A COMPARATIVE ANALYSIS OF AREAL INTERPOLATION METHODS

A Thesis

Presented in Partial Fulfillment of the Requirements for

The Degree Master of Arts in the

Graduate School of The Ohio State University

By

Kevin J. Hawley, B.S.

*****

The Ohio State University
2005

Masters Examination Committee:

Dr. Harold Moellering, Advisor

Dr. Ningchuan Xiao

Dr. Michael Tiefelsdorf

Dr. Carolyn Merry

Approved by

Advisor
Graduate Program in Geography

ABSTRACT

This research implements and compares four different methods of areal interpolation for estimating the population within a set of target zones based on the known population within source zones. The methods include the areal weighting method (Lam, 1983), the pycnophylactic method (Tobler, 1979), an implementation of the dasymetric method (Wright, 1936, Eicher and Brewer, 2001) using remotely sensed data as ancillary information, and an implementation of the network method (Xie, 1995) using the road network as ancillary information. Both the dasymetric and network methods make use of ancillary data, whereas the areal weighting and pycnophylactic methods do not use ancillary data. The interpolation is conducted from source zones (U.S. Census tracts) of lower spatial frequency to target zones (U.S. Census block groups) of higher spatial frequency. The source zones are census tracts and the target zones are census block groups. The methods are implemented in three different counties: Franklin County, Ohio, Hamilton County, Ohio and Jefferson County, KY. Measures of the accuracy of the interpolation methods include the root mean square error (RMSE), an adjusted RMSE measure that normalizes the RMSE based on the population within each target zone and percent error maps for spatial errors. Other visualization methods include scatterplot diagrams and histograms of the errors associated with each of the areal interpolation methods.

This research finds that in terms of the RMSE, the network method produces the most accurate results, followed by the dasymetric method. The areal weighting method produces the least accurate results and the pycnophylactic method is better than the areal weighting method. The adj-RMSE shows more complicated results but mirrors the RMSE results in some cases. Patterns of spatial error are also examined by discussion of percent error maps.

Dedicated to my parents Douglas and Linda Hawley
who have provided a great deal of support throughout my life.

ACKNOWLEDGMENTS

I would like to thank my adviser, Dr. Harold Moellering, for his support, encouragement and and guidance with this research. Dr. Moellering has provided a great deal of inspiration to me throughout my studies in geography and particularly analytical cartography.

I would like to thank the members of my committee, Dr. Ningchuan Xiao, Dr. Carolyn Merry and Dr. Michael Tiefelsdorf, who all provided excellent insights, comments, suggestions and invaluable references to other works. I would also like to thank Dr. Alan Murray who provided earlier comments and suggestions with this research. I would like to express special thanks to Carolyn Merry who granted me access to the Region 1 computing facilities through Dan Vehr (Systems Manager at Region 1) for this research.

VITA

January 18, 1979	Born – Cincinnati, Ohio
2001	B.S. Geography, The Ohio State University
2001 – present	Graduate Teaching Associate The Ohio State University

FIELDS OF STUDY

Major Field: Geography
Specialization: Analytical Cartography and Geographic Information Science

TABLE OF CONTENTS

Page

Abstract ii
Dedication iv
Acknowledgments v
Vita vi

List of Tables ix
List of Figures x

Introduction 1

Chapters:

1. Literature Review 4

1.1 Areal Interpolation Methods without Ancillary Data 4
1.2 Areal Interpolation Methods Using Remote Sensing As Ancillary Data 6
1.2.1 Regression Methods of Areal Interpolation Using Remote Sensing Data 7
1.2.2 Dasymetric Methods of Areal Interpolation Using Remote Sensing Data 8
1.3 Road Network Areal Interpolation Methods 10
1.4 Other Areal Interpolation Methods Making Use of Ancillary Data 13

2. Research Design 17

2.1 Selection of Areal Interpolation Methods 17
2.2 Implementation of Areal Interpolation Methods 19
2.2.1 Areal Weighting Method 19
2.2.2 Pycnophylactic Method 20
2.2.3 Dasymetric Method Using Remote Sensing Data 22
2.2.4 Network Hierarchial Method 23
2.3 Analysis of Results for the Areal Interpolation Methods 24
2.4 Data Collection for Areal Interpolation Research 27
2.5 Facilities Used for the Research 29
2.6 Software Used for the Research 30

3. Results 31

3.1 Intermediate Method Results 31
3.1.1 Intermediate Results for the Pycnophylactic Method 32
3.1.2 Intermediate Results for the Dasymetric Method 34
3.1.3 Intermediate Results for the Network Method 36
3.2 RMSE and adj-RMSE Results 37
3.3 Spatial Visualization of Errors Associated with Areal Interpolation Methods 43
3.3.1 Spatial Errors for Franklin County 44
3.3.2 Spatial Errors for Hamilton County 51
3.3.3 Spatial Errors for Jefferson County 59
3.3.4 Investigation of the High Density Problem 66
3.4 Other Tools for Comparing Areal Interpolation Methods 69

4. Summary and Conclusion 76
4.1 Conclusions 78
4.2 Future Research 79

Bibliography 82

Appendix A: Implementation of Areal Interpolation Methods 86

LIST OF TABLES

Table Page

1.1 Characteristics of Areal Interpolation Methods 16

2.1 Statistics for Counties 29

3.1 Average Spatial Resolution for Source and Target Zones 31

3.2 RMSE and adj-RMSE Results for Franklin County 37

3.3 Adj-RMSE Results for Franklin County without Influential Zone........................ 38

3.4 RMSE and adj-RMSE Results for Hamilton County.. 39

3.5 Adj-RMSE Results for Hamilton County without Influential Zone 40

3.6 RMSE and adj-RMSE Results for Jefferson County 41

3.7 Adj-RMSE Results for Jefferson County without Influential Zone....................... 41

LIST OF FIGURES

Figure Page

2.1 Source and Target Zones for Franklin County 28

2.2 Source and Target Zones for Hamilton County 28

2.3 Source and Target Zones for Jefferson County 29

3.1 Pycnophylactic Surface for Franklin County 33

3.2 Pycnophylactic Surface for Hamilton County 33

3.3 Pycnophylactic Surface for Jefferson County 34

3.4 Residential Land Use Areas 35

3.5 Selected Roads for Network Method 36

3.6 Actual Population Density for Franklin County 44

3.7 Estimated Population Density Maps for Franklin County 45

3.8 Areas of Interest for Percent Error Maps 47

3.9 Percent Error Maps for Franklin County 48

3.10 Actual Population Density for Hamilton County 52

3.11 Estimated Population Density Maps for Hamilton County 53

3.12 Areas of Interest for Percent Error Maps 55

3.13 Percent Error Maps for Hamilton County 56

3.14 Actual Population Density for Jefferson County 60

3.15 Estimated Population Density Maps for Jefferson County 61

3.16 Areas of Interest for Percent Error Maps 62

3.17 Percent Error Maps for Jefferson County.. 63

3.18 Effects of High Density Structures.. 68

3.19 Scatterplots for Franklin County.. 71

3.20 Scatterplots for Hamilton County.. 72

3.21 Scatterplots for Jefferson County... 73

3.22 Error Histograms.. 75

INTRODUCTION

Areal interpolation is the process of estimating the values of one or more variables in a set of target zones, based on known values that exist in a set of source zones. The need for areal interpolation arises when data from different sources are collected in different tessellations of areal units. In the United States, for example, spatial data that has been collected in census zones, such as block groups and tracts, is very common. Many businesses that make use of spatial data will often aggregate their data into zip codes. On the other hand, a useful data source may be aggregated based on natural rather than political boundaries. Trade or service areas may also define a set of zones that may be useful in spatial analysis. Because zones such as zip codes, census tracts, natural boundaries and trade areas are not co-located, and hence incompatible with one another, areal interpolation is necessary to make use of all of this data from various sources. For example, an analyst may have data available for a particular set of zones, but is lacking important information such as the number of people within each zone. Data collected in another set of spatial units contains this information. Areal interpolation could be used to provide an estimate of the number of people within the original set of zones.

Areal interpolation methods may also be implemented for answering geographic questions, such as: How many people are within a particular service area? We may also find uses for areal interpolation in applications involving natural disasters. Natural disasters do not follow political boundaries, and with the estimated boundaries of impact, the number of homes, businesses and people affected need to be estimated. Areal interpolation may also be useful in comparing spatial patterns in different time periods for political units. The boundaries of political units may change over time, and in order to make comparisons between two or more time periods with different boundaries, one needs to normalize the data to account for the differing boundaries. Areal interpolation is one solution to this problem.

There are many different methods of areal interpolation. Each method is unique in its assumptions about the underlying distribution of the data. The more modern methods make use of ancillary data, which can give insight to the underlying distribution of the variable inside the source zone. Having an understanding of the distribution of the data in the source zones can improve the results of the interpolation. The choice of which method to use may be dependent on various factors such as accuracy, ease of implementation, data availability and time. This research will be conducted as a comparative analysis of four different areal interpolation methods using population as the variable of interest. The methods that will be analyzed and implemented include the areal weighting method, the pycnophylactic method, a dasymetric method using remote sensing data as ancillary information, and the road network hierarchial weighted method using the road network as ancillary data. This research makes use of hierarchial census zones. Census tracts serve as source zones and census block groups serve as target zones.

Therefore, the actual population within each of the target zones is already known. The interpolated values within each target zone will be compared to known values within the target zones.

CHAPTER 1

LITERATURE REVIEW

1.1 Areal Interpolation Methods Without Ancillary Data

The literature discussed in this section focuses on areal interpolation methods that do not make use of ancillary data. Lam (1983) distinguishes between two main types of areal interpolation methods: Non-volume-preserving and volume-preserving methods. The "point-based areal interpolation approach" makes use of point interpolation methods, where the points are generally the centroids of the source zones. The main problem associated with this method is that it is not volume preserving. Volume preserving, in the case of areal interpolation, refers to preservation of the total value of the variable within each source zone.

The volume preserving methods discussed by Lam (1983) include the overlay method and the pycnophylactic method. The overlay method is commonly referred to as areal weighting. Intersection zones are created by the overlay of source and target zones. Target zone z values are then estimated based on the values of the source zone and the proportion of the intersection with the source zone using the following formula:

$$V_t = \sum_s \left(\frac{V_s A_{ts}}{A_s} \right) \qquad (1)$$

where V is the value of the variable,
A is the area;
the subscripts s and t refer to source and target zones respectively.

Although this method does preserve volume, it assumes that the variable is homogeneously distributed within the source zones (Lam, 1983).

The pycnophylactic method "assumes the existence of a smooth density function which takes into account the effect of adjacent source zones" (Lam, 1983). This is a method originally proposed by Tobler (1979). This method originally assigns each grid cell the z value of the source zone divided by the number of cells within that source zone. Each cell is then averaged with its neighbors. The predicted z values in each source zone are then compared with the actual z values, and adjusted to meet the pycnophylactic condition. The pycnophylactic condition is defined as follows:

$$\iint_{R_i} Z(x, y)dxdy = H_i \tag{2}$$

where R_i is the *i*th region,
H_i is the value of the variable in region i and
Z(x,y) is the density function

This is an iterative procedure that continues until there is either no significant difference between predicted z values and actual z values within the source zones, or there have been no significant changes of cell values from the previous iteration. Tobler (1979) describes the pycnophylactic method in the following manner where each source zone (state in this case) is represented by a different colored block of clay that is proportional to the population within the zone. The next step is to "sculpt this surface until it is perfectly smooth, but without allowing any clay to move from one state to another and without removing or adding any clay." One of the key assumptions of this method is that unlike the areal weighting method, it assumes a heterogeneous z variable distribution. A heterogeneous variable distribution means that the z variable is not evenly distributed within each source zone whereas a homogeneous variable distribution assumes that the variable is evenly distributed within the source zone.

Rase (2000) has recently refined Tobler's method to incorporate Triangulated Irregular Networks (TIN) rather than the traditional rectangular grid used by Tobler (1979). The reason for using the TIN model as opposed to the rectangular grid is that the original boundaries of the polygons can be maintained. The rectangular grid does not maintain the original boundaries of the polygons, whereas the more flexible TIN model is able to use the points of the geographic boundaries as vertices in the TIN model. This allows for the "error resulting from converting the polygons to a regular grid in the 'classic' version of the pycnophylactic interpolation" (Rase, 2000) to be avoided.

Another areal interpolation method that does not use ancillary data is the point-in-polygon method. The point-in-polygon method transfers the attribute value within a source zone to a target zone, if its “representative point” is inside the target zone (Okabe and Sadahiro, 1997). Although this method requires little computation time, the results using the same dataset can vary greatly based on the locations chosen as the representative points. These points may be centroids, arbitrary points or some type of weighted centers, such as the center of population (Okabe and Sadahiro, 1997).

1.2 Areal Interpolation Methods Using Remote Sensing As Ancillary Data

In recent years, remote sensing data and imagery has become widely available in the public domain. Satellite imagery is very powerful in that its values in different wavelength bands allow one to classify the data into land use types using different spectral signatures for each kind of land use. A land use map of an area can provide better information to an analyst about the distribution of a z variable, such as population. For example, it is quite obvious that no human population exists in a lake or river, and that an urban or suburban area is more densely populated than a woodland area. Therefore, linking this type of ancillary data to an areal interpolation technique can provide for improved results.

1.2.1 Regression Methods of Areal Interpolation Using Remote Sensing Data

Several implementations of areal interpolation methods using remote sensing as ancillary data exist. Langford et al. (1991) have developed a technique that makes use of remote sensing data and term this as "intelligent interpolation." After a satellite image is classified into different land use types (by using different spectral signatures), it is then possible to determine how many pixels of each land use type fall within each source zone. Regression analysis can then be performed based on the source zone populations. The resulting parameters are used to estimate the populations of the target zones. The authors discuss three statistical models: the shotgun model, the focused model, and the simple model. The shotgun model uses all of the land use types as independent variables. The focused model only includes land use types that are known to have population. Also, the authors aggregated these land use types into two variables: dense and residential. The simple model uses only one independent variable, which is a combination of the dense and residential variables used in the focused model. The authors (Langford et. al., 1991) found that although the shotgun model provided the best overall fit of the data, the focused model is preferred because it does not allow negative populations to be predicted in any target zone. In other words, negative parameters could in theory predict a population within an area that is less than zero. Therefore the additive focused model is a better representation of reality. The loss of information from the shotgun model to the focused model is not statistically significant.

1.2.2 Dasymetric Methods of Areal Interpolation Using Remote Sensing Data

An alternative to the regression methods discussed previously when interpolating with remote sensing data, is a dasymetric method. Robinson et al. (1984) describe a dasymetric map as one in which the "mapping unit boundaries are independent of enumeration boundaries." The mapping unit boundaries are also assumed to be relatively homogeneous and are created based on other sources of information. Wright (1936) developed a method to map population densities in his classic article using Cape Cod as the study area. Wright used topographic sheets as ancillary data to create areas that are uninhabitable and areas of differing densities, which are assigned by a principle that he terms "controlled guesswork." The dasymetric map allows the reader to better understand the distribution of the population as compared to the choropleth map. The increased availability of remote sensing data allows dasymetric maps to be created with greater ease than has been previously possible. The general principles of dasymetric mapping can be applied to the areal interpolation problem.

Fisher and Langford (1995) were the first to publish results of areal interpolation using the dasymetric method. The dasymetric method was a variant of Wright's (1936) method in that it uses a binary division of the land use types. A binary division of land use types refers to the presence of absence of population in the area. This method was tested against four others: The areal weighting method, the shotgun model, the focused model, and the simple model. The authors made use of Openshaw's (1977) Monte Carlo simulation for areal units. The algorithm produces a "pseudo-random aggregation of N zones into M zones.....such that all N-zone members of each M zone are spatially contiguous" (Openshaw, 1977). This algorithm is used by Fisher and Langford (1995) to create target zones, that will have a known population, which is the sum of the known N zone populations within each target zone. 250 zone aggregations were created and each method was tested on each zonal system. Their results pointed to the dasymetric method as having the most accurate results. The areal weighting method was the poorest performer. The shotgun model generally provided the best results among the regression methods. Although Fisher and Langford's work is a very thorough study of areal interpolation methods, it is more limited in terms of which methods were compared, versus the research proposed by the author in this work.

Cockings et al. (1997) followed up on previous work (Fisher and Langford, 1995) by suggesting measures for parameterization of areal interpolation errors. A similar Monte Carlo technique was used, with only the areal weighting and dasymetric methods. The results of their research found that geometric characteristics of the target zones were significant factors in the errors associated with the areal weighting method, but not in the errors associated with the dasymetric method. Errors associated with the dasymetric method were correlated with population density.

It is important to understand that like the regression method, the dasymetric method has several forms of implementation. Eicher and Brewer (2001) made use of three dasymetric mapping techniques for areal interpolation. These include the binary method, the three-class method and the limiting variable method. The binary method is the most straightforward and assigns 100 percent of the population to only urban and agricultural/woodland land use types. The three class method uses a weighting scheme to allocate population to urban, agricultural/woodland and forested land use types. The weights are subjectively determined by the analyst and are assisted by knowledge of the study area. The limiting variable method allocates population based on population density constraints that are set for each land use type. Eicher and Brewer (2001) found that the limiting variable method produced the most accurate results.

Although both the regression and dasymetric methods use remotely sensed data as ancillary data, they have a very distinct difference. This difference is that the regression method is global and the dasymetric method is local. The regression method assumes that every pixel of a particular land use type contains the same population throughout the entire study area, regardless of which source zone it is contained within. On the other hand, the dasymetric method is local in that the same land use category within a particular source zone has the same population density. However, two land use zones of the same category in different source zones with the same geometric area will most likely have different populations allocated to them.

1.3 Road Network Areal Interpolation Methods

When dealing with the problem of areal interpolation for population related variables such as raw population or housing units, a road network could certainly be useful as ancillary data. Because residential homes are almost always located on roads, one can usually assume that with a greater density of roads, the greater the population density.

Xie (1995) has developed three algorithms for areal interpolation with the use of the road network as ancillary data. The first and most straightforward of these methods is the network length method. The network length method allocates population to the target zones based on the total length of road within the target zone. This method makes use of the overlay operation. Each source zone is assigned a population per unit length. The target zone populations are determined by the sum of the intersection zone populations.

The network hierarchial weighting method (Xie, 1995) takes additional information about the road segments into account. This method uses Census Feature Class Codes (CFCC's), which are road classification codes (U.S. Bureau of the Census, 1994). For example, interstate highways are assigned CFCC's of A11 through A18. Each road type is weighted based on its CFCC. This offers the possibility of a great improvement over the network length method. For example, no houses are attached to the side of interstate highways. The sum of the weights in each source zone must be 1, and the weights are subjective.

The network housing-bearing method (Xie, 1995) makes use of to and from addresses associated with the line segments of a road network. The basic principle behind this method is that "population can be allocated to houses and houses can be attached to street segments according to the address ranges on each side of the streets in each zone." Therefore, this method uses the additional information of the number of households in each source zone to derive the population per household within each source zone. Xie found that the network hierarchial method provided the best results of the three overlay network algorithms.

Reibel and Bufalino (2004) also made use of the network length method developed by Xie (1995). They tested the network length method against the areal weighting method for interpolating 2000 census data from 1990 census data in Los Angeles. The network length method provided more accurate results than the areal weighting method.

Voss et al. (1999) also experimented with areal interpolation methods that make use of road networks as an ancillary variable. Two methods were explored in his research: the road segment length method and the internal node counts method. The road segment length method is similar to the network length method of Xie (1995). The internal node counts method uses the zero-dimensional node object as the variable of interest. Nodes are geometry-topology (GT) spatial objects (USGS, 1994), which exist in road networks at the beginning and end of each road segment, and at any intersections with other road segments. Population is allocated to target areas based on the portion of the source zones total node count that are located in each intersection zone. These two methods were tested against the point-in-polygon method and the areal weighting method. Their results showed that the road segment length and internal node counts improved upon the methods that do not make use of ancillary data. The internal node counts method proved to be the most accurate of the two road network methods, with the lowest absolute error and mean absolute error.

1.4 Other Areal Interpolation Methods Making Use of Ancillary Data

There have been other statistical techniques developed to improve areal interpolation. These methods make use of ancillary data, but they are more general than the remote sensing or road network methods in that the ancillary information is usually quite flexible. In other words, the methods discussed previously use specific ancillary information such as the road network or land use data.

Flowerdew (1988) developed an areal interpolation method that operates as a Poisson process on a set of binary variables. The binary variable defines the presence or absence of each variable. The parameters are determined by a regression on the source zones, and then applied to the target zones based on the value of the binary variable. In the case of the presence of multiple variables in one target zone, the parameters are weighted based on some measure, such as the areal coverage of this variable (in the case of land cover) or some other meaningful ratio measure. In the case of land use types as the binary variables, this method is similar to that of Langford et al. (1991) discussed previously.

Flowerdew and Green (1991) expanded on this method of areal interpolation by including a piece of theory from Dempster (1977) in the field of statistics. The Expectation-Maximization (EM) algorithm was developed as a statistical technique to deal with missing data. Flowerdew and Green (1991) propose that the areal interpolation problem can be thought of as a missing data problem. The EM algorithm is a two-step iterative procedure. The E-step "computes the conditional expectation of the missing data given the model and observed data." The M-step uses the values computed in the E-step to fit the statistical model by means of maximum likelihood. The EM algorithm uses target zone variables to predict the variable of interest. The authors implemented several interpolations using variables, such as the party winning the constituency votes, number of cars per household, and the percentage of households with more than one person per room. These variables are all available at the target zone level. The results show a significant improvement over the areal weighting method. Flowerdew and Green (1992) later applied the EM algorithm to continuous variables, but were surprised by the results, which showed a poorer fit than with the binary variable.

Goodchild et al. (1993) developed an alternative method that uses ancillary data known as control zones. Control zones are defined as a third set of areal units, which each have constant densities. The analyst is generally able to create these control zones based on local knowledge of the area. By obtaining intersections of both the source and control zones and the target and control zones, control zone densities can be estimated, and target zone populations can be estimated based on the homogeneous nature of the control zones.

As discussed in section 1.1, many of the point interpolation procedures such as inverse-distance weighting and kriging have been applied to the areal interpolation problem, but lack the characteristic of volume preservation. Kyriakidis (2004) discusses area to point interpolation techniques that make use of the geostatistical method of kriging. However, unlike the traditional point interpolation techniques, Kyriakidis is able to preserve the volume of the variable being interpolated within the source zone. This is a promising technique, and kriging allows for the addition of ancillary data.

Deichmann (1996) discusses a method of areal interpolation termed smart interpolation (Deichmann and Eklundh, 1991). Smart interpolation makes use of many different ancillary data sets. These data sets include the location and size of urban areas, the locations of transportation structures, rivers, and any other datasets that relate to the distribution of the population. This is a raster-based interpolation method that creates a series of weights for all of the cells within a source zone. The weights are determined heuristically based on the ancillary information provided. Turner and Openshaw (2001) expanded on the idea of smart interpolation by incorporating neural networks to estimate parameters for their models. Neural networks are "universal approximators capable of learning how to represent virtually any function no matter how complex or discontinuous. (Turner and Openshaw, 2001)." The input into the neural network includes variables, such as distances to parks, distances to main roadways, distances to rivers, distances to urban areas, elevation, and even the population estimate from a pycnophylactic model. They found that their method performs better than the pycnophylactic and areal weighting method.

Table 1.1 compares the methods previously discussed in terms of some of their important characteristics including the assumptions about the variable distribution, ancillary data used, spatial dimensionalities and spatial functions. The variable distribution characteristic describes whether the particular areal interpolation method assumes a homogeneous or heterogeneous variable distribution within a source zone. The ancillary data characteristic refers to the ancillary data (if any) that is used to improve the areal interpolation method. The dimensionalities refer to the original data and any ancillary data that may be included in a particular areal interpolation method. The functions characteristic refers to the main types of spatial operations that are used to implement a particular areal interpolation method.

Method	**Variable Distribution**	**Ancillary Data**	**Dimensionality**	**Functions**
Areal Weighting	Homogeneous	None	2-D	Spatial Overlay
Pycnophylactic	Heterogeneous	None	2-D, 2.5-D surface	Neighborhood Smoothing, Zonal Statistics
Point-In-Polygon	Homogeneous	None	2-D, 0-D	Point-In-Polygon
Regression	Heterogeneous	Land Use/Land Cover	2-D	Regression, Spatial Overlay Classification
Dasymetric	Heterogeneous	Land Use/Land Cover	2-D	Spatial Overlay Classification
Network	Heterogeneous	Road Network	2-D, 1-D	Spatial Overlay
EM	Heterogeneous	Various Variables	2-D	Statistical Algorithm
Control Zones	Heterogeneous	User Defined Zones	2-D	Spatial Overlay Digitizing
Smart	Heterogeneous	Various Variables	2-D, 0-D, 1-D, 3-D	Heuristic Algorithm

Table 1.1. Characteristics of Areal Interpolation Methods

This literature review covers a large number of areal interpolation methods that range from simple methods without ancillary data to very complex methods, which include other valuable ancillary information to better understand the distribution of the variable. The next section will describe the methods of areal interpolation that will be further explored, along with a testing framework used to compare the various methods.

CHAPTER 2

RESEARCH DESIGN

2.1 Selection of Areal Interpolation Methods

The previous section has introduced the reader to a wide variety of areal interpolation methods. The choice of which method to use largely depends on accuracy, data availability, software availability, time and ease of implementation. Each of these methods make certain assumptions regarding the z variable's distribution. The areal weighting method makes the assumption that the z variable of interest is homogeneously distributed. Although this method is quite simple in terms of implementation, rarely is a z variable, such as population, homogeneously distributed in the real world. The pycnophylactic method, which is similar to the areal weighting method in terms of the absence of ancillary data, differs in that it treats the z variable as being heterogeneous. The remote sensing methods assume that population is related to the land use type. One would expect a higher population density in area of suburban or urban land use types than in areas of forested or agricultural land use types. The road network methods assume that population is related to the density of the road network. The network hierarchial method also takes into account the classes of the roads (highway, residential, connecting road). The EM algorithm uses data that are available for target areas, which can assist in determining the variables distribution. The use of control zones assumes that zones of homogeneous population density can be used as a third set of zones to improve the accuracy of the interpolation. The neural network methods incorporate a large variety of ancillary data to improve the accuracy of the interpolation.

Although a variety of methods exist, this research will focus on comparing four different areal interpolation methods. The following methods have been chosen for the analysis:

* The areal weighting method (Lam, 1983)

* The pycnophylactic method (Tobler, 1979; Lam, 1983)

* The limiting variable dasymetric method using remote sensing data (Wright, 1936; Fisher and Langford, 1995; Cockings et al. 1997, Eicher and Brewer, 2001)

* The network hierarchial method (Xie, 1995).

The areal weighting method is chosen because it is used widely and has been the leading method of areal interpolation for many years. The pycnophylactic method is chosen because it is a second method that does not make use of ancillary data, yet it differs from areal weighting in its assumptions about the distribution of the variable. Because remote sensing has become so widely available in the spatial sciences and many implementations of areal interpolation make use of such data, one method from this area has also been chosen. The dasymetric method seems the most promising of the methods that make use of land use data. Finally, one of the more modern methods makes use of the road network as ancillary data. It should be interesting to see how the results of the road network methods compare to the remote sensing methods.

Another interesting property of each of these methods is the spatial dimensionality of the data that is used in the areal interpolation procedure. The areal weighting method is strictly a 2-D procedure, which makes use of least common geographical units (LCGU's) and their areas of intersection. The pycnophylactic method, which like the areal weighting method does not use ancillary data, creates a 2 ½ -D surface from the 2-D zonal data. This method assumes that a continuous distribution of the z variable is a reasonable assumption. The dasymetric method uses a set of 2-D ancillary data that contains valuable information regarding the spatial distribution of the z variable of interest. The network method uses 1-D network data as ancillary data, which also contains valuable information regarding the spatial distribution of the variable of interest. Therefore, these methods provide an interesting analysis that takes into account not only the underlying assumptions of the data, but also the inherent spatial dimensionality of the data.

The z variable of interest in this research is population. As far as the author is aware, these particular methods of areal interpolation being analyzed here have not been tested against each other in any published work.

2.2 Implementation of Areal Interpolation Methods

This section introduces the implementation issues for each of the five areal interpolation methods that will be compared in this research.

2.2.1 Areal Weighting Method

The first step for implementation of the areal weighting method is to overlay the source and target zones. Spatial overlay is a common task in the spatial sciences that can be performed in most GIS software packages. The spatial overlay is performed in ArcGIS 9 as an Intersect Overlay. The key is to identify each source polygon with a source ID and each target polygon with a target ID. After the overlay is performed, an intersection layer is created that contains the areas of each intersection polygon, the source ID and target ID. The source and target ID's from the intersection polygons, allows for the linkage between all three datasets. Target zone population is allocated as follows (Lam, 1983):

$$P_t = \sum_s \left(\frac{P_s A_{ts}}{A_s} \right) \tag{3}$$

where P is the population of the variable,
A is the area;
The subscripts s and t refer to source and target zones respectively

Notice that A_{ts} is the area of intersection of target zone t and source zone s. The interpolation is performed in a program written by the author in VisualBasic using MapObjects.

2.2.2 Pycnophylactic Method

For the implementation of the pycnophylactic method, Tobler's (1979) notation and formulas are used: H_i is the value of the variable (population) in region i, A_i is the area of region i, and R_i is region i. The first step is to assign each cell in a dense grid an identification number (from 1 to M), which identifies the source polygon (R_i) that the cell is contained within. The value of each cell is then initially set as follows.

$$Z_{m,n} = \frac{H_i}{N_i} \tag{4}$$

where N_i is the count of cells within source zone i. For each of the cases discussed in this research, the initial source polygon data are converted to a grid with a resolution of

100 meters (m). Next, each cell is assigned a new population, which is the average of its four neighbors:

$$Z_{i,j} = \frac{1}{4}\left(z_{i,j+1} + z_{i,j-1} + z_{i+1,j} + z_{i-1,j}\right) \tag{5}$$

The adjustments for each region are stored as the sum of the difference between the original value of the cell and the new smoothed value of the cell for each cell in region *i*. This average adjustment for each region is then added to the new smoothed value of each cell in the region (so long as no cell would have a value less than zero). The cumulative population for each region is then computed, and the average population difference is computed as follows:

$$\bar{d}_k = \left(\frac{H_k - H'_k}{A_k}\right) \tag{6}$$

The average population difference is then added to each cell in the region *k*. Tobler's process is repeated until all of the adjustments are less than a specified constraint, or the number of iterations has exceeded a maximum input iteration specified by the analyst. The result of this process is a 2 ½-D population surface. After this process has terminated, the population within each cell will be reallocated to the target zones. The population is reallocated to the target zones by first creating a grid for the set of target zones with a resolution of 100 m. The resulting raster target zone grid contains summary statistics of the pycnophylactic surface. The sum of values within each target zone represent the estimated population within each target zone.

The implementation of the pycnophylactic method is carried out in ArcView 3.2. The first step is to rasterize the source zones, which is done with the spatial analyst extension Convert to Grid. This grid is then used as the input for the pycnophylactic script written by Leopold Riedl of the Technical University Vienna (Riedl, 1998) .

2.2.3 Dasymetric Method Using Remote Sensing Data

The first step in implementing the remote sensing method is to obtain a satellite image of the area and classify the image into land use types. The classification process will utilize a supervised classification approach (Lillesand and Kiefer, 1994), where land use types are selected by the analyst as training areas. The algorithm then classifies all remaining pixels into one of the land use types based on spectral signature clustering of the unclassified pixels as they relate to the training areas. The spectral clustering algorithm makes use of the blue, green, red and near-infrared wavelength bands.

This classified image is then vectorized and finally only the residential land use types are selected. An overlay with the source polygons and land use polygons is then applied. This overlay allows for the identification of residential land use polygons within each source zone. An areal weighted method is applied to each intersection zone *I* as follows:

$$P_i = \left(\frac{P_s a_i}{\sum_{si} a_{si}} \right) \qquad (7)$$

where *P* is the population
a is the area and
subscripts *s* and *i* refer to source and intersect zones respectively

Therefore, any intersection zones within a particular source zone will have the same population density. The next step is to overlay the intersection zones with the target zones. The area of these intersection zones are calculated and areal weighting is applied to allocate the population to the target zones. This should be an acceptable procedure because a dasymetric map makes the assumption that the areas are relatively homogeneous (Robinson et al., 1984).

2.2.4 Network Hierarchial Method

The first step in the network hierarchial method is to overlay the source zones with the target zones to create an intersection layer. The intersection zone is then overlaid with the road network to create the "control-net" (Xie, 1995). A weight matrix needs to be constructed, which assigns a weight to each road class based on residential density. The sum of the weights in each source zone must equal 1. If a road class does not exist in a source zone, its weight should be 0. This implies that the weight matrix will be dynamic and depends on the roads within each source zone. For this particular implementation, a simple hierarchy is created which allocates 75 percent of the population in a source zone to the neighborhood roads and city streets class (CFCC – A4). 15 percent of the population is allocated to connecting road classification (A3). The remaining 10 percent of the population is allocated to the secondary road classification (A2). Other roads, such as interstate highways, dirt roads and jeep trails are not included in the hierarchy. Whenever a road classification(s) is not included in the source zone, the weight of the missing class is evenly distributed between the remaining class(es). The source layer is then overlaid with the network layer. This allows for determining the length of each road class within each source polygon. A weighted length for each source zone is calculated as follows:

$$WeightLen_S = \sum_{SC} \left(L_{SC} * W_{SC} \right) \tag{8}$$

Where L_{sc} and W_{sc} are the length and weight of class c in source zone s. The population-length of each source zone can then be calculated as follows:

$$PopLen_S = \frac{Pop_S}{WeightLen_S} \tag{9}$$

A similar procedure is then calculated throughout the intersection zones, which yields a WeightLen2_i. WeightLen2_i is then multiplied by PopLen_s to obtain population estimates for intersection zones. The population for the target zones is calculated by summing the population estimates in the intersection zones. This implementation makes use of the left and right polygon topology in the output overlay layers.

2.3 Analysis of Results for the Areal Interpolation Methods

To evaluate the accuracy of each of the areal interpolation methods, this research will make use of the nested hierarchy of census reporting units. Census data are collected in blocks. Blocks are then aggregated into block groups, which are further aggregated into tracts and so on. The source zones are census tracts. The target zones are census block groups. In terms of spatial frequencies, this research is interpolating from a lower spatial frequency (tracts) to a higher spatial frequency (block groups). Within each of the target zones the populations are already known (to a certain degree of accuracy). This allows for the comparison between the predicted populations and the actual populations.

The accuracy of each method will be evaluated based on the root mean square error (RMSE), an adjusted root mean square error (Adj-RMSE) (Gregory, 2000) and the percent error. The root mean square error measures will provide an overall indication of the global performance of the areal interpolation method. The root mean square error and adjusted root mean square error are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} \tag{10}$$

$$Adj\text{-}RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - \hat{z}_i}{z_i}\right)^2} \tag{11}$$

where z is the actual population for zone i
$\hat{z}$ is the predicted population for zone i
and n is the number of zones

The percent error values for each method are mapped as a choropleth map. The resulting maps may show spatial patterns of error, which could help determine which methods work best under certain circumstances (such as rural areas or urban areas). The percent error for each zone is calculated as follows:

$$PE_i = \left(\frac{\hat{z}_i - z_i}{z_i}\right)*100 \tag{12}$$

Estimated population density maps will also be created for each areal interpolation method tested. The spatial patterns presented on these maps may be compared to the actual population density maps for each set of target zones.

Scatterplots will also be used to visualize the errors associated with each of the methods. The scatterplots show the estimated population on the x-axis and the actual population on the y-axis. However, due to the large variation in population size throughout the study area, it may be appropriate to transform the population and estimated population counts. A transformation of the variable will allow for the heteroscedasticity of associated with the variation in population counts to be stabilized.

In this research the interpolation is being conducted from a lower to higher spatial frequency. Tobler (2000) created an equation to calculate the average spatial resolution of spatial objects in the irregular domain, called Resels, which is an abbreviation of resolution elements. In the regular domain, the spatial resolution is consistent throughout. However, when dealing with polygons whose boundaries are defined by political means or any other form, the average spatial resolution can give more insight to resolution of these Resel polygons. Tobler's new Resel equation works in any dimension and is calculated as follows:

$$\overline{R} = \left(\frac{km^{d}}{N} \right)^{1/d} \qquad (13)$$

where $\overline{R}$ is the Resel
km is the total study area in kilometers
d is the spatial dimension of the area
and *N* is the number of units of dimension *d*

This Resel measurement approximates the average resolution of geographic data in an irregular domain. In the regular domain one could refer to Landsat imagery as having 30 m resolution. This means each grid cell has a length and width of approximately 30 m, containing an area of 900 m^2. The Resel measurement takes the *d*th root of the average area (in the 2 dimensional irregular domain) and expresses it in the context of an average length or width similar to the measurement used in the regular domain.

2.4 Data Collection for Areal Interpolation Research

This research uses 2000 US Census population and geometric data for Franklin County, Ohio (U.S. Bureau of the Census, 2000 (a,b,d,e)), Hamilton County, Ohio (U.S. Bureau of the Census, 2000 (f,g,i,j)) and Jefferson County, Kentucky (U.S. Bureau of the Census, 2000 (k,l,n,o)). Franklin County, Ohio contains the city of Columbus, Hamilton County, Ohio contains the city of Cincinnati and Jefferson County, Kentucky contains the city of Louisville. Each county's census tracts are used as source zones. Each county's census block groups are used as target zones. Satellite images or land use data for each of the three counties are used as ancillary data for the dasymetric method. For the Franklin County case, the satellite image is classified into land use categories using a supervised classification approach. The satellite image was obtained from the OhioLINK Landsat 7 Satellite Image Server (OhioLINK, 2000). The supervised classification uses 49 training areas defined by the author. These areas were defined based on local knowledge of the area, the use of aerial photographs and the use of topographic maps. Originally, the classified image contained residential, commercial, agricultural, water, forest and grasslands classifications. The classified image was later processed to only include residential areas. In the Hamilton County and the Jefferson County cases, a Land Use Land Cover (LULC) dataset compiled by the United States Geological Survey (USGS, 2001 (a-b)) is used as the land use dataset. This dataset is for the year 2001 and was created using Landsat imagery from 1999, 2000 and 2001. At the time of this research, a current LULC dataset was not available for Franklin County. A TIGER line file (U.S. Bureau of the Census, 2000 (c,h,m)) will be used as ancillary data for the network hierarchial method. Figure 2.1 shows the source and target zones for Franklin

County, Ohio. Figure 2.2 shows the source and target zones for Hamilton County, Ohio. Figure 2.3 shows the source and target zones for Jefferson County, Kentucky. In this research the interpolation is being conducted from a lower to higher spatial frequency.

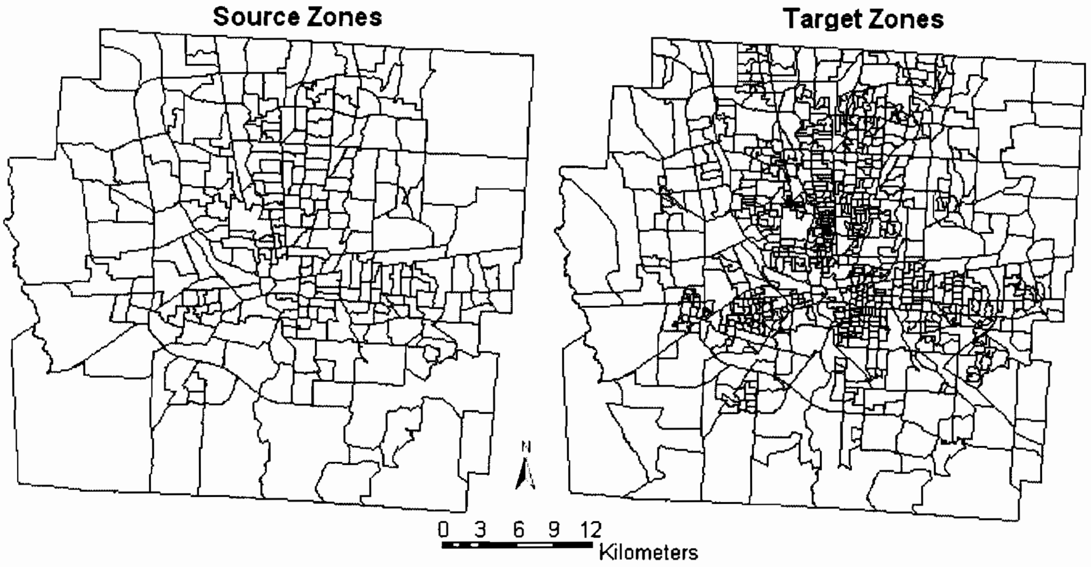


Figure 2.1 – Source and Target Zones for Franklin County, Ohio

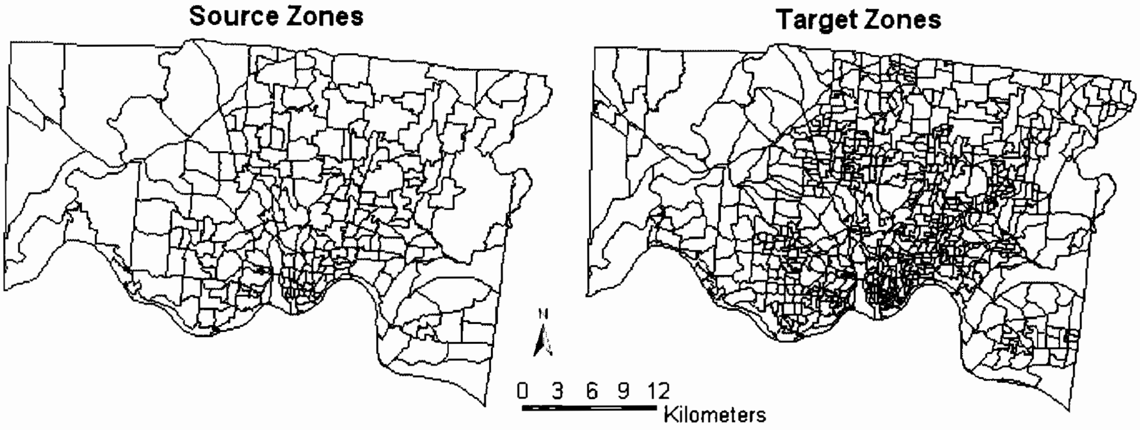


Figure 2.2 – Source and Target Zones for Hamilton County, Ohio

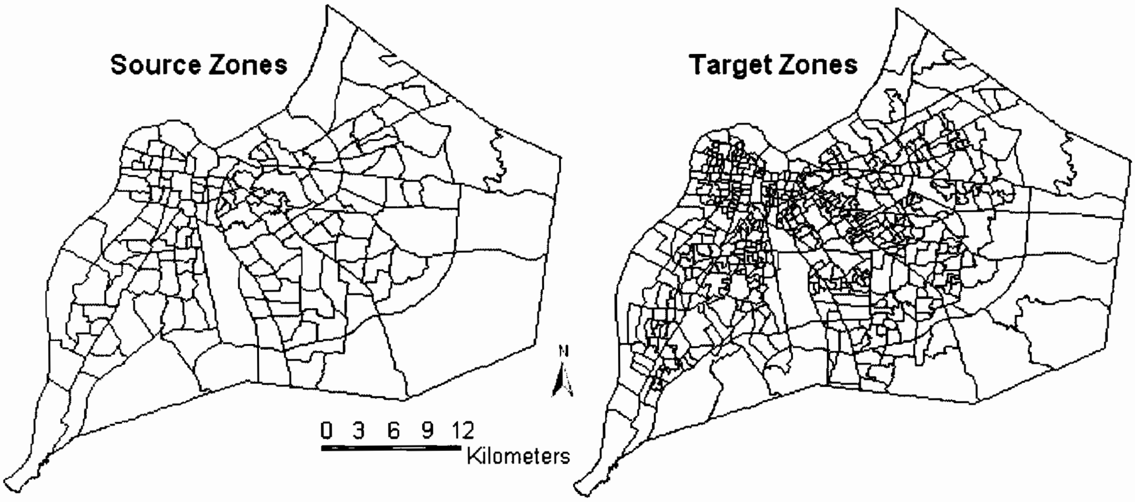


Figure 2.3 Source and Target Zones for Jefferson County, Kentucky

Table 2.1 shows the population, area, number of census tracts and number of census block groups for the three counties.

County	**Population** (no. of persons)	**Area (km^2)**	**Tracts**	**Block Groups**
Franklin	1,068,978	1407.137	264	883
Hamilton	845,303	1068.982	230	736
Jefferson	693,604	1032.259	170	556

Table 2.1 – Statistics for the Counties

2.5 Ohio State University Facilities to be Used for the Research

This research was conducted in the Numerical Cartography Laboratory in the Department of Geography, and the Region 1 Computing Facilities in the Department of

Civil and Environmental Engineering and Geodetic Science. Access to the Region 1 Facility has been graciously granted by Dr. Carolyn Merry and Dan Vehr (Systems Manager).

2.6 Software Used for the Research

This research was performed using a variety of software systems. The overlay procedures were performed in ArcGIS. The areal weighting method has been implemented by the author using VisualBasic with MapObjects. The pycnophylactic method was implemented using ArcView with an ArcView script created by Leopold Riedl. The dasymetric methods makes use of ERDAS Imagine for classification techniques. The interpolation of the dasymetric method was implemented by the author in VisualBasic with MapObjects. The network hierarchial method was also implemented by the author in VisualBasic with MapObjects.

CHAPTER 3

RESULTS OF AREAL INTERPOLATION METHODS

3.1 Intermediate Method Results

The differing spatial resolutions of the source and target zones in this research is of particular importance. Equation 13 provides the method for calculating the average spatial resolution for a set of irregular Resel polygons. Table 3.1 shows the average spatial resolution for the source and target zones for each of the three counties in this research.

County	**Source Zone** $\overline{R}$	**Target Zone** $\overline{R}$
Franklin	2.308	1.262
Hamilton	2.156	1.205
Jefferson	2.464	1.363

Table 3.1 – Average Spatial Resolution (km) for Source and Target Zones

From the values in Table 3.1, one can see that the interpolation is being performed from a lower to higher spatial resolution because all of the source zones have a larger spatial wavelength (resolution) than the target zones. It should also be noted that in the case of the nested hierarchy, the square of the source zone Resel divided by the square of the target zone Resel is equivalent to the average number of target zones contained within a source zone. This is shown in equation 14 where $\overline{N}_t$ is the average number of target zones within a source zone.

$$\overline{N}_t = \frac{\overline{R}_s^2}{\overline{R}_t^2} \qquad (14)$$

Several of these areal interpolation results have intermediate results. The pycnophylactic method creates a continuous population surface based on the populations within the source zones. This resulting surface is used to estimate populations for the target zones. The land use polygons are created from the supervised classification of a satellite image or the vectorization of a raster LULC dataset. The following sections provide these intermediate results.

3.1.1 Intermediate Results for the Pycnophylactic Method

The bulk of theory associated with the pycnophylactic method is in the creation of the population surface. This surface is created from the areas of the source zones and the populations contained within the source zones. The output surface is a raster grid with 100 m resolution. Figure 3.1 shows the pycnophylactic surface created for Franklin County. The algorithm converged after 46 iterations. Figure 3.2 shows the pycnophylactic surface created for Hamilton County. The algorithm converged after 48 iterations. Figure 3.3 shows the pycnophylactic surface created for Jefferson County. The algorithm converged after 76 iterations. The three figures show the source zone boundaries overlaying the surface. The values of each of the cells is shown as a

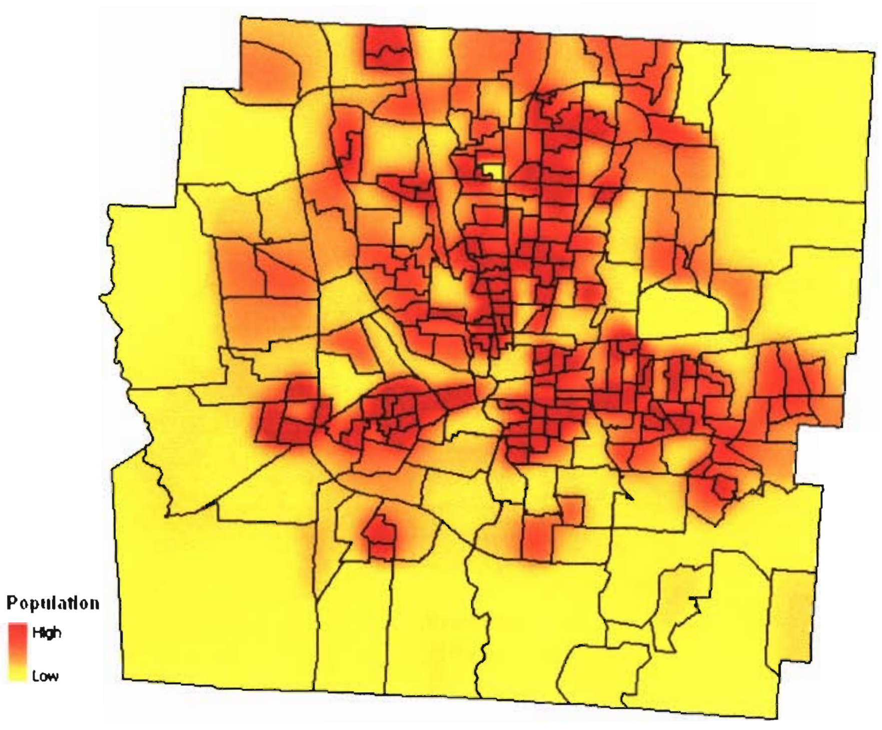


Figure 3.1 Pycnopylactic Surface for Franklin County

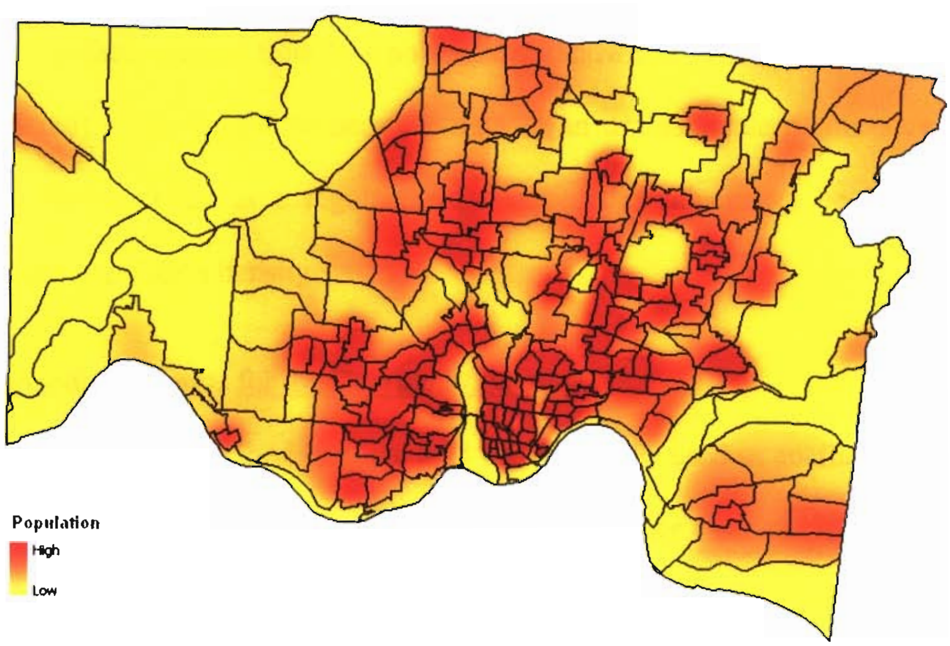


Figure 3.2 Pycnophylactic Surface for Hamilton County

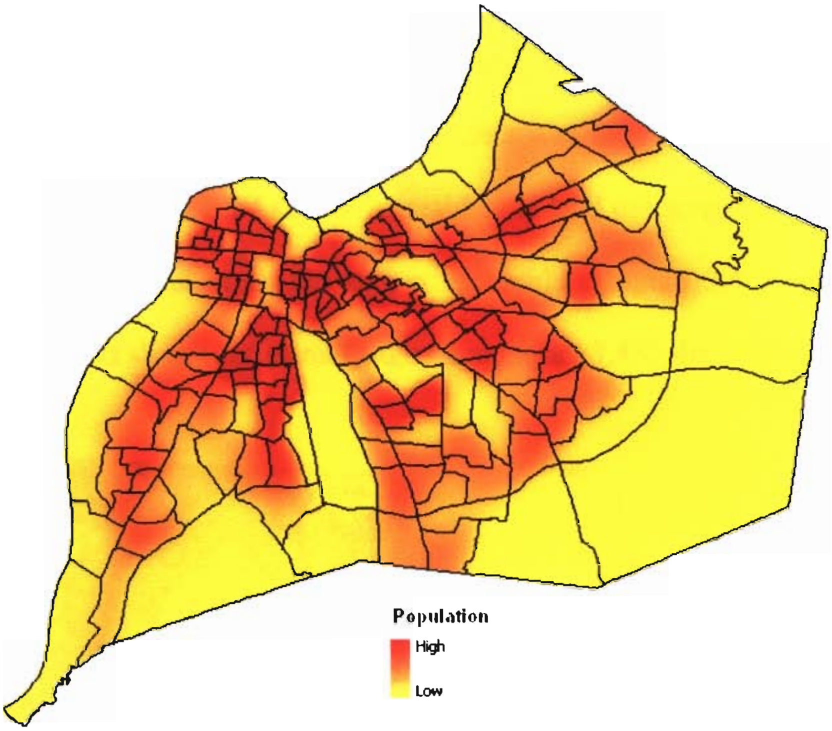


Figure 3.3 Pycnophylactic Surface for Jefferson County

continuous surface ranging from low populations (shown in yellow) to high populations (shown in red). The legend reduces the data to an ordinal scale of measurement for simplicity. Because this is a regular grid surface, the values can be interpreted as population or population density.

3.1.2 Intermediate Results for the Dasymetric Method

The dasymetric method requires land use data to be used as ancillary information for the interpolation of population. For the Franklin County case, the land use data was created by a supervised classification of a 30 m resolution Landsat image. The resulting classification was then smoothed using a nominal smoothing filter and then vectorized.

The land use data for Hamilton and Jefferson Counties were available as LULC data from the USGS. This dataset was in raster form, and was vectorized to create land use polygons. Each of the land use polygon datasets was then queried to select only the residential polygons. Figure 3.4 shows the residential land use areas for each of the three counties. The residential areas are shown in purple. Figure 3.4 A shows Franklin County, Figure 3.4 B shows Jefferson County and Figure 3.4 C shows Hamilton County.

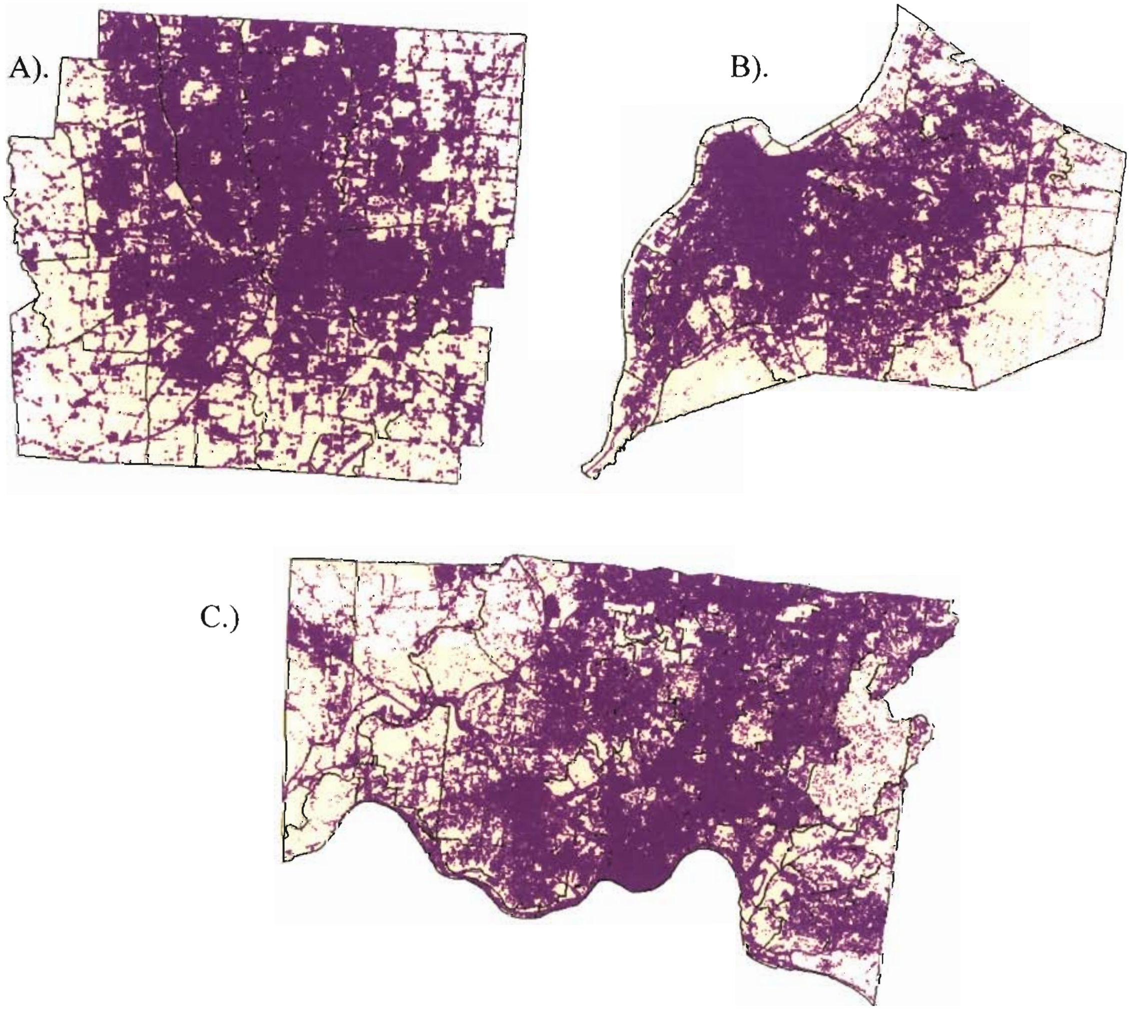


Figure 3.4 Residential Land Use Areas

3.1.3 Intermediate Results for Network Method

The network method uses the road network in the study area as ancillary information for the interpolation. However, in the case of the network hierarchial method, only particular road segments are used. For this research, the CFCC classes of roads used are the A2, A3 and A4 types. Figure 3.5 shows the A2, A3 and A4 roads for each of the counties. Each of the counties uses greater than 95% of the original road network. The interstate highways make up a large percentage of the data that is not being used.

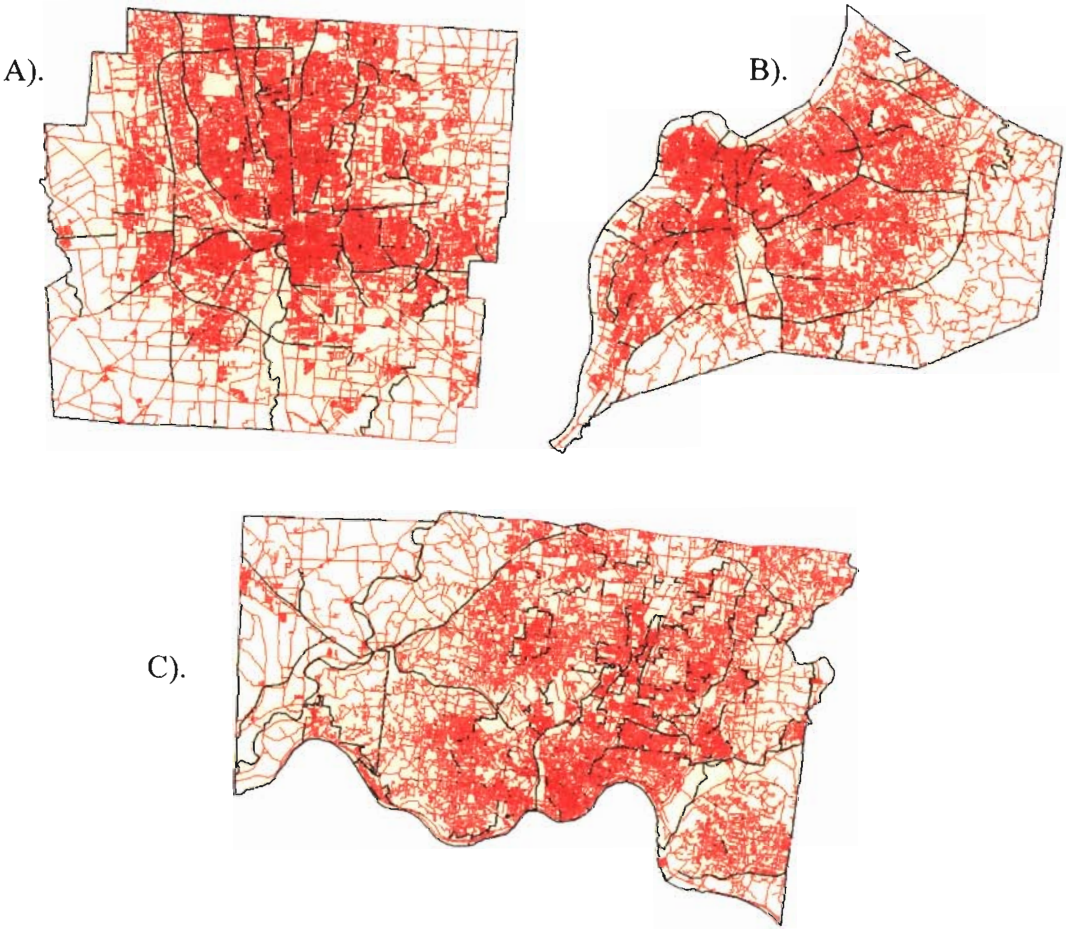


Figure 3.5 Selected Roads for Network Method

3.2 RMSE and Adj-RMSE Results

The RMSE and adj-RMSE serve as numerical indices of the global results of each of the areal interpolation methods, as defined in equations 10 and 11. The RMSE uses the absolute differences between the actual population and the estimated population within each of the target zones. The adj-RMSE normalizes this difference by the actual population within each target zone. In other words, a small absolute error may have a small net effect on the RMSE, but a huge effect on the adj-RMSE. The adj-RMSE is a measure of the percent error, whereas the RMSE is a measure of the absolute error.

As shown in Table 3.2, for the Franklin County case, the network method performed the best (in terms of RMSE), followed by the dasymetric method, then the pycnophylactic method, and then the areal weighting method. Notice that the word "best" when referring to these measures means the smallest value. The smaller the measure, the smaller the error associated with the particular method. In terms of the adj-RMSE, the Franklin County case shows that the network method is again the best, followed by the pycnophylactic method, then the dasymetric method, and finally the areal weighting method.

Method	RMSE	Rank	Adj-RMSE	Rank
Areal Weighting	751	4	1.956	4
Pycnophylactic	695	3	1.633	2
Dasymetric	540	2	1.792	3
Network	450	1	1.508	1

Table 3.2 – RMSE and Adj-RMSE Results for Franklin County

Although the network method provided the best results in terms of RMSE and adj-RMSE for the Franklin County case, and the areal weighting method provided the worst results, it was unexpected that the pycnophylactic and dasymetric methods would show similar results in terms of adj-RMSE. After further investigation into these results, it was evident that 1 particular target zone was the single largest contributor to the adj-RMSE in all of the methods. This location of this zone is highlighted in red in the graphic appearing next to Table 3.3. A car rental service and a hotel are located in this zone. In fact, this single zone with an actual population of 9 people was contributing to more than 50% of the adj-RMSE error associated with each of the methods. These estimated populations ranged from 286 persons to 441 persons for the various methods, which corresponds to percent errors of 3077 – 4800%. If this particular target zone's contribution is taken out, the results show that the network method performs best, followed by the dasymetric method. The pycnophylactic and areal weighting methods have very similar results, which is expected due to their similar RMSE measures. These results are shown in Table 3.3.

Method	Adj-RMSE	Rank
Areal Weighting	1.166	3
Pycnophylactic	1.181	4
Dasymetric	0.771	2
Network	0.626	1

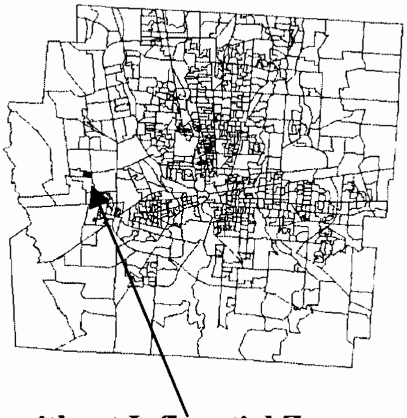


Table 3.3 – Adj-RMSE Results for Franklin County without Influential Zone

For the Hamilton County case, the RMSE and adj-RMSE results are very similar to those of the Franklin County case. Table 3.4 shows the results of these measures. In terms of RMSE, the areal interpolation method rankings are the same, with the network method performing best, followed by the dasymetric method, the pycnophylactic method, and the areal weighting method. The adj-RMSE results were also rather similar, by showing the network method as the best performer, the areal weighting method as the worst, and the pycnophylactic method performing better than the dasymetric method.

Method	RMSE	Rank	Adj-RMSE	Rank
Areal Weighting	472	4	15.08	4
Pycnophylactic	438	3	11.76	2
Dasymetric	399	2	14.16	3
Network	329	1	7.87	1

Table 3.4 – RMSE and Adj-RMSE Results for Hamilton County

Once again, it was suspected that a particular target zone was the main influence on the adj-RMSE results. This suspicion was validated, with the single largest contributor to each method's adj-RMSE attributed to one particular target zone. The adj-RMSE results with the exclusion of the influential zone are shown in Table 3.5. This influential zone is the location of a sewage disposal plant. In the Hamilton County case, this zone was accounting for more than 99% of the error measurement for each of the methods. The influential target zone in this county has an actual population of 2 persons, with estimates ranging from 428 to 819 persons. The target zone is highlighted in red in the graphic next to Table 3.5. After taking out the influential target zone and recomputing the adj-RMSE, the adj-RMSE follows the same ranking as the RMSE measure. The areal weighting method is the worst interpolator and the network is the best. The pycnophylactic and dasymetric have similar results, with the pycnophylactic performing a little better.

Method	Adj-RMSE	Rank
Areal Weighting	0.538	4
Pycnophylactic	0.519	2
Dasymetric	0.522	3
Network	0.357	1

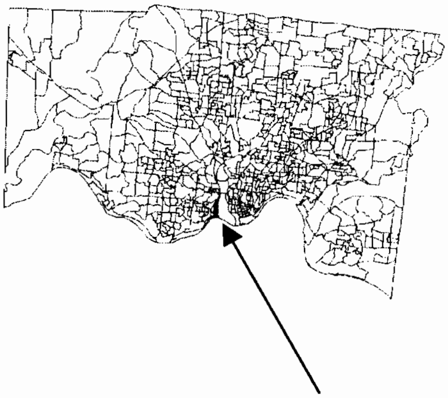


Table 3.5 – Adj-RMSE Results for Hamilton County without Influential Zone

The RMSE results for the Jefferson County case are consistent with the results of the Franklin and Hamilton County cases. As shown in Table 3.6, the network method performed best with the lowest RMSE and the areal weighting method performed the worst. The dasymetric method performed better than the pycnophylactic method. In terms of the adj-RMSE, the pycnophylactic method actually performed best, followed by the network method. The dasymetric method had the worst adj-RMSE results.

Method	RMSE	Rank	Adj-RMSE	Rank
Areal Weighting	596	4	16.55	3
Pycnophylactic	557	3	7.527	1
Dasymetric	447	2	18.846	4
Network	373	1	9.471	2

Table 3.6 – RMSE and adj-RMSE Results for Jefferson County

Similar to the Franklin and Hamilton cases, the Jefferson case also contained a single target zone that was contributing more than 99% of the error associated with the adj-RMSE for each of the methods. This particular target zone has a population of 5 persons, and contains the Louisville Airport, and the County Fairgrounds. This influential zone is shown in the graphic appearing next to Table 3.7. Estimated populations for each of the methods ranged from 887 to 2224. The results of the adj-RMSE with the exclusion of this particular influential target zone are shown in Table 3.7.

Method	Adj-RMSE	Rank
Areal Weighting	0.571	3
Pycnophylactic	0.686	4
Dasymetric	0.480	2
Network	0.446	1

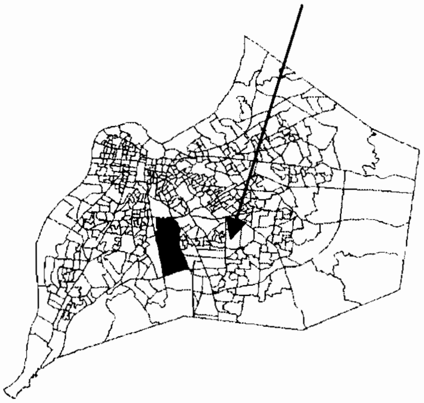


Table 3.7 – Adj-RMSE Results for Jefferson County without Influential Zone

With the exclusion of the influential zone, these results show that the network method has the best adj-RMSE, followed by the dasymetric method. The pycnophylactic method performs the worst with the exclusion of the influential zone. This is an interesting finding in that the pycnophylactic method had the best adj-RMSE measure overall with all of the target zones used to calculate the measure.

The Franklin, Hamilton and Jefferson County cases show consistent results in terms of the RMSE measure. In all cases the network method performs best and the areal weighting method performs worst. In all cases the dasymetric method performs better than the pycnophylactic method. These results also show the improved performance of the methods that make use of ancillary data. Unfortunately, this is not as clear in terms of the adj-RMSE measure. Although the network method is still the best and the areal weighting method is the worst in both the Franklin and Hamilton County cases, the pycnophylactic and dasymetric methods actually reverse in terms of rank with regards to the adj-RMSE. However, in the Jefferson County case, the pycnophylactic method had the best results followed by the network method. It was, however, found that the adj-RMSE, much like the mean, is an easily influenced measure, with one target zone accounting for more than 99% of the adj-RMSE in the Hamilton and Jefferson County cases. After removing the most influential zone from the Franklin County case, the adj-RMSE ranks were equivalent to those of the RMSE ranks. However, in the Hamilton County case, the pycnophylactic and dasymetric adj-RMSE's were reversed compared to the RMSE ranks, although they are very similar (0.003 difference). In the Jefferson County case, the network and dasymetric methods followed rank, but the areal weighting and pycnophylactic methods were reversed compared to the RMSE ranks, after the exclusion of the influential case. The fact that one particular target zone could have such a large effect on the outcome of the adj-RMSE measures is the reason that two sets of results are presented here.

3.3 Spatial Visualization of Errors Associated with Areal Interpolation Methods

This section examines the spatial errors associated with each of the areal interpolation methods. This section is subdivided by county. For each county, the actual population density map is shown along with estimated population density maps for each of the methods. A map of percent error is also shown for each of the methods. The map symobolization and classification schemes are similar throughout to make meaningful comparisons between the methods. Each population density map uses a green to blue span hue (Eastman, 1986) color sequence and consists of 10 classes. The classes are defined by the Jenk's Natural Breaks classification of the actual population density data for each county. The percent error maps use a red to blue double ended color sequence with nine classes. The red hues represent positive percent errors and increase in chroma and decrease in value as the magnitude of the error increases. The blue hues represent negative percent errors and increase in chroma and decrease in value as the magnitude of the error increases. The middle class of the distribution is a light gray and represents target zones with absolute errors of less than 5 percent. The next class on each side of the distribution represents target zones with absolute errors between 5 and 10 percent. Another class is for absolute errors of 10 to 25 percent, followed by a class for absolute errors of 25 – 50 percent, and finally a class for absolute errors greater than 50 percent.

3.3.1 Spatial Errors for Franklin County

The actual population density map for the target zones of Franklin County is shown in Figure 3.6. The areas of highest population density are near the Ohio State University area (just above the center of the map). This is a very dense residential area where most of the undergraduate student body lives. The low density southern and western areas are mostly agricultural. Figure 3.7 shows the estimated population density maps for each of the areal interpolation methods for Franklin County.

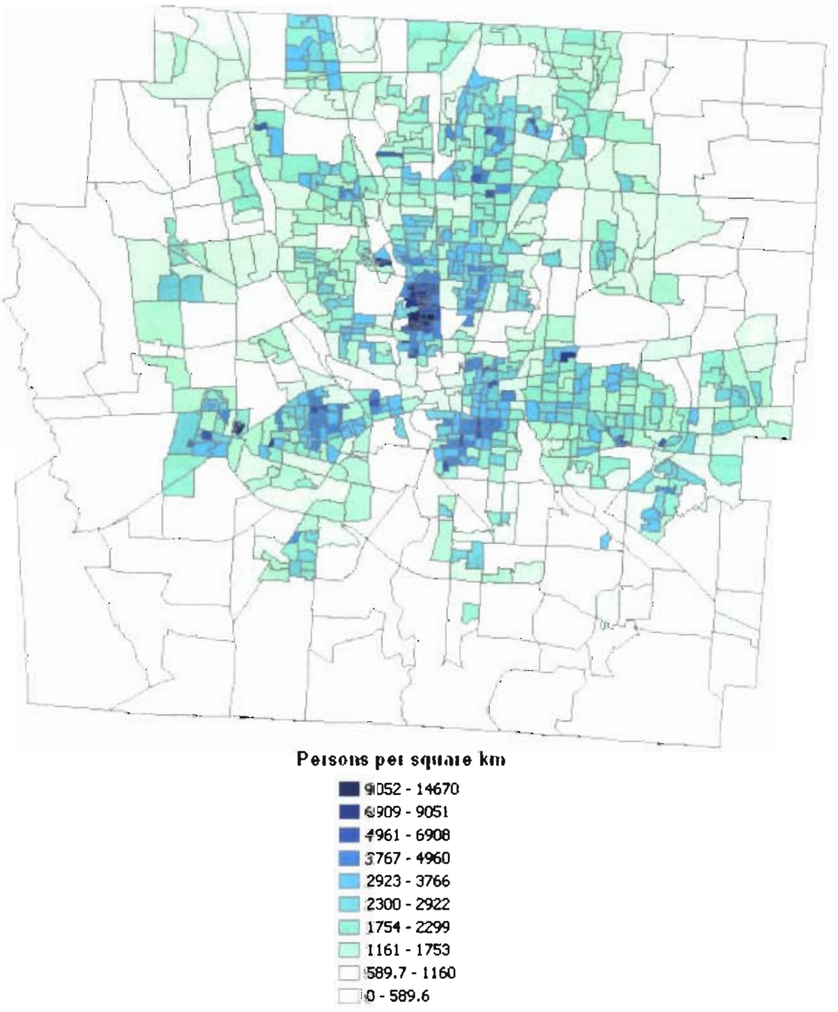


Figure 3.6 Actual Population Density for Franklin County

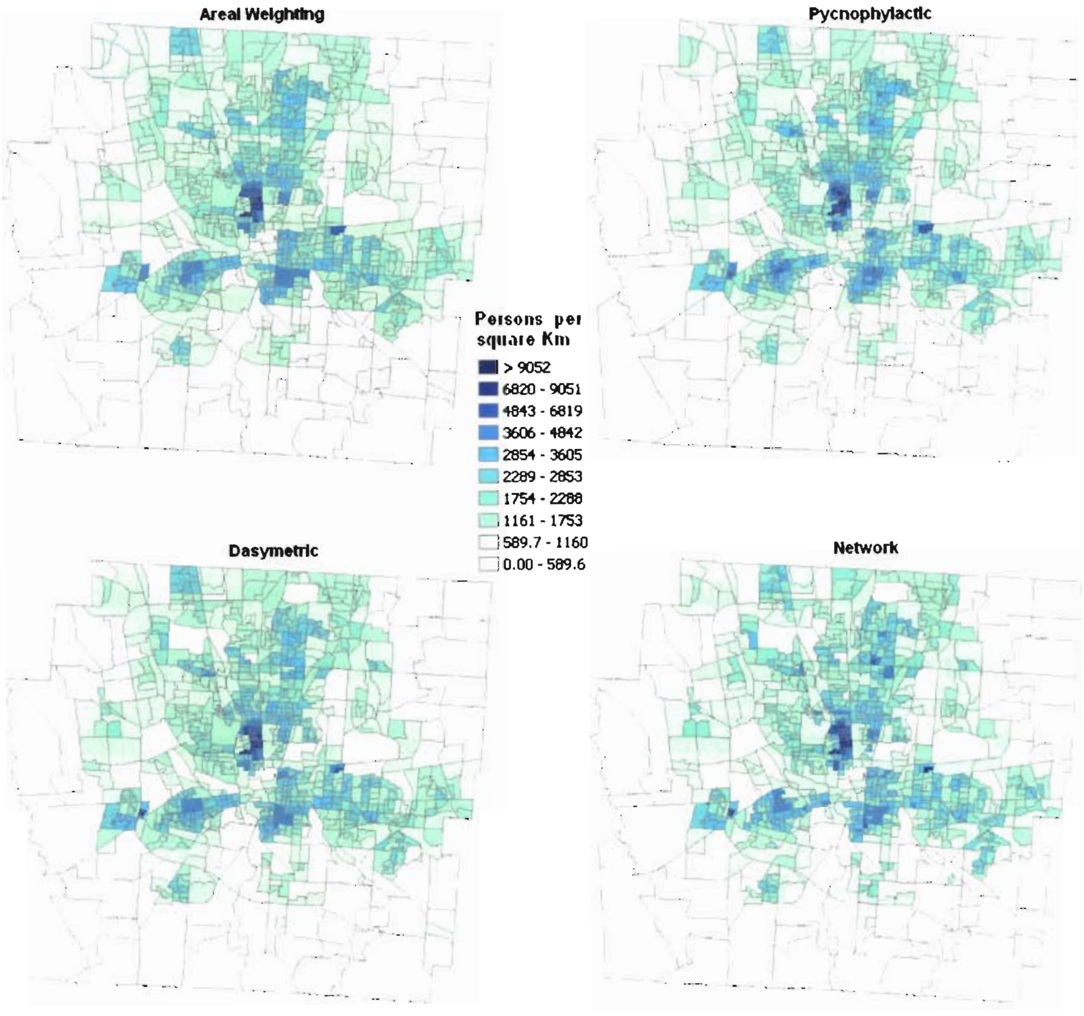


Figure 3.7 Estimated Population Density Maps for Franklin County

One interesting pattern that can be seen in the population density map for the areal weighting method is the constant densities estimated for target zones that fall within the same source zone. Because the census units are a nested hierarchy, all target zones within a particular source zone will have the same population density. These groupings of constant population density can easily be seen by observing the map. This process also affects the values of population density associated with the areal weighting method.

Target zones of smaller actual population density are allocated higher population densities, and target zones of higher actual population density are allocated lower population densities by the interpolation process. The highest estimated population density for the areal weighting method is 9803 persons/km^2. However, in Figure 3.6, one can see that the highest population density is actually 14,670 persons/km^2. Similar patterns of low and high density are maintained with the areal weighting method, however, there is no differentiation between areas from the same source zone.

Like the areal weighting method, the pycnophylactic method does not make use of ancillary data. However, in terms of the estimated population density map, it appears to do a better job of estimating the population density. This method is able to break the constant density pattern found in the areal weighting method. It performs better than the areal weighting method in the higher density areas.

The dasymetric method identifies the general pattern of the spatial distribution quite well. However, it shows grouping patterns in certain areas that are identical to those predicted for the areal weighting method. This occurs in areas of relatively homogeneous residential land use. Because the dasymetric method is an areal weighting method applied on ancillary data, areas of residential homogeneity will produce similar population densities.

The network method also does a nice job of identifying the general spatial population density patterns. In terms of local patterns (patterns within a source zone or small group of source zones), it does a better job of replicating these patterns. However, like the other methods, there are several areas where the network method is unable to reproduce the patterns shown on the actual population density map.

On can also find interesting spatial patterns by analyzing the percent error maps for each of the areal interpolation methods. The percent error for each target zone is calculated from equation 12. To more adequately describe some of the spatial patterns present in the percent error maps, a series of areas are defined in Figure 3.8.

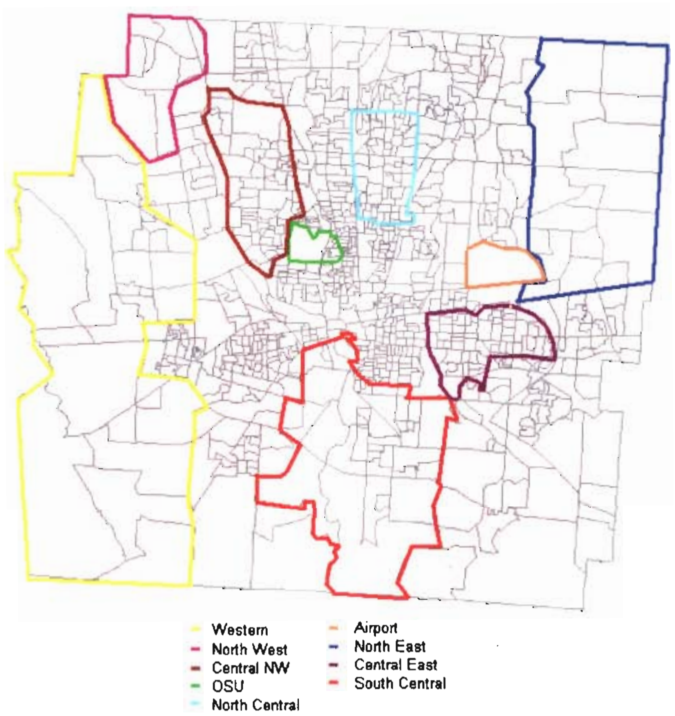


Figure 3.8 Areas of Interest for Percent Error Maps

When discussing the patterns in the percent error maps, each area will be referred to by the name presented in the legend of Figure 3.8. Figure 3.9 shows the percent error maps for Franklin County. These areas were defined by the author because they exhibit certain spatial patterns in the various methods of interpolation. Some areas show consistent findings throughout all of the methods, while others show varying patterns from method to method. They also contain a mixture of varying population distributions, such as urban, rural and suburban areas.

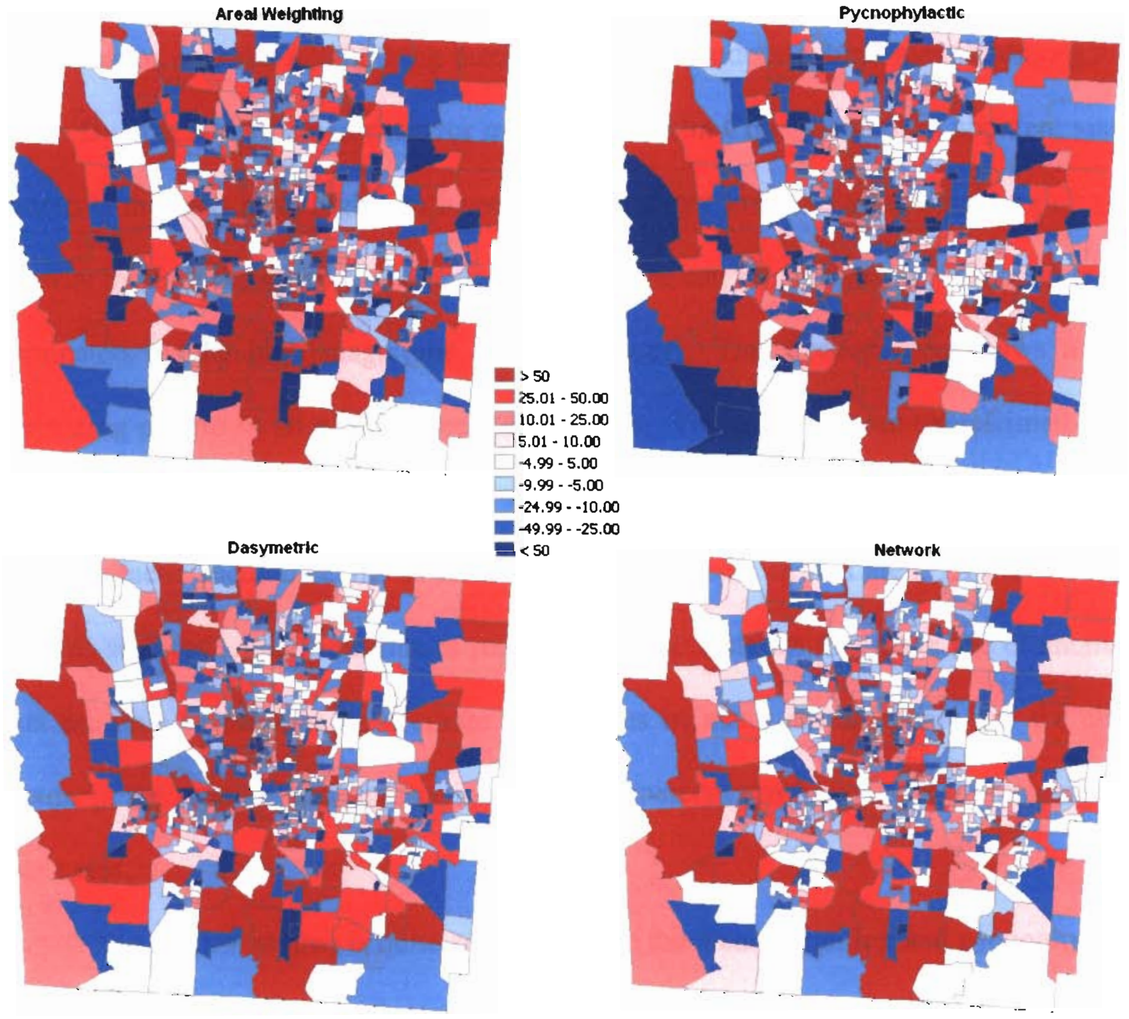


Figure 3.9 Percent Error Maps for Franklin County

Due to the nested nature of the source and target zones, patterns of negative spatial autocorrelation can be seen in all of the percent error maps. This is due to the zero sum constraint that is a characteristic of all volume preserving methods. The sum of all model residuals of target zone estimations from the actual target zone populations within a particular source zone will be equal to zero. In other words, if a source zone contains two nested target zones and one of the target zones produces a high magnitude positive percent error, then its neighboring target zone must produce a high magnitude negative error. By observing the percent error maps for Franklin County, the most apparent visual pattern is the shift from higher to lower absolute errors from the areal weighting map all the way through the series to the network map. One can see that the network map is less dominated by the high magnitude reds and blues. The western area shows a very interesting pattern in all of the percent error maps. The areas that are overestimated and those that are underestimated are very similar from method to method. The pycnophylactic method seems to have the most problems in this area. Although these patterns are very similar in terms of red and blue, the methods that make use of ancillary data are able to make more accurate estimations in many cases. There are six block groups in the western area that are in the highest percent error overestimation class in all four cases.

The northwest area is quite similar in both the areal weighting and pycnophylactic maps. The dasymetric method provides the most accurate estimations in this area followed by the network method. The central NW area shows that the network method provides the best results in this area. Most of the errors are within 25 percent. The areal weighting, pycnophylactic and dasymetric methods show similar over and under estimation patterns, however, of varying magnitudes. The large northeastern target zone in the central NW area has a high magnitude overestimation in all of the percent error maps. This area is the Ohio State University airport. Unlike the airport area shown in Figure 3.8, this target zone's source zone is made up of several other block groups. The airport area shown in Figure 3.8 has a source zone with identical geometric boundaries.

Because these target and source zone boundaries are geometrically identical, and the each of the areal interpolation methods is volume preserving, the predicted population is equal to the actual population in this area.

The OSU (Ohio State University) area is another area that shows similar error patterns in all of the areal interpolation methods. The large target zone in this area is west campus. West campus contains academic buildings and agricultural areas. The east side of campus is the area where the dormitories are located. Dormitories are areas of very high population density. The population for the west campus target zone is consistently overestimated because it contains the largest area, larger areas of built up land, and a greater number of roads than its eastern counterparts. Because these target zones are from a common source zone, more of the population is allocated to the western campus area than the eastern areas. Living structures, such as dormitories and apartment buildings, can create problems for all of the methods. Section 3.3.4 provides a more detailed discussion on this issue.

The north central area was chosen as an area of interest because it appears that all of the methods performed fairly well in the central portions of this area. By observing the source zones and the actual population densities, one will notice that the areal weighting method performed well in this area because the population is fairly homogeneously distributed in these areas. The northern portions of this area show large absolute errors in the areal weighting, pycnophylactic and dasymetric methods with improved results in the network method. The central east is another area where all methods performed fairly well.

The north east area shows nearly identical patterns in all methods in terms of over and underestimation. The magnitudes do differ from method to method. The major difference is the underestimation in the areal weighting and pycnophylactic methods and the overestimation in the dasymetric and network methods in the target zone located on the east side, just north of the center of the map.

The south central area is an example of a set of target zones that is problematic for all of these methods of interpolation. The linear area of red that extends from the south into the downtown area of Columbus in the areal weighting and pycnophylactic methods is particularly interesting. This area serves as the corridor for Route 23, 104 and Interstate 71. Most of these areas have very small populations, so a small overestimation can contribute to a large percent error.

There are complex spatial patterns that occur with the various areal interpolation methods. The general trend shows smaller magnitudes of error in the methods that make use of ancillary data. One also notices that certain areas are problematic for all of the methods, and other areas are interpolated with little error in all of the methods.

3.3.2 Spatial Errors for Hamilton County

Figure 3.10 shows the actual population density map for Hamilton County at the target zone (block group) level. This map shows that the western and eastern areas of the county are areas of the lowest population density, along with parts of the northeast. The northern area of downtown Cincinnati (south central area on the map) shows the areas of highest population density. These areas of high population density extend northward to

Clifton, which is the location of the University of Cincinnati. Other areas of relatively high population density can be seen west of downtown, separated by a very low density area. This low density area contains a train station with many lines of track.

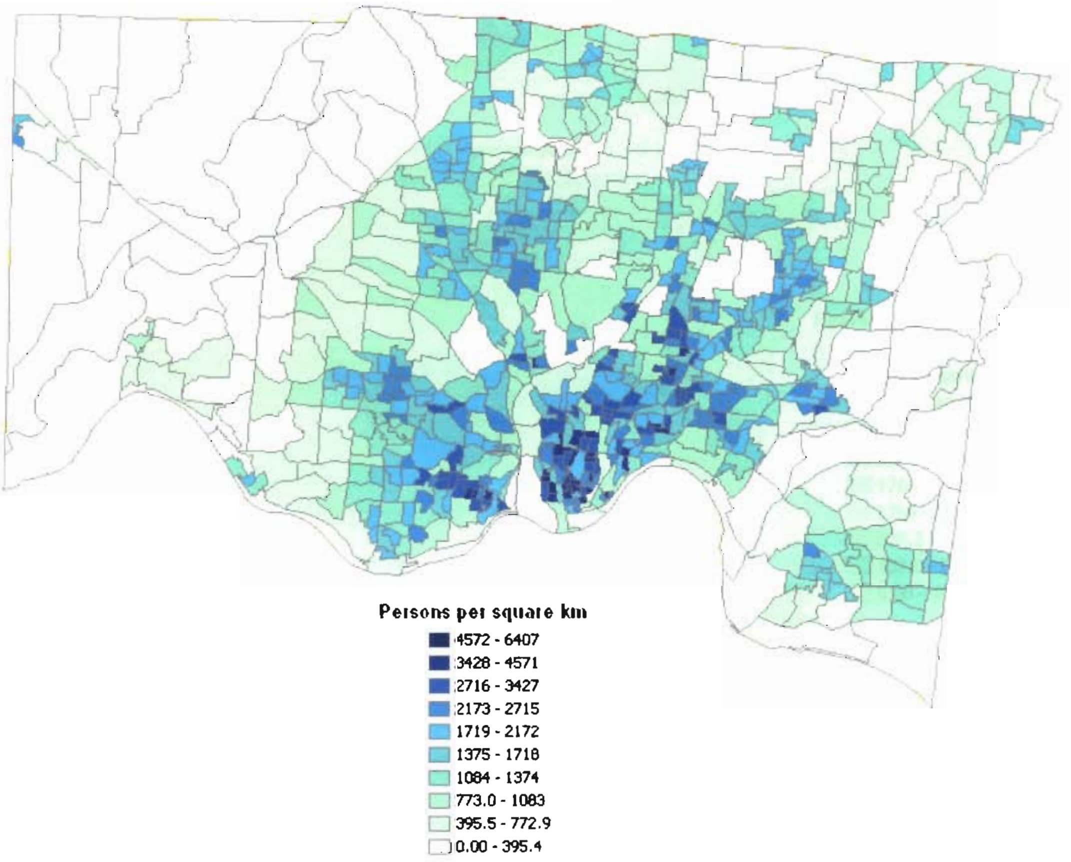


Figure 3.10 Actual Population Density of Hamilton County

Figure 3.11 shows the estimated population density maps for each of the interpolation methods. Each of the methods is able to produce a fairly accurate representation of the general patterns of high and low population density. Once again the grouping of target zones within the same source zone can be seen in the areal weighting

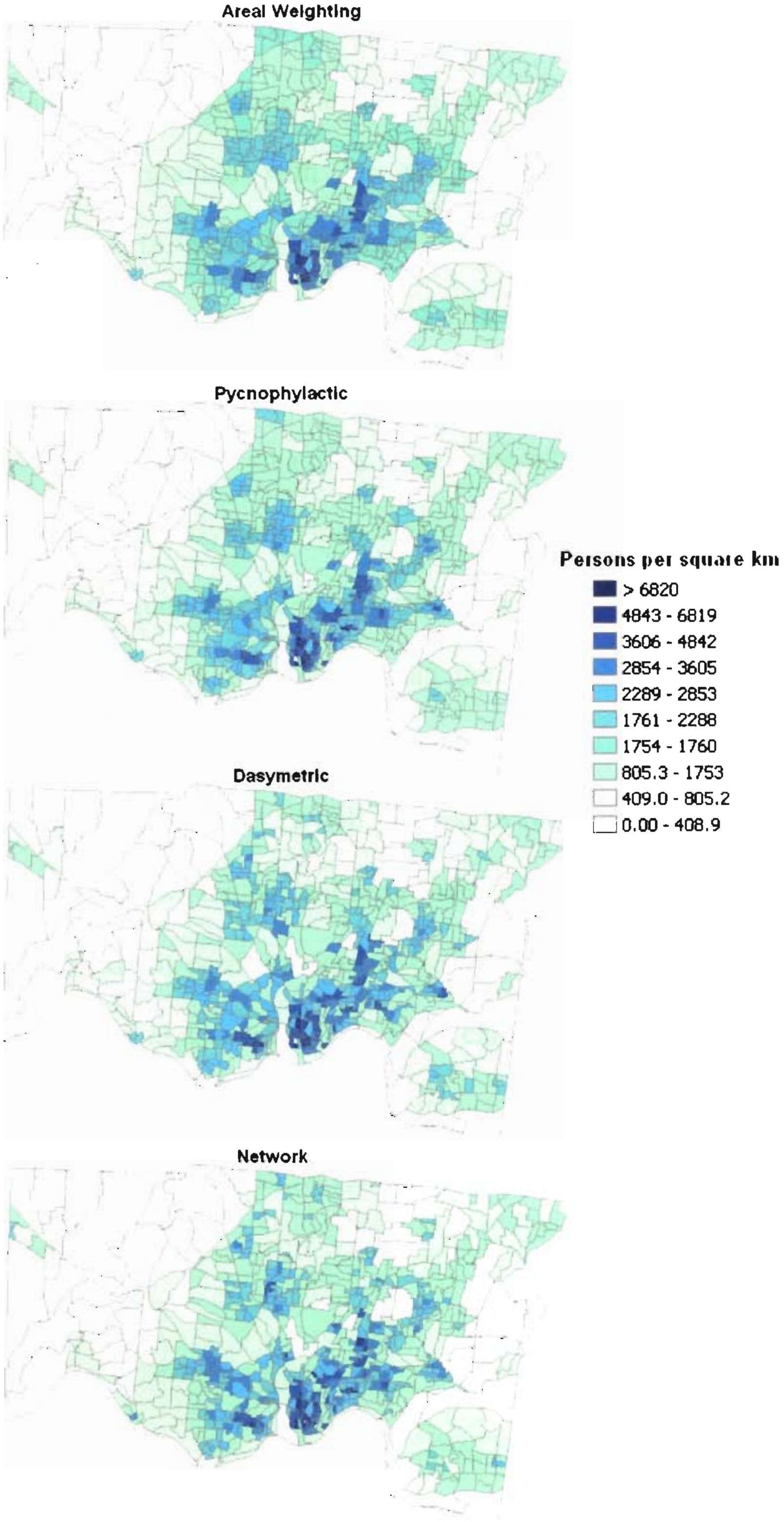


Figure 3.11 Estimated Population Density Maps for Hamilton County

method. This constant density makes for poor population density estimation in areas of heterogeneity. The population density patterns that are seen north east of downtown in Figure 3.10 cannot be represented with the areal weighting method. The pycnophylactic method improves in terms of dealing with the heterogeneity in some of these areas. However, the central northern areas of this map cannot capture the heterogeneity of the area. These areas appear very grouped on the pycnophylactic and areal weighting method, when in fact the area does show quite a bit of variation on the actual population density map. The dasymetric map is able to capture the variation of the area west of the railroad station quite nicely. It also performs well in the area north of downtown, and is able to capture more variation than the pycnophylactic and areal weighting methods in the central northern areas.

For the percent area maps, Figure 3.12 will be used to refer to specific areas of Hamilton County. Figure 3.13 shows the percent error maps for Hamilton County. The areas defined in Figure 3.12 exhibit particular spatial patterns, which can be compared from method to method. Some of the areas have been chosen because they provide consistent patterns between methods, while others demonstrate varying patterns from method to method. The defined areas also show variations of the population distribution, such as urban, suburban and rural areas.

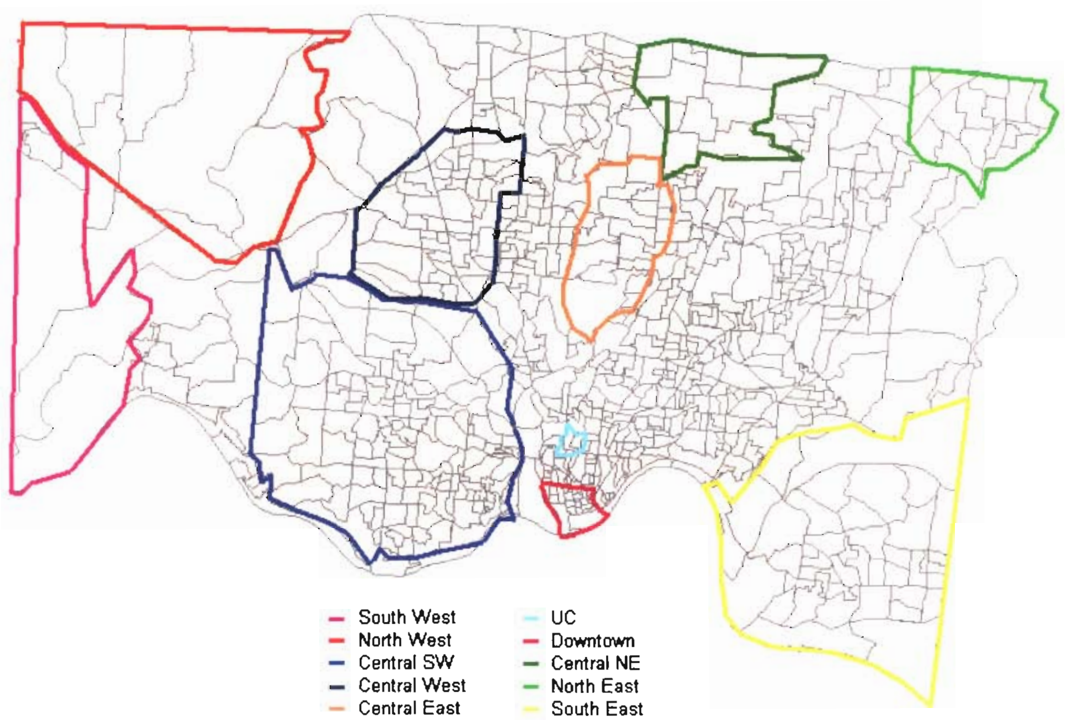


Figure 3.12 Areas of Interest for Percent Error Maps

By looking at Figure 3.13, one is again able to see the general pattern of more accurate results as one moves from the areal weighting percent error map (at the top left) to the network percent error map (at the bottom right). The south west area of the maps show the highest magnitude of overestimation in the areal weighting and pycnophylactic maps. The dasymetric method has the most accurate results in this area. The network method reduces the error in one of these three target zones. The network method outperforms the others in the northernmost target zones in this area.

In the north west area, the network method outperforms all of the other methods. However, the areal weighting method seems to be more accurate than the pycnophylactic or the dasymetric methods. The large target zone in this area, which is greatly

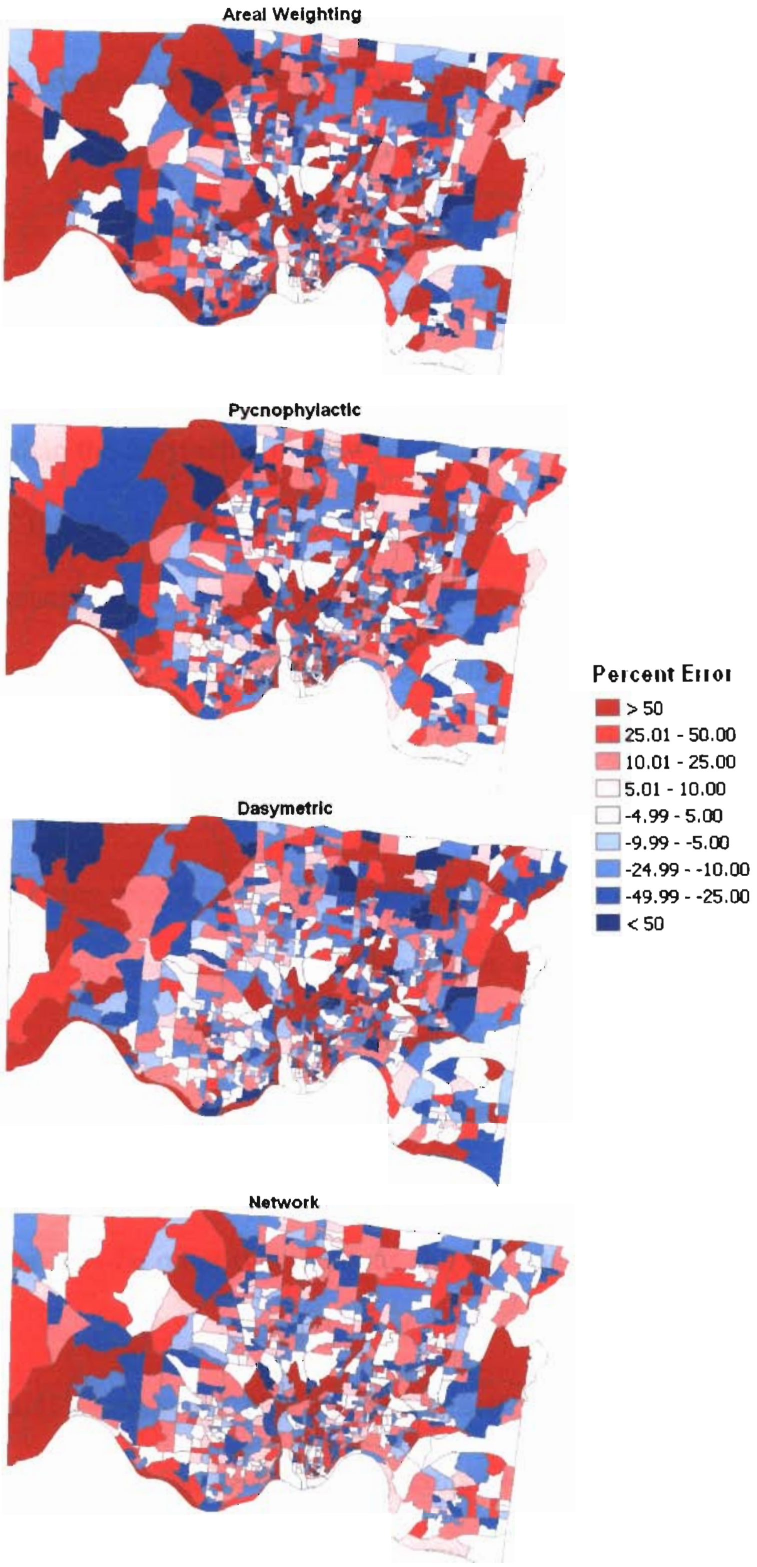


Figure 3.13 Percent Error Maps for Hamilton County

overestimated in the areal weighting and dasymetric method, overestimated in the network method and underestimated in the pycnophylactic method, and contains the Miami Whitewater Forest.

The population in the central SW area is most accurately predicted by the dasymetric method. The network method also performs fairly strong in this area. The pycnophylactic and areal weighting methods perform very well in certain areas, but show stronger absolute errors than the dasymetric or network methods.

The east side of the central west area shows very strong performance from the areal weighting and pycnophylactic methods. The network method also shows small errors in this area. In the center of this area, all of the methods show similar patterns in the context of red and blue, however, they show differing magnitudes from method to method. The central east is another area that shows very similar patterns of over and underestimation with differing magnitudes. The southernmost target zone in this area is predicted with no error by all methods because the target zone and source zone are geometrically identical. The downtown area shows similar patterns in the areal weighting, pycnophylactic and dasymetric methods. The network method is similar to these, except in the northernmost target zones where the network method shows a mix of over and underestimation and the other two methods show all underestimation for these areas.

In the UC area (University of Cincinnati) we see a similar pattern of over predicting population on the west side and underpredicting population on the east side in all methods. However, this relationship is not as strong as it was in the Ohio State University area of Franklin County. In fact in this case, the areal weighting method and dasymetric method have the smallest errors. The pycnophylactic and network method have similar errors. Having attended this University, the dormitories are more evenly distributed throughout all areas of the campus than they are at Ohio State. The L-shaped target zone (overpredicted in all cases) includes several dormitories on the southern area that forms the base of the L. If this area were not included in this target zone, results similar to the OSU case are quite possible.

In the northwestern portion of the central NE, the network method outperforms all of the others, with similar patterns of over and underprediction. This area is interesting in that it is very heterogeneous in terms of the population distribution. There are large areas that contain parks, cemeteries and shopping malls. I believe that the large errors associated with the dasymetric method in this area are due to the mall having similar spectral characteristics to residential areas, such as a mixture of grass, trees, cement, and automobiles. The northeastern portion of this area has the smallest errors associated with the areal weighting and dasymetric methods. The southern portion of this area has consistent red and blue patterns throughout all of the methods, with the network method performing best.

The northeast area shows very small errors associated with the network method. This is a suburban area. The areal weighting and dasymetric methods also perform quite well in this area. The dasymetric method performs the worst in this area. The interior portion of the southeast area is also a suburban area. The dasymetric method performs very well in this area. The network method also performs well in this area. The areal weighting and pycnophylactic methods show similar patterns in this area.

3.3.3 Spatial Errors for Jefferson County

The actual population density for Jefferson County at the target zone level can be seen in Figure 3.14. The areas of highest population density are located in the downtown area near the University of Louisville and the western downtown area. Moving from the downtown area toward the suburbs, the population density decreases. The low population density areas in the southwestern portion of the county contains the Jefferson County Memorial Forest. The low population density areas on the east side of the county contain forested areas and rolling hills. The large target zone of low population density south of the downtown area contains the Kentucky Fair and Exposition Center, the airport and a Ford assembly plant. Figure 3.15 shows the population density estimates for each of the areal interpolation methods. As with the Franklin and Hamilton County cases, similar patterns can be seen with the population density maps. The areal weighting method produces a lot of blocking patterns, and the methods making use of ancillary data are better able to capture the heterogeneity of the population distribution. The areal weighting and dasymetric methods both are lacking target zones of population density greater than 4572 persons/km^2. Therefore, these two maps do not have any observations in the highest population density class.

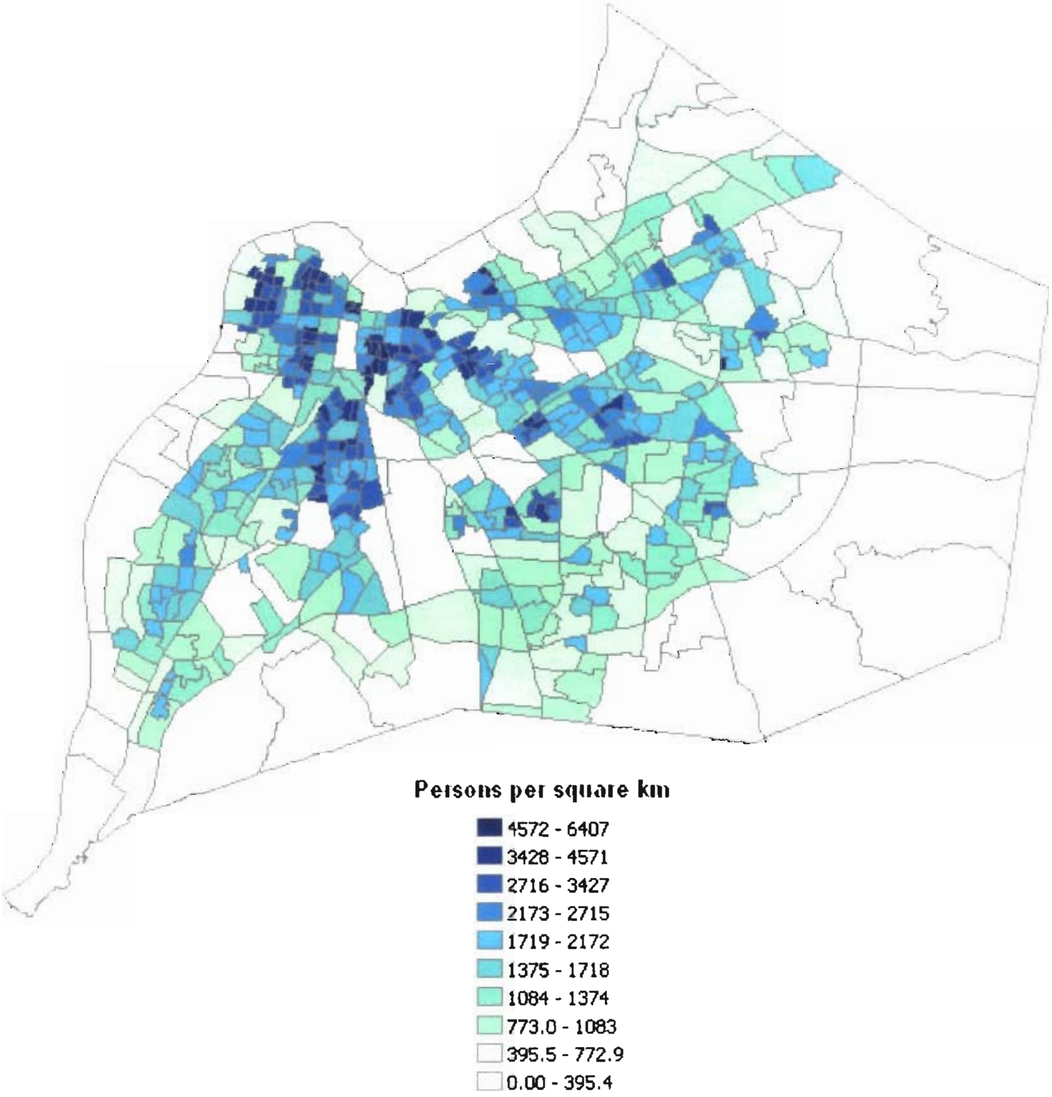


Figure 3.14 Actual Population Density for Jefferson County

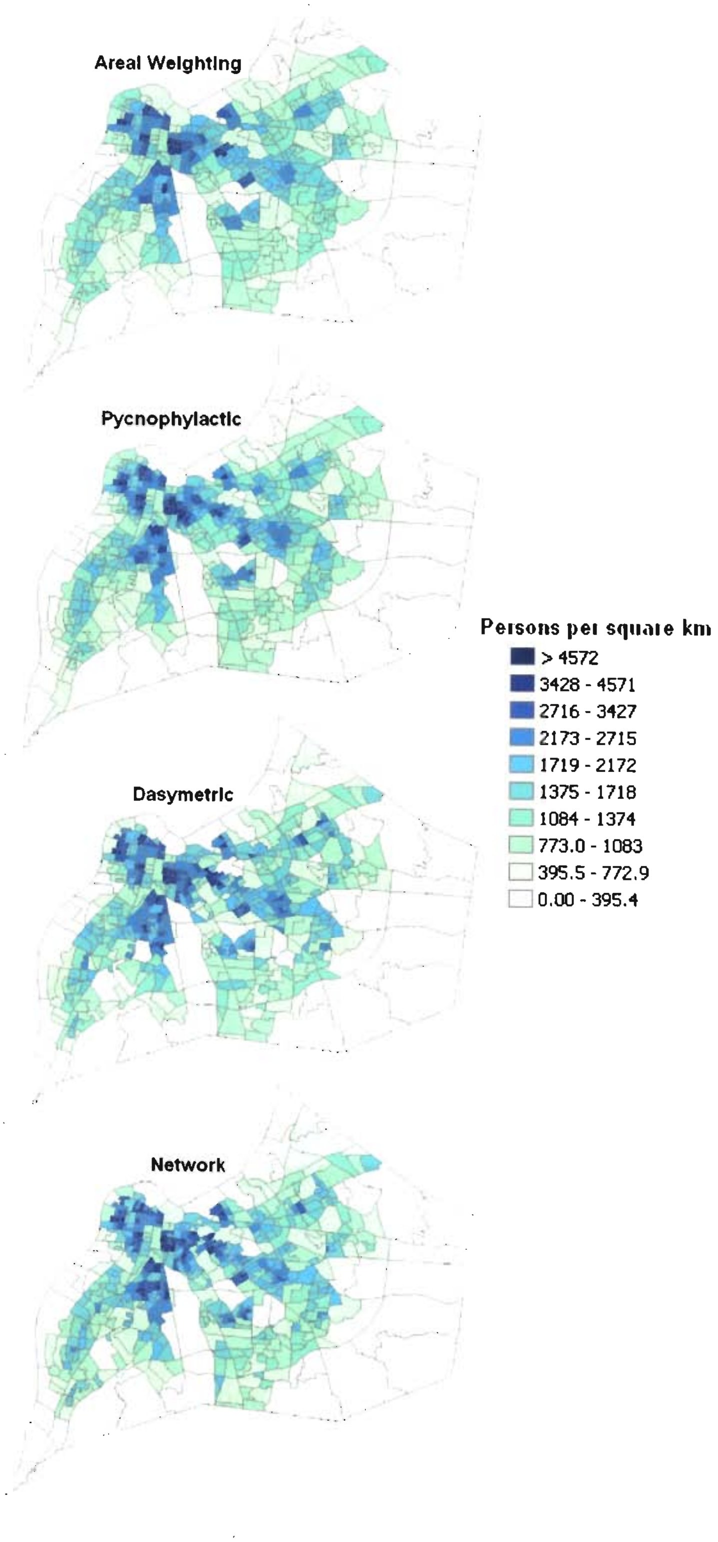


Figure 3.15 Estimated Population Density Maps for Jefferson County

Figure 3.16 shows particular areas that will be referred to when discussing the percent error maps for Jefferson County that are shown in Figure 3.17. As for the areas defined in the previous maps, Figure 3.16 shows areas that depict varying or similar spatial patterns throughout all of the various interpolation methods.

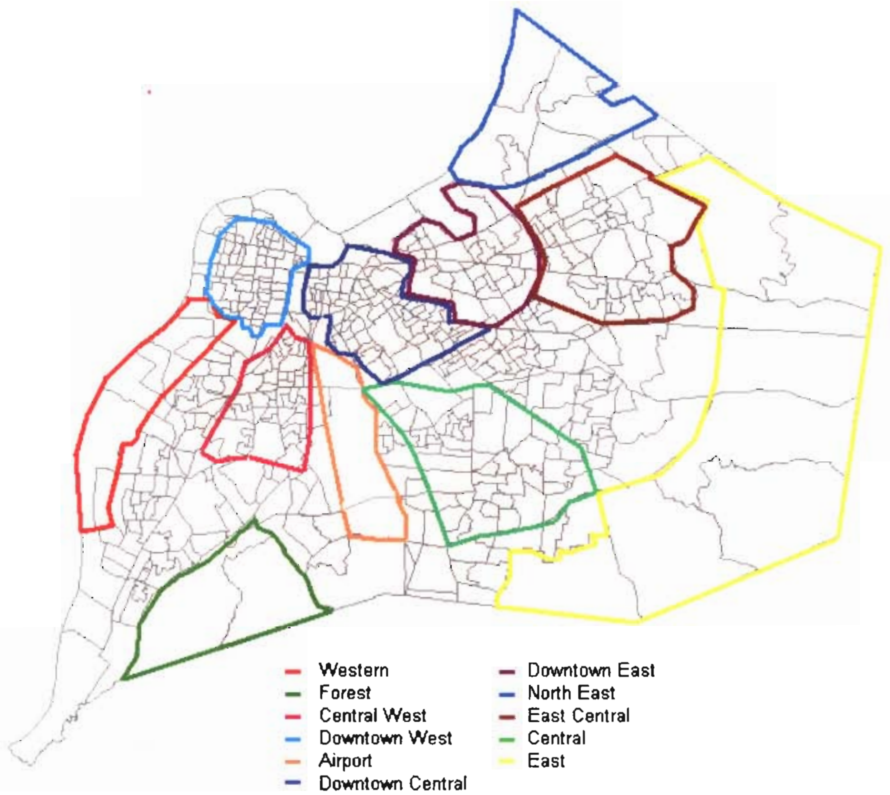


Figure 3.16 Areas of Interest for Percent Error Maps

The western area of Jefferson County shows a lot of variation between the various percent error maps. In the northern portion of this area the dasymetric method performs best. All but one of these areas show less than 5 percent error with the dasymetric method. The southern portion of this area, however, shows the smallest errors in the network method.

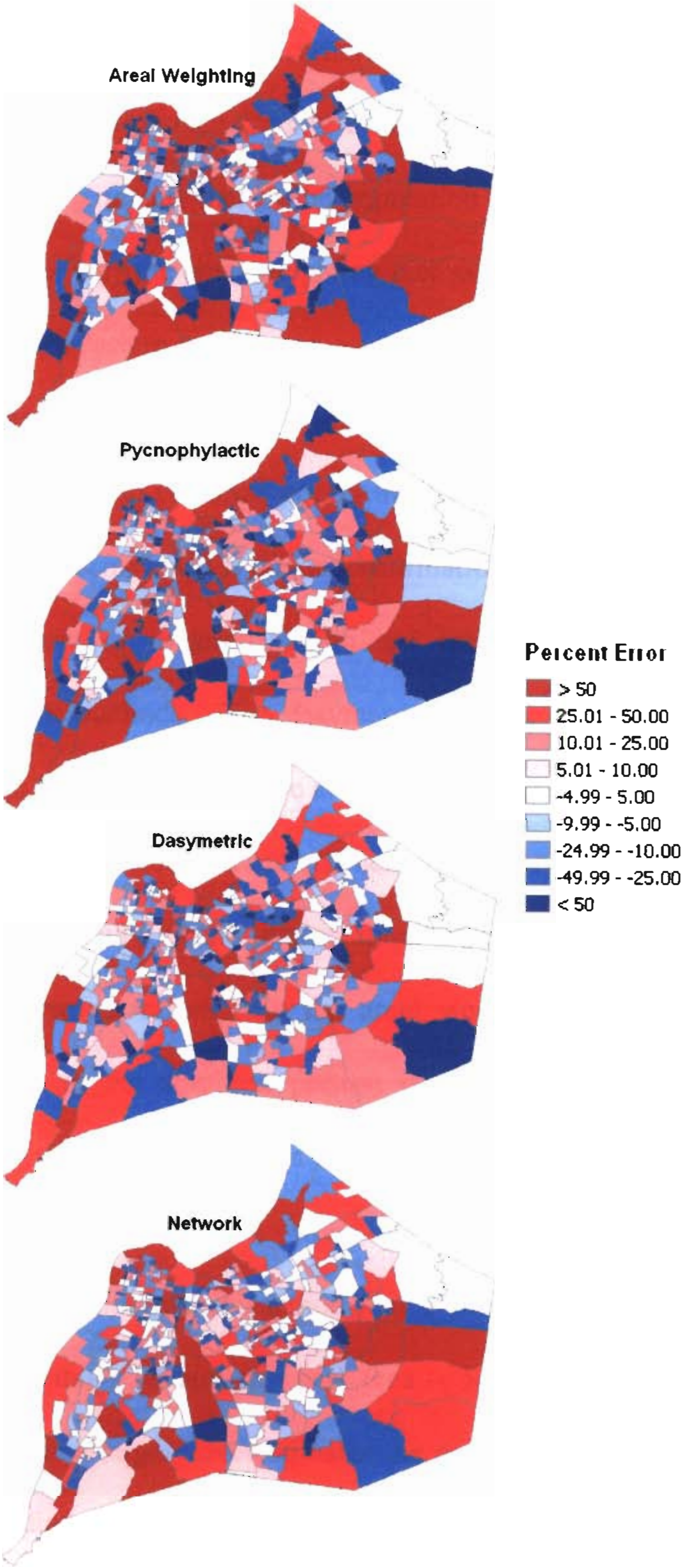


Figure 3.17 Percent Error Maps for Jefferson County

The forest area also shows a lot of variation between the percent error maps. Most of this area is within the Jefferson County Memorial Forest. The western target zone actually does have approximately half of the population of the entire area and about half of the total area. This is why the areal weighting method performs fairly well on the western side. The small northeastern target zone contains more population than the large southeastern zone. This is why the areal weighting method performs poorly in this tract. However, the network method also performs poorly in this target zone. The one pattern that is seen in all four percent error maps is the underestimation for the small northeastern zone. Large percent errors of underestimation are seen in the areal weighting, pycnophylactic and network methods. The dasymetric method shows absolute errors between 25 and 50 percent throughout the entire area.

In the central west area, the dasymetric and network methods show how ancillary data are able to improve the results of the interpolation. These methods perform much better here in terms of absolute errors than the areal weighting and pycnophylactic methods. The large tract in the southern portion of this area that is consistently overestimated by all methods contains Iroquois Park, which is a large recreational area. It is easy to understand why the areal weighting and pycnophylactic methods overestimated this area. The dasymetric method had a much smaller overestimation compared with the other methods because the park area is not classified as a residential land use. The network method largely overestimates the population in this area because the park contains long winding roads, which are classified as city or neighborhood roads.

The airport area shows the same patterns with all methods in terms of magnitude and over/underestimation. The southernmost target zone in this area contains the residential areas. The northernmost target zone in this area contains the fair and exposition center, and the central area contains the airport. The airport zone only has a population of 5 persons, so a predicted population of 8 people for this area would produce a percent error of greater than 50 percent. However, all methods greatly overpredicted the population in this area.

The downtown west area is a high population density area and shows complex patterns on all of the maps. It is clear that the areal weighting method performs the worst in this area. One pattern that occurs in all methods is the red target zones that begin in the northern part of this area and run down in a straight line to the south southeast. This pattern was further investigated and it was found that this is the path of the interstate. The northernmost of these target zones also contains a large cemetery on its eastern side. Therefore, these tracts have smaller populations due to their proximity to the interstate.

The downtown central area is a great example of an area where the methods making use of ancillary data perform better than those that do not. The downtown eastern area shows the best results for the network method and probably the worst results for the dasymetric method. The dasymetric, areal weighting and pycnophylactic methods all have similar patterns in this area, however, the dasymetric method show larger absolute errors.

The dasymetric and network methods produce the best results in the northeast area. The pycnophylactic method has less than 5 percent error in the northernmost target zone in this area. There is actually no population in an area of about 1 km inland from the Ohio River in this target zone and the target zone to the west. The pycnophylactic method inherently distributes the population away from this border and toward higher population areas inland. The pycnophylactic method has very large negative and positive errors in other parts of the northeastern area. The areal weighting method does very poorly in this area relative to the results of the other methods.

The east central area shows the most accurate results for the network method. The dasymetric method also performs fairly well in this area. Similar patterns of error are shown in all four methods, however, the dasymetric and network methods produce smaller absolute errors than the areal weighting and pycnophylactic methods in most cases.

The eastern area shows some very interesting results. The northern portion of this area is predicted with small absolute error (actually none) by all methods because the source and target zones are geometrically identical. The southern portion of this area shows very similar results between the network and areal weighting methods. However, the network method shows smaller absolute errors in most of the target zones. The dasymetric method does best overall in the eastern area. This area has large portions of forested coverage and contains some rolling hills topography. The roads are fairly evenly distributed, but the population is not. This is why the dasymetric method is able to show improvement over the network method, and the areal weighting method performs very poorly.

3.3.4 Investigation of the High Density Problem

In general, the percent error maps show that the methods making use of ancillary data show better results than the methods that do not make use of ancillary data. As discussed, there are exceptions to this general pattern, and certain areas show very similar patterns of positive and negative errors. One particular real world situation that causes problems for all of the areal interpolation methods is the case of high density residential structures, such as dormitories and high rise apartment complexes.

In order to better illustrate the complications that such areas can cause with respect to the areal interpolation problem, Figure 3.18 shows the example of the OSU area. All OSU dormitories have been geocoded and are represented with a pictographic symbol. The yellow circles on the map represent the population contained in each area. Notice the high populations in the smaller eastern target zones, compared with the low population in the larger western zones. The higher populations in the eastern zones are due to the clustered distribution of the dormitories. However, each of the areal interpolation methods consistently greatly overestimates the population in the western areas and greatly underestimates the population in the eastern area. Of course such a situation would be expected with the areal weighting method and perhaps the pycnophylactic method. The methods making use of ancillary data do not perform much better in these areas. The network method allocates population to the road network, and the dasymetric method allocates population to the urban/residential areas. Because all of these target zones are contained within the same source zone, the street network will essentially be allocated the same population per unit length among all roads within the entire area, and the urban/residential areas will all be allocated the same population density within the entire area. The western contains a greater total length of roads than any of the eastern areas, and also a greater urban/residential area than the eastern target zones. Therefore, the population allocated to the western area is greater than the population allocated to the eastern area. This is why similar patterns of over and underestimation are seen in each of the four areal interpolation methods. For large scale areal interpolation, knowledge of the locations of such structures could improve results of the interpolation.

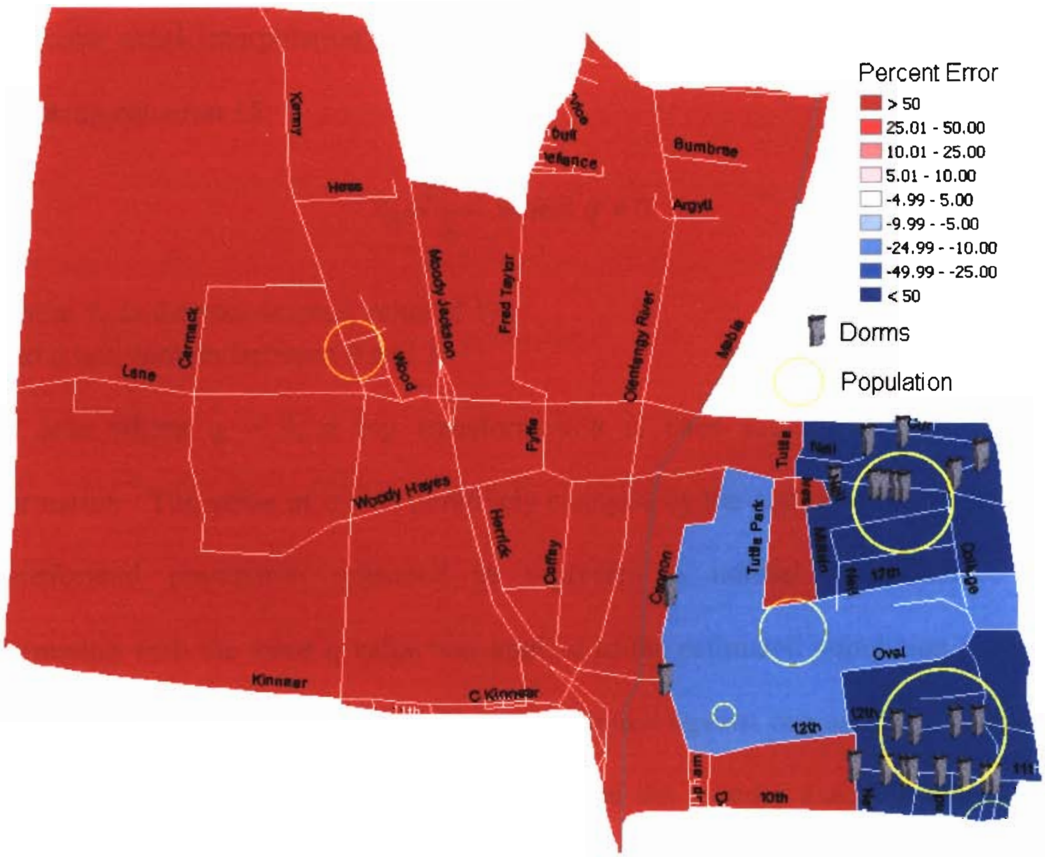


Figure 3.18. Effects of High Density Structures.

3.4 Other Tools for Comparing Areal Interpolation Methods

One useful tool for comparing the efficiency of the various areal interpolation methods is the use of scatterplots. Scatterplots are graphs of point symbols that represent the value of one variable on the y-axis and another variable on the x-axis. In this case, a Box-Cox transformation (Hamilton, 1992) of the actual population of the target zone is the y-axis. The x-axis shows the Box-Cox transformation of the estimated population for the particular areal interpolation method. The Box-Cox transformation is defined as follows using equation 15:

$$Y_{Ti} = \frac{Y_i^q - 1}{q} \quad \text{where } q \neq 0 \tag{15}$$

where Y_{Ti} is the transformed value of Y_i
and q is a number between 0 and 1

In the case where q = 0, a log transformation is used rather than the Box-Cox transformation. The value of q was iteratively changed by the author until the histogram of transformed population appeared to represent a normal distribution. The transformation with the same q value was applied to the estimated population values for each method and the transformed variables were plotted against one another. A diagonal red line represents transformed actual population on the x and y axes. In other words, this line represents a perfect interpolator. The population variables have been transformed in order to stabilize the heterogeneity that is due to the large range of populations occurring in the target zones. Each point on the scatterplot shows the relationship between the actual and estimated populations for a particular target zone. Figure 3.19 shows the scatterplots for each of the interpolation methods for Franklin

County. Figure 3.20 shows the scatterplots for Hamilton County. Figure 3.21 shows the scatterplots for Jefferson County. Because the red line represents a perfect interpolator, the crowding of the points around this line indicate the efficiency of the interpolation method. In all of the figures, a similar trend is noticed. The areal weighting method always has the greatest spread about the red line. The network method always has the tightest spread around the red line. The dasymetric scatterplots are more tightly concentrated around the line than the pycnophylactic method. These scatterplots agree with the RMSE rankings discussed in section 3.2. The scatterplots allow for a nice visualization of the results that are easy to interpret. However they lack a spatial element in the visualizations shown in this research. Of course in an interactive virtual system, these plots could be linked with the spatial data. This would allow for selection of interesting points in the scatterplots which would in turn select the geographic features that are linked to the points selected in the scatterplots. Such a system would allow for easy identification of outliers in the scatterplot distribution.

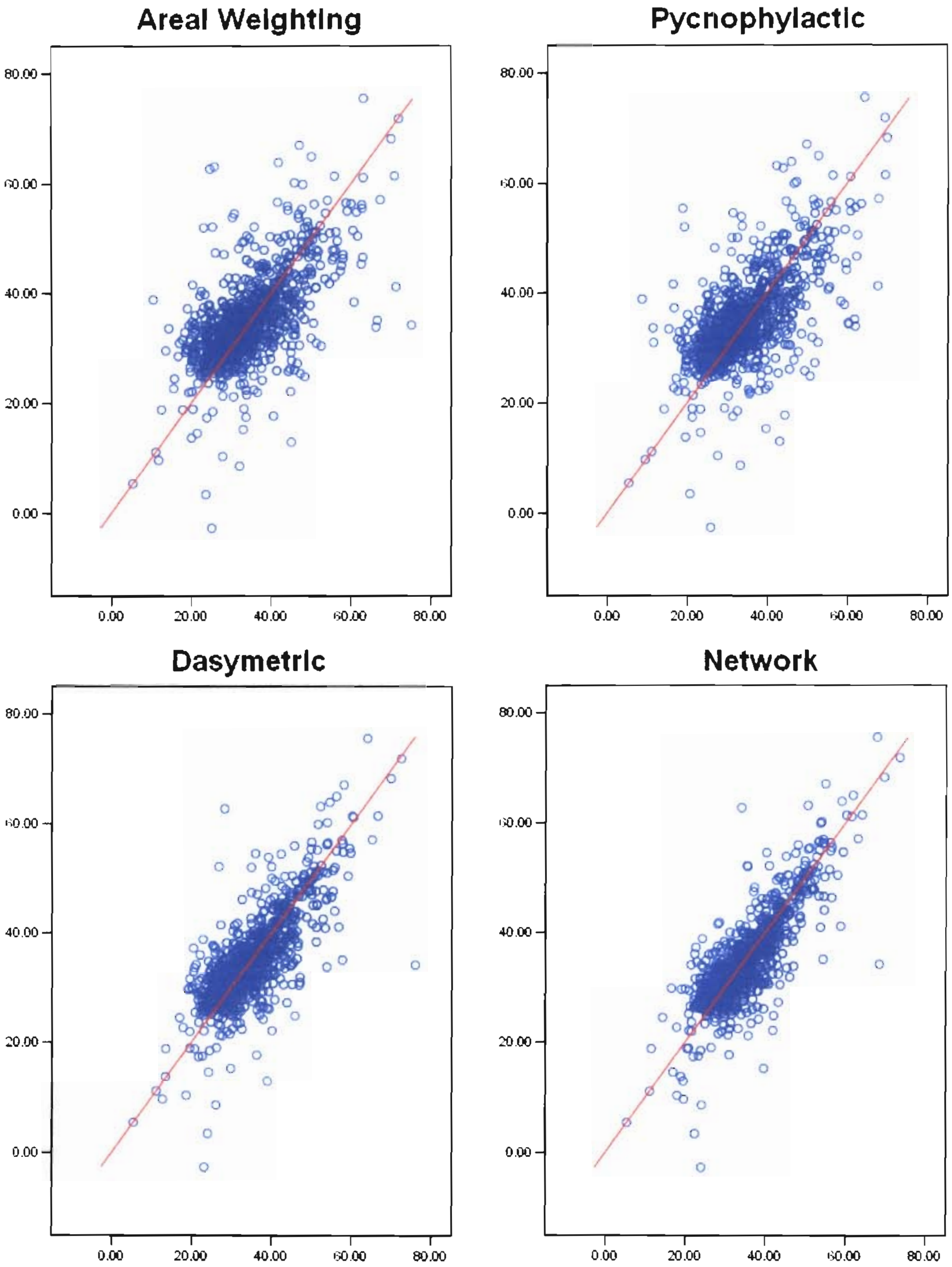


Figure 3.19 Scatterplots for Franklin County

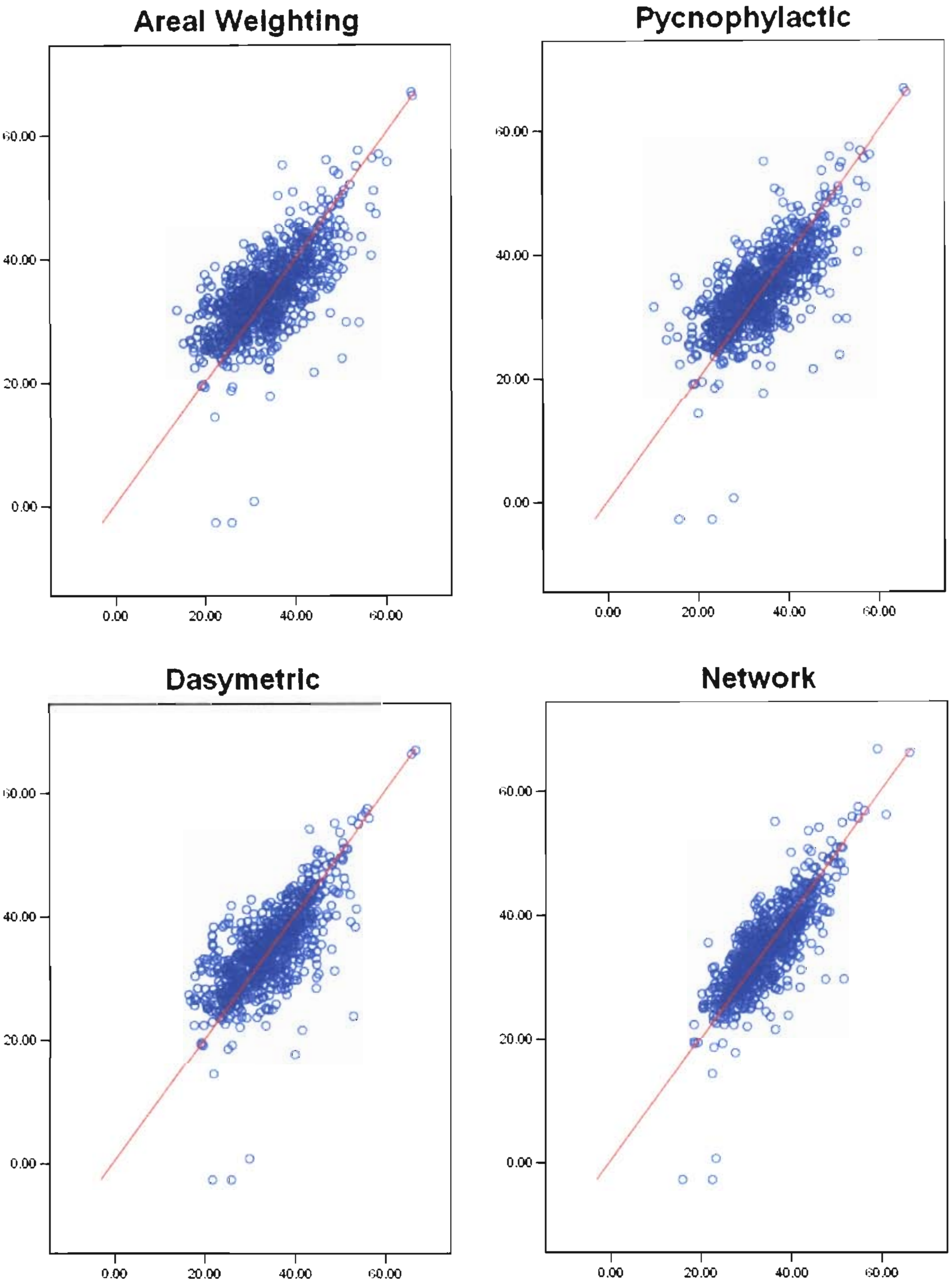
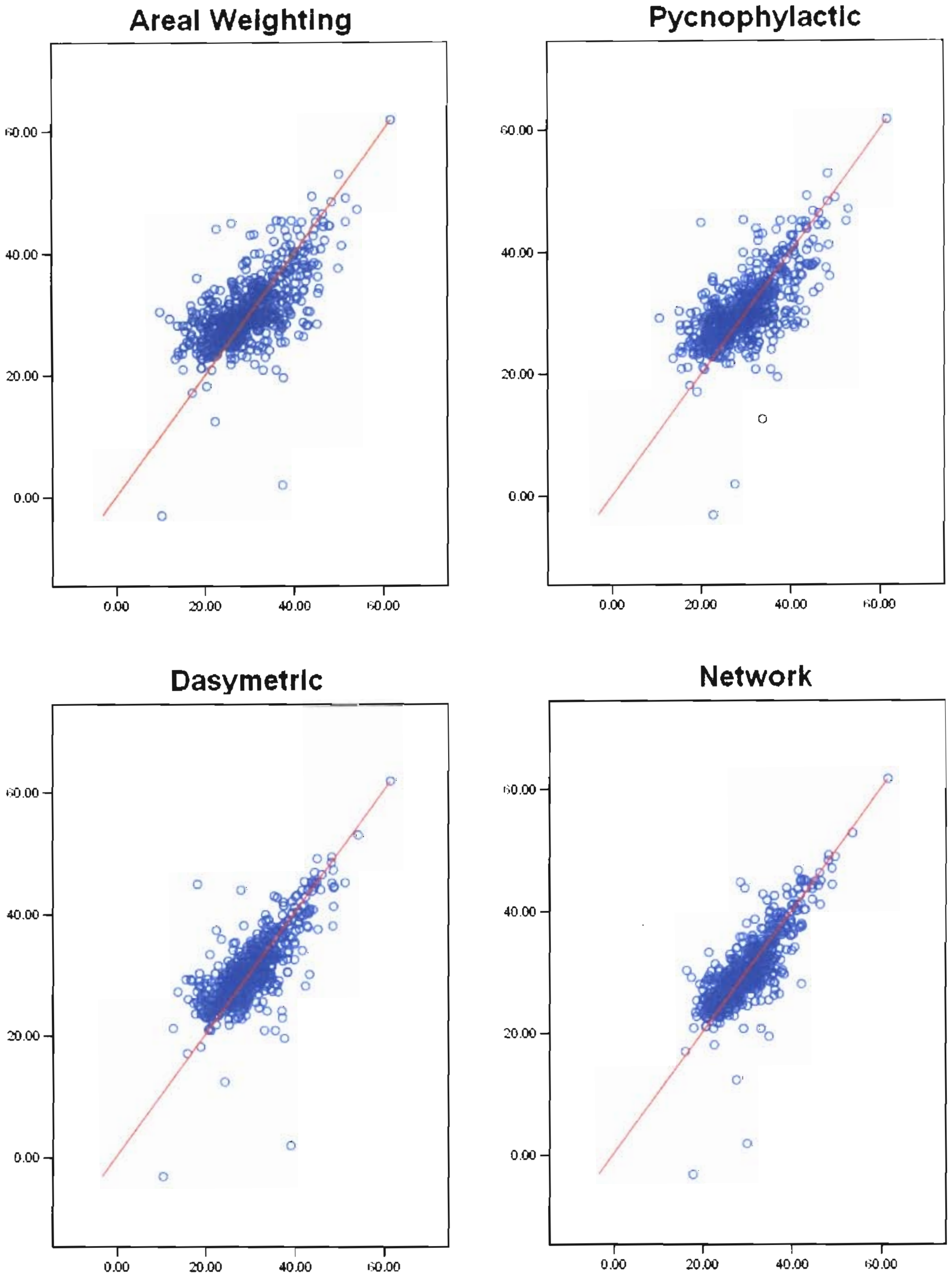


Figure 3.20 Scatterplots for Hamilton County



Areal Weighting

Pycnophylactic

Dasymetric

Network

Figure 3.21 Scatterplots for Jefferson County

Like the scatterplots, histograms of the errors associated with each of the interpolation methods can also be viewed. Figure 3.22 shows the error histograms for each of the interpolation methods in each county. The y axis represents the number of observations within each histogram bin. The x axis represents the value of the error. Each of the histograms for a particular county have the same x and y axes, making comparison between the histograms possible. It is also important to add that the standard deviation associated with the error distribution shown in the histograms is identical to the RMSE values shown in the tables of Section 3.2.

The general trend of the histograms can be seen by reading them from left (areal weighting) to right (network). This movement shows higher peaks and tighter error distribution. The areal weighting method has the widest error distribution of all of the methods in all cases. The network method has the tightest error distribution of all of the methods in all cases. One exception to the rising peaks of the histogram can be seen in Jefferson County where the peak is higher in the areal weighting method compared with the pycnophylactic method. However, the distribution of the errors is tighter in the pycnophylactic method.

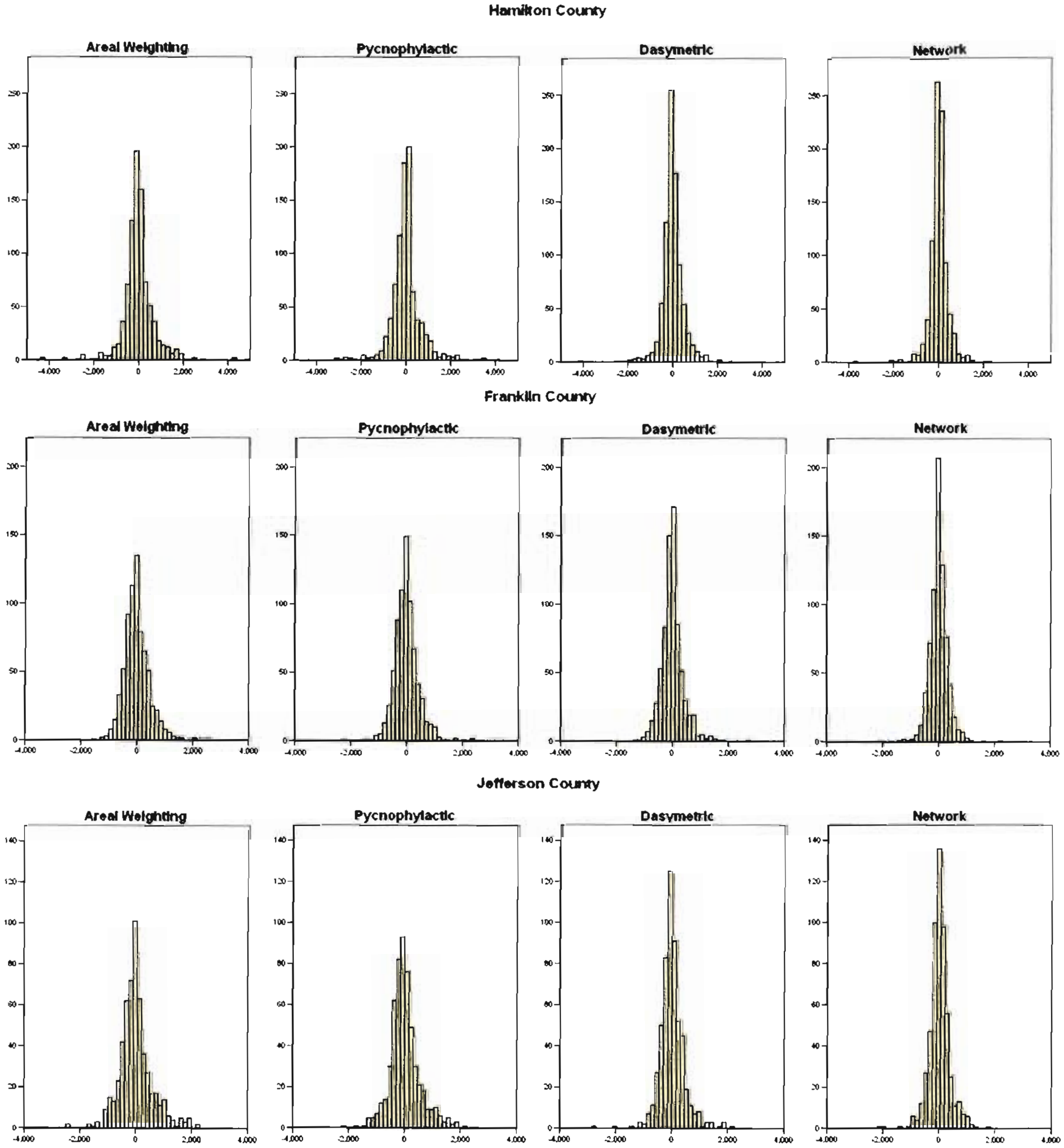


Figure 3.22 Error Histograms of Areal Interpolation Methods

CHAPTER 4

SUMMARY AND CONCLUSIONS

Three of the four areal interpolation methods analyzed in this research produced some form of intermediate results. In the case of the pycnophylactic method, the intermediate results were the creation of smooth population surfaces. These surfaces are mathematically-derived from population values assigned to grid cells based on source zone populations and the population values of neighboring cells. The dasymetric method makes use of satellite image data or land use data. The intermediate results in this case are the areas that are classified as residential areas. These areas create a nice visualization of how the population is distributed throughout the study area. The network method also has intermediate results in that it is only making use of particular roads that are likely to link to places of residence. These maps can also create a nice visualization of the population distribution throughout the study area.

This research has shown fairly consistent results with four different methods of areal interpolation for three different spatial situations. In terms of the RMSE, the network method provided the most accurate results in each of the three counties. The dasymetric method provided the second most accurate results, followed by the pycnophylactic method. The areal weighting method provided the least accurate results in all three cases. Because the RMSE results are in the same units as the analysis variable (population), on average, the network method is either over or underestimating the population in a particular target zone by a smaller amount than any of the other methods.

The adj-RMSE results showed some inconsistencies with the RMSE results. In the Franklin and Hamilton County cases, the network method provided the most accurate results and the areal weighting method had the worst results. However, in the Jefferson County case, the pycnophylactic method had the smallest adj-RMSE and the dasymetric method performed worst. Due to the highly influential characteristics of one target zone in each of the counties, alternative measures of adj-RMSE were shown to account for the large influence of a single target zone. With the exclusion of the influential zone in each county, the network method provided the best results. Either the areal weighting or pycnophylactic method provided the least accurate results.

The spatial patterns presented in this research show very complex patterns of negative spatial autocorrelation in the percent error maps due to the nested nature of the target zones. Specific areas in each of the counties were discussed to provide some insight into the spatial patterns appearing on these maps. In general, it was found that the methods making use of ancillary data were able to minimize the errors found in many of the areas. However, some areas cause problems for all of the areal interpolation methods implemented in this research. These include very high density areas, such as apartment buildings or dormitories, and some of the very low density rural areas. In general, all of the methods were able to capture the overall patterns of high and low population density. This is due in part to the underlying structure of the areal interpolation methods implemented in this research. All of the methods are volume preserving, which is an important attribute of an areal interpolation method. This means that if inverse interpolation were performed (or in the case of nested units, summation of the target zones was performed), predicted population values of the source zones would be identical to actual population values of the source zones.

It was also found that non-spatial visualization techniques, such as scatterplots and error histograms, can further complement the RMSE results. The scatterplots show very noticeable differences of spread about the line of perfect interpolation between the methods making use of ancillary data and those that do not use ancillary data. A similar pattern can be detected in the error histograms, which show tighter error histograms in the methods making use of ancillary data.

4.1 Conclusions

This research has implemented a consistent research design that allows various areal interpolation methods to be tested against one another. The use of source and control zones in the defined U.S. Census zone hierarchy that allows comparison of the estimated population values to the actual population values provides the framework for this research. The four methods tested in this research have not been tested against one another in previously published work.

The results of this research point to the network method as the most promising method of areal interpolation in this context. Improvements in the accuracy of areal interpolation can also be seen in the dasymetric method. However, in terms of data availability and data preparation, road network data are easily available for most areas, and the data preparation is not as involved as the dasymetric method, which may involve classification of the dataset. Based on this research, with a lack of ancillary data, the pycnophylactic method may be a better choice of methods as compared to the areal weighting method. Although the areal weighting method is much simpler to perform, the improved accuracy of the pycnophylactic method may justify implementation of the pycnophylactic method in the case where no ancillary information is available. However, the results of this research cannot be extrapolated to all areas. The areas tested in this research all have similar structures, including a major city, rural, suburban and urban areas. These areas were chosen because they show variation in the population structure, yet all have a similar geographical structure. Further research in areal interpolation will create a greater understanding of the effects of spatial setting and scale.

4.2 Future Research

In terms of future research, it would be interesting to examine the performance of these methods at different resolutions or scale levels. As mentioned previously, this research focuses on the interpolation from lower to higher spatial resolutions. Would we see similar rankings of these methods if the interpolation is performed from higher to lower spatial frequencies, or to and from zones of similar spatial frequencies? Results of such research could give insights to the scale dependency of the methods. Monte Carlo simulation could be used to create many datasets of varying spatial resolutions from random aggregations of smaller enumeration units, such as census blocks or census block groups. The larger samples for each scale level would also allow for measures of statistical significance.

It would also be interesting to look at the performance of these methods in many different spatial settings. Although this research implements areal interpolation methods in three different counties, they all have similar spatial settings. They all contain a large US city with suburban areas surrounding the city and rural areas surrounding the suburbs. What would happen in a spatial setting such as Manhattan, or a low populated area in Montana? A research design that could cover such issues of place and scale would surely provide interesting results.

The spatial errors presented in this research focus on positive and negative percent errors. However, due to the nested nature of the source and target zones, very complex patterns of negative spatial autocorrelation are seen. It may be beneficial to map absolute percent errors in addition to the positive and negative percent errors. This may give a better indication of which methods perform better under certain spatial circumstances.

It is also important to point out that these methods are only a subsection of the many areal interpolation methods. The more modern and complex methods that make use of neural networks and several different layers of ancillary information may likely be the future of areal interpolation. This research could not address all methods of areal interpolation, however, it provides a diverse sampling of methods that differ in terms of their assumptions about the underlying distribution of the population, the ancillary data used, and the dimensionalities of the ancillary data.

Research should also be conducted for determining appropriate weights for the network method. These weights are subjectively, not scientifically derived, and a method for determining these weights could improve the results. The dasymetric method implemented in this research uses a binomial weighting. This is one variation of the dasymetric method. Using different weights for the various land use classes and even the different residential classes could improve the results associated with this method. Also, using higher resolution remotely sensed data may improve the results of the dasymetric method. Remote sensing data is currently available at resolutions of 0.6 m. This means that objects greater than 1.2 m can be resolved. Therefore, residential houses could be classified, which should improve the results of the dasymetric method. This implies that the quality of the ancillary information can affect the results of the areal interpolation methods that make use of ancillary data. It should be noted that the results provided in this research are not valid across all spatial settings and scales, and can only be discussed in terms of the three cases presented in this research.

REFERENCES

Bloom, L., P.J. Pedler, and G.E. Wragg. 1996. Implementation of Enhanced Areal Interpolation Using MapInfo. *Computers & Geosciences.* 22(5): 459 – 456.

Cockings, S., P. Fisher, and M. Langford. 1997. Parameterization and Visualization of The Errors in Areal Interpolation. *Geographical Analysis.* 29(4): 231 – 232.

Deichmann Uwe. 1996. A Review of Spatial Population Database Design and Modelling. *Technical Report 96-3 for National Center for Geographic Information and Analysis.* 1-62.

Eastman, R. 1986. Opponent Process Theory and Syntax for Qualitative Relationships In Quantitative Series. *The American Cartographer.* 13(4): 324 - 333.

Eicher, C. and C.A. Brewer. 2001. Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science.* 28(2): 125 – 138.

Fisher, P., and M. Langford. 1995. Modeling the errors in areal interpolation between The zonal systems by Monte Carlo simulation. *Environment & Planning A.* 27: 211 – 224.

Flowerdew.R, and M. Green. 1991. Data integration: Statistical methods for transferring Data between zonal systems. In *Handling geographical Information.* Edited by I. Masser and M.B. Blakemore. Longman Scientific and Technical. 38 – 53.

Flowerdew, R. and M. Green. 1992. Developments in Areal Interpolation Methods And GIS. *Annals of Regional Science.* 26(67): 67 – 78.

Flowerdew, R. and M. Green. 1994. Areal Interpolation and Types of Data. In *Spatial Analysis and GIS.* Edited by S. Fotheringham and P. Rogerson. London, U.K. Taylor and Francis. 121 – 145.

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

Gregory, I.N. 2000. An evaluation of the accuracy of the areal interpolation of data for the analysis of long-term change in England and Wales. *Paper at Geocomputation 2000*

Hamilton, L. 1992. *Regression With Graphics: A Second Course in Applied Statistic.* Brooks/Cole Publishing Company.

Kyriakidis, P. 2005. A Geostatistical Framework for Area-to-Point Spatial Interpolation. *Geographical Analysis* 36: 259 - 289.

Lam, N.S. 1983. Spatial Interpolation Methods: A Review. *The American Cartographer.* 10(2): 129 – 149.

Langford, M., D.J. Maguire and D.J. Unwin. 1991. The Areal Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Framework. In *Handling Geographic Information: Methodology and Potential Applications,* edited by I. Masser and M. Blakemore (New York: Longman Scientific & Technical): 55 – 77.

Lillesand, T., and R.W. Kiefer. 1994. *Remote Sensing and Image Interpretation.* 4th Edition. Wiley Text Books.

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. *Geographical Information Science.* 11(1): 93 – 106.

Openshaw, S. 1977. Algorithm 3: A procedure to generate psuedo-random aggregations of N zones into M zones, where M is less than N. *Environment & Planning A.* 9: 1420 – 1426.

Rase, W-D. 2001. Volume-preserving Interpolation of a Smooth Surface from Polygon Related Data. *Journal of Geographical Systems.* 3, 199 - 203.

Reibel, M. and M.E. Bufalino. 2004. Street Weighted Interpolation Techniques For Demographic Count Estimation in Incompatible Zone Systems. *Environment & Planning A (forthcoming).*

Riedl, L. 1998. Pycnophylactic Interpolation Program for ArcView 3.2.

Robinson, A., J. Morrison, P. Muehrcke, A. Kimerling, and S. Guptill. 1984. *Elements of Cartography.* New York, New York: John Wiley & Sons.

Tobler, W. 1979. Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association.* 74(367): 519 –530.

Turner, A. and S. Openshaw. 2001. Dissaggregative Spatial Interpolation. *Paper Presented at GISRUK 2001 in Glamorgan Wales.*

United States Geological Survey. 1994. *Spatial Data Transfer Standard. Part 1: Logical Specifications.* FIPSPUB 173-1.

Voss, P., D.D. Long and R.B. Hammer. 1999. When Census Geography Doesn't Work: Using Ancillary Information to Improve the Spatial Interpolation of Demographic Data. *Environment & Planning A.*

Wright, J. 1936. A Method of Mapping Densities of Population: With Cape Cod as An Example. *Geographical Review.* 26(1): 103 –110.

U.S. Bureau of the Census. 1994 Tiger/Line Files. Appendix E: Census Feature Class Codes.

Xie, Y. 1995. The Overlaid Network Algorithms for the Areal Interpolation Problem. *Computer, Environment and Urban Systems.* 19(4): 287 – 306.

Data References:

OhioLINK Landsat 7 Image Server. 2000. Path 19, Row 32. http://dmc.ohiolink.edu/GEO/LS7/

U.S. Bureau of the Census. 2000 (a). Franklin County, Ohio Census Tracts.

_______________ 2000 (b). Franklin County, Ohio Census Block Groups.

_______________ 2000 (c). Franklin County, Ohio Roads.

_______________ 2000 (d). Franklin County, Ohio Census Tracts Demographics: SF1.

_______________ 2000 (e). Franklin County, Ohio Census Block Groups Demographics: SF1.

_______________ 2000 (f). Hamilton County, Ohio Census Tracts.

_______________ 2000 (g). Hamilton County, Ohio Census Block Groups.

_______________ 2000 (h). Hamilton County, Ohio Roads.

_______________ 2000 (i). Hamilton County, Ohio Census Tracts Demographics: SF1.

______________ 2000 (j). Hamilton County, Ohio Census Block Groups Demographics: SF1.

______________ 2000 (k). Jefferson County, Ohio Census Tracts.

______________ 2000 (l). Jefferson County, Ohio Census Block Groups.

______________ 2000 (m). Jefferson County, Ohio Roads.

______________ 2000 (n). Jefferson County, Ohio Census Tracts Demographics: SF1.

______________ 2000 (o). Jefferson County, Ohio Census Block Groups Demographics: SF1.

United States Geological Survey. 2001 (a). Land Use Land Cover Data 2001: Hamilton County, Ohio.

United States Geological Survey. 2001 (b). Land Use Land Cover Data 2001: Jefferson County, Kentucky.

APPENDIX A:

Implementation of Areal Interpolation Methods

This appendix is provided as a more detailed description of the actual steps taken by the author to implement the various areal interpolation methods. It provides an explanation of functionality used in commercial GIS software packages, along with the functionality programmed by the author.

Areal Weighting Method

1). The source and target zones were overlaid in ArcToolbox (in ArcMap 9) using the Analysis Tools --> Overlay --> Intersect command. Attributes from both the source and target zones were joined to the intersection layer.

2). Areas are calculated for the intersection zones.

3). A program using VisualBasic 6.0 with MapObjects 2.1 was written by the author. This program allows the user to add the intersect layer and the target zone layer to the program through use of a file menu. The user is then able to choose the variable to be interpolated, the target ID, the source ID, the source area and the intersection area from a set of combo boxes.

4). The user clicks the 'Interpolate' button, and the interpolation algorithm computes the areally weighted population values for the target zone. The computed value is written directly to a field named 'EstPopAW' that is contained in the target zone layer.

Pycnophylactic Method

1). The source zones are rasterized by using the 'Convert to Grid' function in ArcView 3.2. The output cell size is specified as 100 m, and the attribute table is joined to the grid.

2). The pycnophylactic interpolation script written by Leopold Riedl is run. The script asks the user for the ID that defines the zones. Next, the script asks the user for the variable to be interpolated and whether this variable should be normalized. The variable is population and it needs to be normalized. Finally the interpolation iterates, and the pycnophylactic interpolation is completed.

3). The target zones are rasterized by using the 'Convert to Grid' function in ArcView 3.2. The output cell size is specified as 100 m.

4). The 'Summarize by Zones' function in ArcView 3.2 is used to obtain the sum of the grid cell values within each target zone. The output of this function is an attribute table.

5). The resulting attribute table is joined to the target zone polygon layer by the unique target ID.

Dasymetric Method

1). The source zones and land use zones are overlaid using the ArcToolbox Intersect function. Attributes from both the source zones and land use zones are joined to the resulting layer.

2). Areas are calculated for the resulting intersection layer.

3). A program using VisualBasic with MapObjects was written by the author. The program allows the user to input a source zone layer and an intersection zone layer as input through the file menu. The user is then able to choose the variable to be interpolated, the source ID, the source area and the intersection area through combo boxes.

4). The user clicks the 'Interpolate' button and the algorithm calculates population values for each of the intersection zones. This is a modified version of the areal weighting method. The calculated population values are directly written to the a field called 'EstPop' in the intersection zone layer.

5). The intersection zones and target zones are overlaid using the ArcToolbox Intersect function.

6). Areas are calcuated for the resulting intersection layer.

7). The new intersection layer and the target zone layer become input files for the areal weighting method program written by the author. However, in this case, the final population estimates for the target zones are written to a field in the target zone layer called 'DasPop'.

Network Method

1). The source zones, target zones and road network layers are all converted to ESRI coverages. This will allow for the topology to be displayed as attributes.

2). The source zone coverage and the target zone coverage are overlaid using the ArcToolbox --> Coverage Tools --> Overlay --> Intersect function. This is referred to as the 'intersection zone coverage'.

3). The intersection zone coverage is overlaid with the road network coverage using the Intersection Function. The resulting layer is called the control-net.

4). The source zone coverage is overlaid with the network coverage using the Intersection function. The resulting layer is called the source-network layer.

5). A program using VisualBasic with MapObjects was written by the author. The program allows the user to input the source zone layer, the target zone layer, the intersection zone layer, the control-net layer and the source-network layer.

6). The user clicks the 'Interpolate' button, and the program computes population for the target zones. The computed population for the target zones is written to a field called 'EstPopNt' in the target zone layer.