |
| SEM3 | 894 | 743 | 501 | 381 |
| QR14 | 428 | 319 | 428 | 319 |
| QR3 | 430 | 324 | 430 | 324 |

*Note:* MAE = mean absolute error; OLS = ordinary least squares; QR = quantile regression; RMS = root mean square; SEM = spatial error model.

Figure 8 compares how the estimated population values for QR14 compare against the actual values when performing one type of analysis, choropleth mapping. For the purposes of choropleth mapping, the population counts are converted first to population density. Figure 8A is a choropleth map of the actual population density

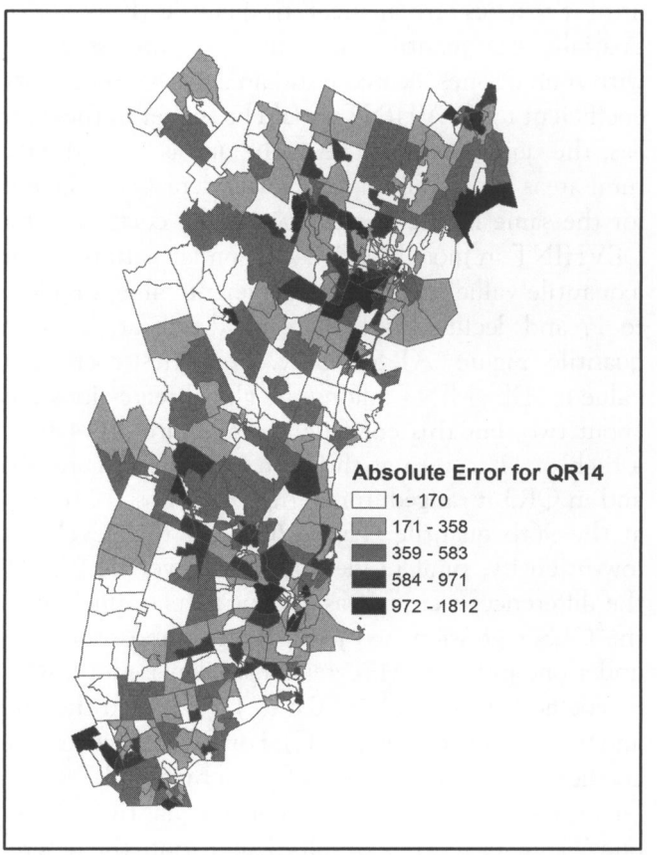


**Figure 7.** The spatial distribution of absolute deviation error for the Q14 model.

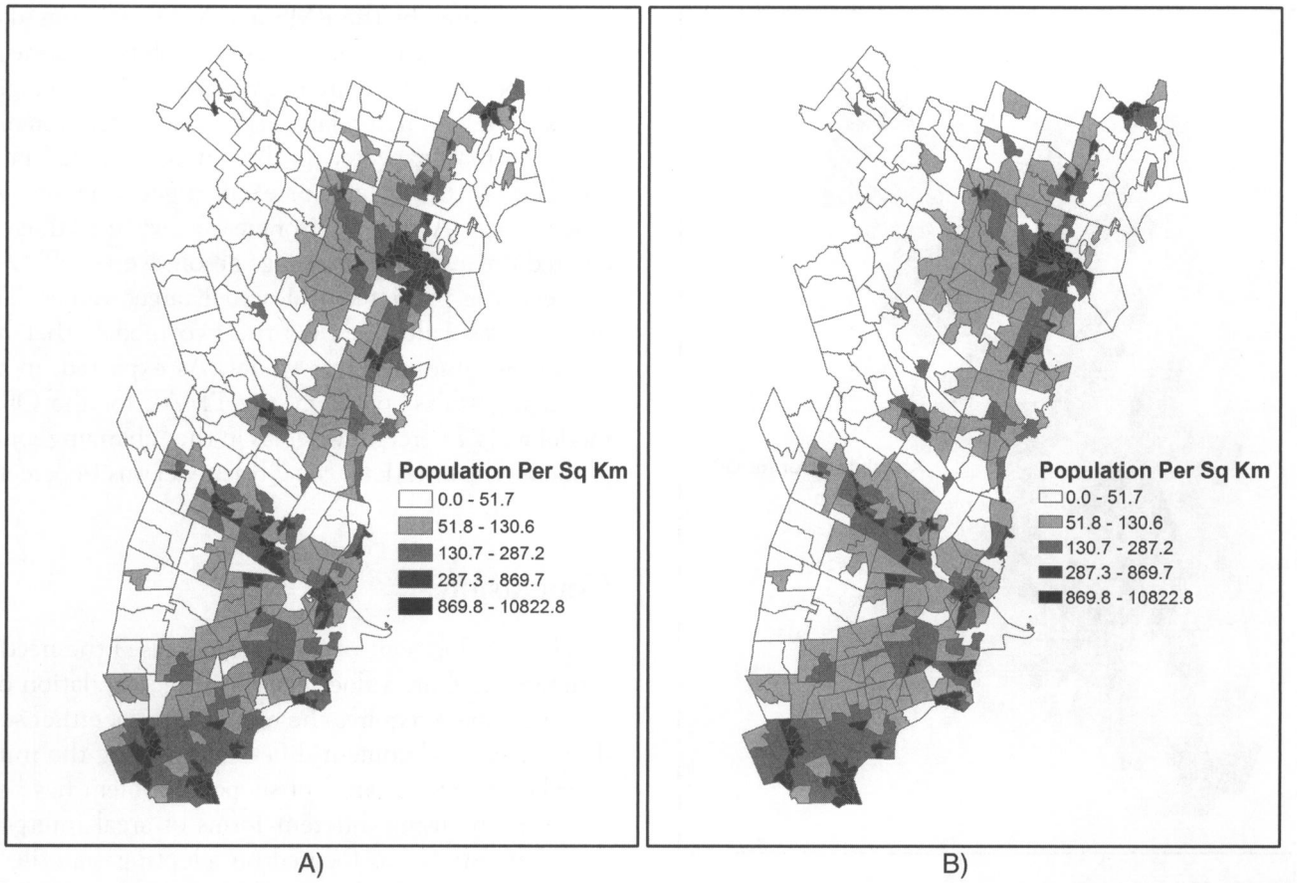


**Figure 8.** Comparison of choropleth maps for actual versus estimated block group population density: (A) A quintile map of actual block group values; (B) A quintile map of estimated block group values.

at the block group level using a quintile classification. This classification was used so that there would be approximately the same number of observations in each interval resulting in a more complex pattern. Figure 8B is a choropleth map of the estimated population density based on the QR14 model using the same interval breaks. Overall, the visual complexity of Figure 8B is similar to that of Figure 8A. In Figure 8A there were 122, 124, 122, 122, and 122 observations in the five successive intervals; in Figure 8B there were 108, 144, 116, 130, and 116 in the same successive intervals.

For the alternative geography analysis, the same regression models at the tract level were used to estimate population density per pixel as in the first analysis; the difference is that these values are applied to the intersection zones and the population totals for the intersection are then summed to the alternative tracts. In this analysis, the quantile models are still guaranteed to have the pycnophylatic property because, like the block groups, the land cover pixels at the intersection level also sum to the total land cover pixels at the tract level. Table 4 provides the accuracy measures for the estimators for the alternative tracts. This time, the QR3 model was marginally better than the QR14 and both quantile models were again more accurate than any other

**Table 4.** Results of different areal interpolation models for estimating alternative tract populations

| | Estimates before scaling | | Estimates after scaling | |
|---|---|---|---|---|
| Interpolator | RMS value | MAE value | RMS value | MAE value |
| Areal weighting | 1,347 | 992 | 1,347 | 992 |
| Dasymetric | 723 | 568 | 723 | 568 |
| OLS14 | 1,750 | 1,244 | 832 | 622 |
| OLS3 | 1,773 | 1,275 | 885 | 620 |
| SEM14 | 3,071 | 2,591 | 899 | 698 |
| SEM3 | 2,926 | 2,454 | 1,167 | 682 |
| QR14 | 672 | 525 | 672 | 525 |
| QR3 | 659 | 521 | 659 | 521 |

*Note:* MAE = mean absolute error; OLS = ordinary least squares; QR = quantile regression; RMS = root mean square; SEM = spatial error model.

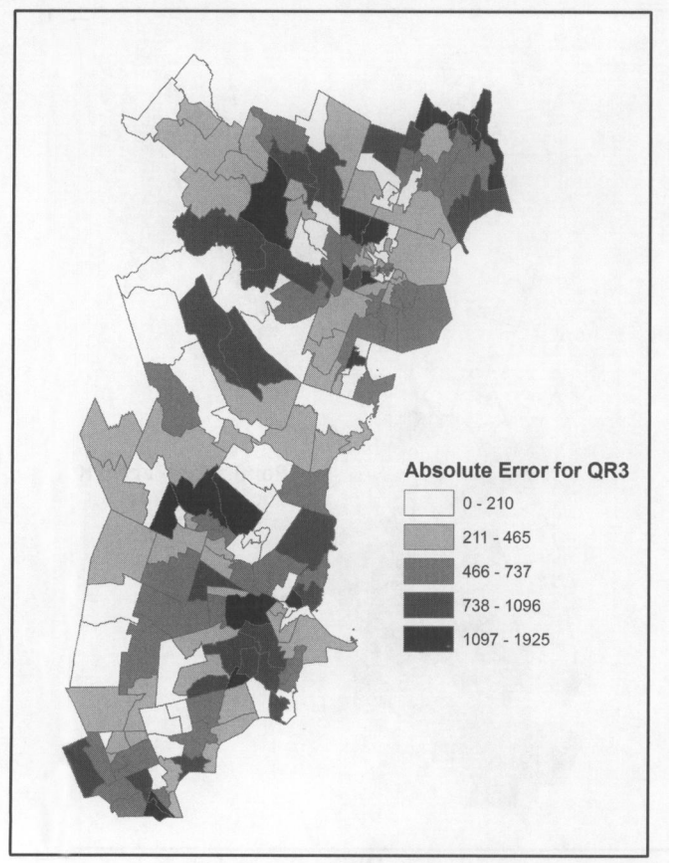


**Figure 9.** The spatial distribution of absolute deviation error for the Q3 model.

interpolator. Figure 9 displays the absolute error for the QR3 model using a natural breaks classification. As was the case for the "changing scales" problem, the pattern of error is fairly random.

Figure 10 again compares how the estimated population values for the most accurate model (QR3 in this case) compare against the actual values when performing choropleth mapping. Figure 10A is again a choropleth map of the actual population density using a quintile classification and Figure 10B is the equivalent estimated population density based on the QR3 model. In Figure 10A there were 36, 36, 36, 36, and 35 observations in the five successive intervals; in Figure 10B there were 34, 35, 37, 38, and 37 observations in the same successive intervals. Overall, QR3 areal interpolation produced a more accurate classified alternative tract map with only 21 out of 179 observations (12 percent) being misclassified, whereas 118 out of 612 observations (19 percent) were misclassified in the block group map.

Finally, comparing the results for the "changing zones" against the "changing scales" problem, the "changing scales" problem has lower absolute error values as measured by the RMS and MAE statistics. This is expected as the overall values for each target zone are smaller because there are more units in the "changing scales" problem. In a relative sense, however, a smaller relative error is expected for the "changing zones" problem because the values for each target zone are now much higher. A relative error measure was used that calculated the average percentage absolute error (PMAE) between the predicted and actual target values. This measure was calculated for the two models that had the lowest absolute error values. As expected, in the "changing scales" problem, the PMAE for the QR14 model was 27 percent, whereas in the "changing zones" problem, the PMAE for the QR3 model was 14 percent.

## Conclusions

The development of GIS has increased the need to estimate attribute values because the compilation of a single database requires the rectification of either scale differences or alignment differences among the initial data layers. The change of support problem has been addressed by using different forms of areal interpolation. This study has focused on adopting quantile regression as the underlying technique in the estimation process. Quantile regression-based areal interpolation has by definition the volume-preserving property necessary for this type of interpolation. Similar to earlier attempts by Yuan, Smith, and Limp (1997) and Mennis and Hultgren (2006), this approach permits modeling of the spatial variability within categories of ancillary data. Quantile regression is somewhat more flexible, however, because no a priori threshold levels or regional delineation of the data within the study area is needed. It also has been shown that the traditional binary dasymetric interpolator is itself equivalent to a quantile regression interpolator with one independent variable.

Earlier evaluative studies (Fisher and Langford 1995; Langford 2006) have found dasymetric methods to be more accurate than statistical methods. The case study situated in northern New England found that quantile regression-based regression was superior to areal weighting, binary dasymetric, and OLS and SEM regression-based interpolators although it was only 6 to 9 percent better than the binary dasymetric. Because this is only one case study and it is based on an underlying census geography, the question remains whether these results would be generalized to other study areas. The ability to disaggregate the land cover correlate into more cate-

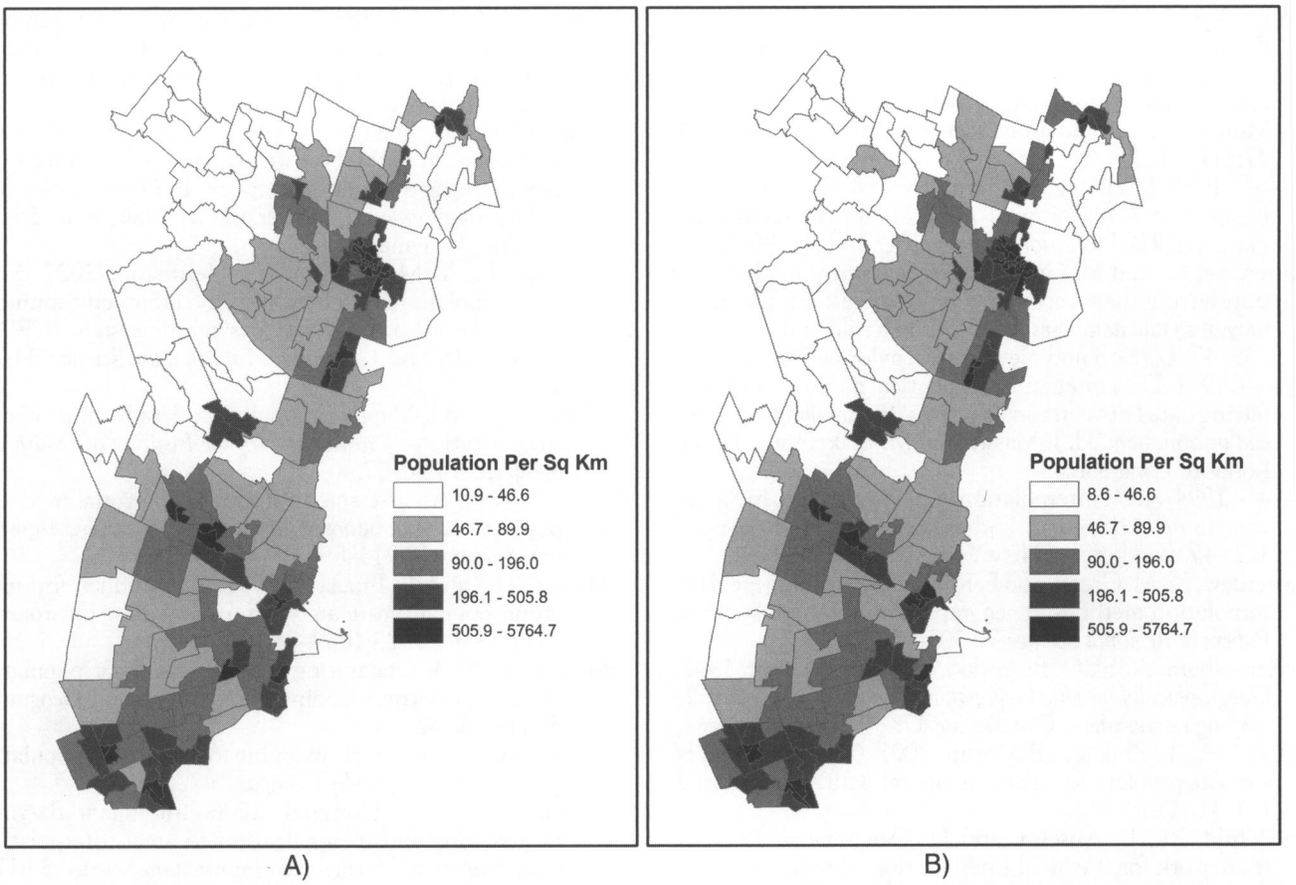


**Figure 10.** Comparison of choropleth maps for actual versus estimated alternative tract population density: (A) A quintile map of actual alternative tract values; (B) A quintile map of estimated alternative tract values.

gories than is possible in the binary daysmetric method (populated vs. unpopulated) should improve population density estimates. This method also permits more combinations of possible correlates. For example, land cover could be overlaid against road buffers to differentiate residential or commercial areas near roads versus the same categories located farther from a road. It might be better to use quantile regression if higher interpolation accuracy is required, but given the increased complexity of this method, a binary dasymetric approach might be sufficient otherwise.

Improving estimates can be accomplished by either acquiring more accurate ancillary data or by developing more robust methods. Continuously collecting higher quality data is an expensive proposition, whereas an improved method, once developed, can be incorporated into software systems. As GIS software continues to offer more analytical capabilities, quantile regression has several positive characteristics. Besides being used for areal interpolation, it can be used as an alternative to OLS for any regression analysis and can be solved using general linear programming techniques that themselves could also be used for other types of optimization problems. Future research will investigate the trade-offs between using data with different accuracy levels versus different methods and the integration of this type of interpolation into existing GIS software.

## References

Anselin, L. 2003. Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review* 26:153–66.

Bloom, L., P. Pedler, and G. Wragg. 1996. Implementation of enhanced areal interpolation using MapInfo. *Computers and Geosciences* 22:459–66.

Bracken, I. 1991. A surface model approach to small area population estimation. *Town Planning Review* 62:225–37.

Bracken, I., and D. Martin. 1989. The generation of spatial population distributions from census centroid data. *Environment and Planning A* 21:537–43.

Eicher, C., and C. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28:125–38.

ESRI. 2009. *ArcGIS Desktop: Release 9.3.1*. Redlands, CA: Environmental Systems Research Institute.

Fisher, P., and M. Langford. 1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A* 27:211–24.

———. 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., and M. Green. 1989. Statistical methods for transferring data between zonal systems. In *The accuracy of spatial databases*, ed. M. Goodchild and S. Gopal, 239–47. London and New York: Taylor & Francis.

———. 1991. Data integration: Statistical methods for transferring data between zonal systems. In *Handling geographical information*, ed. I. Masser and M. Blakemore, 38–54. London: Longman.

———. 1994. Areal interpolation and types of data. In *Spatial analysis and GIS*, ed. S. Fotheringham and P. Rogerson, 121–45. London and New York: Taylor & Francis.

Flowerdew, R., M. Green, and E. Kehris. 1991. Using areal interpolation methods in geographic information systems. *Papers in Regional Science* 70:303–15.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 1992. *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, UK: Wiley.

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

Goodchild, M., L. Anselin, and U. Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* 25:383–97.

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

Green, M. 1990. Statistical models for areal interpolation. In *EGIS '90: Proceedings, First European Conference of Geographical Information Systems*, ed. J. Harts, H. Owens, and H. Scholten, 392–99. Utrecht, The Netherlands: EGIS Foundation.

Griffith, D., R. J. Bennett, and R. Haining. 1989. Statistical analysis of spatial data in the presence of missing observations: A methodological guide and application to urban census data. *Environment and Planning A* 21:1511–23.

Hadley, G. 1964. *Linear programming*. Reading, MA: Addison-Wesley.

Hao, L., and D. Naiman. 2007. *Quantile regression*. Los Angeles: Sage.

Holt, J., C. P. Lo, and T. Hodler. 2004. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science* 31:103–21.

Koenker, R. 2005. *Quantile regression*. Cambridge, UK: Cambridge University Press.

Koenker, R., and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica* 46:33–50.

Kyriakidis, P. 2004. A geostatistical framework for the area-to-point spatial interpolation. *Geographical Analysis* 36:41–50.

Lam, N. 1983. Spatial interpolation methods: A review. *American Cartographer* 10:129–49.

Langford, M. 2006. Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems* 30:161–80.

Langford, M., D. Maguire, and D. Unwin. 1991. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In *Handling geographical information*, ed. I. Masser and M. Blakemore, 55–77. London: Longman.

Maantay, J., A. Maroko, and C. Herrmann. 2007. Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science* 34:77–102.

Martin, D. 1989. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers* 14:90–97.

———. 1996. An assessment of surface and zonal models of population. *International Journal of Geographic Information Systems* 10:973–89.

Martin, D., and I. Bracken. 1991. Techniques for modelling population-related raster databases. *Environment and Planning A* 23:1069–75.

Mennis, J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55 (1): 31–42.

———. 2009. Dasymetric mapping for small area population estimation. *Geography Compass* 3:727–45.

Mennis, J., and T. Hultgren. 2006. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science* 33:179–94.

Merwin, D., R. Cromley, and D. Civco. 2009. A neural network-based method for solving "nested hierarchy" areal interpolation problem. *Cartography and Geographic Information Science* 36:347–65.

Moxey, A., and P. Allanson. 1994. Areal interpolation of spatially extensive variables: A comparison of alternative techniques. *International Journal of Geographical Information Systems* 8:479–87.

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

Mugglin, A., and B. Carlin. 1998. Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics* 3:111–30.

Mugglin, A., B. Carlin, and A. Gelfand. 2000. Fully model based approaches for misaligned spatial data. *Journal of the American Statistical Association* 95:877–87.

Mugglin, A., B. Carlin, L. Zhu, and E. Conlon. 1999. Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems. *Environment and Planning A* 31:1337–52.

Okabe, A., and Y. Sadahiro. 1997. Variation in count data transferred from a set of irregular zones to a set of regular zones through the point-in-polygon method. *International Journal of Geographical Information Science* 11:93–106.

Pace, R., and O. Gilley. 1997. Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics* 14:333–40.

Rase, W. 2001. Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems* 3:199–213.
Reibel, M., and A. Agrawal. 2007. Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review* 26:619–33.
Tapp, A. 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science* 37:215–28.
Tobler, W. 1979. Smooth pycnophylatic interpolation of geographical regions. *Journal of the American Statistical Association* 74:519–30.
Tobler, W., and S. Kennedy. 1985. Smooth multidimensional interpolation. *Geographic Analysis* 3:251–57.
White, M. 1979. A survey on the mathematics of maps. *Proceedings of Auto-Carto IV* 1:82–96.
Worboys, M., and M. Duckham. 2004. *GIS: A computing perspective*. London and New York: CRC Press.
Wright, J. 1936. A method of mapping densities of population. *The Geographical Review* 26:103–10.
Xie, Y. 1995. The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems* 19:287–306.
Yoo, E.-H., P. Kyriakidis, and W. Tobler. 2010. Reconstructing population density surfaces from areal data: A comparison of Tobler's pycnophylactic interpolation method and area-to-point kriging. *Geographical Analysis* 42:78–98.
Yuan, Y., R. Smith, and W. Limp. 1997. Remodeling census population with spatial information from LandSat TM imagery. *Computers, Environment and Urban Systems* 21:245–58.

*Correspondence:* Department of Geography, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269–4148, e-mail: robert.cromley@uconn.edu (Cromley); dean.hanink@uconn.edu (Hanink); george.bentley@uconn.edu (Bentley).