University of Redlands

**Testing Kriging-Based Areal Interpolation for
Census-Based Socioeconomic Data**

A Major Individual Project submitted in partial satisfaction of the requirements
for the degree of Master of Science in Geographic Information Systems

by

Brian D. Carson

Douglas Flewelling, Ph.D., Committee Chair

Russell Weaver, Ph.D.

August 2013

Testing Kriging-Based Areal Interpolation for
Census-Based Socioeconomic Data

Copyright © 2013

by

Brian D. Carson

The report of Brian D. Carson is approved.

A handwritten signature in dark ink, appearing to read 'Russell Weaver', written over a horizontal line.

Russell Weaver, Ph.D.

A handwritten signature in dark ink, appearing to read 'Douglas M. Flewelling', written over a horizontal line.

Douglas Flewelling, Ph.D., Committee Chair

August 2013

Acknowledgements

I'd like to thank Debbie Riley for being there for me whenever I had an issue or just needed someone to talk to. I'd like to thank my advisor, Douglas Flewelling, Ph.D. I would also like to thank Kevin Butler, Ph.D., for encouraging me to challenge myself through research in both my undergraduate and graduate education. Also, for helping me to stay grounded during my time in Redlands. I'd like to thank my girlfriend, Kimberly Graham, for being patient during the course of my master's degree and sticking by my side even though we were 2,300 miles apart; my parents, George and Shelley Carson, for encouraging me to follow my dreams, and showing me the value of education even though it was difficult at times. I'd also like to thank all the members of Cohorts 21, 22, and 23 for being true friends and family during the last year.

Abstract

Testing Kriging-Based Areal Interpolation for Census-Based Socioeconomic Data

by
Brian D. Carson

This paper statistically tests the accuracy and efficacy of areal interpolation functionality introduced in esri's ArcGIS 10.1 Geostatistical Wizard. Esri has implemented a modified kriging-based technique for areal interpolation in the Geostatistical Wizard which is intended to simplify the overall process of interpolation. However, since kriging assumes data stationary, the feasibility of this method is questionable when applied to census-based socioeconomic data. To test the accuracy and efficacy of kriging-based areal interpolation, this project tested the efficacy of esri's methodology using nine variables representing continuous, count and ratio measurement scale data in the Combined Statistical Area of Kansas City, Kansas, Kansas City, Missouri, and Topeka, Kansas. The testing methodology employed both a Wilcoxon signed ranks test and a 90% confidence interval to determine whether kriging-based areal interpolation estimates generated by esri's Geostatistical Wizard would reproduce results considered acceptable by the U.S. Census Bureau.

Table of Contents

Chapter 1 – Introduction.....	1
1.1 Client.....	1
1.2 Problem Statement.....	1
1.3 Proposed Solution.....	1
1.3.1 Esri’s Objectives as the Driver for Research.....	2
1.3.2 Goals and Objectives.....	2
1.3.3 Scope.....	2
1.3.4 Methods.....	3
1.4 Audience.....	3
1.5 Overview of the Remainder of the Report.....	3
Chapter 2 – Previous Work	5
2.1 Smooth Pycnophylactic Areal Interpolation.....	5
2.2 Overlay.....	6
2.3 Overlaid Network.....	8
2.4 Modified Kriging-Based Areal Interpolation.....	9
2.5 Summary.....	11
Chapter 3 – Systems Analysis and Design	13
3.1 Problem Statement.....	13
3.2 Requirements Analysis.....	13
3.3 System Design.....	13
3.4 Project Plan.....	15
3.5 Summary of Chapter.....	15
Chapter 4 – Database Design	17
4.1 Conceptual Data Model.....	17
4.2 Logical Data Model.....	17
4.3 Data Sources.....	19
4.4 Data Collection Methods.....	19
4.5 Data Scrubbing and Loading.....	19
4.6 Summary.....	20
Chapter 5 – Implementation	21
5.1 Areal Interpolation via the Geostatistical Wizard.....	21
5.2 Statistical Testing.....	24
5.3 Chapter Summary.....	25
Chapter 6 – Results and Analysis.....	27
6.1 Kriging Model.....	28
6.2 Continuous Data – Average Household Income.....	28
6.2.1 Continuous Data - Average HH Income: HHr 25 – 34.....	32
6.2.2 Continuous Data – Per Capita Income.....	34
6.3 Count Data – Total Population.....	37
6.3.1 Count Data – Total Households.....	41

6.3.2	Count Data – HHr 25-34/Income <\$15,000	45
6.4	Ratio Data – Percent of Population 65 and Older	48
6.4.1	Ratio Data – Population-Single Race: Black Percentage	51
6.4.2	Ratio Data – Percent Multigenerational Households.....	55
6.5	Residuals	57
6.6	Local Analysis	57
Chapter 7 – Conclusions and Future Work		61
Works Cited		65
Appendix A. Kriging Parameters		67

Table of Figures

Figure 2-1: Pycnophylactic Interpolation (National Center for Geographic Information and Analysis).....	6
Figure 2-2: Overlaid Interpolation.....	7
Figure 2-3, Network Length Interpolation. Reprinted from “The Overlaid Network Algorithms for Areal Interpolation Problem,” by Y. Xie, 1995, 19(4), 295. Copyright 1996 Elsevier Science Ltd. Reprinted with permission ..	9
Figure 2-4: Kriging-based Areal Interpolation (ESRI, 2012).....	10
Figure 3-1. Topeka, Kansas and Kansas City, Kansas-Missouri Combined Statistical Areas.....	14
Figure 4-1, Conceptual Database Model	17
Figure 4-2, Logical Data Model.....	18
Figure 5-1, Areal Interpolation	21
Figure 5-2, Interavtive Variography Window in the Geostatistical Wizard Areal Interpolation Tool.....	22
Figure 5-3, Initial Prediction Surface.....	23
Figure 5-4, Cross-validation Results.....	24
Figure 6-1, Study Area – Kansas City, Kansas-Missouri and Topeka, Kansas, Combined Statistical Areas. Inset map showing the location of study area on the Kansas-Missouri Border.	27
Figure 6.2. Average Household Income	29
Figure 6-3. Block Groups Within Plus or Minus 5% of True Value	31
Figure 6-4. Average HH Income: HHr 25-34	33
Figure 6-6. Per Capita Income	35
Figure 6-7. Block Groups Within Plus or Minus 5% of the True Value	36
Figure 6-8. Total Population.....	38
Figure 6-8. Block Groups within +/- 5% of True Value.....	40
Figure 6-9. Total Households	42
Figure 6-10. Block Groups within +/- 5% of True Value.....	44
Figure 6-11. HHr 25-34/Income < \$15,000.....	46
Figure 6-13. Percent of Population Aged 65+	49
Figure 6-14. Block Groups within +/- 5% of True Value.....	51
Figure 6-15. Percent Black	52
Figure 6-16. Block Groups within +/- 5% of True Value.....	54
Figure 6-17, Percent Multigenerational Households	56

List of Tables

Table 1.	Kriging model validation results are used to illustrate that the kriging models used in the interpolation of the tracts to the block groups were valid.....	28
Table 2.	Number of Block Groups Which Fell Within the American Community Survey's Margin of Error by Area	58
Table 3.	Block Groups Within American Community Survey Margin of Error Based on Whether They are on the Edge or Not on the Edge of the Study Area	59
Table 4.	Percentage of Block Groups Falling Within the Margin of Error based on Urban or Rural Designation	60
Table 5.	Table of Results for Kriging model (Mean Standard Error and Root Mean Square Error) and Statistical Comparison of True and Predicted Block Groups (p-value, Number of Block Groups Within Confidence Interval, and Percent Within +/- 5% of Confidence Interval). * Derived Variables not Included in the analysis Because Margin of Error not Available.....	62

Chapter 1 – Introduction

The purpose of this project is to test the accuracy and efficacy of esri's kriging-based areal interpolation methodology that was introduced in the Geostatistical Wizard in ArcGIS 10.1. This chapter provides an overview of the project, discusses the problem identified by the client, and gives a brief overview of the analysis methods used.

1.1 Client

The client for this project was Kevin Butler, Ph.D., a Product Engineer on the Geoprocessing Team at esri. Dr. Butler was interested in investigating and assessing a methodology that esri added to the Geostatistical Analyst toolbar for areal interpolation. Specifically, Dr. Butler wanted to evaluate how well the approach to areal interpolation would work when applied to census-based socioeconomic data.

1.2 Problem Statement

Areal interpolation, the process of aggregating and disaggregating existing data to new areal units, is by definition a process of estimation. Accordingly, it is only as accurate as the algorithms created to accomplish the task. In ArcGIS 10.1, an areal interpolation tool was introduced to facilitate the process of disaggregating and reaggregating data from one set of polygons to a second set of polygons. Esri's new tool employs a modified kriging technique to accomplish the areal interpolation for continuous, ratio and count data (Krivoruchko, Gribov, & Krause, 2011). Kriging is a geostatistical estimator that allows one to infer the value of a field at an unobserved location based on known values at sample locations. Esri's methodology greatly changes previous methods of using a ratio of the new land area to the old as described in Hart's seminal article, "The Perimetropolitan Bow Wave" (Hart, 1991). While kriging has been used in the realm of physical geography for some time, it is just emerging in the field of human geography. However, concerns have been raised as to the accuracy of the new functionality when applied to demographic and socioeconomic phenomena.

1.3 Proposed Solution

The Krivoruchko et al. (2011) modified kriging technique for the disaggregation of continuous and count data is a novel approach to areal interpolation. However, as with all innovative statistical methods, thorough testing is needed to determine the appropriateness of the method when applied to known data. The goal of this project is to illustrate the accuracy and efficacy of esri's tool. In order to accomplish this goal, a U.S. Census Combined Statistical Area, covering both urban and rural landscapes, was used as the study area in order to determine the applicability of esri's kriging-based areal interpolation methodology for census-based socioeconomic data. Results of the interpolation were evaluated using a Wilcoxon signed ranks test to compare the median of the known and predicted values. The predicted values were further analyzed to determine if they fell within 10%, the U.S. Census Bureau's acceptable margin of error, of the known values. The project produced this report that details the statistical support

for the areal interpolation tool's accuracy and efficacy when applied to continuous, count, and ratio data.

1.3.1 Esri's Objectives as the Driver for Research

In early 2012, esri released new functionality in version 10.1 of the Geostatistical Analyst extension to solve a classic problem in GIS; the reallocation of data from one set of polygons to a different set of polygons. Esri's implementation, which is based on kriging, was a strict departure methodologically from existing algorithms present in the literature. All previous methods had either an assumption of homogeneity across the study area or required ancillary variables to control for the non-homogeneity. Also, esri explicitly supported averaged, rate counts, and event data. While esri was confident in the statistical robustness of their implementation, they had concerns regarding its operational feasibility across their diverse client base. Their implementation of areal interpolation requires the use of interactive variography and kriging. The group of users arguably most interested in areal interpolation, social scientists and planning professionals, were likely unfamiliar with these techniques. Given that esri had implemented a new areal interpolation algorithm, expanded its application to data captured on different measurement scales and implemented it through the Geostatistical Wizard motivated the following objectives which serve as the driver for this research:

1. Assess the operational feasibility of esri's areal interpolation workflow for a user new to kriging and interactive variography.
2. Determine the accuracy of the predicted values generated by esri's implementation of areal interpolation using a real-world case study for continuous, rate and count socio-economic data.
3. Determine if there are spatial patterns in the predicted values that might inform best practices for using areal interpolation with census based areal units.

1.3.2 Goals and Objectives

The goal of this project is to assess the accuracy and utility of ArcGIS 10.1's areal interpolation tool when applied to census-based socioeconomic data. To achieve this goal, a database was constructed with census data for the Combined Statistical Areas of Kansas City, Kansas, Kansas City, Missouri, and Topeka, Kansas. Data were obtained from esri's Community Analyst[®], which is based on the American Community Survey. Through a statistical analysis of the tool's predictive output against the actual data, this report was created which illustrates the accuracy and efficacy of esri's methodology through the use of maps and tables. In addition, the project developed a sustainable testing methodology for testing the tool's results.

1.3.3 Scope

This project focuses on developing and implementing a testing methodology for evaluating the accuracy and efficacy of esri's areal interpolation functionality. A database was created to manage the data used for the analysis, as well as organizing the tables summarizing the testing results so that they may be statistically compared to the true data

values for accuracy. A technical report detailing the findings of the analysis and how to repeat the testing procedure appears at the end of this report.

1.3.4 Methods

The process of determining the accuracy and efficacy of esri's areal interpolation functionality involves several steps. First, data were obtained for the study area at both the census tract and block group levels. The data were then organized into feature datasets in a file geodatabase.

After the database was constructed, the areal interpolation tool was applied to census-based socioeconomic data. The data were first disaggregated to new areal units, from the tracts to block groups. During this process, the statistics created during the interactive variography process were documented. Upon completion of the interactive variography, the Geostatistical Wizard output, geostatistical layers, were then used to predict the data values from the tracts to the block groups. During this process, a standard error, a representation of prediction uncertainty, was given and documented to be used in comparing the target polygons to the available ground truth data.

1.4 Audience

The primary audience for this project is GIS professionals. The project will serve as a resource for informed decision making on the appropriateness of kriging-based areal interpolation for socioeconomic data. The GIS professionals targeted for the project potentially represent non-governmental organizations, city, county, and regional planning organizations, consultants, and non-profit groups. Also, academics and other research personnel might be interested in the results of the project.

1.5 Overview of the Remainder of the Report

In Chapter Two, this report provides background information on areal interpolation, how it has been previously used and a comparison of how the old and new methodologies differ. Chapter Three covers the systems analysis and design. Chapter Four describes the structure and development of the database. Chapter Five discusses how the project was implemented. Chapter Six presents the results and analysis. Conclusions and ideas for future work are presented in Chapter Seven.

Chapter 2 – Previous Work

Since 1812, when the U.S. Census of Population and Housing was expanded to collect information not only on total population but also on social and economic indicators, areal interpolation of census-based socioeconomic data has been a common problem in geography (U.S. Census Bureau, 2012). Census data are limited in that it is only available at certain geographies or summary levels. Unfortunately, certain variables, such as median household income and educational attainment, are useful to municipalities and planning organizations at many geographies, but fall into the aforementioned category. For data on these variables to be of use at smaller geographies, they must be disaggregated to boundaries that suit the needs of the organization doing the analysis. Since data come from a variety of sources, users are often presented with areal units which are incompatible (Xie, 1995). Local governments and planning organizations need census data at the boundaries that are most suitable for question that they are trying to answer, e.g., traffic analysis zones, school districts, wards, etc. In response to this problem several methods have been proposed to date, including, smooth pycnophylactic areal interpolation, overlaid interpolation, overlaid network interpolation, and kriging-based areal interpolation.

2.1 Smooth Pycnophylactic Areal Interpolation

Tobler (1979) stated that methodologies for areal interpolation had not changed in 100 years. Tobler outlined an approach to smoothly aggregate a spatial distribution based on data that had been obtained at a specified areal unit. Data are available from the U.S. Census Bureau at a variety of areal units. The data provided by the U.S. Census Bureau are used by organizations to help understand the particular socioeconomic characteristics of an area. However, census data often are not available at the areal unit that an organization requires due to privacy constraints. Thus, a method was needed for creating a density surface for the reaggregation of the original data. Tobler devised a technique which he referred to as smooth pycnophylactic interpolation (1979). Pycnophylactic indicates that the method is mass preserving, or that each of the smaller areal units that fall within the source polygon add up to the initial value (Tobler, 1979). Tobler described the method by looking at blocks of clay. He visualized it as blocks of different colored clay, with each block's mass representing the population, laid out in the shape of the area in question. The next step was to shape the clay into a perfectly smooth surface without allowing any of the clay to move from one block to the next—thus maintaining the pycnophylactic condition (Tobler, 1979, p. 520).

His technique used a grid (Figure 2-1) to approach the problem (Tobler, 1979). The grid is laid over the study area, ensuring the grid size is small enough so that ideally multiple points fall within each polygon. From there, density can then be calculated for each point. Once the density surface was created, the density values could be categorized at the intersections of the lattice based on which areal unit they fell within. This disaggregated the original data to the new areal unit (Tobler, 1979).

Figure 2-1: Pycnophylactic Interpolation (National Center for Geographic Information and Analysis)

One issue with the overlaid grid was how to address the values at the edge of the study area, or the edge effect. The edge effect, or the values assigned to the outside of the study area, impacts not only the overall smoothness of areal interpolation at the edge, but also throughout the study area (Tobler, 1979). Based on the fact that Tobler was using an elliptical partial differential equation as part of his method, he created boundary conditions (1979, p. 523). Tobler theorized that there were two types of boundary specifications, the Dirichlet condition and the Neumann condition (1979). In the Dirichlet condition grid cells along the edge of the study area are assigned a numerical value, while in the Neumann condition, the rate at which the densities change across the boundary must be specified (Tobler, 1979). Tobler ascertained that his method preserved volume, and believed that it improved on previous methods (Tobler, 1979).

2.2 Overlay

In 1980, Goodchild and Lam revisited Tobler's (1979) pycnophylactic approach. Believing that their variation was a logical extension of Tobler's approach, Goodchild and Lam aimed to avoid the point interpolation outlined by Tobler (1979). They approached the problem using a polygon overlay of the source and target polygons (Figure 2-2). Upon completion of the overlay Goodchild and Lam then defined a matrix representing the areas of intersection between the source and target polygons. A key assumption of the method was that of homogeneity within the source zones (Goodchild & Lam, 1980).

Census Blocks

National Park Boundary

Overlay of Census Blocks
and National Park Boundary

Figure 2-2: Overlaid Interpolation

While Goodchild and Lam believed that their process was intuitively simple; there was still a possible flaw. This process could be easily reversed (p. 301). By using the predicted values as the inputs it is possible, using the same interpolation method, to revise the values of the source polygons (Goodchild & Lam, 1980). Goodchild and Lam stated that in, *Lattice Tuning* (1979), Tobler argued towards using the difference, between the source and target polygons, as a measure of the overall performance of areal interpolation.

In 1980, the Goodchild and Lam overlay method for areal interpolation was too computationally difficult and expensive to calculate exactly for each polygon. As a result, they suggested using an approximation. By using approximation, or assigning the source value to the target with the greatest number of intersections to the source, Goodchild and Lam stated that it would be less computationally intensive. “These methods [approximation] have considerable operational advantages; since the longest intersections, and the non-zero intersections can both be identified visually one avoids the need to compute overlays and measure areas, and multiply large numbers of terms” (Goodchild et al, p. 308).

The degree of homogeneity within the source zone has a significant effect on the accuracy of the overlay method (Goodchild and Lam, 1980). For example, with Goodchild and Lam’s methodology, the correlation between the source and target zones increased, but the percent error deteriorated once criteria for compactness and size were included (1980). Comparing Goodchild and Lam’s approach with that of Tobler, both introduce a pycnophylactic condition. Goodchild and Lam concluded that the best approach to areal interpolation would be one that finds a balance between the pycnophylactic condition and homogeneity (1980).

2.3 Overlaid Network

Xie (1995) approached the problem in a manner that differed both from Tobler (1979) and Goodchild and Lam (1980). Xie believed, that while much literature had been written dealing with the issue of areal interpolation, there was still not a satisfactory solution (1995). Existing methods were limited due to the fact that they dealt with either a single variable or an estimate (Xie, 1995). As a result, Xie incorporated GIS techniques to include more relevant and directly related information (Xie, 1995). By leveraging the power of GIS and advances in computing, Xie developed a multivariate methodology which uses a series of algorithms based on a street network (1995).

Xie hypothesized that TIGER/Line[®] files containing street segments would be useful for the areal interpolation process. Since people typically live in houses along streets, Xie hypothesized that the street network would be closely related to the population’s distribution over an area (1995, p. 294). Based on the source layer and information included in the TIGER/Line file[®], Xie (1995) determined that population could be distributed in three ways based on a series of overlays; by street segment length, by street segment classes and by the number of houses on street segments (Figure 2-3).

Figure 2-3, Network Length Interpolation. Reprinted from “The Overlaid Network Algorithms for Areal Interpolation Problem,” by Y. Xie, 1995, 19(4), 295. Copyright 1996 Elsevier Science Ltd. Reprinted with permission

According to Xie (1995), previous methods produced a crude spatial interpolation. The use of his overlaid network algorithms illustrated new ways of aggregating population values. Through the use of modern GIS, analytical methods, and the availability of digital information, Xie developed a method that would produce a more accurate interpolation than previously possible (1995).

2.4 Modified Kriging-Based Areal Interpolation

Krivoruchko, Gribov and Krause (2011) described a new implementation of areal interpolation which used a kriging-based technique for the disaggregation of data

measured on three different scales: binomial, ratio and continuous data. Their method created an areal interpolation model that overcame the issue of disproportionately sized overlapping or disjointed polygons (Krivoruchko et al., 2011) (Figure 2-4). Their method uses interactive variography and a series of models to create a prediction surface from the source data through the use of an interactive wizard. Variography is a geostatistical method used to measure spatial continuity based on the assumption that observations which are spatially close will have similar values due to being generated under similar conditions (Isaaks & Srivastava, 1989). Upon completion of the interactive variography, the output prediction surface can then be reaggregated to the target polygons thus completing the areal interpolation process.

Figure 2-4: Kriging-based Areal Interpolation (ESRI, 2012)

While the technique introduced by Krivoruchko et al. seems to have improved the speed at which data are interpolated, the main assumption of kriging, stationarity, was not explicitly addressed. Stationarity means that as you move from one region to the next, the average value of the points is relatively constant. For further explanation of the concept of stationarity, the concept of fields versus objects in geography must be explained.

In geography, geographic phenomena have been divided into two main ontologies, object-based and field-based. Within the object-based ontology the world is described as a space filled with discrete units, or objects. Objects, which are typically important in the realm of human geography may include: people, places, and other phenomena which may be measured with discrete units. Field-based data are common in the realm of physical geography and are more commonly used with kriging. Field-based data include things such as temperature, elevation, or anything which can be continuously measured at any point. On the other hand, objects, or things that occupy a specific space, which are typically important in the realm of human geography, are often nonstationary. The difference between field-based data, which are commonly used with kriging, compared to object-based data which are not, was a distinction which was not addressed by Krivoruchko et al.

2.5 Summary

The concept of areal interpolation has been in use in the fields of geography and demography for some time. Tobler (1979) laid out a method for areal interpolation which focused on maintaining the pycnophylactic condition and creating a smooth surface. Goodchild and Lam (1980) revisited Tobler's (1979) approach with the aim of finding a balance between maintaining the pycnophylactic and homogeneity qualities. Xie (1995) aimed to relate the population to something that was physically measureable through the use of overlaid networks. In doing so, Xie's approach advanced the process by taking the spatial structure of an area into account. Krivoruchko et al. (2011) brought the statistical robustness of a kriging-based technique to the world of GIS. With the assumption of stationarity that comes with kriging, concerns remain as to the efficacy of the technique when used with census-based socioeconomic data which are often nonstationary.

Chapter 3 – Systems Analysis and Design

This chapter will discuss the requirements that were specified by the client at the start of the project. In Section 3.1, the problem will be reiterated. An analysis of the client's specifications follows in Section 3.2. Section 3.3 outlines the general system design as well as the timeline for the project. The project plan follows in Section 3.4 and includes more detail as to why certain choices were made throughout the project lifecycle.

3.1 Problem Statement

Determine if esri's implementation of a kriging-based methodology for areal interpolation provides acceptable results when applied to census-based socioeconomic data.

3.2 Requirements Analysis

The primary requirement for this project was to design and implement a process for testing the overall accuracy of esri's adopted kriging-based areal interpolation method when applied to census-based socioeconomic data. To ascertain the utility of esri's method, a set of acceptance criteria was outlined. This included meeting the U.S. Census Bureau's 10% margins of error, statistical testing to measure the overall difference between the median values of true and predicted target polygons, and determination of the validity of the kriging models used.

Five functional requirements were defined by the client. The first functional requirement was to test the accuracy esri's kriging-based method when used with census-based socioeconomic data. This functional requirement is important to the client due to esri's method not being comprehensively tested with census-based socioeconomic data. The client wanted to determine how well a kriging-based methodology, which assumes stationarity, would work with data that are often nonstationary. In order to test the accuracy and efficacy of esri's new method, the client specified that a study area be chosen that contained a mix of rural and urban areas with a population of two million or greater. Statistical evidence of whether or not the method would work with the data types in question was also requested. Lastly, the client specified that three different variables be tested for each of three data types that are available for use in the Geostatistical Wizard.

3.3 System Design

This section will illustrate the general design structure for the project requested by Dr. Butler. Dr. Butler requested that a testing program be designed to aid in the determination of the accuracy and efficacy of esri's adopted kriging-based approach to areal interpolation, when applied to census-based socioeconomic data. Because esri's adopted technique for areal interpolation is based on kriging and therefore assumes stationarity, there is reason to doubt its ability to accurately disaggregate and reaggregate nonstationary census data. This testing protocol aids in demonstrating the effectiveness of the areal interpolation methodology to esri customers, as well as to help to remove some of the concerns surrounding the kriging-based approach.

The data required for this project were all collected from the U.S. Census Bureau. The data included both demographic and socioeconomic variables which were obtained through esri's Community Analyst[®]. Using Community Analyst[®], as opposed to obtaining all data directly from the U.S. Census Bureau, expedited the data acquisition process. Data were obtained at both the tract and block group levels. As the final deliverable for the project is the development of a testing protocol for esri's areal interpolation functionality, the same variables were obtained at both summary levels. This allows for a baseline at the block group level to be created. This baseline was then used to compare to the output from the Geostatistical Wizard's areal interpolation tool.

The study area for this project was the Combined Statistical Area of Kansas City, Kansas-Missouri and Topeka, Kansas (Figure 3-1). The Kansas City, Kansas-Missouri Combined Statistical Area is made up of fifteen counties: Bates, Caldwell, Cass, Clay, Clinton, Franklin, Jackson, Johnson, Lafayette, Leavenworth, Linn, Miami, Platte, Ray, and Wyandotte. The Topeka, Kansas Combined Statistical Area is comprised of five counties: Jackson, Jefferson, Osage, Shawnee, and Wabaunsee. Falling in the middle of the study area is the metropolitan statistical area of Douglas County, Kansas. This area was excluded from the study area because it is not part of either Combined Statistical Area. The total population for the twenty counties within the study area as of the 2010 U.S. Census was 2,576,878.

Figure 3-1. Topeka, Kansas and Kansas City, Kansas-Missouri Combined Statistical Areas

This project was completed using ArcGIS 10.1, IBM SPSS 20 and Microsoft Office 2010 software package. Within the ArcGIS 10.1 suite, ArcMap, ArcCatalog, Community Analyst®, and the Geostatistical Analyst extension were utilized. The analysis database was built in ArcCatalog. Community Analyst® facilitated the data collection process. ArcMap and the Geostatistical Analyst extension were used in conjunction with the Geostatistical Wizard to perform the areal interpolation across the study area. IBM SPSS was used to carry out statistical testing.

3.4 Project Plan

The goal of the project was to assess the ability of esri's kriging-based areal interpolation tool to reproduce known census-based socioeconomic data at various levels of disaggregation. Because population demographics tend to be non-stationary, the accuracy and efficacy of esri's kriging-approach is in question. Since socioeconomic data are often nonstationary, a kriging-based approach can be problematic due to the violation of the method's assumptions. By conducting statistical analyses of the effectiveness of a kriging-based approach to areal interpolation when applied across different areal units, city models, and levels of homogeneity, the report will produce empirical evidence for readers to assess the utility of esri's approach, in the context of non-stationary data.

The major components of the project included the acquisition of data, interactive variography to create the kriging model, reaggregation of the prediction surfaces to target polygons, and the comparison of the true and predicted values. The deliverable for this project is this report which outlines the accuracy and efficacy of kriging-based areal interpolation when applied to census-based socioeconomic data.

When it came to identifying the critical success factors for the project, the sole factor determining the overall success of the project was the final report. The report fulfills the clients requirement to design and complete an analysis of the accuracy and efficacy of the kriging-based approach that esri has taken towards areal interpolation in regard to typically non-stationary data.

3.5 Summary of Chapter

Overall, the project proceeded with few alterations to the initial project plan. However, the timeline was changed due to the analysis portion of the project taking much longer than expected. This was due mostly to limited help documentation for the esri tool, and the limitations of only being able to perform geoprocessing in a 32-bit environment. While the technical documentation does provide an outline for how to complete the interactive variography, it lacks sufficient detail as to how to best create a valid kriging model. Also, the interactive variography used to create the kriging model has been implemented solely through a wizard. It does not allow the user to leverage the computing power of 64-bit background geoprocessing for use with large datasets.

Chapter 4 – Database Design

In order to facilitate the development of the project's statistical testing methodology, it was important to ensure that the database design was simple and intuitive. After developing the conceptual model, the logical model needed to be robust enough to handle both the original input datasets and the tool's output products. Once the database was constructed, the data were imported from a variety of sources including Community Analyst and Tiger\Line files®.

4.1 Conceptual Data Model

When designing the conceptual data model for this project the primary goal was to keep it simple and well organized. The basic structure that was to be used for the database was to separate the elements out by their areal unit (Figure 4-5). For this project, the database's sole purpose was to maintain organization of the data throughout the testing process. Accordingly, the database structure is relatively simple. Within the database, there are two data sub groups: tracts and block groups. The block group data are further subdivided into known block groups and predicted block groups. By organizing the database in this manner, it maintains the requirement of ease of duplication of the testing procedure as set forth by the client.

Figure 4-1, Conceptual Database Model

4.2 Logical Data Model

The conceptual database model was implemented using a file geodatabase built in ArcCatalog 10.1 (Figure 4-2). The basis for the creation of the logical database model was to have something that could be recreated from an XML file as an empty shell. This was critical as it allows future users to replicate the data structure used for the input and output files from the Geostatistical Wizard.

Figure 4-2, Logical Data Model

Within the geodatabase, there are four feature datasets and one raster catalog. Feature datasets were used to organize initial raw data, predicted outputs, cross validation results and study area input data. The raster catalog was used to organize the prediction surfaces which were output by the Geostatistical Wizard. The first feature dataset, “Raw_Shapefiles”, includes initial data collection, the study area boundaries, and the “Predicted_Block_Groups” table. The second feature dataset, “Interpolated_Data”, was used for the predicted polygons. The third feature dataset, “CrossValidationResults”, was used for the output cross validation layers from the Geostatistical Wizard. The final feature dataset, “TopekaKS_KansasCityMOKS”, was used to hold the master data files for the true values for tracts and block groups.

Organizing the geodatabase in this manner provided a logical structure for testing esri’s areal interpolation methodology. Due to the large number of outputs from the process, it was important to have a separate feature dataset for the different types of output layers. This facilitated organizing the results from the areal interpolation process, and provided a method for tracking what had been completed during the process.

4.3 Data Sources

Data for the project came from both the U.S. Census Bureau and from esri's Community Analyst. The data from the U.S. Census Bureau were TIGER/Line® Shapefiles. These layers included the boundaries for the study area which was the combined statistical area of Kansas City, Kansas-Missouri and Topeka, Kansas, the counties which fell within the study area, the state boundaries, and the boundaries for census tracts and block groups for the 2010 decennial census. The socioeconomic data used were sourced from esri's Community Analyst. The socioeconomic data obtained from Community Analyst was compiled from the 2010 American Community Survey. The use of Community Analyst provided for a faster data collection method than would have been possible through American FactFinder. American FactFinder limits the amount of data in which the user can download at one time, and does not allow for download based on a study area. Through using Community Analyst, the time spent obtaining and scrubbing data was reduced.

4.4 Data Collection Methods

The shapefiles used for the study area boundary were obtained from the U.S. Census Bureau's TIGER/Line® shapefiles. The data was delivered as a .zip file, which was extracted and loaded into the "Raw_Shapefiles" feature dataset.

The process of obtaining the socioeconomic data from esri's Community Analyst, however, was not as simple. First, the Community Analyst extension had to be downloaded and added to ArcMap. Once the extension was added to ArcMap, it could be opened through a toolbar icon. The Community Analyst extension provides access to compiled demographic and socioeconomic data. It can be used to either generate a formatted report containing user-identified information about an area of interest, or to automatically join the data to the shapefile which the user set as area of interest. While using Community Analyst to obtain the data was quicker than using The New American Fact Finder from the U.S. Census Bureau, it does have some limitations. Community Analyst limits the user to 100 tuples at a time when joining data to the shapefile. While this does add time to the overall data collection process, it was still ultimately faster and more user friendly than going directly through the U.S. Census Bureau.

4.5 Data Scrubbing and Loading

The process of scrubbing and loading the data for this process into the geodatabase was all handled within ArcCatalog 10.1. The raw shapefiles obtained from the U.S. Census Bureau's TIGER/Line® Shapefiles were loaded into the "Raw_Shapefiles" feature dataset within the geodatabase via the Feature Class to Geodatabase tool. Once the shapefiles were in the feature dataset data could be joined 100 tuples at a time in Community Analyst. As the data were joined to the study area, each join had to be exported to a new shapefile. Once the shapefile for each set of 100 polygons was generated, the files were merged to create a single feature class for both the tracts and the block groups within the study area. Upon completion of the merge, the Delete Fields tool was used to remove all but the variables of interest from the shapefiles. Finally, the

completed tract and block group datasets were moved to the TopekaKS_KansasCityMOKS Feature Dataset.

4.6 Summary

The structure of the database was designed specifically for the use of the areal interpolation functionality of the Geostatistical Wizard. Through the use of separate feature datasets and a raster catalog, the three outputs which are produced are able to be organized for future use. The database design also allowed the user to track the areal interpolation tool's process based on the generated outputs. Using one feature dataset to organize the raw data during data collection, facilitated tracking which data tuples had been obtained from Community Analyst.

Chapter 5 – Implementation

The design and implementation of a testing program for kriging-based areal interpolation of census-based socioeconomic data was the primary goal of this project. This chapter will discuss the process of implementing esri's new areal interpolation functionality from the Geostatistical Wizard, as well as a statistical approach for assessing the method's suitability for use with census-based socioeconomic data. In section 5.1, esri's new approach to areal interpolation is discussed. Section 5.2 outlines the process for statistically testing the predicted values against true values for the final areal unit. Section 5.3 provides a summary of the entire chapter.

5.1 Areal Interpolation via the Geostatistical Wizard

Kriging-based areal interpolation involves interpolating from source polygons to target polygons. This is a five step process implemented using the Geostatistical Wizard from the Geostatistical Analyst toolbar (Figure 5-1). After selecting areal interpolation in the wizard, the dialog box shows the input data options. The wizard allows the user to choose from three input data types: average or continuous data such as income, rate or ratio data such as percent of a population, and event or count data such as total population.

Figure 5-1, Areal Interpolation

After the user identifies the input data and data type, the user must perform interactive variography. Interactive variography is the process of fitting the model to the data in order to develop a valid kriging model (Figure 5-2). During interactive variography, the wizard presents the user with a graphical view of the model that can be changed to better fit the model to the data. Next, the user adjusts the lag size, or the

sample distance used to group or bin pairs of points, and the number of lags, or the number of bins, to align the mean values, as represented by the blue plus signs (Figure 5-2), to fall within the confidence intervals, as represented by the red bars. Once the model has been fit to the data, the user can then proceed to step three of the wizard.

Figure 5-2, Interactive Variography Window in the Geostatistical Wizard Areal Interpolation Tool

The third step of the wizard provides an initial picture of the prediction surface based on the model created thus far. It also provides the user with the opportunity to examine the predicted values for a given point by moving a cursor (+ sign) within the circle (Figure 5-3). Once the user is satisfied with the prediction surface, they can proceed to Step Four of the wizard. However, should the user see issues with their prediction surface, they can return to Step Two and adjust the model.

Figure 5-3, Initial Prediction Surface

Step Four of the wizard allows the user to view the cross-validation results for the prediction surface. Cross-validation is used to assess how the results of the kriging model will generalize when predicted to a new areal unit. During this step, the user can examine measured and predicted values, error, standard error, standardized error, and normal values. The cross-validation results can also be used to assess the validity of the kriging model. To determine the validity of the kriging model, the mean standardized error, and the root mean square standardized errors need to be examined. For the model to be considered valid, the mean standardized error should be close to 0, and the root mean square standardized error (RMSE) should be close to 1. The mean standardized error is used to measure the overall validity of the kriging model, while the root mean square error is used to determine if the variance in the predictions had been over estimated (value below 1), under estimated (value above 1), or close to 1 if the predictions are valid (Bivand, Pebesma, & Gómez-Rubio, p. 225). It also creates plots for predicted, error, standardized error, and a normal QQPlot should the user be interested in viewing those. In addition to being able to view the cross-validation results, the user is also provided with the option to export the results to a file

Figure 5-4, Cross-validation Results

Step Four, completes the areal interpolation process. After the wizard completes, the prediction surface is added to the map document as a Geostatistical Analyst layer. At this point, the user can right-click on the prediction surface and select “Predict to Polygons”. This opens a dialog box allowing the user to specify target polygons, and in turn, push (i.e. reaggregate) the predicted values to those polygons.

5.2 Statistical Testing

After the prediction surfaces were created, the project sought to statistically compare the true values for the study area with the predicted values. Predicted values were tract values predicted to block group polygons. True values were the actual block group values reported by the U.S. Census. To compare the true and predicted values, the project utilized a nonparametric test. A nonparametric test was necessary because the census data selected for this project are not normally distributed. Had the data been normally distributed, it would have been acceptable to use a two-difference of means t test. However, when the data are not normally distributed, the median is a better measure of centrality. Accordingly, a nonparametric test was used to analyze median differences.

The statistical testing was completed using IBM’s SPSS Statistics, Version 20. The Wilcoxon signed rank test, which makes an inference regarding the median difference between two populations in a matched pairs setting, was used. The test was accessed in SPSS through the two related samples option, in the nonparametric test dialog window. The Wilcoxon signed ranks test includes the sample size, negative and positive ranks, mean rank for both the negative and the positive values, sum of ranks, a z score and a p-value. For the purpose of testing the accuracy of kriging-based areal interpolation, the only statistic of interest was the p-value, measured on a scale of 0 - 1. The p-value was used to either reject or not reject the null hypothesis. The null hypothesis is that the

medians of the predicted and true samples are equal. The alternate hypothesis is that the medians of the predicted and true values are significantly different. If $p > 0.05$, we fail to reject the null hypothesis at a 95% level of confidence. However, if the p-value is less than 0.05, the null hypothesis is confidently rejected, and one can conclude the medians are significantly different.

Using this approach, the p-values can be used as an indicator of whether the medians of the true (U.S. Census block group estimates) and predicted values (results from areal interpolation) are statistically significantly different. However, this is a global measure of kriging-based areal interpolations efficacy with census-based socioeconomic data. In order to answer that question from a spatial perspective, it is important to quantify both the number of predicted values that fell within 10%, 5% above and below, of the true values and their spatial pattern. A 10% confidence interval was used because it is the same margin of error used by the U.S. Census Bureau for American Community Survey data (U.S. Census Bureau, 2012). For testing purposes, the margin of error, provided with American Community Survey data, was used as the confidence interval. Predicted values were classified based on whether the difference from the true value fell within the confidence interval. This was accomplished by selecting by attribute, all of the polygons whose values fell within 5% above or below the true value. Based on the selection, polygons were then classified as not acceptable, 0, or acceptable, 1. This helps quantify the number of block groups where kriging-based areal interpolation estimates significantly differed from true values. More importantly, this enabled the exploration of the spatial distribution of the method's accuracy.

5.3 Chapter Summary

This chapter outlined a process for statistically testing esri's kriging-based areal interpolation method. The interactive variography portion of the Geostatistical Wizard is, by far, the most time-consuming aspect of the process. This is due in part to limited tool documentation, and every data set being different. Therefore, there are not set base parameters. The statistical testing portion is very straight forward within SPSS.

Chapter 6 – Results and Analysis

For the purpose of determining the accuracy and efficacy of esri's kriging-based approach to areal interpolation, nine variables were selected for analysis within the study area (Figure 6-1).

Figure 6-1, Study Area – Kansas City, Kansas-Missouri and Topeka, Kansas, Combined Statistical Areas. Inset map showing the location of study area on the Kansas-Missouri Border.

The nine variables represent continuous, count, and ratio data. The continuous data were average household income, average household income for head of household reporting age 25 to 34, and per capita income. The count data were total population, total households, and households reporting head of household was aged 25 to 34 with income less than \$15,000. The ratio data were percent of total population which are black, percent of total population over age 65, and percent of households reported as multigenerational. Of the nine variables selected for analysis, all were used in the global comparison of the median values between the known and predicted values; however, six were selected for additional local analysis. Those six were average household income, per capita income, total population, total households, percent greater than age 65, and percent of population reporting one race: black.

6.1 Kriging Model

In order to fit the model to the data, an interactive variography process was used through the Geostatistical Wizard. Prior to determining the accuracy and efficacy of esri's new areal interpolation methodology, the validity of the kriging model used needed to be assessed. To determine the validity of the kriging model, the mean standardized error, and the root mean square standardized errors were examined (Table 1). For the model to be considered valid, the mean standardized error should be close to 0, and the root mean square standardized error (RMSE) should be close to 1. The mean standardized error is used to measure the overall validity of the kriging model, while the root mean square error is used to determine if the variance in the predictions had been over estimated (value below 1), under estimated (value above 1), or close to 1 if the predictions are valid (Bivand, Pebesma, & Gómez-Rubio, p. 225).

Table 1. Kriging model validation results are used to illustrate that the kriging models used in the interpolation of the tracts to the block groups were valid.

	Variable	Mean Standardized Error	Root Mean Square Error
Continuous	Avg HH Income	0.04	2.52
	Avg HH Income HHr 25 – 34	0.04	2.48
	Per Capita Income	0.12	5.98
Count	Total Population	- 0.05	0.26
	Total Households	0.04	3.21
	HHr 25 – 34 / Inc < \$15,000	0.002	0.91
Ratio	Percent 65+	-0.38	2.67
	Percent Black	0.79	3.03
	PCT Multigenerational HH	0.02	1.25

6.2 Continuous Data – Average Household Income

The results varied when looking at the use of esri's kriging-based areal interpolation methodology for continuous data. The first variable examined was average household income. It had a mean value of \$64,501 across 592 tracts within the study area. The true values for the block groups, which were used for statistical testing of the predicted values, had a mean value of \$64,401 across 1,739 block groups. When kriging-based

areal interpolation was used to predict the average household income from the tracts to the block groups, the mean of the predicted values was \$63,865 (Figure 6-2).

Figure 6.2. Average Household Income

The true and predicted values for the block groups were compared in SPSS to determine if there was a statistically significant difference between the median values using a

Wilcoxon signed ranks test. The results of the test showed a p-value of 0.7, indicating that the median values for the true and predicted block groups were not significantly statistically different. However, the Wilcoxon signed ranks only addresses the difference between the median values. In order to fully measure the accuracy and efficacy of esri's kriging-based areal methodology, a 90% confidence interval was calculated for each of the block group true values. The 90% confidence interval provided the 10% range considered to be an acceptable margin of error by the U.S. Census Bureau. The 10% range was then divided to illustrate the 5% above and below the true value that the U.S. Census Bureau uses as a margin of error (U.S. Census Bureau, 2012). In the case of average household income, it was found that 1,218 of the 1,739, or 70%, fell within the U.S. Census Bureau's acceptable margin of error (Figure 6-3); of the 1,218 which fell within the confidence interval, 621 were over-predicted, and 597 were under-predicted. The scatterplot (Figure 6-3) shows that the lowest residuals were seen in the areas which had lower average household incomes, with the residuals increasing as average household income increased

Average Household Income

Predicted Values Falling Within
Acceptable Confidence Interval

Figure 6-3. Block Groups Within Plus or Minus 5% of True Value

When comparing kriging-based areal interpolation estimates of average household income to their true values, the highest average household incomes remained on the outskirts of Kansas City and Topeka, as they were in both the tracts, and the true block groups. However, even though the p-value, generated by the Wilcoxon signed ranks test, indicated no significant difference between the median of the true and predicted values, a relatively localized spatial analysis revealed that 70% of the 1,739 block groups meeting the U.S. Census Bureau's margin of error (Figure 6.2, 6.3).

6.2.1 Continuous Data - Average HH Income: HHr 25 – 34

Next, average household incomes for households reporting that the head of household was aged 25 to 34 were evaluated. The average household income for households reporting that the head of household was aged 25 to 34 for all tracts within the study area was \$60,628. Similarly to the average household income, the highest values for average household income for households reporting that the head of household was aged 25 to 34 are distributed on the outskirts of both Kansas City and Topeka with the lowest values occurring both within and around the Kansas City metro area (Figure 6.4). When compared to average household income for households reporting that the head of household was aged 25 to 34 for the true block groups, the average value was found to have increased to \$62,197, while still showing a similar pattern to the tracts.

However, when the average value for the predicted block groups was examined, it was found to be closer to the average of the tracts, than the average of the true block groups. The spatial pattern within the predicted block groups did remain similar to that of both the tracts and the true block groups. However, the concentration of low values, visible in the southwestern corner of the study area, was reduced (Figure 6-4).

Average Household Income Head of Household Reported Age 25 - 34

Tracts

Predicted Block Groups

True Block Groups

Figure 6-4. Average HH Income: HHr 25-34

With the difference between the mean values of both the true, and predicted values for average household income for households reporting the head of household was aged 25 to 34 being just under \$2,000 across the study area, the assumption could be made that there would be little difference between the medians. However, when compared with a Wilcoxon signed ranks test, the medians had a p-value of < 0.01 , which indicates the values to be statistically different.

When examining the results of esri's kriging-based areal interpolation methodology for average household income for households reporting the head of household was aged 25 to 34, the predicted block groups with the highest values were still located on the outskirts of Kansas City and Topeka as they were in both the tracts, and the true block groups. groups that fell within the confidence interval still had to be.

Based on the p-value of < 0.001 it can be concluded, based on the statistically significant difference in the median, that in the case of average household income for households reporting head of household was aged 25 to 34 that kriging-based areal interpolation does not produce estimates that are sufficiently similar to known values.

6.2.2 Continuous Data – Per Capita Income

The next continuous variable tested was per capita income. As with the other two continuous variables tested at the tract level, the highest per capita income values were seen on the fringe of both Topeka and Kansas City. The mean per capita income for the tracts within the study area was \$26,049. When the tracts were compared with the true values for the block groups, which had a mean value of \$25,882, the spatial pattern was similar. However, when the true block groups were compared with the predicted block groups, which had a mean of \$25,875, the spatial pattern remained the same. However, when compared with a Wilcoxon signed ranks test, the median values for the study area were found to not be significantly statistically different with a p-value of 0.2 (Figure 6-6).

Per Capita Income

Tracts

Predicted Block Groups

True Block Groups

Figure 6-6. Per Capita Income

Of the 1,739 block groups within the study area, 1,297 block groups which represent 74.6% of the total, met the U.S. Census Bureau's acceptable margin of error; of the 1,297 which fell within the confidence interval 651 were over-predicted and 646 were under-predicted. The block groups falling within the acceptable margin of error were primarily

located within the city centers, the edge of the cities, and along the edge of the study area. The scatterplot shows that the lowest residuals were seen in the areas which had lower per capita incomes, with the residuals increasing as per capita income increased (Figure 6-7).

Per Capita Income

Predicted Values Falling Within Acceptable Confidence Interval

Figure 6-7. Block Groups Within Plus or Minus 5% of the True Value

With a statistically significant difference in the medians, and 74.6% of the predicted block groups falling within the U.S. Census Bureau's acceptable margin of error, careful consideration regarding how accurate the results need to be should be taken before using kriging-based areal interpolation.

6.3 Count Data – Total Population

When looking at the use of kriging-based areal interpolation for count data, such as population or number of households, the results were varied. The first variable that was analyzed was total population. Mean population was 3,833 across the 592 tracts within the study area. The true values for the block groups, which were used for statistical testing of the predicted values, had a mean population of 1,304 across 1,739 block groups. When esri's kriging-based areal interpolation methodology was used to predict the total population from the tract level to the block group level, the mean population was 1,419, a difference of only 115 (Figure 6-8).

Total Population

Tracts

Predicted Block Groups

True Block Groups

Figure 6-8. Total Population

The true and predicted values for the block groups were compared in SPSS with a Wilcoxon signed ranks test to determine if there was a statistically significant difference between the median values. The results of the test showed a p-value of < 0.001 , indicating that the median values for the true and predicted block groups were

statistically significantly different. However, the Wilcoxon signed ranks test only addresses the difference between the median values. Once again, in order to more comprehensively evaluate the accuracy and efficacy of esri's kriging-based areal interpolation methodology, a 90% confidence interval was calculated for the true value of the block groups. In the case of total population, it was found that only 930 of the 1,739, or 53.5%, fell within the U.S. Census Bureau's acceptable 10% margin of error (Figure 6-8); of the 930 which fell within the confidence interval, 352 were over-predicted and 579 were under predicted. The scatterplot shows that the lowest residuals were seen in the areas which had lower populations, with the residuals increasing as population increased.

Total Population

Predicted Values Falling Within
Acceptable Confidence Interval

Figure 6-8. Block Groups within +/- 5% of True Value

When examining the results of esri's kriging-based areal interpolation of average household income for the predicted block groups, the highest values were still located on the outskirts of Kansas City and Topeka as they were in both the tracts, and the true block groups (Figure 6-7). However, the results of the Wilcoxon signed ranks test, $p < 0.001$, indicated that the difference between the median of the true and predicted values was significantly statistically different across the study area, and with only 53.5% (Figure 6-8) of the 1,739 block groups meeting the U.S. Census Bureau's 10% margin of error kriging-based areal interpolation may not be an appropriate choice.

6.3.1 Count Data – Total Households

Next, total households were explored. The mean number of households over the tracts was 1,510. The highest household concentrations were distributed on the outskirts of both Kansas City and Topeka with the lowest values occurring both within, and around the Kansas City metro area. The total households for the true block groups had a mean of 514. The true block groups had a similar spatial pattern to the tracts. However, the true block groups also had high concentrations of low values on the edge of the study area.

The spatial pattern within the predicted block groups was not similar to the true block groups. First, the mean number of households per block group increased to 599. In addition, the high concentrations of households seen in the true block groups were concentrated at the edge of the study area, with the lowest number of households being

seen in the cities (Figure 6-9).

Total Households

Figure 6-9. Total Households

Upon comparison of the medians through the use of the Wilcoxon signed ranks test, it was found that there was a statistically significant difference between the true and predicted block groups with a p-value of < 0.001 . When predicted block groups were

further examined as to how many fall within the U.S. Census Bureau's 10% margin of error, it was determined that, of the 1,739 block groups, only 640, or 36.8% were within the acceptable range (Figure 6-10); of the 640 which fell within the confidence interval, 353 were over-predicted and 288 were under-predicted. The scatterplot shows that the lowest residuals were seen in the areas which had least number of households.

Total Households

Predicted Values Falling Within
Acceptable Confidence Interval

Figure 6-10. Block Groups within +/- 5% of True Value

Based on both the p-value of < 0.001 , and that only 36.8% fell within the 10% confidence interval, kriging-based areal interpolation would not be an appropriate choice, for the reaggregation of count data in this instance.

6.3.2 Count Data – HHr 25-34/Income <\$15,000

The third count variable tested was the total number of households where the head of the household was reported to be aged between 25 to 34 and had income less than \$15,000. At the tract level, the mean number of households per block group which met these criteria was 27. The tracts which contained the highest number of households meeting this criteria were found in the urban core. The tracts with the lowest number of households meeting this criteria were found on the edge of Topeka and Kansas City. Low values were seen throughout the study area. The highest concentration of households reporting the head of household was aged between 25 to 34 were found in the urban core, and the mean number of households across the study area was nine. When predicted values for the block groups were examined, there were still some areal units within the urban core that exhibited the highest concentrations of low income households headed by people aged 25 to 34. However, the majority of the edge of the study area had high values, which were not identified in the true block groups (Figure 6-11).

HHr 25-34/Inc <\$15000

Tracts

- 0 - 17
- 18 - 37
- 38 - 63
- 64 - 103
- 104 - 187
- Kansas City
- Topeka

Predicted Block Groups

- 0 - 9
- 10 - 25
- 26 - 61
- 62 - 139
- 140 - 282
- Kansas City
- Topeka

True Block Groups

- 0 - 6
- 7 - 15
- 16 - 29
- 30 - 59
- 60 - 133
- Kansas City
- Topeka

Figure 6-11. HHr 25-34/Income < \$15,000

When the true and predicted block groups were compared using a Wilcoxon signed ranks test, a statistically significant difference in the median value was not found at the 95% level of confidence (p-value of 0.06). However, we can reject the null hypothesis at a 90% level of confidence. Once it was determined that there was not a statistically

significant difference in the median value between the the true and predicted block groups, the highest concentration of areal units that were acceptable were found in the urban cores.

Based on the p-value of 0.06 it can be concluded that kriging-based areal interpolation is not appropriate for this count data.

6.4 Ratio Data – Percent of Population 65 and Older

The final portion of esri's kriging-based areal interpolation methodology which had to be tested was for ratio data. The first ratio data variable tested was the percent of population aged 65 or older. While the mean percentage per tract was 12.2% and there were some tracts within the urban core with values greater than 22%, the majority of the tracts exhibiting high percentages were seen outside of the cities and on the edge of the study area. Upon examining the percent greater than age 65 in the true block groups, the true block groups had a mean of 12.7%. At the smaller areal unit provided by the block groups, a high concentration of block groups exhibiting high percentages of people aged 65 and older could be seen on the edge of the Kansas City metro area. However, the lowest values were still seen within the urban core. When the mean of the predicted block groups was examined, they were found to have a mean value of 12.2%, 0.5 percentage points lower than that of the true block groups. However, the high values, seen along the edge of the study area, became more prevalent throughout the entire study area. The areas with low percentages seen in the urban core moved towards the edge of Kansas City, and were replaced by areas of high percentage (Figure 6-13).

Percent of Population 65+

Tracts

Predicted Block Groups

True Block Groups

Figure 6-13. Percent of Population Aged 65+

When the medians of the true and predicted block groups were compared using a Wilcoxon signed ranks test, no statistically significant difference was found with a p-value of 0.54. When the predicted block groups were examined to determine the percentage that falling within the U.S. Census Bureau's acceptable 10% margin of error,

only 1,443, or 82.9%, of the 1,739 block groups were found to be in the acceptable range (Figure 6-14); of the 1,443 which fell within the confidence interval, 737 were over-predicted and 706 were under-predicted. The scatterplot shows that the lowest residuals were seen in the areas which had lower percentage of the population over age 65, with the residuals increasing as percentage over age 65 increased.

Percent of Population 65+

Predicted Values Falling Within
Acceptable Confidence Interval

Figure 6-14. Block Groups within +/- 5% of True Value

The results of a Wilcoxon signed ranks test indicated that there was not a statistically significant difference in the median value for the percent of the population aged older than 65. With 82.9% of the predicted block groups falling within the U.S. Census Bureau's margin of error, one must proceed with caution when using kriging-based areal interpolation for this ratio data depending on the level of accuracy needed.

6.4.1 Ratio Data – Population-Single Race: Black Percentage

Next, the analysis addressed the percent of the population which reported black as their only race. The mean percentage per tract was 16%, and the majority of the tracts exhibiting high percentages were within the urban core of the cities. The lowest values fell on the edge of the study area and the urban core. Upon examining the percent black in the true block groups, they were found to have a mean of 14.5%. At the smaller areal unit of the block groups, a high concentration of block groups exhibiting high percentages of the population self-identifying as black could be seen in the urban core of both the Kansas City and Topeka metro areas. However, the lowest values were still located on the edge of both cities and the study area. When the mean of the predicted block groups was examined, they were found to have a mean value of 13.6%, which is 0.9 percentage points lower than that of the true block groups. However, the high values seen in the urban core spread out throughout both cities. The low values, which extended from the edge of the city to the edge of the study area, did not change (Figure 6-15).

Percent Black

Tracts

Predicted Block Groups

True Block Groups

Figure 6-15. Percent Black

When the percent black medians of the true and predicted block groups were compared, a Wilcoxon signed ranks test developed a p-value of < 0.01 . Therefore, it was determined that there was a statistically significant difference between block groups. When the predicted block groups were examined as to the percentage that fell within the

U.S. Census Bureau's acceptable 10% margin of error, 1,417, or 81.5%, of the 1,739 block groups were found to be in the acceptable range (Figure 6-16) of the 1,417 which fell within the confidence interval, 434 were over-predicted and 983 were under-predicted. The scatterplot shows that the lowest residuals were seen in the areas which had lower percentages of the population reporting that they were black, with the residuals increasing as the percentage increased.

Percent Black

Predicted Values Falling Within
Acceptable Confidence Interval

Figure 6-16. Block Groups within +/- 5% of True Value

While the results of a Wilcoxon signed ranks test indicated that there was a statistically significant difference in the median value for the percent of the population

reporting that they were black, and 81.5% of the predicted block groups falling within the U.S. Census Bureau's margin of error. Caution should be used when considering the use of kriging-based areal interpolation for this ratio data depending on the level of accuracy required.

6.4.2 Ratio Data – Percent Multigenerational Households

The final variable tested was percent of households reporting as multigenerational. The mean percentage per tract was 3.5%. The majority of the tracts with high percentages were seen within the cities and on the edge of the study area. The percent of multigenerational households within the true block groups had a mean of 3.6%. At the smaller areal unit of the block groups, high concentrations of block groups with high percentages of multigenerational households were identified in the Kansas City metro area. However, values falling closest to the mean could be seen across the entire study area. When the predicted block groups were examined, the mean value was calculated to be 3.5%, which is 0.1 percentage points lower than that of the true block groups. However, the high values, which were scattered in the urban core, became more concentrated. The areas with low percentages seen in the urban core moved towards the edge of Kansas City, and were replaced by areas which fell closer to the mean (Figure 6-17).

Multigeneration HHs Per

Tracts

Predicted Block Groups

True Block Groups

Figure 6-17, Percent Multigenerational Households

When the medians of the true and predicted block groups were compared with a Wilcoxon signed ranks test, which returned a p-value of 0.77, it was determined that there was not a statistically significant difference between them.

The Wilcoxon signed ranks test results indicate a lack of statistically significant difference in the median value for the percent of households reported as multigenerational. Therefore, it can be concluded, based on the medians that the use of kriging-based areal interpolation may be an appropriate choice for this ratio data.

6.5 Residuals

The residuals, or the predicted minus the known values, were explored to determine if there was a distinct spatial pattern of the high values or the low values. For the nine variables used in the testing of esri's kriging-based areal interpolation methodology there was not a distinct pattern showing concentrations of high or low residuals next to major roads. Instead, the pattern was random, variable to variable, with a mixture of high and low residuals dispersed throughout the study area.

6.6 Local Analysis

To determine where esri's kriging-based areal interpolation was effective, further local analyses were performed on average household income, per capita income, total population, total households, percent of the population over age 65, and percent of population reporting one race which was black. The local analyses were used to determine if there were any spatial patterns in the block groups which fell outside of the margin of error used by the U.S. Census Bureau for American Community Survey data.

For the first local analysis, block groups were classified based on their area in square miles using a Jenks / natural breaks classification. The Jenks classification was used to find the natural breaks in the area in square miles, and then to group the block groups based on those breaks. The data breaks were small, which included block groups between 0 and 26 square miles, medium, block groups between 26.1 and 102.5 square miles, and large, or block groups between 102.5 and 262.2 square miles. There were 1,599 block groups in the small class, 126 in the medium class and 14 in the large class. Once the block groups were broken into separate classes the number of block groups which fell into each class which fell within the American Community Survey's margin of error were identified (Table 2).

Table 2. Number of Block Groups Which Fell Within the American Community Survey's Margin of Error by Area

	Variable	Small	% Small	Medium	% Medium	Large	% Large
Continuous	Average Household Income	1,100	68.8%	105	83.3%	13	92.9%
	Per Capita Income	1,180	73.8%	105	83.3%	12	85.7%
Count	Total Population	849	53.1%	73	57.9%	9	64.3%
	Total Households	634	39.6%	7	5.6%	0	0.0%
Ratio	Percent Greater than Age 65	1,313	82.1%	117	92.9%	13	92.9%
	Percent Black	1,300	81.3%	107	84.9%	10	71.4%

Based on the results of the analyses, block groups which had small land area were best predicted with ratio data; percent greater than age 65 had 1,313, or 82.1% matched, and percent black had 1,300, or 81.3% matched. The next best results were with continuous data; per capita income had 1,180, or 73.8% matched, and average household income had 1,100, or 68.8% matched. Count data provided the least accurate results; total population had 849, or 53.1% matched, and total households had 634, or 39.6% matched.

Block groups which had medium land area were best predicted with ratio data; percent greater than age 65 had 117, or 92.9% matched, and percent black had 107, or 84.9% matched. The next best results were with continuous data; both per capita income and average household income had 105, or 83.3% matched. Count data provided the least accurate results; total population had 73, or 57.9% matched, and total households had 7, or 5.6% matched.

Block groups which had large land area were best predicted with continuous data; average household income had 13, or 92.9% matched, and per capita income had 12, or 85.7% matched. The next best results were with ratio data; percent greater than age 65 had 13, or 92.9% matched, and percent black had 10, or 71.4% matched. Count data provided the least accurate results; total population had 9, or 64.3% matched, and total households had 0 matched.

The next local analysis which was performed was to identify whether the kriging-based areal interpolation method was influenced by an edge effect, or in other words are predictions on the edge of the study area more inaccurate. To accomplish this task block groups within the study area were divided into two groups, those which touched the edge of the study area (including the island of Douglas County in the middle of the study area), and all others. The percentage of block groups falling within the American Community Survey margin of error were identified from each group (Table 3). Of the 1,739 block groups within the study area, 80 were found to be touching the edge of the study area boundary and 1,659 were not touching.

Table 3. Block Groups Within American Community Survey Margin of Error Based on Whether They are on the Edge or Not on the Edge of the Study Area

	Variable	Edge		Non-edge	
		Number	% Within Margin of Error	Number	% Within Margin of Error
Continuous	Average Household Income	67	83.8%	1,151	69.4%
	Per Capita Income	66	82.5%	1,231	74.2%
Count	Total Population	49	61.3%	882	53.2%
	Total Households	1	1.3%	640	38.6%
Ratio	Percent Greater than Age 65	65	81.3%	1,367	82.4%
	Percent Black	63	78.8%	1,354	81.6%

Based on the analysis, the results varied based on data type. For continuous data, the highest number of matched block groups were on the edge of the study area; average household income had 67, or 83.8% matched, and per capita income had 66, or 82.5% matched. With the non-edge block groups, per capita income had 1,231, or 74.2% matched, and average household income had 1,151, or 69.4% matched. For count data, total population had a greater number matched on the edge of the study area, 49, or 61.3%, than it did on block groups identified as non-edge, 882, or 53.2% matched. Total households saw a greater number matched on block groups identified as non-edge, 640, or 38.6% matched, compared to 1, or 1.3% of block groups which touched the edge falling within the margin of error.

The final local analysis performed was to identify block groups falling within the margin of error based on whether they were in an urban or rural area. To determine what was urban and what was rural, the U.S. Census Bureau's designation of urban was used. There were 1,349 block groups within urban areas and 390 in rural areas. This local analysis was used to determine if kriging-based areal interpolation was more accurate in highly density urban areas or in low density rural areas (Table 4).

Table 4. Percentage of Block Groups Falling Within the Margin of Error based on Urban or Rural Designation

		Urban		Rural	
	Variable	Number	% Within Margin of Error	Number	% Within Margin of Error
Continuous	Average Household Income	937	69.5%	281	72.1%
	Per Capita Income	1,004	74.4%	293	75.1%
Count	Total Population	686	50.9%	245	62.8%
	Total Households	582	43.1%	59	15.1%
Ratio	Percent Greater than Age 65	1,112	82.4%	331	84.9%
	Percent Black	1,086	80.5%	331	84.9%

Table 4 shows little difference in the areal interpolation results between urban and rural block groups for the continuous and ratio variables. However, noticeable differences can be seen in the count variables total population and total households.

Chapter 7 – Conclusions and Future Work

The primary goal of this project was to provide the client with an assessment of the accuracy and efficacy of esri's kriging-based areal interpolation methodology when applied to census-based socioeconomic data. Through the analysis of data aggregated through the use of the areal interpolation portion of the Geostatistical Wizard, the comparison of the median values of the true and predicted block groups with a Wilcoxon signed ranks test, and the identification of the percentage of predicted block groups which fell within the U.S. Census Bureau's acceptable margin of error, this project has met the client's request.

In the case of this project, it appears, based on this analysis, that the accuracy and efficacy of esri's kriging-based areal interpolation methodology for the reaggregation of census-based socioeconomic data varies (Table 3).

Of the nine variables tested, five; average household income, per capita income, households reporting the head of household was aged 25 to 34 with income less than \$15,000, percent over age 65, and percent multigenerational households had p-values indicating that the median values of the estimated and true were not statistically significantly different at the block group level. While this demonstrates kriging-based areal interpolations ability to maintain the median value of a dataset, it alone cannot be used to measure the accuracy and efficacy of the method. However, based on the mean standard error for each variable it is clear that while the kriging model used was valid, in the majority of the variables subjected to a local analysis, the model overestimated the variability of the predictions. Three of the variables, total population, head of household reported aged 25 to 34 with income less than \$15,000, and percent multigenerational households, had both a mean standardized error close to 0.0, and an root mean square error close to 1.0, which was indicative of both a good kriging model and low variability in the predictions. However, even with a good kriging model, and low variability, the highest number of block groups predicted within the U.S. Census Bureau's 10% margin of error was 82.9%. Therefore, based on the statistical analysis of the predictions, it can be concluded that the results of esri's kriging-based areal interpolation methodology vary for the reaggregation of census-based socioeconomic data.

Table 5. Table of Results for Kriging model (Mean Standard Error and Root Mean Square Error) and Statistical Comparison of True and Predicted Block Groups (p-value, Number of Block Groups Within Confidence Interval, and Percent Within +/- 5% of Confidence Interval). * Derived Variables not Included in the analysis Because Margin of Error not Available.

	Variable	Mean Standard Error	Root Mean Square Error	p-value	Number of Block Groups Within Confidence Interval	Percent Within +/- 5% of Confidence Interval
Continuous	Avg HH Income	0.03	2.52	0.703	1,218	70%
	Avg HH Income HHr 25 - 34	0.04	2.48	0.000	*	*
	Per Capita Income	0.12	5.98	0.182	1,297	74.6
Count	Total Population	-0.05	0.26	0.000	930	53.5%
	Total Households	0.04	3.21	0.000	640	36.8%
	HHr 25 – 34 / Inc < \$15,000	0.002	0.91	0.061	*	*
Ratio	Percent 65+	-0.38	2.67	0.535	1,443	82.9%
	Percent Black	0.79	3.03	0.000	1,417	81.5%
	PCT Multigenerational HH	0.02	1.25	0.774	*	*

Further research could be done on how best to fit the kriging model to the data that is being input. The interactive variography portion of areal interpolation through the Geostatistical Wizard was, by far, the most time consuming portion of the analysis. While esri's help documentation and tutorial were helpful, the only advice they gave on fitting the kriging model to the data was that all but one of the averages needed to fall within the confidence intervals. Clear advice on the determination of the number of lags, how to determine the correct lag size, and what types of data the models available should be used with would have been very beneficial. In addition, it would be interesting to see a comparison of how the results from Empirical Bayesian Kriging would have compared to those of kriging-based areal interpolation. The advantage of Empirical Bayesian Kriging is that it allows for moderately nonstationary data to be used, whereas the kriging-based technique implemented as part of the Geostatistical Wizard makes an assumption of

stationarity. Also, since Empirical Bayesian Kriging is implemented as a stand-alone tool in ArcGIS, it is able to take advantage of the ability to perform 64-bit background geoprocessing. This is something that could be done quickly through using the data and results from this analysis, and comparing them with the results from empirical Bayesian kriging.

While esri's approach to areal interpolation, through a kriging-based methodology, is a novel approach, its' assumption of stationarity, and the fact that it uses one kriging model for an entire study area are limiting factors to its' accuracy and efficacy for use with census-based socioeconomic data. Although, in some cases, there was not a statistically significant difference between the true and predicted values for the block groups, when the six of the variables were subjected to the acceptable 10% margin of error, which is used by the U.S. Census Bureau, the highest number of block groups matched was 82.9%. Caution should be used when considering esri's kriging-based areal interpolation methodology, available through the Geostatistical Wizard, depending on the level of accuracy required for analysis.

Works Cited

- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (Second ed.). Springer.
- ESRI. (2012). *ArcGIS Resources*. Retrieved October 8, 2012, from ArcGIS Help 10.1: http://resources.arcgis.com/en/help/main/10.1/index.html#/What_is_areal_interpolation/0031000000q8000000/
- Goodchild, M. F., & Lam, N. S.-N. (1980). Areal Interpolation: A Variant of the Traditional Spatial Problem. *Geo-Processing*(1), 297-312.
- Hart, J. F. (1991, January). The Perimetropolitan Bow Wave. *Geographical Review*, 81(1), 33-51.
- Krivoruchko, K., Gribov, A., & Krause, E. (2011). Multivariate Areal Interpolation for Continuous and Count Data. *Procedia Environmental Sciences*, 3, 14-19.
- Tobler, W. R. (1979, September). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*, 74(367), 519-530.
- U.S. Census Bureau. (2012). *History*. Retrieved October 2, 2012, from US Census Bureau: <http://www.census.gov>
- Xie, Y. (1995). The Overlaid Network Algorithms for Areal Interpolation Problem. *Computer, Environment, and Urban Systems*, 19(4), 287-306.

Appendix A. Kriging Parameters

	Variable	Lag Size	Number of Lags	Model Type	Lattice Spacing
Continuous	Avg HH Income	11,703.66	5	Spherical	5,022.72
	Avg HH Income HHr 25 - 34	10,703.66	5	Spherical	5,022.72
	Per Capita Income	10,703.66	5	Spherical	5,022.72
Count	Total Population	1,425	80	K-Bessel	5,022.72
	Total Households	46,703.66	5	Spherical	5,022.72
	HHr 25 – 34 / Inc < \$15,000	46,703.66	5	Spherical	5,022.72
Ratio	Percent 65+	750	12	Spherical	5,022.72
	Percent Black	500	12	Spherical	5,022.72
	PCT Multigenerational HH	65,703.66	5	Spherical	5,022.72