Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources

Anna F. Tapp

ABSTRACT: This research contributes to the body of literature on areal interpolation and dasymetric mapping by introducing algorithms that make use of local ancillary data sources. The address weighting (AW) and parcel distribution (PD) methods are based on county address points and parcels. The algorithms can be effectively applied in rural and transitional areas where geographies are large and population counts are low. These new methods were compared to existing algorithms that use nationally available land cover and street network datasets. Compared with existing methods, the new methods yielded significant improvement in reducing estimate error for the study areas. Both new methods succeeded in maintaining high accuracy in both urban and rural areas. The research presents opportunities for increasing the accuracy of both areal interpolation and dasymetric mapping in areas where accurate local data are available.

KEYWORDS: Areal interpolation, dasymetric mapping, address points, cadastral data

Introduction

Geographers use a variety of methods for visualizing and analyzing population distributions across space. When the geometry of a research area differs significantly from census boundaries, areal interpolation techniques are employed to approximate true population. Areal interpolation is the process of taking aggregate data arranged in specific geographic zones and using algorithms to re-aggregate the data into new geographic zones. As Geographic Information Systems (GIS) continue to advance, researchers are testing previously conceptual methods for the effectiveness of their application. Promising algorithms are being fine-tuned for more accurate performance.

Enumeration districts (EDs) in rural areas pose aggregation difficulties due to their large geographic size. According to the U.S. Census Bureau (2009), the optimum size of a census tract is 4000 people but can include up to 8000 people. There are 546 counties in the United States which reported less than 8000 people in the 2000 Decennial Census. Many of them are comprised of only one census tract. In these situations the confidentiality of respondents takes precedence over geographic precision. The only opportunity in such cases to generate more specific population distributions from the aggregate data is through the use of ancillary data sources.

This research contributes to the body of literature on areal interpolation by introducing methods that can be effectively applied in rural areas where geographies are large and population counts are low. Accurately estimating population distributions in rural areas has a number of spatial applications for both commercial and social services. Service area applications range from medical facilities, to retail stores, to libraries. The methods introduced here can be used to estimate block-size data from the regular countywide updates of the American Community Survey (ACS). The purpose is to identify an effective algorithm for detailed population analyses in rural and transitional areas. In contrast, the majority of previous areal interpolation research has focused on urban areas. Therefore in order to effectively compare the results of this study to previous literature, a combination of rural and urban populations is included in the analysis. The article begins with a survey of foundational literature, including previous patterns of error observed in rural and transitional areas. Two local government data sources are introduced as promising ancillary datasets for population re-aggregation. The statistical errors of the new methods are compared to the errors of two of the most applicable methods presented in the literature. Finally, conclusions are drawn as to the

Anna F. Tapp, Center for Geographic Information Science, The University of North Carolina at Greensboro, Greensboro, North Carolina 27412. Tel: (336) 334-3916. E-mail:<aftapp@uncg.edu>.

feasibility of using these algorithms in rural areas across the country.

Literature Review

The choropleth map is the timeless convention of representing population density from census counts. The earliest known choropleth map was drawn by Baron Pierre Charles Dupin, showing the distribution of literacy in France (Gillespie 1970-1990). The choropleth map has distinct boundaries drawn independently of the collected measurements. The probability of population density variation within an enumeration district is unknown since the boundaries are meant to facilitate the information-gathering process, not to represent a line of homogeneity (Mark and Csillag 1989). The census EDs visualized through choropleth maps present the ecological fallacy that all observations within an area are similar to the average for the area. In other words, choropleth maps would have the reader assume that all places within a unit have equal population densities. This assumption is often untrue.

Areal Weighting and Point Interpolation

The problem arises of how to create new zonal boundaries from aggregate census data based on research-driven regions of homogeneity. Two widespread solutions are areal weighting interpolation and point interpolation. Simple areal weighting interpolation is a long-held and computationally straightforward way to redistribute count data to new zonal boundaries (Wright 1936; Goodchild and Lam 1980; Lam 1983). This map overlay algorithm calculates the proportion of each source zone that lies within each target zone. The appropriate fractions of the source zones are assigned to each target zone (Goodchild and Lam 1980; Lam 1983). Those associated proportions are summed for a population count of the target zone. This method does not derive a population surface. Instead, the procedure transitions immediately from source to target vector data. It possesses the obvious drawback of the homogeneity assumption (Goodchild and Lam 1980), a variation of the ecological fallacy. There is a possibility that the area selected from the source zone has a different population density than the average population density of the source zone. However in the absence of ancillary data, it remains a reasonable solution (Xie 1995).

In contrast to areal weighting, point interpolation assigns each ED a control point. The population density of the corresponding ED is assigned to that point. A surface is created from one of a variety of algorithms, including ordinary Kriging, Polynomial Trend Surface Analysis, and Moving Average Analysis (Lam 1983; Xie 1995). There have been two major drawbacks to point interpolation. First, the placement of the control point has a significant impact on the resulting surface. In some cases, the geometric centroid lies outside the boundary of the polygon, generating a questionable result (Schmid and MacCannell 1955; Xie 1995). Second, the total volume of each census division may not be preserved on the interpolated surface. Tobler introduced pycnophylactic algorithms to overcome the latter difficulty (Tobler 1979). Known as local smoothing algorithms, one begins with a raster version of the choropleth map. A local smoothing algorithm is used to blend the boundaries between zones. After each smoothing iteration, the values of all the cells within each zone are adjusted uniformly to preserve volume. Smoothing and adjusting continues until all EDs are within an acceptable pycnophylactic threshold (Tobler 1979). Although Tobler's smoothing algorithm is not necessarily employed in current research, the concept of preserving volume remains an important constraint of each new interpolation algorithm.

Dasymetric Mapping

Each of the above interpolation methods spreads the data across the area of interest. Such methods assume that the probability is greater than zero that one will find a person living in every place on the population density map. Despite the popularity of population density surfaces, some scholars have long disagreed with the notion that population should be conveyed as a continuous phenomenon. In 1936, John Wright asserted that this sort of mapping was unrealistic. He argued that unlike topography, population is not a continuously observed phenomenon. Many areas with no observed population abruptly transition to settlement areas. Wright preferred the concept of the dasymetric map, calling it controlled guesswork. He uses a United States Geological Survey (USGS) quadrangle map as his ancillary data source in his population distribution of Cape Cod. First, he eliminates all areas known to be uninhabited and recalculates the population density based on the remaining area. Next, he takes known populations of vil-

lages and distributes them appropriately in the populated areas marked on the USGS map. The remaining population is redistributed across the inhabitable countryside. This method of taking known densities to balance the population surface between different land use classes has been called the limiting variable dasymetric method (Eicher and Brewer 2001).

As ancillary data sources became widely available in the 1990s, dasymetric mapping gained momentum. Many assert that smoothing algorithms, although cartographically valid, are not appropriate for advanced spatial analysis (Yuan et al. 1997; Eicher and Brewer 2001; Mennis 2003; Langford 2007). The newer overall goal of creating a population surface is not simply to create a cartographic impression of reality, but to free the spatial analyst from census-derived boundaries entirely (Moon and Farmer 2001). The expectation is that a properly created population density surface can be the foundation for calculating the population of any number of target zones drawn for any number of purposes. Overcoming the ecological fallacy and maintaining the pycnophylactic principle remain two important issues that every new technique must address.

The most common application of dasymetric mapping is the use of a binary mask in which all the areas known to be uninhabited are removed from the population density surface (Langford and Unwin 1994; Yuan et al. 1997; Eicher and Brewer 2001; Mennis 2003; Langford 2007). Following the binary mask operation, predictor variables are commonly used to distribute the population across the remainder of the surface. These predictor variables are often derived from remote sensing image classification schemes (Fisher and Langford 1995; 1996; Yuan et al. 1997; Eicher and Brewer 2001; Mennis 2003; Langford 2007). Statistical regression models help determine an average population density value associated with each land cover class. The advantage of this method is the ability to customize the regression coefficients of certain areas of the population surface. For example, Yuan et al. (1997) found that residential densities close to an urban center are higher than residential densities farther from an urban center. Finally, after an error analysis, the researcher analyzes the residuals in this way and modifies the coefficients locally.

A drawback of the statistical model is its tendency toward negative intercept values. These negative intercepts result in negative population predictions in sparsely populated areas (Yuan et al. 1997). Researchers circumvent this problem by either forcing a zero intercept into their regression equations or by scaling the final data upward until all values are non-negative (Yuan et al. 1997).

Rural Population Densities

Most of these modern interpolation and dasymetric methods involve the use of small census divisions to train the model and maximize accuracy. They assume that one has a full decennial dataset from which to initially base the population distribution. Furthermore, they take for granted that there will be a sufficient number of census divisions within an area on which to base appropriate density estimates for distinct land cover classes. There has been little research conducted where county-level population counts are the only census data used in the algorithm. An algorithm that can accurately distribute county-level population density without sub-county census data would be a tremendous asset in rural applications. Since 2005, the U.S. Census Bureau's American Community Survey (ACS) has been annually updating geographies with greater than 65,000 people (U.S. Census Bureau 2008). Starting in 2008, they began generating multiyear estimates for geographies with more than 20,000 people. All other areas will have access to five-year estimates, to be published in 2010 (U.S. Census Bureau 2008). In each decennial census there is tremendous population growth in many areas classified as rural, yet located on the urban fringe. A methodology to model such areas with a higher temporal resolution would be invaluable in a variety of planning applications.

Riebel and Buffalino (2005) took a street network approach to this problem with their street weighting method. Analyzing the distribution of people in Los Angeles County census tracts, they developed an algorithm which estimates how many people live along each street segment in the study area. Each street is assigned a weight to preserve the pycnophylactic property. The weight combined with the total count of the census tract population determines the population for the street segment. The model performed better for housing counts than for population counts. Although it did well in the urban center where streets are regularly spaced, it tended to fail in rural areas where streets are farther apart and residences are located at irregular intervals.

The methodology of Eicher and Brewer (2001) could also be used to approach this rural aggregation difficulty. They model population densities across 159 counties in Pennsylvania, West Virginia,

Author(s)	Year	Algorithm	Study Area	Ancillary Data	Accuracy Assessment
Wright	1936	Dasymetric map	Cape Cod, MA	USGS Quad Map	No
Tobler	1979	Smooth pycnophylactic interpolation	Ann Arbor, MI	n/a	No
Goodchild & Lam	1980	Simple areal weighting	London, UK	n/a	Yes
Xie	1995	Overlaid network	Amherst, NY	Street TIGER lines	Yes
Fisher & Langford	1996	Dasymetric map using binary mask delineating residential areas	Charnwood, Leicester, Oadby & Wigston, UK	LANDSAT TM	Yes
Yuan et al.	1997	Dasymetric map using statistical regression	Faulkner, Lonoke, Pulaski & Saline Counties, AR	LANDSAT TM	No
Eicher & Brewer	2001	Limiting variable dasymetric map	Pennsylvania, Maryland, District of Columbia, West Virginia & Virginia	USGS LULC dataset	Yes
Mennis	2003	Dasymetric map using weighted urban densities	Philadelphia & Southeast PA	Urban density classes	No
Riebel & Buffalino	2005	Street weighting areal interpolation	Los Angeles County, CA	Street TIGER lines	Yes

Table 1. Summary of major contributions to areal interpolation and dasymetric mapping.

Maryland, Virginia, and the District of Columbia. In addition to testing their method across a wide variety of population densities and land cover classes, they only use smaller census divisions in the error assessment phase. In a visualization of percent error values, they notice a pattern of error around metropolitan areas as they transition between urban and agricultural/forest areas. The agricultural/forest areas immediately outside the urban border are under-predicted. This pattern of error corroborates the results noticed by Yuan et al. (1997), in which residential densities close to urban areas are higher than residential densities farther from urban areas. The differing densities could be attributed to a variety of factors ranging from zoning restraints to land values. The anomaly indicates a necessity to train the model further for increased accuracy.

Previous research conducted by Hawley and Moellering (2005) has compared street network and dasymetric methods for urban and suburban areas. Those results demonstrate a slightly higher performance by the street network method. There are also indications that the inclusion of network ancillary data provides the resolution necessary to estimate small target zones from much larger source zones. Since the study area was limited to urban counties, it is still unknown how the results will hold in rural locales. Table 1 outlines some of the major contributions of previous research to areal interpolation and dasymetric mapping methodology.

Methods

The current research improves the accuracy of previous methods by testing modifications to the street weighting (SW) and limiting variable (LV) algorithms. These modifications are called address weighting (AW) and parcel distribution (PD). The AW method, like Riebel and Buffalino's street weighting algorithm, is entirely vector based. In contrast, the PD method results in a raster dasymetric map that could be used to compute the population of any shape or size vector target zone. Both are designed for application in rural areas, while successfully generating small target zones out of large source zones. They are tested against the SW and LV algorithms presented in the literature.

The Street Weighted (SW) Method

According to Riebel and Buffalino's method, each street segment is assigned a weight according to its length. In order to facilitate the weighting process, all the roads were segmented at the intersection of the boundary of a census block.

This segmentation prevents any road from being counted twice or eliminated entirely from the weighting process. The following equations were used to assign a weight to each street segment in the study area:

$$W_i = L_i / L_t \quad (1)$$

$$P_i = W_i \times P_t \quad (2)$$

$$P_b = \sum P_i \quad (3)$$

The weight assigned to the ith street segment in the county (W_i) is equal to the length of the ith street (L_i) divided by the total length of all street segments in a county (L_t). The population of the ith street segment (P_i) is found by multiplying the weight of that segment (W_i) with the total population of the county (P_t). Finally the population of a census block (P_b) is calculated by summing each of the street segment populations contained within that block.

The process is repeated for each block within each county in the study area. In this way, a county-wide population count is distributed across each block in the county while preserving the pycnophylactic principle.

The Limiting Variable (LV) Method

The 2001 National Land Cover Dataset (NLCD) was used as the ancillary data source in the LV method. Fifteen of the twenty-one NLCD classes were present in the three counties. These fifteen classes were consolidated into four classes according to Eicher and Brewer's methodology: Unpopulated, Urban, Agricultural/Woodland and Forest. Bodies of water represented unpopulated areas.

The areas classified as unpopulated were eliminated from the dasymetric map altogether. Next, the county population was evenly distributed across the remainder of the county. In the second iteration of the LV method, the population of the forested areas was limited to 15 people per square km (39 per square mile). The remaining population was evenly distributed across the Agricultural/Woodland and Urban areas. The third pass limits the population of the Agricultural/Woodland areas to 50 people per square km (130 per square mile). The final remaining population is distributed across the urban areas.

The Address Weighted (AW) Method

The study conducted by Riebel and Buffalino (2005) in the Los Angeles area performed well in high-density settlements and poorly in the rural hills surrounding the Los Angeles valley. The authors indicated that the relative unpredictability of home locations along rural roads was a major contributor to the high error. The AW method circumvents this problem by using address points instead of street centerlines as the ancillary data source. Many counties across the United States use address point locations for dispatch of 911 first responders. Even rural counties with fewer GIS resources digitize addresses due to the difficulty of quickly locating residences in an emergency. According the North Carolina GIS Inventory, half the counties in the state report having a 911 address point dataset (NC One Map 2009). Address points eliminate the necessity of relying on less accurate street centerline geocoding to route emergency vehicles. This study takes advantage of available address point data and models the distribution of population by using the known locations of homes along rural roads.

The AW method is computationally simpler than the SW method. Each county has one address weight assigned to each address within that county.

$$W_c = 1 / N_a \quad (4)$$

$$P_a = W_c \times P_t \quad (5)$$

$$P_b = N_{ab} \times P_a \quad (6)$$

The weight of addresses in a county (W_c) is equal to one divided by the total number of addresses in the county (N_{ac}). The average population of each address (P_a) is the product of the address weight and the total population of the county (P_t). Finally, the population of a block (P_b) is the number of addresses lying within the block (N_{ab}) multiplied by the average population at each address.

It is hypothesized that the AW method will yield smaller errors than the SW method because there is no ecological fallacy assuming that addresses are evenly distributed along each and every street in the county.

The Parcel Distribution (PD) Method

The PD method is a blend of the limiting variable method (Eicher and Brewer 2001) and the binary mask method (Langford and Unwin 1994). In contrast to land cover statistical methods used in previous literature, the PD method takes advantage of local parcel data. Every

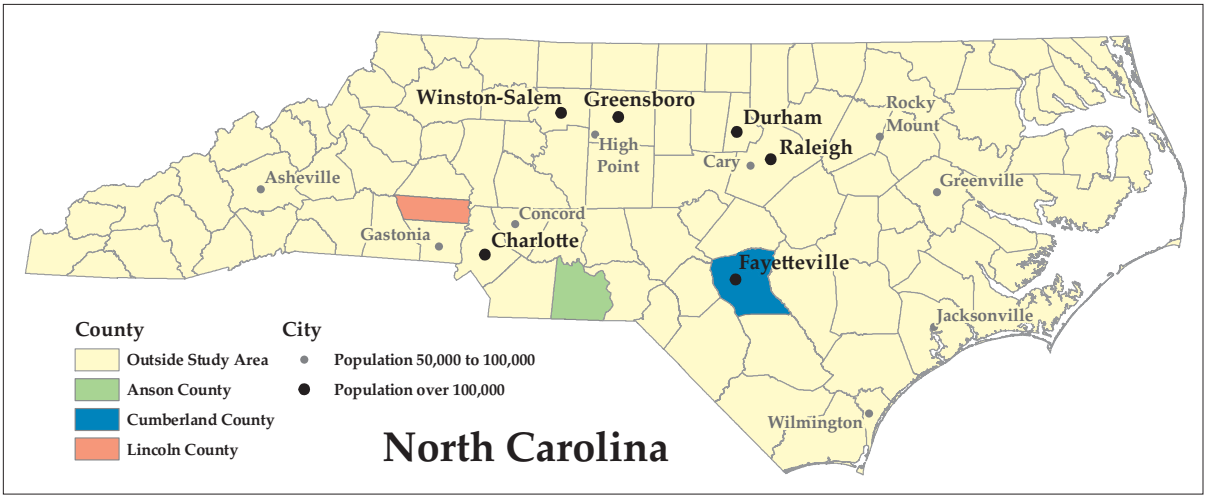

Figure 1. Study area counties.

year more and more counties in the United States digitize their cadastral data. The Federal Geographic Data Subcommittee on Cadastral Data reports that in 2005, 68 percent of parcels had been digitized nationwide (Stage and von Meyer 2006). Nineteen states had digitized at least 80 percent of their parcels, with Delaware, Hawaii, Oregon, and Wyoming having 100 percent of their parcels in geospatial format. Other states leading the way with 95 percent or more were Florida, Kentucky, Montana, New York, and North Carolina (Stage and von Meyer 2006). These results demonstrate that even in rural areas, many geographies in the United States have parcel data which could be used for dasymetric mapping.

Parcel data provides enhanced resolution and accuracy in rural areas in contrast to the use of land cover classes. The favored source of imagery for land cover classification is LANDSAT. This satellite offers 30 m resolution with 900 square meter (9600 square foot) cells. Since most housing in the United States is well below 900 square meters in ground area, LANDSAT imagery does not effectively detect individual residences. Finding the location of a single house in a forested area presents difficulties. Using LANDSAT data is much more effective for delineating settlement groups that exceed 900 square meters in area. Despite this lack of precision, dasymetric mappers using LANDSAT as their ancillary data must assume that people live in these agricultural or forested areas even if the density is much less than that of urban settlements. The result has been an imprecise approximation of rural population distribution.

The limiting variable of the PD method is based on average household size. The average household size according to the census was always higher than the average household size reflected by the county ancillary data. This discrepancy occurred because the ancillary data were unable to detect vacant homes. Despite this known source of error, in order to preserve the pycnophylactic principle, the average household size used in the PD method was taken from the ancillary data. The sum of all the address points present in the residential parcels was divided by the 2000 total population for the county to determine an appropriate limiting variable benchmark.

Each parcel containing one address point had one average house population assigned to that parcel. Each parcel containing more than one address point had the proportional number of people assigned to that parcel according to the following equation:

$$P_p = A \times N_a \tag{7}$$

where:

P_p = the population of the parcel,
A = average number of people per house hold, and
N = number of addresses within the parcel.

When converting the parcels to raster, it is important that the pixel size of the output dasymetric map be smaller than the size of any residential parcel in order for each parcel to be represented by at least one pixel. The pixel size was set to 10 m (32.8 feet), or 100 square meters (1076 square feet, 0.02 acres), in order to remain sufficiently small. The area of the parcels in square meters was divided by 100 square meters to determine the number of pixels that would be present in each parcel. The number of people assigned to

each parcel (Pp) was divided by the number of parcel pixels. The parcels were then converted to raster, where each 10-m pixel lying within a parcel was assigned the value of a fraction of a person. In this way, the population of each parcel was evenly distributed across the parcel area. All areas outside a residential parcel received a pixel value of zero.

This method contrasts with the LV algorithm in two fundamental ways. First, the areas designated as uninhabited are broadened from bodies of water to also include areas with a land use other than residential. Second, the population density of each parcel is individually calculated, allowing for density changes as parcel size changes. The hypothesis is that the PD method will outperform the LV method by more precisely constraining the population to residential geographies, and by incorporating the population density impact of differing parcel sizes. By accounting for parcel size in population distribution, this algorithm may counteract the error trend noticed by Eicher and Brewer, where clusters of inaccuracy resided in transition zones immediately outside urban areas (Eicher and Brewer 2001; Yuan et al. 1997).

Results

The study area is located within the state of North Carolina and represents a range of population densities for ready comparison with previous results. Two goals are accomplished:

- Begin with a county-level population count and disperse the population according to predicted settlement patterns, and
- Generate acceptable accuracy in the form of a low root mean square error (RMSE), using the smaller census divisions as accuracy indicators.

Data from the U.S. 2000 decennial census is combined with appropriate ancillary data from the same time frame. Acceptable accuracy is judged by the coefficient of variance derived from the RMSE of blocks within block groups.

The study area is comprised of three North Carolina Counties—Anson, Lincoln, and Cumberland (see Figure 1). Anson County is the most rural in the study area, with an average population density of 18.5 people per square kilometer (48 per square mile). Centrally located along the border with South Carolina, in 2000 its population was 25,275 (U.S. Census Bureau 2000). Lincoln County is located 46 km (14 miles) northwest of downtown Charlotte, North Carolina. With 63,780 people in 2000, the population of Lincoln County is expected to double between the years 2000 and 2030 from the growth of the Charlotte Metro Area. It is a prime example of a growing micropolitan area (Lincoln County Planning Department 2007; U.S. Census Bureau 2002). Cumberland County is the most densely populated of the study. With 179 people per square kilometer (464 per square mile), the county includes the Fayetteville metropolitan area and the Fort Bragg U.S. Army Base. For the purpose of this research, Fort Bragg was eliminated from the study area. Cumberland County does not offer ancillary data for the military facility, and the area is subject to frequent variations in population due to transient military personnel. Fayetteville is the only metropolitan area included the study. In contrast to Lincoln County, Cumberland is not experiencing noticeable growth.

Address points, parcel boundaries, and associated tables were acquired from each county's GIS department. The parcels were filtered for residential parcels having at least one residence that was built at or before the year 2000. The address points had no attributes that could be used to date the residence. Due to insufficient attribution, the addresses were joined by either street address or parcel ID to the parent parcel in order to filter for residential addresses that existed at or before the year 2000.

The population data comes from the 2000 decennial census. Population counts at the county, block group, and block levels were downloaded from the U.S. Census Bureau along with their associated geometries.

Root Mean Square Error

In order to statistically compare the accuracy of different methods, the root mean square error (RMSE) of each block group was calculated according to the assessment principles of previous research (Fisher and Langford 1995; 1996; Eicher and Brewer 2001; Riebel and Buffalino 2005; Hawley and Moellering 2005). A vector census block dataset was overlaid on each of the four outputs. For the SW and AW methods, each block was assigned the sum of the street or address population contained within that block. For the LV and PD methods, zonal statistics were used to compute the sum of the pixel values contained within each block. The actual population counts of the 8087 census blocks in the study area were compared to the predicted values of each of the models to produce a raw error score for each census block. The distribution of

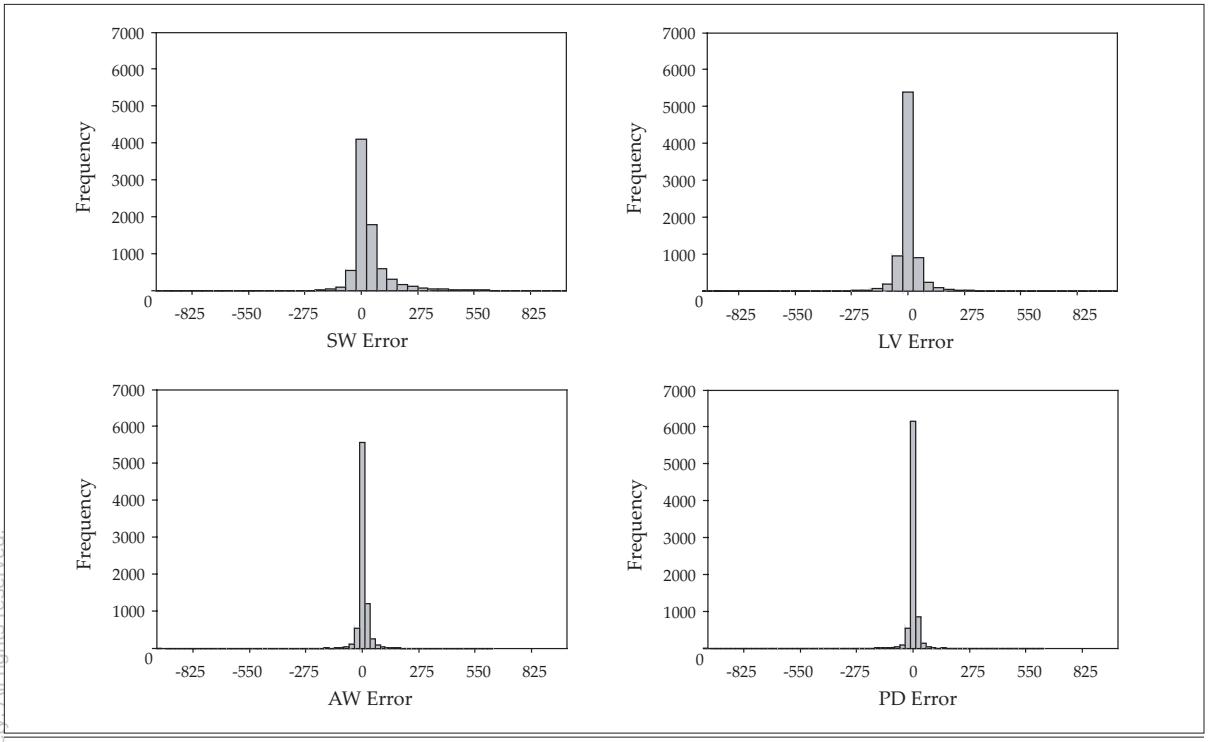

Figure 2. Histograms of raw error scores for each block in the study area.

the raw error scores for each block is shown in Figure 2.

Each raw error was squared, after which a mean squared error for each of the 228 block groups was calculated. The square root of each mean squared error was computed to produce an RMSE for each block group (see Fisher and Langford 1995 for a more detailed description of RMSE application to areal interpolation). The PD method yielded the fewest errors while the SW method yielded the most frequent errors. The mean RMSE value for each method is shown in Table 2, while the mean RMSE values by county are shown in Table 3. Anson County contains 21 block groups, Lincoln County contains 44, and Cumberland County contains 163.

Coefficient of Variance

What constitutes a good RMSE value varies depending on the size of the areal unit. To compare variability across different population groups, the coefficient of variance (CV) is often used (Ott and Longnecker 2001). The CV is the RMSE divided by the average areal unit. In this case, the CV is the block group RMSE divided by the average block population within that block group. Table 3 shows the mean CV of each model type. The CVs not only differed according to the method, but also according to location. Table 4 specifies the mean CVs for each method by county.

Method	Mean RMSE	Mean CV	Median CV
SW	88.10	1.46	1.20
LV	61.74	1.08	1.03
AW	34.83	0.56	0.52
PD	32.75	0.39	0.48

Table 2. Initial accuracy results of the four methods.

The PD method performed the best in Anson and Lincoln Counties. The AW method scored a slightly higher CV in Cumberland County. The discrepancy in Anson County between PD and AW could be due to inaccuracies in the address point file as to the location of the home within the parcel. In each county, at least one of the new methods outperformed both the methods from the literature.

Analysis of Variance (ANOVA)

In order to determine if different models yielded statistically different results, an analysis of variance was performed on the data. To normalize the results, the natural log transformation was performed on the CVs. After the transformation, equality of variance was tested, yielding a Levene's stat of 6.37 with a p-value of < 0.0001. This result indicated a strong possibility that

Method	Anson		Lincoln		Cumberland	
	Mean RMSE	Mean CV	Mean RMSE	Mean CV	Mean RMSE	Mean CV
SW	48.95	3.00	69.59	1.52	98.07	2.33
LV	29.05	1.84	55.98	1.22	67.47	1.48
AW	38.76	1.87	19.90	0.49	38.33	0.66
PD	0.465	0.02	20.42	0.51	40.19	0.72

Table 3. Initial accuracy results for the four methods, by county.

	SW	LV	AW
LV	0.31 0.11 – 0.51		
AW	0.96 0.76 – 1.16	0.65 0.47 – 0.84	
PD	1.32 1.07 – 1.57	1.01 0.77 – 1.25	0.36 0.12 – 0.60

Table 4. Mean difference between groups with 95 percent confidence interval.

variances were unequal. To compensate for unequal variances, Welch's ANOVA was run with the posthoc Tamhane's T2 test. Both are considered conservative and robust statistics appropriate for groups with unequal variances. Welch's ANOVA showed a significant difference between groups, while Tamhane's T2 test demonstrated that differences existed between all four models. Welch's test statistic was 95.11, and the p-value was < 0.0001. Table 4 shows the mean difference results of the ANOVA tests. In each case, the p-value of mean difference was significant at less than 0.001.

In order to determine additional trends in the data, the same ANOVA procedure was run on each county separately. The results proved more complex than the initial aggregate ANOVA analysis. In Anson County, the most sparsely populated of the study, only the PD method was significantly more accurate than all other methods. In Lincoln County, PD and AW yielded very similar accuracy in comparison with the LV and SW methods. The LV and SW methods in Lincoln County were not significantly different. Finally, in the urban Cumberland County, PD and AW behaved similarly, while LV and SW were each significantly different from the rest. In each case the SW method was the least accurate. Each of these similarities and differences is outlined in Table 5 and graphically shown in Figure 3.

Correlations and Pattern Analysis

Two trends were recorded in the literature indicating the possibility of patterns in the error results. Reibel and Buffalino (2005) found that the sparser the population, the greater the error. Eicher and Brewer (2001) and Yuan et al. (1997) speculated on the possibility of spatial patterns of error in transition areas between urban and rural.

In order to test the possibility of a relationship between population density and accuracy, a

County		SW	LV	AW
Anson N = 21 Pop = 25,275	**LV**	0.788 -0.12 – 1.76		
	AW	0.168 -0.40 – 0.74	0.621 -1.49 – 0.25	
	PD	4.61* 4.03 – 5.21	3.83* 2.95 – 4.71	4.45* 4.17 – 4.74
Lincoln N = 44 Pop = 63,780	**LV**	0.207 -0.06 – 0.47		
	AW	1.20* 0.88 – 1.51	0.989* 0.69 – 1.29	
	PD	1.16* 0.85 – 1.47	0.953* 0.66 – 1.25	0.036 -0.30 – 0.38
Cumberland N = 163 Pop = 271,172	**LV**	0.270* 0.03 – 0.51		
	AW	0.994* 0.76 – 1.23	0.724* 0.52 – 0.93	
	PD	0.933* 0.69 – 1.17	0.663* 0.45 – 0.87	0.061 -0.14 – 0.26

Table 5. Mean difference between groups with 95 percent confidence interval, by county. [* Significant with alpha < 0.05.]

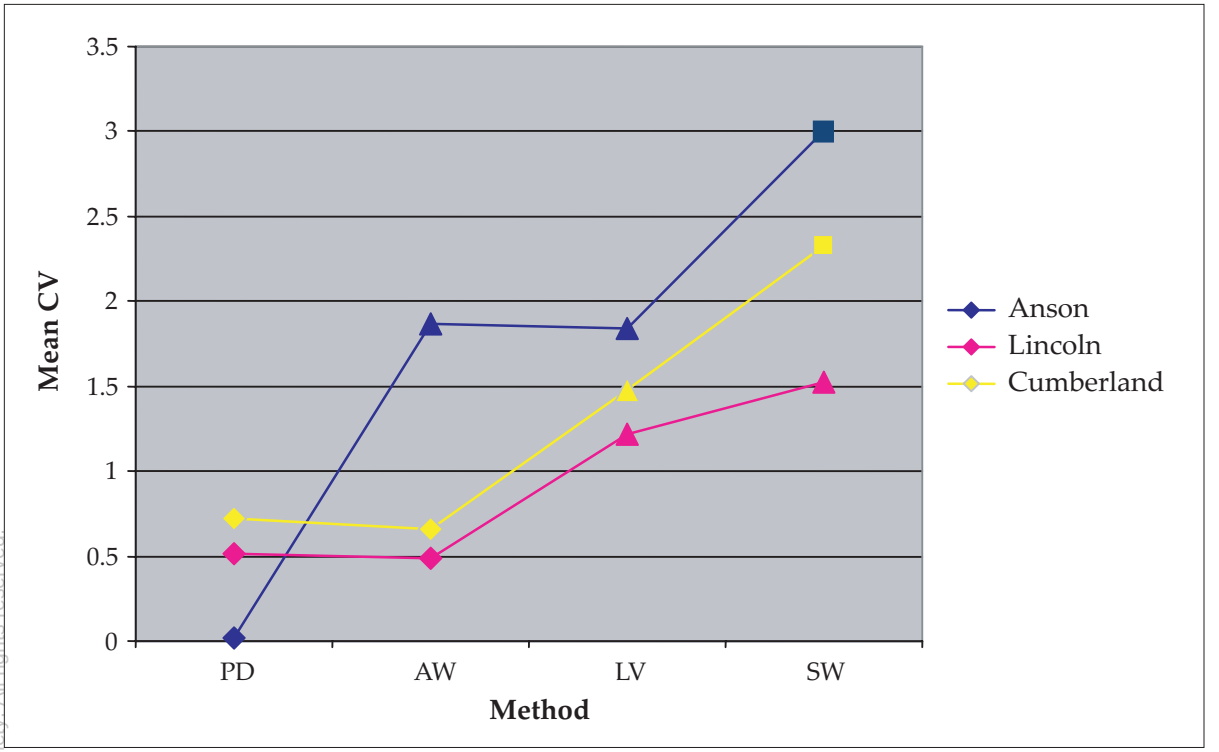

Figure 3. ANOVA chart, mean CV of each method.

Method	Correlation Coefficient	P-value
SW	-0.469	< 0.0001*
LV	-0.343	< 0.0001*
AW	-0.129	0.052
PD	0.104	0.119

Table 6. Population density and CV correlation results. [* Significant with alpha < 0.05.]

correlation procedure compared the CV to the population density of each block group. The SW and LV methods both demonstrated significant correlations between population density and error (see Table 6). Both these models tend to be more accurate in high-density areas. The AW method showed a similar weak correlation, with a p-value of 0.052. The errors in the PD method were not significantly correlated to population density. This finding indicates that the PD method maintains statistically similar accuracy in both high- and low-density populations.

In order to determine the extent to which the errors exhibit spatial patterns, the Anselin Local Moran's I was computed for each method in each county in the study area. Polygon contiguity was the spatial relationship tested in the procedure. The raw error score was chosen for this analysis as it contained information related to the direction of the error. In this way, clusters of over- and under-prediction could be detected. Each county was analyzed separately due to their geographic separation.

Table 7 shows the results of the pattern analysis. According to the Anselin Local Moran's I, the AW output contained the most diffuse errors, with an overall percentage of only 2.4 percent clustered blocks. The SW method demonstrated the most error clusters, with 11.0 percent of blocks located within a cluster. The LV method landed in the middle with 8.0 percent clustered, while the PD method demonstrated relatively diffuse errors with only 3.6 percent of blocks clustered.

Figure 4 shows the Anselin Local Moran's I results graphically. The contrast between higher and lower population densities is most striking with the SW maps. In each county, the most populated areas are diffuse, while the rural areas demonstrate clustered errors. Similar to Eicher and Brewer's (2001) findings, the LV method yields errors on urban fringes. This effect is less noticeable for Lincoln County, where the errors begin to cluster very close to the county seat of Lincolnton. The AW and PD methods, while demonstrating less clustering overall, were not immune to transition zone errors.

	Anson		Lincoln		Cumberland		TOTAL	
Method	Number	Percent	Number	Percent	Number	Percent	Number	Percent
SW	173	12.0	173	12.0	546	10.5	892	11.0
LV	115	8.0	138	9.6	390	7.5	643	8.0
AW	43	3.0	60	4.2	135	2.6	238	2.4
PD	88	6.1	69	4.8	130	2.5	287	3.6

Table 7. Anselin local Moran's I results, number of block errors significantly clustered.

Discussion and Conclusions

Most previous areal interpolation and dasymetric mapping research has focused on aggregating smaller or similar-sized source zones into target zones, while little time has been spent creating smaller target zones out of larger source zones. This research used two established methods that could be appropriately applied to the latter, while testing two new methods that make use of common local ancillary data sources.

The new methods yielded significant improvement in reducing estimate error, while maintaining a low CV even in sparsely populated rural areas. The AW and PD methods were very similar, yet, the ANOVA proved the errors to be significantly different, with PD performing the best overall. These methods stand out against previous methodologies in their use of local ancillary data as opposed to federal government datasets. The high-resolution parcel and address point data were easy to manipulate; they vastly improved the accuracy of both the AW areal interpolation and the PD dasymetric map.

The lower performance of the SW and LV methods was due to their insensitivity to a variety of factors, including parcel size and land use. Neither rural housing spacing nor residential land use can be accurately determined from either a LANDSAT image or a street centerline file. As a result, the SW and LV methods were not able to accurately pinpoint the majority of uninhabited areas. On the other hand, the AW and PD methods demonstrated their own sources of error. The local data lacked information related to vacancy rates, making empty residences a prime source of error in the AW and PD methods. Yet another confounding factor, most applicable in urban areas, was multi-family housing. There were likely many cases where multiple households lived on the same parcel yet shared an address. If there were no distinct addresses, then it was assumed that the parcel was occupied by one household. Since the average household size was used in all AW and PD calculations, the vacancies and multi-family errors were propagated into the entire dataset. These sources of error therefore contribute to the entire result and not just the areas directly affected by the discrepancies.

Whatever the method is used to determine the location of the population, it is important to keep in mind that the output of each of these models is a representation of probability and not reality. If one uses a street centerline file or a LANDSAT image, one would have to assume that there is an equal probability of finding a residence along the entire road or across an entire land cover class. The use of address point and parcel data significantly narrows the geography of the residences, increasing the probabilities in certain locations and decreasing the probabilities in others. The more precise the ancillary data and the more sophisticated the density predictions, the more useful the model was in predicting the population of the target zones.

This principle of probability is what differentiates AW from PD. The former is vector-based, while the latter is raster-based. Their differing data formats resulted in fundamental differences between the treatments of housing located along the boundaries of target zones. The AW method put the whole weight of probability of a home location at one discrete point. In contrast, the PD method spread out the weight of probability across an entire parcel. Whenever a parcel straddled two target zones, the AW method would pick one or the other target zone, while the PD method would spread the household population between the two zones. The PD method was statistically more accurate overall. The drastically higher accuracy of the PD method over the AW method in Anson County was a result of the nature of the address points in combination with very large parcels. Closer examination of the address points with orthophotography showed that the county frequently placed the points at driveway entrances rather than at

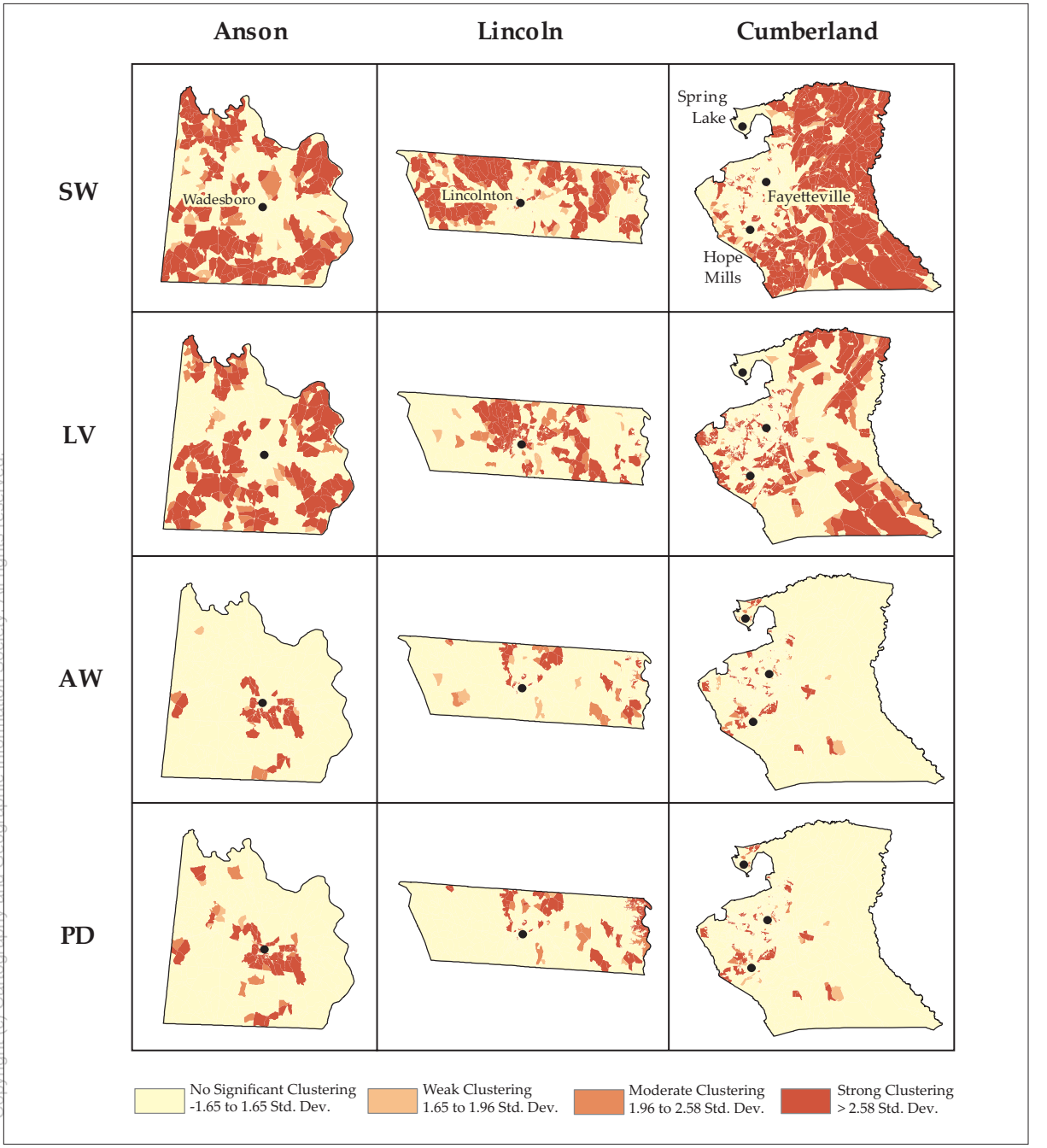

Figure 4. Local Moran's I results by county.

home locations. Since this dataset is maintained for the benefit of emergency response, it is likely that many other sparsely populated counties would follow a similar procedure. Therefore in very rural areas, it would be wise to investigate the placement of the address points before deciding which method to employ.

The new methods are significantly less prone to error clusters than the previous methods. Including the local ancillary data seems to partially overcome the previously observed tendency for errors just outside urban centers. Although there is no consistent error on the urban fringe, as seen by Yuan et al. (1997) and Eicher and Brewer (2001), the strongest errors in the AW and PD methods were found just outside the population centers. These errors are most likely due to the impact of rapid housing development on the new algorithms. When agricultural or natural land transitions to a residential community, the developer commissions

a surveyor to create a plat, dividing the land into several single family parcels. This parcel division precedes the building of the structure. After the street is paved and the structure built, emergency management wants an address point incorporated into their routing database. A family moving into the home is the last step of the process. Therefore in rapidly developing areas, one can expect to find many vacant addresses and even more empty parcels. Both Anson and Lincoln counties encountered more error clusters with the PD method than with the AW method. Cumberland showed similar error clusters between the two methods, with AW having slightly more clusters than PD. This result is consistent with the population growth patterns of each of the counties. As long as vacancies remain undocumented, discrepancies in transition areas will perpetuate.

Although consistent with previous dasymetric mapping studies, a limitation to this research was the use of a LANDSAT-based land cover raster for the LV method. Imagery with higher resolution than LANDSAT is increasingly available to the public. This new imagery has an enhanced ability to identify buildings, particularly in rural areas. It is likely that the use of such imagery would eliminate the problems of delineating individual structures from their surrounding land. However, imagery by nature is unable to make a firm distinction of different human uses of structures. One might be able to pick out a building clearly but not necessarily determine if the building is a business or a residence. Since parcel data are maintained for tax purposes, the data not only contain attribution regarding the existence of a structure, but also the use of that structure. Logically, this additional information ought to provide greater accuracy than a dasymetric map based on high resolution imagery. Additional research would need to take place to compare the two.

These new methods can be readily applied throughout the country wherever appropriate ancillary data are available. In particularly rural areas of the West and Midwest, the use of highly accurate address points would yield the most accurate results. Care would need to be taken that the points are located at residences, as opposed to at mailboxes or driveway entrances. In the latter case, where the address points are long distances from the actual residences, the PD method would be more appropriate. In contrast, the AW method would be preferable in rapidly developing transition areas where empty parcels are common. Since parcel data are more readily available throughout the country, the use of parcel boundaries would be an acceptable substitution to address points in any area of interest.

This research generated target zones with much greater detail than the inputted census data with accuracy greater than the literature. The methods can be applied to geographies that lack detailed information even through the decennial census. Furthermore, they can be used to create timely population distributions in geographies that are only updated by the American Community Survey on the county level. The AW and PD methods should be further tested in different landscapes to ensure that the accuracy remains reliable under a wide variety of conditions. More research focusing on the target zone boundaries could further investigate the sources of the differing errors between the two methods.

ACKNOWLEDGMENTS

The author would like to thank Rick Bunch and the editors and reviewers of this journal for their comments and suggestions on early drafts of this paper. This research was supported by the Center for GISc at the University of North Carolina at Greensboro, and is related to the author's ongoing dissertation work with the Geography Department of the University of North Carolina at Greensboro.

REFERENCES

Eicher, C.L., and C. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28(2): 125-38.

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

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

Gillespie, C.C. (ed.). 1970-1990. *Dictionary of scientific biography*. New York, New York: Scribner.

Goodchild, M.F., and N.S.-N. Lam. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-processing* 1: 297-312.

Hawley, K., and H. Moellering. 2005. A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science* 32(4): 411-23.

Lam, N.S.-N. 1983. Spatial interpolation methods: A review. *American Cartographer* 10(2): 129-49.

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., and D. Unwin. 1994. Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal* 31: 21-6.

Lincoln County Planning Department. 2007. Land use plan. [http://ww.lincolncounty.org/DocumentView.asp?DID=283; accessed April 25, 2009].

Mark, M., and F. Csillag. 1989. The nature of boundaries on 'area-class' maps. *Cartographica* 26: 65-79.

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

Moon, Z.K., and F.L. Farmer. 2001. Population density surface: A new approach to an old problem. *Society & Natural Resources* 14: 39-49.

One Map, N.C. 2009. NC GIS inventory. [http://www.nconemap.com/GISInventory/tabid/288/Default.aspx; accessed August 1, 2009].

Ott, R.L., and M. Longnecker. 2001. *An introduction to statistical methods and data analysis*. Pacific Grove: Thomas Learning, Inc.

Riebel, M., and M. Buffalino. 2005. Street-weighted interpolation techniques for demographic count estimates in incompatible zone systems. *Environment & Planning A* 37: 127-39.

Schmid, C.F., and E.H. MacCannell. 1955. Basic problems, techniques, and theory of isopleths mapping. *Journal of the American Statistical Association* 50(269): 220-39.

Stage, D., and N. von Meyer. 2006. An assessment of parcel data in the United States, 2005 survey results. Federal Geographic Data Subcommittee on Cadastral Data. [http://nationalcad.org/data/documents/Cadastral_Inv_2005_v2.pdf; accessed April 25, 2009].

Tobler, W.R. 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74(367): 519-30.

U.S. Census Bureau. 2000. Census tracts: Cartographic boundary files descriptions and metadata. [http://www.census.gov/geo/www/cob/tr_metadata.html; accessed April 24, 2009].

U.S. Census Bureau. 2002. Guide to the economic census. [http://www.census.gov/econ/census02/guide/g02gmaps.htm; accessed April 25, 2009].

U.S. Census Bureau. 2008. American community survey. [http://factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_pageId=sp1_acs&_submenuId=; accessed April 21, 2009].

U.S. Census Bureau. 2009. History. [http://www.census.gov/history/; accessed October 15, 2009].

Wright, J.K. 1936. A method of mapping densities of population with Cape Cod as an example. *Geographical Review* 26: 103-10.

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

Yuan, Y., R.M. Smith, and W. Frederick Limp. 1997. Remodeling census population with spatial information from LANDSAT TM imagery. *Computers, Environment and Urban Systems* 21(3/4): 245-58.