
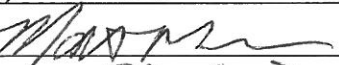
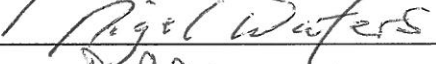
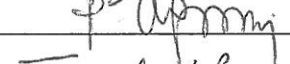
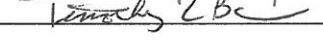



EVALUATING THE ERRORS ASSOCIATED WITH ZIP CODE
POLYGON WHEN EMPLOYED FOR SPATIAL ANALYSES

by

Tunaggina Subrina Khan
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geography and GeoInformation Science

Committee:

	Dr. Kevin M. Curtin, Thesis Director
	Dr. Matt Rice, Committee Member
	Dr. Nigel Waters, Committee Member
	Dr. Peggy Agouris, Department Chairperson
	Dr. Timothy L. Born, Associate Dean for Student and Academic Affairs, College of Science
	Dr. Vikas Chandhoke, Dean, College of Science

Date: 08/24/2012

Fall Semester 2012
George Mason University
Fairfax, VA

Evaluating the Errors Associated with Zip Code Polygon When Employed for Spatial Analyses

A thesis submitted in partial fulfillment of the requirements for the degree of Master of
Science at George Mason University

by

Tunaggina Subrina Khan
Bachelor of Science
University of Dhaka, Bangladesh, 2006

Director: Kevin M. Curtin, Professor
Department of Geography and GeoInformation Science

Fall Semester 2012
George Mason University
Fairfax, VA



This work is licensed under a [creative commons attribution-noncommercial 3.0 unported license](https://creativecommons.org/licenses/by-nc/3.0/).

DEDICATION

This is dedicated to my husband Arif, my parents and to all of my friends who always give me moral and intellectual support in every step of my life.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this thesis.

My utmost gratitude to my Thesis Director Dr. Kevin M. Curtin who gave me the idea of this thesis work, the primary data and guided me with his valuable comments, suggestions, patience and steadfast encouragement to complete this thesis. Dr. Kevin M. Curtin has been my inspiration as I hurdle all the obstacles in the completion this research work.

I would also like to thank Dr. Nigel Waters for his unfailing support and advice as my Thesis committee member. His comments on every detail of my thesis helped me to improve the thesis and inspired me to think in depth.

Dr. Matt Rice, my Thesis committee member, for his supervision and intellectual guidance and for giving me important references for this thesis work.

Chris Oxendine, PhD student in the Earth Systems and GeoInformation Science at George Mason University, for sharing his valuable insight and helping me with proof reading.

Thanks go out to the Fenwick Library for providing a well organized and gigantic reference repository. The universities who provided me the reference I needed through the inter-library loan system.

Finally, my parents and my husband for their support and encouragement every step of the way.

TABLE OF CONTENTS

	Page
List of Tables	viii
List of Figures	ix
List of Equations	xi
Abstract	xii
Section 1	
Introduction	1
Section 2	
Background and Literature review	5
Section 3	
Data: Multiple ZIP Code Polygon Representations	11
Section 4	
Descriptive Comparisons: Exploratory Comparative Analysis	16
4.1 Methods of Comparative Analysis	16
4.2 Results of Comparative Analysis	19
Section 5	
The Influence of Polygon Representation on Spatial Statistical Results	28
5.1 Methods for spatial autocorrelation and cluster analysis	28
5.2 Results	32
Section 6	
The Influence of Polygon Representation on Network Analytical Results	35
6.1 Network-analytic Literature Survey	36
6.2 Study area and data	38
6.3 Methods	40
6.4 Results of network comparisons	42
6.4.1 Comparison of network distances within and between ZIP Code maps	42
6.4.2 Deviations in network distance from the USPS ZIP Code map	49
Section 7	
Segregation of Hispanic population	53

7.1 Literature survey	53
7.2 Data and Study area.....	56
7.3 Methods.....	57
7.3.1 Segregation index	57
7.3.2 Detailed implementation of the segregation index	60
7.4 Results	62
Section 8	
Ranking of ZIP Codes.....	69
8.1 Literature survey	71
8.2 Study area and data	74
8.3 Method	75
8.3.1 Estimating market value of houses	75
8.3.2 Estimating ranking of ZIP Codes	76
8.4 Results	79
Section 9	
Accessibility to Emergency room in Fairfax County	95
9.1 Literature review:	97
9.2 Study area and data	100
9.3 Method	101
9.4 Results	109
Section 10	
Statistical Similarity.....	112
10.1 Data	112
10.2 Methods.....	112
10.2.1 Test for normality	112
10.2.2 Choice of parametric or non-parametric test for independence	114
10.2.3 Test for independence of ZIP Code maps	115
10.2.4 Linear regression	119
10.3 Results	121
10.3.1 Tests of independence	121
10.3.2 Linear regression	125
Section 11	
Discussion	129

Appendix.....	139
Index	143
References.....	144

LIST OF TABLES

Table	Page
Table 1: Differences (in km ²) in ZIP Code areas within datasets compared with the USPS polygon dataset	21
Table 2: Areas (km ²) of the ZIP Codes within a dataset that do not exist within the USPS dataset (using Symmetric Difference).	24
Table 3: Percentages of areas of the ZIP Codes within datasets that do not exist within the USPS dataset	25
Table 4: Deviation of centroid locations of ZIP Codes in different polygon datasets compared to the USPS dataset (in km).	26
Table 5: Percent change in segregation level within the datasets compared to the USPS dataset	65
Table 6: Pair-wise comparison among datasets regarding the level of segregation of Hispanic population.	66
Table 7: Percentages of the ZIP Codes that switched to a different quantile class across datasets.	67
Table 8: Pair-wise comparison of the ZIP Codes that changed ranking classes across the datasets	87
Table 9: Model Summary and Parameter Estimates for linear regression.....	120
Table 10: Model Summary and Parameter Estimates for linear regression.....	120
Table 11: Outliers in a simple linear regression analysis on the USPS and the Census ZIP Code datasets.	127

LIST OF FIGURES

Figure	Page
Figure 1: Study area of Northern Virginia	12
Figure 2: Spatial extent of the datasets	14
Figure 3: Flow chart of the methodology	17
Figure 4: Symmetric difference (C) from USPS (A) and Clarke (B) ZIP Code 20135....	19
Figure 5: Deviation of size, shape and position of ZIP Codes in different datasets.	20
Figure 6: Total areas (km ²) of the corresponding ZIP Codes that do not match in the spatial extent or do not overlap within the USPS and an individual dataset measured by Symmetric Difference.....	23
Figure 7: Calculating population for USPS ZIP Code 20176 based on intersected area with Census ZIP Codes	30
Figure 8: Methodologies for determining populaiton and spatial pattern.....	31
Figure 9: Results of Local Moran's I for population within ZIP Codes.....	33
Figure 10: Study area of Fairfax County	39
Figure 11: Unmatched ZIP Code that are excluded from network analysis	40
Figure 12: ZIP Code 22124 in different datasets has the same route to the Inova Fair Oak Hospital.	43
Figure 13: The route between ZIP Code 22066 and Reston Surgery Center is 12 km long in the Census (route 1) and 12.5 km long in the USPS, Sammamish and Fairfax County data (route 2).....	44
Figure 14: The difference in network distance between ZIP Code 20120 and Reston Surgery Center in Sammamish and USPS datasets	46
Figure 15: The routes that vary by a km or more across datasets in centroid-hospital network distance calculation.....	47
Figure 16: Comparison of network distance between ZIP Code maps.....	48
Figure 17: Comparison of network distance within ZIP Code maps.....	49
Figure 18: Number of routes that change by different percentages within datasets in centroid-hospital network analysis	51
Figure 19: Study area of Fairfax County	57
Figure 20: Implementation of Areal Interpolation method and isolation index of segregation.	61
Figure 21: Segregation of Hispanic population in Fairfax County at ZIP Code level for Census, Fairfax County, Sammamish and USPS datasets.....	63
Figure 22: Detail methodology of ZIP Code ranking based on the ranking of available schools within ZIP Code polygon boundary	79
Figure 23: Ranking of ZIP Codes based on school rankings available within ZIP Codes	81

Figure 24: ZIP Code 22033 is ranked as ‘high’ in the Census dataset but ‘medium high’ in Sammamish, Fairfax County and USPS datasets	82
Figure 25: Influence of school ranking on ZIP Code 20041(in USPS) or 22091 (in Fairfax County) and ZIP Code 20171 in different data sources.....	84
Figure 26: Percent of the total ZIP Codes within the USPS dataset that alter quantile classes across datasets.....	86
Figure 27: Average Property prices (in Dollar) within ZIP Code polygon boundary in Different data sources.	88
Figure 28: Overlay of ZIP Code 22091 and 20041 with census tracts	89
Figure 29: Inconsistency of ZIP Code area creating variable property price	91
Figure 31: Procedures of implementing 2SFCA in ArcGIS.	108
Figure 32: Comparison of accessibility to emergency room across datasets.....	110
Figure 33: Normality plots of the Census dataset.....	115
Figure 34: Curve fit for the Fairfax County dataset.....	121
Figure 35: Results from pair-wise non-parametric tests (selected by the software) on the datasets.	122
Figure 36: Results from the Friedman non-parametric tests on datasets	123
Figure 37: Results from the Wilcoxon Signed Ranks non-parametric tests on datasets.	124
Figure 38: Significance of regression model on the datasets.....	126

LIST OF EQUATIONS

Equation	Page
$A \Delta B = A \setminus B \cup B \setminus A$	Equation 1..... 18
$A \Delta B = A \cup B \setminus B \cap A$	Equation 2..... 18
$R_j = \sum_{i=1}^n \frac{H_i}{H_{total}} \times \frac{H}{T_i}$	Equation 3..... 59
$V_j = \sum_{i=1}^n \frac{A_i}{A_{total}} \times V_i$	Equation 4... ..88
$R_j = \sum_{i=1}^n \frac{A_i}{A_{total}} \times R_i$	Equation 5.....77
$R_j = \frac{S_j}{\sum_{k \in \{d_{kj} \leq d_o\}} P_k}$	Equation 6..... 104
$A_i^F = \sum_{j \in \{d_{ij} \leq d_o\}} R_j = \sum_{j \in \{d_{ij} \leq d_o\}} \left(\frac{S_j}{\sum_{k \in \{d_{kj} \leq d_o\}} P_k} \right)$	Equation 7 104
$d_j = x_{1j} - x_{2j}$	Equation 8..... 117
$\bar{d} = \frac{\sum_{j=1}^n dj}{n}$	Equation 9..... 118
$S_d = \sqrt{\frac{\sum_{i=1}^n (dj - \bar{d})^2}{n-1}}$	Equation 10.....118

ABSTRACT

EVALUATING THE ERRORS ASSOCIATED WITH ZIP CODE POLYGON WHEN EMPLOYED FOR SPATIAL ANALYSES

Tunaggina Subrina Khan, MS

George Mason University, 2012

Thesis Director: Dr. Kevin M. Curtin

ZIP Codes have traditionally been represented cartographically as polygon features. Polygon-based representations of ZIP Codes are derived from point features employing interpolation techniques. There has been an increasing understanding in the literature that these interpolations can introduce error into the results of spatial analyses that employ ZIP Code polygon representations. This research uses multiple ZIP Code datasets which have been collected from eleven different sources. First the study seeks to determine if the ZIP Code representations are identical as they are purported to be. Comparisons are made based on several spatial characteristics; specifically area, level of overlap, and centroid location. It has been determined that frequently there are considerable, in some cases statistically significant, differences in these measurements across polygon representations. The consequences of these differences for the results of spatial analyses that employ ZIP Code polygons are explored through typical applications. These applications include a test for spatial autocorrelation, an examination

of network distances, a ranking of ZIP Codes by school quality and availability, an examination of segregation level, and a description of accessibility to emergency rooms. Finally, a statistical comparison of ZIP Code polygon datasets is made, demonstrating that the representations are not identical and in many cases are not even statistically similar. Conclusions regarding best practices for ZIP Code polygon use and suggestions for future research are provided.

SECTION 1 INTRODUCTION

A Zoning Improvement Plan (ZIP) Code is a number assigned to every address in the United States. ZIP Codes are maintained and assigned solely at the discretion of the United States Postal Service (USPS). They are free to change, edit, combine, split, or otherwise alter ZIP Codes whenever, and for whatever purpose, they need in the service of the efficient delivery of mail. The USPS is not obligated to report changes to ZIP Codes in any formal way. The ZIP Codes are assigned to address points by the USPS, and these points are then assigned to a particular post office and route for delivery.

The USPS does not maintain or release the geographic boundaries of the ZIP Codes, though some USPS facilities create their own ZIP Code area maps for public interest. However, different vendors and GIS users have created ZIP Code maps by interpolating polygon boundaries between occurrences of ZIP Codes attributes as assigned to either point or line features (US Bureau of the Census 2001).

The fact that many different representations of ZIP Code polygons have been generated – none of which are from the authoritative source (the USPS) – leads to critical potential problems for those conducting scientific spatial analysis. For example, if different ZIP Code polygon representations have different areas, then any calculation that employs the area of the polygon will generate different results based on the choice of the input ZIP Code dataset. The same is true for other measures based on the spatial

definitions of the ZIP Codes, such as centroid locations, perimeters, or distances between ZIP Codes and other locations. Perhaps most importantly, since the great majority of the vendors and users who generate their own ZIP Code polygons maintain no information regarding the interpolation methods they employ to create the polygon maps or changes they make over time, there may be no way to reliably repeat the scientific experiments using ZIP Code polygons. This is a fundamental requirement of the scientific method, and is confounded by the current situation with regard to ZIP Code polygon databases.

This thesis provides a research overview designed to quantify the extent of the differences in the spatial components of ZIP Code polygons across a set of representations.

Section 1 gives an overview of the problem and outlines important issues with ZIP Codes in spatial analyses.

Section 2 discusses the use of ZIP Codes for spatial analyses in the literature and section 3 discusses the sources of data that have been used in this thesis in detail.

Section 4 demonstrates the disparity among ZIP Code polygon representations through measures of the areas of the ZIP Codes, the centroid locations of the polygons, and the level of overlap with the USPS polygon data as a reference. The results suggest that area values can differ by as much as 325 square km, and inter-ZIP Code distances can vary by more than 200 km. These initial results suggest that the use of the spatial components of ZIP Code polygons for spatial analytic research raises questions as to the validity of the results of the analyses.

Section 5 employs tests of spatial autocorrelation to reveal whether the ZIP Codes from different data sources can create different outcomes due to variations in ZIP Code areas and shapes within datasets.

Section 6 discusses the problems that can arise in network analysis using the centroids of ZIP Code polygons for measuring distances between the centroids and other points of interest. The shortest route distance is used to determine network distances between ZIP Code centroids and hospitals using ZIP Code polygon maps from different data sources. Network distances are measured and compared within and between the ZIP Code representations. Deviations in network distance from the USPS ZIP Code representation are also examined. Results indicate that network distances between a ZIP Code and a location of interest can be considerably different in ZIP Code maps from different data sources.

Section 7 demonstrates how the level of segregation of the Hispanic population within Fairfax County varies when using different ZIP Code polygon maps. An isolation index is employed to determine the segregation at the ZIP Code level. An areal interpolation method is used to estimate the population at the ZIP Code level from the population data of census 2010. Comparisons of the results from the segregation measurement across datasets indicate that the segregation level can vary based on ZIP Code representations. These results indicate that the use of different ZIP Code polygons for measuring segregation can create different outcomes across different data sources.

Section 8 presents a hypothetical situation in which ZIP Codes are ranked for providing real estate information. The ZIP Codes are ranked according to the influence of

the quality of schools available within the ZIP Code area. The rankings of ZIP Codes are compared across a set of ZIP Code maps and it is concluded that these rankings differ when ZIP Code representations from different data sources are used for the analysis.

Section 9 provides an example of using ZIP Codes in accessibility measurement from ZIP Code centroids to emergency rooms using two step floating catchment area methods. Results of the level of accessibility using different datasets also demonstrate variable results.

Perhaps most importantly, section 10 presents a statistical comparison of the areas of the ZIP Codes across the datasets to determine whether these datasets are significantly dissimilar. The results show that in 15% to 20% of the pair-wise comparisons, the ZIP Code datasets (which are largely presumed to be identical) cannot be shown to be statistically similar. That is, in many cases different ZIP Code polygon representation for the same area are, for the purposes of spatial-analytic research, totally different spatial domains.

SECTION 2

BACKGROUND AND LITERATURE REVIEW

A ZIP Code is a 5-digit number where each number identifies geographic areas with increasing specificity. The first digit is assigned to address points in a broad geographic region of the U.S. The second and third digits represent Sectional Center Facilities, and the fourth and fifth digits represent Post Offices or postal zones from which mail is delivered to address points. Address points may represent individual homes, or a group of apartments, or alternatively an individual high-volume receiver of mail (Roberts 2007). For even greater spatial detail, the USPS uses a mail sorting machine to determine the correct ZIP+5 Code from the address along with a specific delivery point. ZIP Codes were first implemented in 1963 as a part of the national zoning plan and were developed for the purpose of mail delivery. It is not remarkable that in the early development of ZIP Codes it was not imagined that the codes and their spatial representations would become a common observation platform for spatial analysis specifically, and for scientific research more generally. Therefore, the potential problems that may arise from utilizing these ZIP Codes in spatial analysis could not have been foreseen.

In terms of the spatial representation of ZIP Codes, many people envision ZIP Codes as polygons and/or as attributes along the sides of streets. Since the USPS does not provide any such representations of ZIP Codes in a formal way, other entities have

designed them for their own purposes. For example, the US Census Bureau provides ZIP Codes as attributes along the streets in the TIGER Line database. Due to public interest in having statistics tabulated by ZIP Codes, the Census Bureau also created areas called the ZIP Code Tabulation Areas (ZCTAs) first developed for the 2000 census. ZCTAs are statistical entities developed for tabulating summary statistics from the census data built from census blocks and the Census Bureau believe these boundaries overcome the difficulties in precisely defining the land area covered by a ZIP Code. The Census Bureau defined ZCTAs as small, relatively permanent statistical subdivisions of a county designed to be relatively homogeneous with respect to population characteristics, economic status, and living conditions (US Bureau of the Census 2000; US Bureau of the Census 2001; US Bureau of the Census 2011). Many other federal and private organizations also are producing their own polygon datasets, sometimes for very limited purposes.

Although ZIP Codes in urban areas may often resemble spatial areas since they are comprised of spatially clustered street ranges, there are many spatial cases where individual polygons simply cannot topographically be drawn around all points in a ZIP Code. The example of the two ZIP Codes that apply to the Sears Tower (2 ZIP Codes cover the same block) is just one example. Further, in rural areas ZIP codes can be collections of rural delivery routes that in reality do not look much like a closed spatial area. The areas that do not require any mail delivery (e.g. deserts, mountains, lakes, parks) have no defined ZIP codes.

The three most important issues with using ZIP Codes for spatial analyses are: the uncertainty of the boundaries interpolated from ZIP Codes; mismatches of these boundaries collected from different data sources; and temporal changes of ZIP Codes. As the USPS ZIP Codes change with time, so do the ZCTAs. As described in the ZCTA technical documentation, the ZCTA boundaries are based on the ZIP Codes that are available at the time of a census and thus these boundaries are subject to change (US Bureau of the Census 2000). Another important issue with ZIP Codes is the addition or deletion of ZIP Codes with time. Krieger et al. (2002) mentioned mismatches where 91% of the total cancer incidents were geocoded to ZIP Codes that did not exist when the data was first collected.

The practice of multiple users generating their own ZIP Code polygons can cause considerable error in spatial analyses and distributional interpretations. There is a significant amount of research employing ZIP Code polygon centroids as the basis for geocoding (McElroy et al. 2003), for finding geographic distance between services (Beyer et al. 2011), for examining results within ZIP Code boundaries or to determine levels of spatial autocorrelation or measures of clustering with point pattern statistics (Matisziw, Grubestic, and Wei 2008). These practices may create flawed outcomes as the point data produced from these features are both poor substitutes for known address locations, and more importantly are inconsistent since the polygons themselves differ. ZIP Code polygons differ both in size and shape when they are collected from alternate data sources. When other polygons are aggregated to form ZIP Code polygon

representations (such as census tracts) the method of aggregation can also influence analytic results (Fotheringham, Rogerson, 2009).

Grubestic (2008a) provides an overview of the problems and prospects of utilizing ZIP Codes for spatial analysis through some important issues like spatial contiguity, data aggregation, and boundary definitions and concluded that ZIP Codes are not strong geographic units for spatial or statistical analyses. Grubestic, and Matisziw (2006) discussed the challenges of the use of Census ZCTA boundaries in spatial analysis by comparing local concentration of non-native street segments within a ZIP Code and larger area. McElroy et al. (2003) attempts to employ multistep iterative geocoding processes within ZIP Codes whereas Shi (2007) discovers the errors associated with Census ZCTA boundaries compared to ZIP Code point datasets. Using a restricted Monte Carlo approach, he tried to evaluate uncertainty in cluster analysis caused by imprecise polygon level addresses. Post Office Box addresses also cause problems when these employ ZIP Codes for geocoding rather than street addresses (Hurley et al. 2003). Grubestic (2008b) points out how the interpretation of spatial distribution can be significantly wrong in delineating service areas for broadband communication if ZIP boundaries are used as the basis of analyses. The study combines ZIP Codes collected from a private company and the census block boundaries with the service coverage polygons around each telephone exchange Central Offices to estimate the percentage of households within the coverage range of DSL service for each ZIP Code area. Since ZIP Code representations can vary according to different interpolation methods across different data sources, the accuracies of these analyses are not known. It is thus important

to compare the datasets being used in spatial research when these data are collected from different sources and quantify the extent of how much contrast is possible while working with ZIP Code polygons from variable sources.

In recent years there has been discussion about which geographic unit is more reliable for geocoding locations and area based socio-economic measurement in spatial research. Several studies mentioned the problems of inconsistency between ZCTA and ZIP Code boundaries (Grubestic and Matisziw 2006; Dai 2010). For example: in Dai, 2010, ZCTA boundaries were inconsistent with the scale of breast cancer data that was collected at ZIP Code level. As the boundaries did not match between two ZIP Code representations, ZCTA boundaries were not used in accessibility measurement. Studies also approached the topic of uncertainty with USPS ZIP Code boundaries as these are arbitrary boundaries around addresses with a specific ZIP Code number and changes over time (Schultz, Beyer, and Rushton 2007; Cudnik et al. 2012; Grubestic and Matisziw 2006). Even if some researchers consider ZIP Code as a reliable geographic unit for accessibility measurement and integrating other socio-economic factors, studies in this thesis have found that ZIP Code itself can have different boundaries over datasets and is capable of creating different results. Krieger et al. (2002) compares different geographic units of observation such as ZIP Codes, census tracts and blocks to identify socio-economic gradients in health research and concluded that the ZIP Code is the least desirable geographic unit. Multiple outcomes showed no gradient or very low gradient compared to census tracts and blocks.

However, those studies mainly discussed the inconsistency between ZIP Code and ZCTA boundaries. Some prefer ZIP Code polygons whereas some other studies argue for ZCTA boundaries but the fact that, ZIP Codes may have different boundaries as collected from different data sources, has received less attention in literature. Moreover, there is no known comprehensive quantitative examination of the kinds and magnitudes of differences between available ZIP Code datasets or the errors that those differences may generate in the results of spatial analyses.

This thesis will quantify the errors and evaluate the uncertainty associated with ZIP Codes and thus will help researchers to acknowledge the facts about the uncertainty in results in spatial analyses employing ZIP Code polygons. Different sections of this thesis primarily focus on various measures showing the differences between datasets, examples of the changes in results based on the choice of ZIP Code dataset, and most importantly the generation of measures by which it can be determined if two ZIP Code datasets are significantly different from one another.

SECTION 3

DATA: MULTIPLE ZIP CODE POLYGON REPRESENTATIONS

This section discusses the data collection and compilation process. In order to make comparisons across ZIP Code polygon datasets a number of ZIP Code polygons have been collected for several Northern Virginia counties from a number of different data sources: the USPS, the US Bureau of the Census, Sammamish Data Systems Inc. and other local/county representations (Alexandria, Arlington, Clarke, Fairfax County and Fairfax City, Frederick, Loudoun and Shenandoah) were collected. The USPS, Census Bureau and Sammamish datasets provide the ZIP Codes for nearly the entire Northern Virginia study area, whereas all other county datasets include only the ZIP Codes that fall within that particular county (or County Equivalent Area). Figure 1 represents the location of the study area within the state of Virginia. Given that some datasets were only available as of an update date of 2000; all other ZIP Code data used in the analyses of this thesis work is collected for the year of 2000 even if more recent data are available. This allows more reasonable comparisons across datasets.

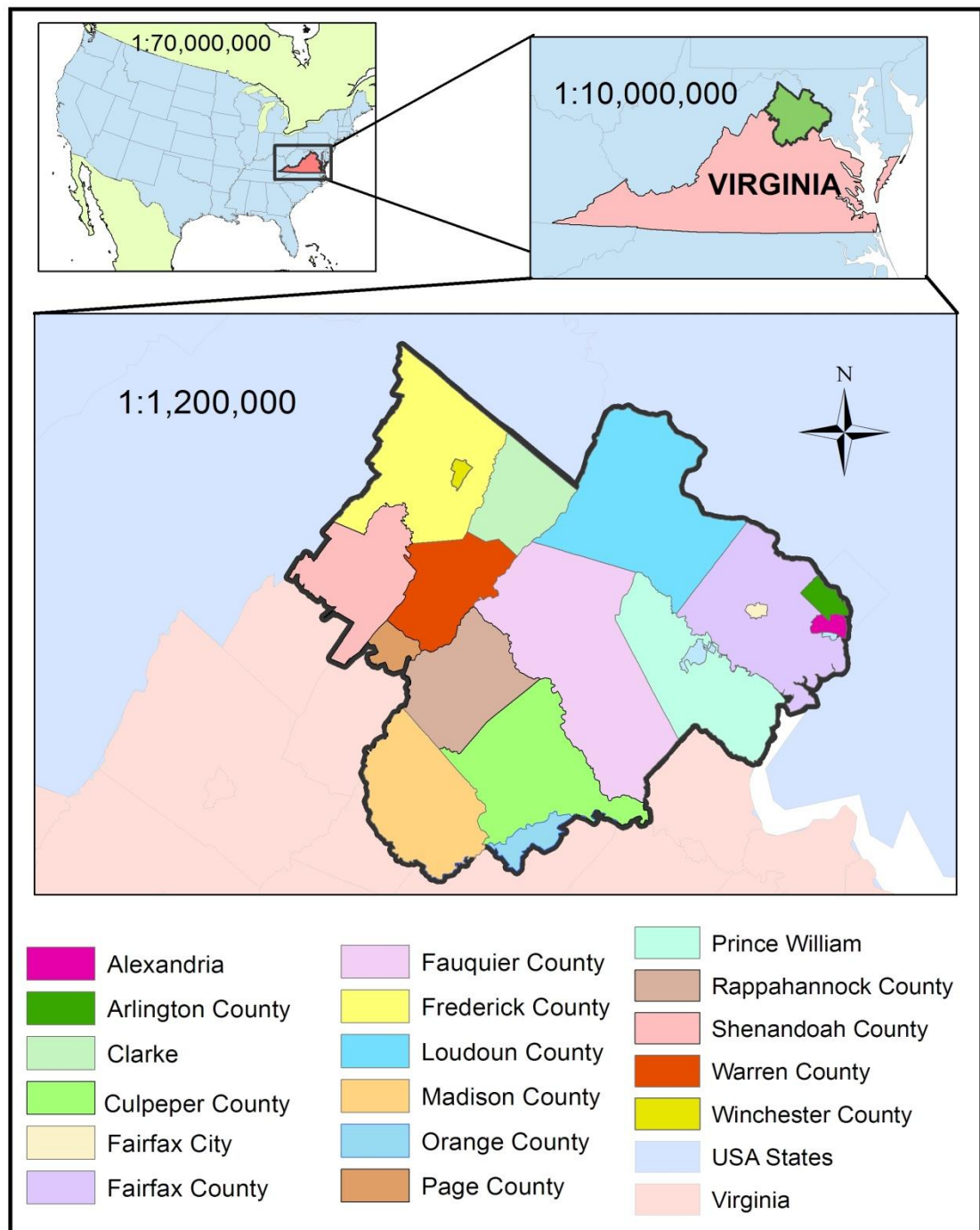


Figure 1: Study area of Northern Virginia

In this research many comparisons between ZIP Code datasets are made. In some cases pair-wise comparisons are made between all datasets. In other cases each dataset is compared to one definitive source. Given that the USPS is the authoritative source for ZIP Code assignment, the polygon dataset that they have provided for the study area is considered to be the “true” ZIP Code map. However, in reality the polygon representation from the USPS is not truly a representative for ZIP Codes. It has been used in this research just for the purpose of emphasis the facts about ZIP Code polygons. The USPS dataset contains 166 ZIP Codes for the study area whereas the Sammamish and the Census datasets provide 156 and 160 ZIP Codes respectively. The Census and the Sammamish datasets cover almost the entire study area whereas other individual county datasets provide different numbers of ZIP Codes (the Alexandria, Arlington, Clarke, Fairfax City, Fairfax county, Frederick, Loudoun, and Shenandoah datasets have 8, 11, 6, 3, 46, 11, 22, and 7 ZIP Code records, respectively). 3 digit ZIP Codes that include water features are avoided because the Census Bureau assigns ZCTAs based on three digit ZIP Codes to some undeveloped areas which have no MAF (Master Address File) address (US Bureau of the Census 2011) . When population data are used for comparisons across ZIP Code datasets, this information is taken from the Census Bureau for the year 2010. Figure 2 demonstrates the spatial extent of the datasets.

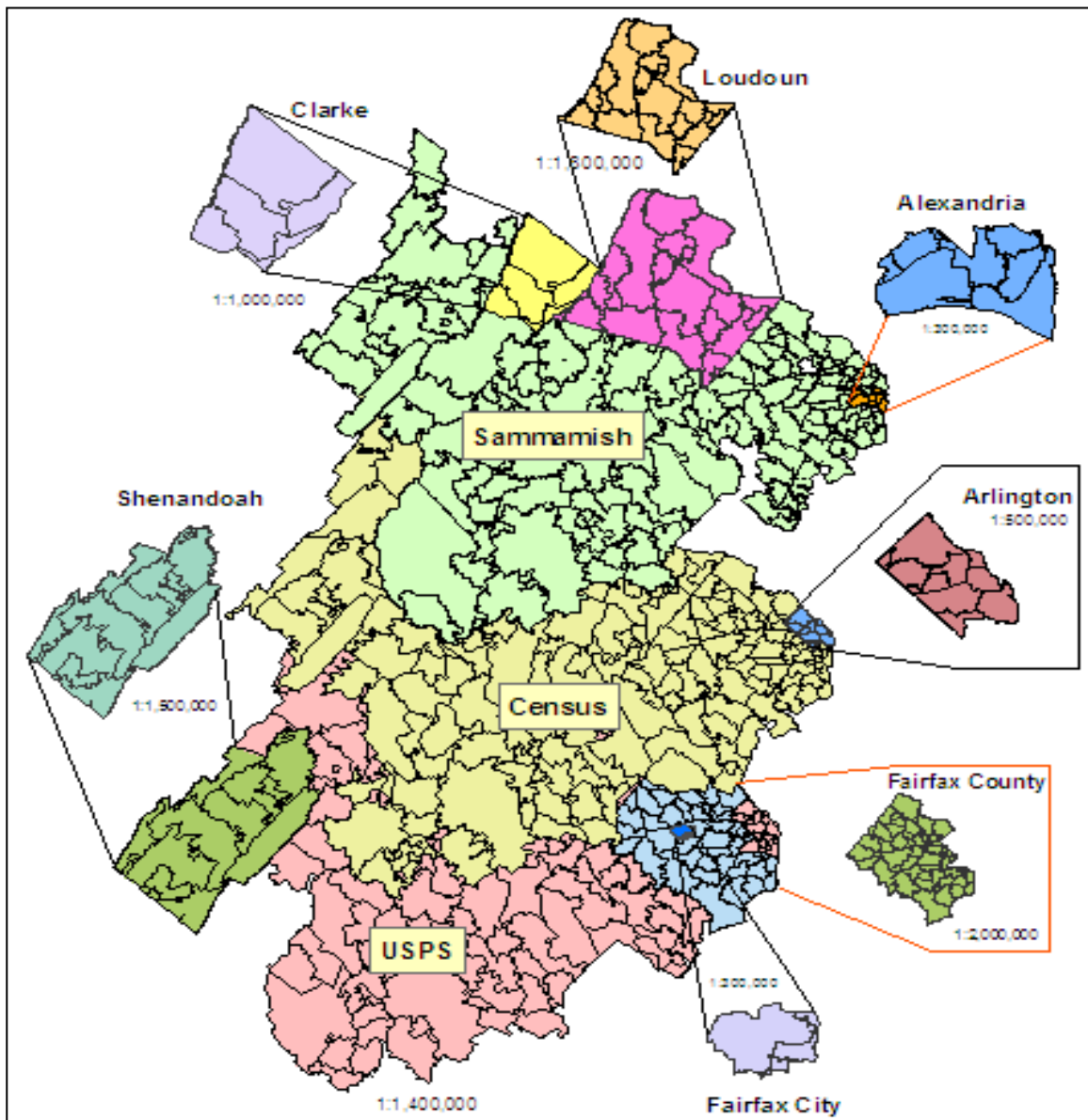


Figure 2: Spatial extent of the datasets

All of the data layers are projected to USA Contiguous Albers Equal Area Conic co-ordinate system to preserve area for areal differentiation and Azimuthal Equidistant Projection for measuring the distances within and among ZIP Codes. To avoid multiple

entry of a ZIP Code, i.e. when a ZIP Code area is split into multiple polygons, all records of that particular ZIP Code are merged together into a single feature in the database, although the spatial separation is maintained. Population is assigned to USPS, Loudoun and Sammamish ZIP areas from the Census data.

The street centerline data of Fairfax County that has been used for the building road network in Section 6 and Section 9 was collected from the Fairfax County Government (<http://www.fairfaxcounty.gov>). The locations of hospital used to measure network distances in Section 6, and school location and attendance areas utilized for ranking the ZIP Codes in Section 7 are also obtained from the Fairfax County Government. School ranks based on the overall school performance or test scores are downloaded from the website <http://www.greatschools.org/>. Hispanic population data for examining the level of segregation at ZIP Code level are collected from the US Census Bureau website (<http://www.census.gov>). To measure accessibility of the ZIP Codes to emergency room services Fairfax County emergency room data are collected from the online website <http://www.yellowpages.com/fairfax-va/emergency-room>.

SECTION 4

DESCRIPTIVE COMPARISONS: EXPLORATORY COMPARATIVE ANALYSIS

This section measures the areas and centroid locations for ZIP Code polygons and compares the differences across different datasets. Areas and centroid locations are first examined to demonstrate the extent of the difference among ZIP Code representations. Comparisons of the areal extent of ZIP Codes are also made using an overlay operation termed Symmetric Difference.

Section 4.1 describes the methodologies used in the study for determining differences in area boundaries and geographic center locations and comparisons of the results across the data sources. Segment 4.2 discusses results from the study in detail followed by a discussion of the consequences of using ZIP polygons in spatial analyses.

4.1 Methods of Comparative Analysis

Several geoprocessing and spatial statistics tools have been utilized for measuring areas and identifying centroid locations of the ZIP Codes. For comparison purposes, the individual ZIP Code boundaries are mapped as separate layers for all the datasets. Figure 3 shows a flow chart of the important steps that have been utilized in this research.

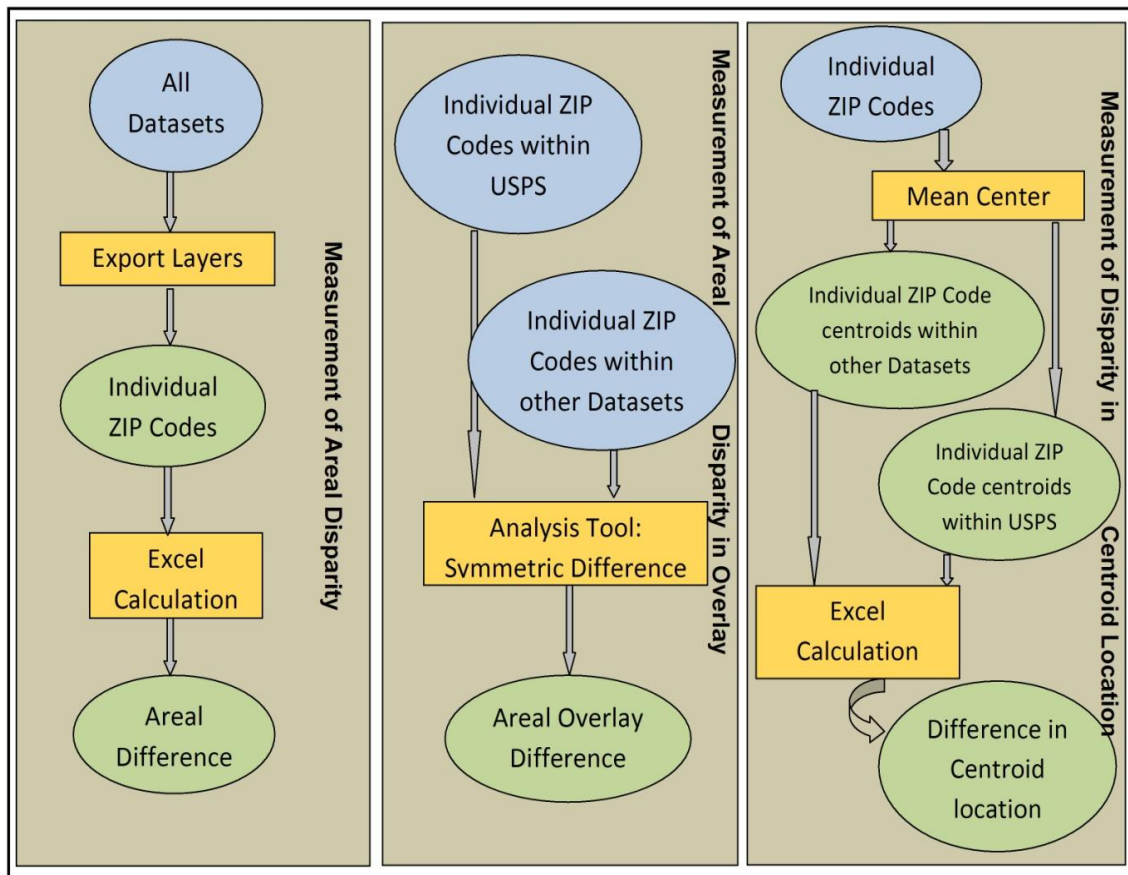


Figure 3: Flow chart of the methodology

Since areal measurements are commonly used in spatial analyses using ZIP Codes as the unit of observation, the areas of the individual ZIP Codes from all datasets are compared to the USPS ZIP Code polygon areas in order to evaluate the extent of area differences across the datasets. An esri ArcGIS analysis tool ‘Symmetric Difference’ - is used to further investigate the mismatched areas between a dataset and the USPS dataset. The symmetric difference of two sets is the set of elements which are in either of the sets and not in their intersection. The symmetric difference of the sets *A* and *B* is commonly

denoted by $A \Delta B$. The symmetric difference is equivalent to the union of both relative complements (Equation 1) or the union of the two sets subtracting their intersection (Equation 2).

$$A \Delta B = (A \setminus B) \cup (B \setminus A) \quad \text{.. Equation 1}$$

$$A \Delta B = (A \cup B) \setminus (B \cap A) \quad \text{.. Equation 2}$$

Figure 4 gives an overview how this process works. The non-matched portion of USPS (A) and Clarke (B) ZIP Code 20135 is clearly visible in the symmetric difference result (C). The centroid location of ZIP Code polygons are identified using the spatial statistics tool ‘Mean Center’. To find out how the geographic characteristics of ZIP Codes deviate across data sets, difference (in area, centroid locations, and overlay) are calculated and compared in MS Excel.

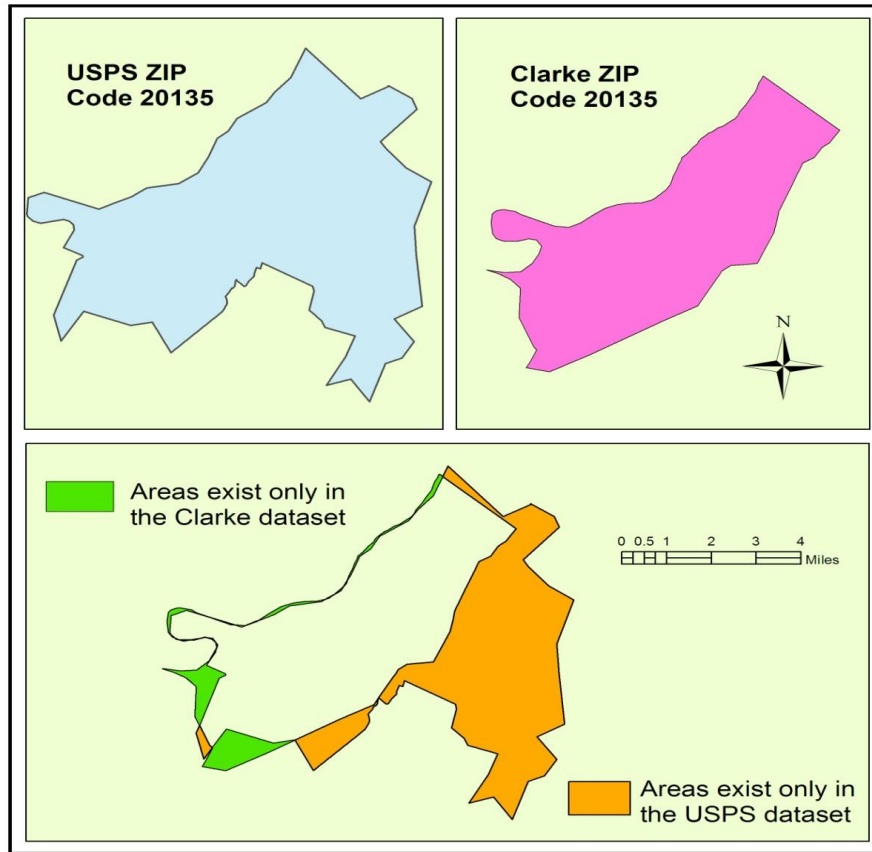


Figure 4: Symmetric difference (C) from USPS (A) and Clarke (B) ZIP Code 20135

4.2 Results of Comparative Analysis

For some cases, the spatial extents of ZIP Codes within one dataset can be very different from other datasets. Figure 5 presents some examples of the deviations in size, shape and position of the ZIP Codes across datasets.

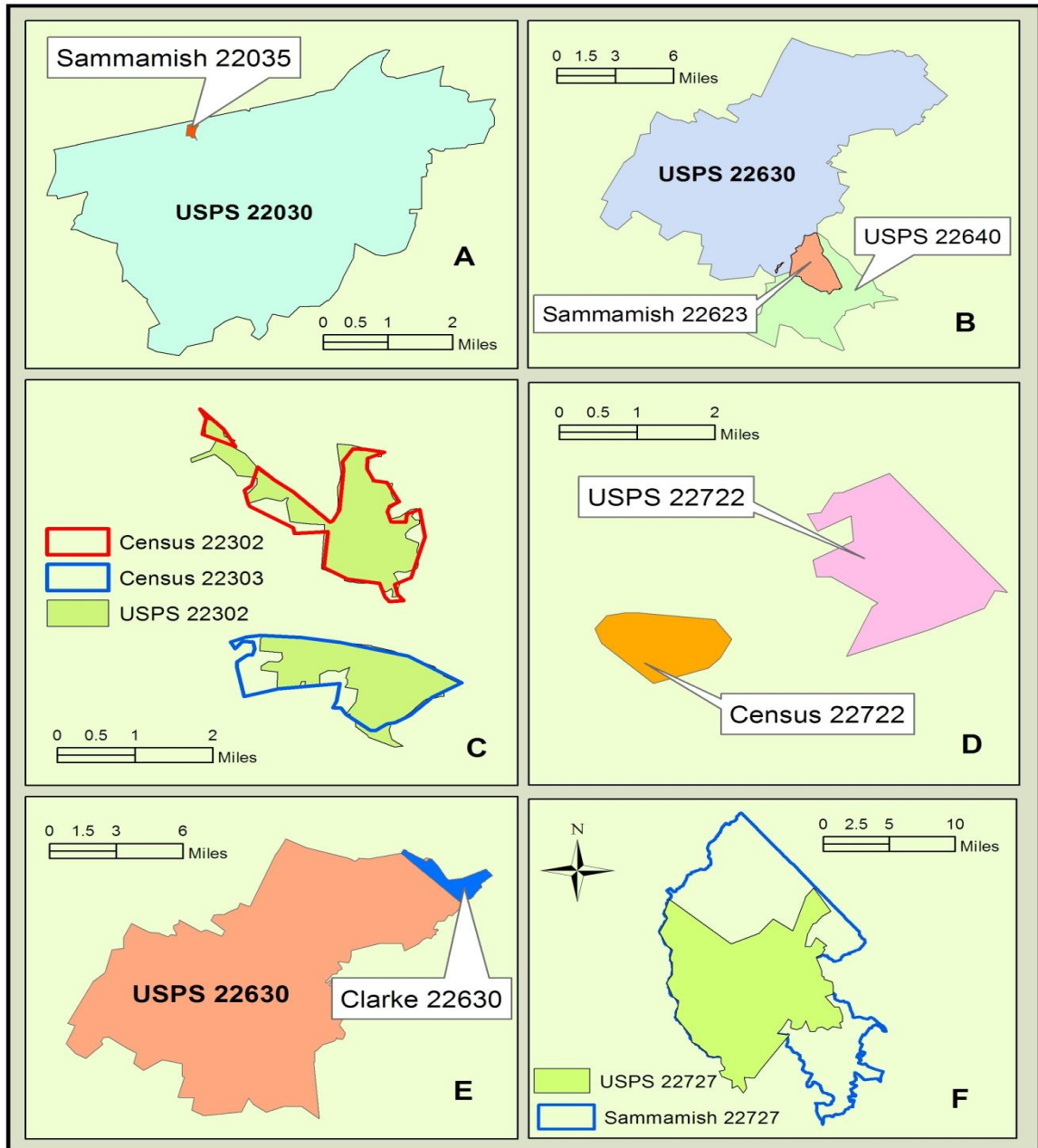


Figure 5: Deviation of size, shape and position of ZIP Codes across datasets.

In Figure 5E ZIP Code 22630 within the USPS dataset is almost 40 times larger than the corresponding ZIP Code area within the Clarke dataset. In Figure 5D, ZIP Code representations of 22722 are in completely different locations within the USPS and the

Census dataset. Thus, in some cases, the ZIP Code area within one dataset has no correspondence within the other dataset. More than 300 km² of the Sammamish ZIP Code 22727 does not exist within the USPS ZIP Code 22727 (Figure 5F). In Figure 5A Sammamish ZIP Code 22035 is completely within USPS ZIP Code 22030 and there is no corresponding 22035 ZIP Code polygon in the USPS dataset. Therefore, a ZIP Code based spatial analysis will obtain a result for ZIP Code 22035 when using the Sammamish Dataset but will miss any value for this ZIP Code when the USPS dataset is employed. The same type of problem is seen for Sammamish 22623 which actually is a part of USPS 22630 and 22640 (Figure 5B). Figure 5C shows the USPS ZIP Code 22302 which is split into two different polygons. These polygons are regarded as the ZIP Codes 22302 and 22303 in the Census dataset.

Table 1: Differences (in km²) in ZIP Code areas within datasets compared with the USPS polygon dataset

	Alexandria	Arlington	Census	Clarke	Fairfax City	Fairfax County	Frederick	Loudoun	Sammamish	Shenandoah
Maximum Difference	12	64	101	325	38	16	91	67	251	79
Minimum Difference	0.24	0.02	0.03	1.11	16	0.01	0.001	0.26	0.01	0.44
Average Difference	5	6	9	68	24	1	18	15	10	18

The area calculation reveals considerable areal contrast between different data sources (Table 1). This difference can be as small as 0.001 km^2 (Frederick ZIP Code 22601) to as high as 325 km^2 (Clarke ZIP Code 22630 in Figure 5E). The ZIP Codes within the Clarke dataset have an average difference of about 70 km^2 which is the highest average areal difference among all datasets compared to the USPS. On average, all the datasets deviate by 17.5 km^2 from the USPS dataset regarding ZIP Code areas.

The results from the symmetric difference overlay analysis complement these results and reveal the extent of overlap disparity one might confront while working with these datasets. Table 2 and Figure 6 provide the results from the area calculations of the ZIP Codes using symmetric difference. Figure 6 summarizes the total areas of corresponding ZIP Codes that do not match with its spatial extent within the USPS and the individual datasets being compared. These are the areas of ZIP Codes that are not in common and do not overlap between the two datasets being compared. The unmatched areas range from 0.009 km^2 (ZIP Code 22601 for USPS-Frederick pair) to 336 km^2 (ZIP Code 22630 for USPS-Clarke pair). There is an average of 21 km^2 of total unmatched area for any USPS-dataset pair. Consider that this average value is higher than the average amount of mismatch when area itself is compared. This suggests that even when area values are similar the polygons themselves are not spatially coincident.

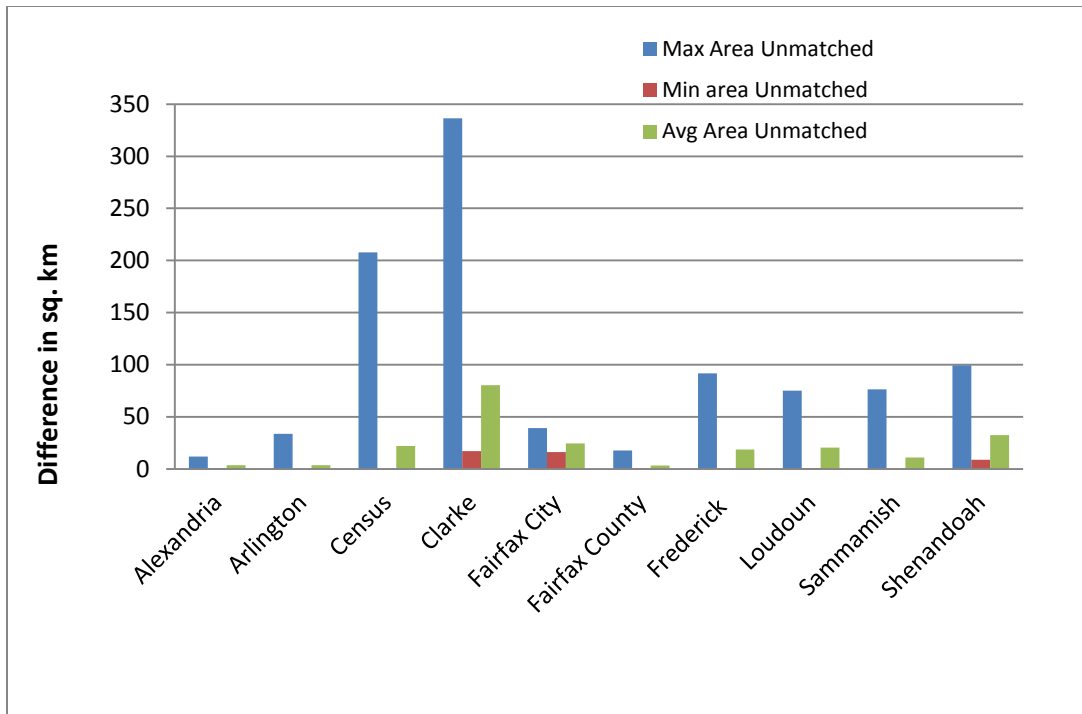


Figure 6: Total areas (km²) of the corresponding ZIP Codes that do not match in the spatial extent or do not overlap within the USPS and an individual dataset measured by Symmetric Difference.

Table 2 summarizes the areas of the ZIP Codes that are unique for individual datasets. These numbers represent the total areas of the ZIP Codes for individual datasets that are not present within the USPS dataset. More than 300 km² of the Sammamish ZIP Code 22727 does not exist within the corresponding ZIP Code area within the USPS dataset (Figure 6F). Almost 135 km² area of the Census ZIP Code 22134 does not exist within the USPS ZIP Code 22134. On average, the nonexistent areas within the Sammamish, Shenandoah, and the Census datasets are about 12, 11, and 10.7 km² respectively.

Table 2: Areas (km²) of the ZIP Codes within a dataset that do not exist within the USPS dataset (using Symmetric Difference).

	Alexandria	Arlington	Census	Clarke	Fairfax City	Fairfax County	Frederick	Loudoun	Sammamish	Shenandoah
Maximum Difference	0.79	1	135	10	0.76	17	1	16	305	21
Minimum Difference	0.00	0.07	0.00	2.72	0.09	0.04	0.01	0.00	0.03	4.21
Average Difference	0.39	0.30	11	6	0.34	1	0.24	4	12	11

In Table 3 the unmatched areas are presented as percentages of the total ZIP Code areas within individual datasets. These are the areas of ZIP Codes within a dataset that are not found in the corresponding ZIP Code areas within the USPS dataset. If the USPS and another dataset are referred to as A and B respectively then the area will be the complement of the USPS relative to that other dataset and thus will refer to the $B \setminus A$ in $A \Delta B = (A \setminus B) \cup (B \setminus A)$.. Equation 1. Some datasets have relatively small differences; for example, the Frederick dataset has the lowest percentage of mismatched ZIP Code areas (5% maximum change). On average the areas of ZIP Codes within the Frederick dataset differ by less than 1% compared to the USPS dataset. In other words, this dataset has the highest percentage of similarity among all the datasets with the USPS dataset.

Table 3: Percentages of areas of the ZIP Codes within datasets that do not exist within the USPS dataset

	Alexandria	Arlington	Census	Clarke	Fairfax City	Fairfax County	Frederick	Loudoun	Sammamish	Shenandoah
Maximum Difference	17	100	100	72	12	18	5	63	100	42
Minimum Difference	0.50	1	0.94	4	6	1	0.02	1	7	1
Average Difference	7	14	17	22	8	4	0.77	12	14	16

However, the rest of the datasets have high percentages of ZIP Code areas that are not found within the USPS dataset. The entire areas for several ZIP Codes within the Arlington, Census and Sammamish datasets cannot be found in the corresponding ZIP Code areas within the USPS dataset. On average, 22% of the area of each ZIP Code within the Clarke dataset does not exist within the corresponding ZIP Code areas of the USPS dataset. Finally, several ZIP Codes within the Arlington, Census and the Sammamish datasets have areas which are completely absent within the USPS dataset. The Clarke, Loudoun and Shenandoah datasets also have large ZIP Code areas with little

or no correspondence within the USPS dataset. Large errors in average difference and missing data also have been found that can lead to potential errors in any spatial analysis.

Table 4: Deviation of centroid locations of ZIP Codes in different polygon datasets compared to the USPS dataset (in km).

	Alexandria	Arlington	Census	Clarke	Fairfax City	Fairfax County	Frederick	Loudoun	Sammamish	Shenandoah
Maximum Difference	8.95	4.03	11.41	16.22	3.57	6.55	6.57	14.64	12.73	7.62
Minimum Difference	0.04	0.03	0.01	0.39	2.05	0.02	0.00	0.02	0.01	0.43
Average Difference	2	0.46	1	4	3	0.29	1.40	1.73	0.88	1.83

Table 4 summarizes the difference in centroid locations of ZIP Codes polygons. The maximum difference is found for the Clarke dataset; where a centroid is more than 16 km from the corresponding USPS centroid location (ZIP Code 22630). The Frederick dataset has a minimum difference of 21 meters which is associated with the ZIP Code 22601. The ZIP Codes within the Clarke dataset differ by 4 km on average from the USPS ZIP Code centroid locations. The Alexandria, Fairfax City, Loudoun, and the Shenandoah datasets also have large average differences of 2, 3, 1.7 and 1.8 km

respectively in centroid locations compared to the USPS dataset. The average difference in ZIP Code centroid locations between the USPS and all other datasets is 1.6 km

Large differences in areas, centroid locations and overlaps across the datasets suggest that any spatial analysis can have very distinct outcomes depending on which spatial characterization of ZIP Codes is employed for that analysis.

SECTION 5

THE INFLUENCE OF POLYGON REPRESENTATION ON SPATIAL STATISTICAL RESULTS

Many common spatial statistics are employed on polygon datasets, including ZIP Code datasets. Once again, if there is no definitive ZIP Code dataset to use as a basis for these analyses, the results could be questionable. This section examines the extent to which the results of spatial statistical analyses can be altered due to changes in the underlying polygon datasets. Global Moran's I and Anselin's Local Moran's I are used to provide examples of how the geographic distribution of a spatial event or incident can be different based on the dataset used for the analysis.

This study concentrates on Loudoun County of Northern Virginia. As was demonstrated in viii, the spatial structure (especially the size, shape, and position of ZIP Code polygons) varies greatly across different data sources. This section seeks to determine if those differences produce diverse spatial correlation values for a variable over space. ZIP Code population is used as an example value to illustrate variations that may occur when using different ZIP Code datasets.

5.1 Methods for spatial autocorrelation and cluster analysis

Since there is no direct population associated with USPS ZIP Codes, population values have been determined based on the area that intersects with the Census ZIP Codes. The method of obtaining values from other feature layers is known as the 'Areal

Interpolation' method. This can be done in ArcGIS by a series of summing and joining feature attributes. Figure 7 illustrates an example of how this method works. In the figure USPS ZIP Code 20129 has parts that are located within four distinct Census ZIP Codes (20129, 20158, 20176, and 20197). First the intersecting parts of each of these four polygons have been determined using the overlay analysis tool 'Intersect'. The population is calculated for the areas that intersect with that particular ZIP Code in the Census. Finally these four populations are added together to determine the population for the whole area of ZIP Code 20129 within the USPS dataset. The same process is employed for all 20 ZIP Codes taken for analysis from the USPS, Sammamish, Census and Loudoun datasets. It is not possible to test autocorrelation on all the counties at the same time, because each has specific ZIP Codes that are not present in other counties.

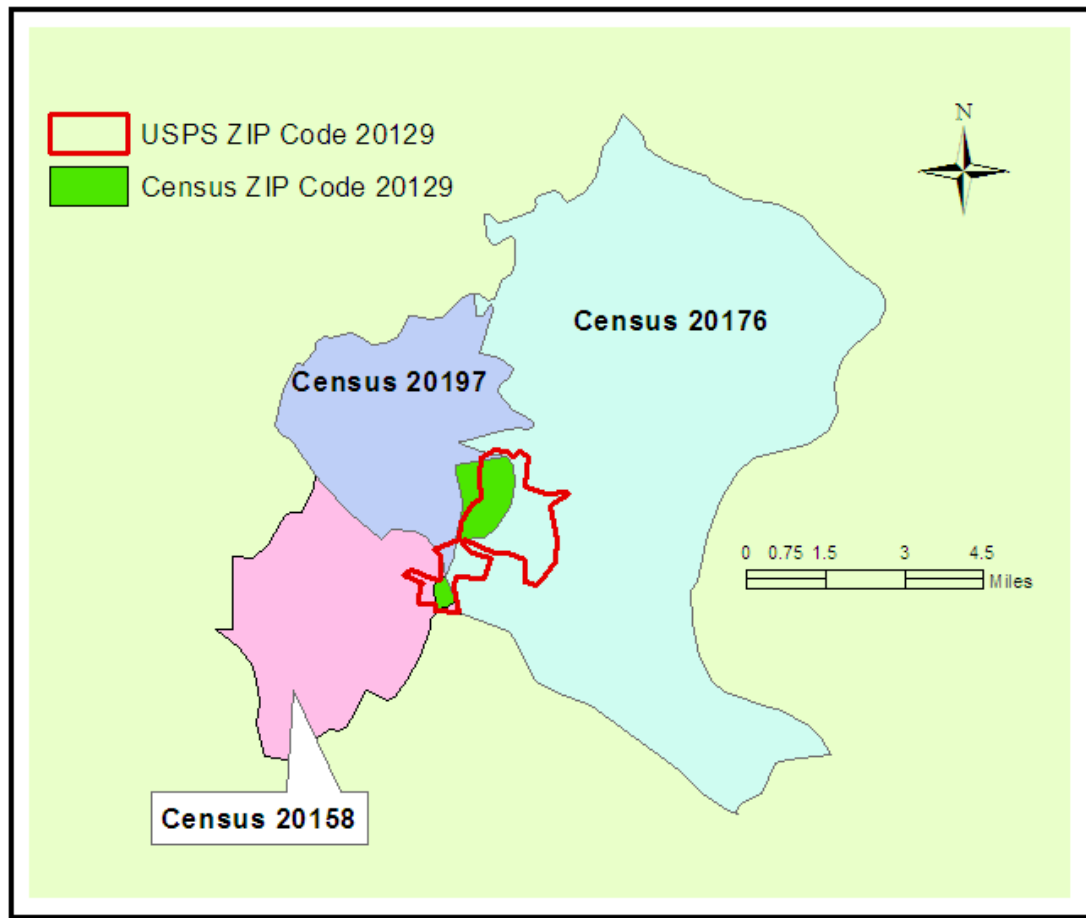


Figure 7: Calculating population for USPS ZIP Code 20176 based on intersected area with Census ZIP Codes

This type of calculation for population could be very different from the actual value but it ensures the most similar population distribution possible. Problems with estimating population in this way have been well documented (Wilson and Mansfield 2010; Eicher and Brewer 2001) particularly due to the assumption of even population distribution. However, since the focus of this research is not focused on the areal interpolation problem, we accept this assumption in this case.

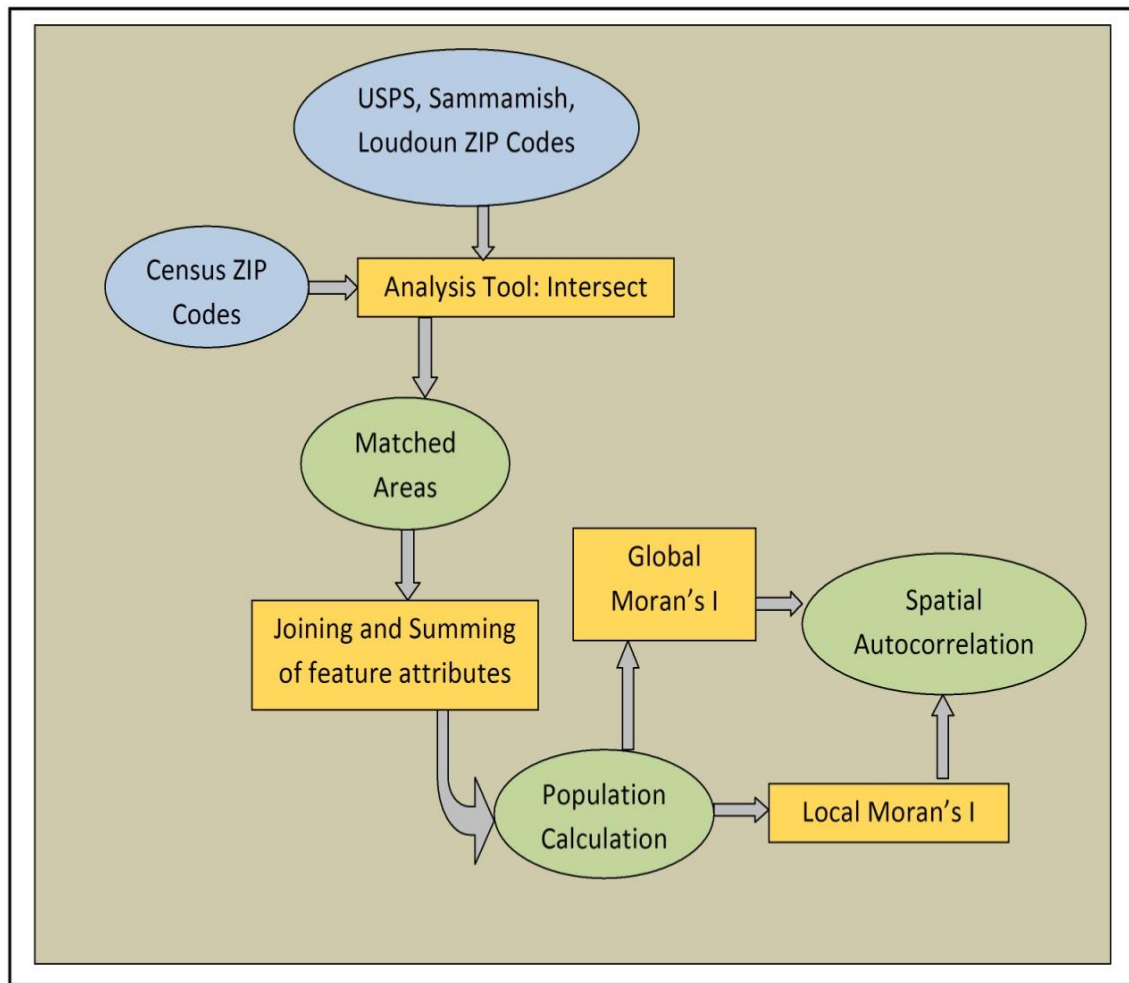


Figure 8: Methodologies for determining populaiton and spatial pattern.

After calculating the population, Global Moran's I is executed to obtain a value of overall spatial autocorrelation of population among ZIP Codes. This analysis is performed on the four different datasets. Local Moran's I is similarly tested in order to examine how the adjacent ZIP Codes are correlated according to their population values. This test reveals whether the same types of values are close to or far from each other. Figure 8 presents an overview of the methods employed to determine population.

5.2 Results

The Moran's I comparisons reveal moderate clustering at the 0.05 significance level for Census, Sammamish and USPS data but very high clustering for Loudoun at the 0.01 significance level. Although the overall result shows similarity when testing autocorrelation through Local Moran's I, Some ZIP Code areas within the Sammamish dataset could not be calculated because of mismatches with other datasets.

Figure 9 compares the results of Local Moran's I across datasets. All four datasets display high levels of clustering for ZIP Codes 20164 and 20165. However, in the Census and Sammamish datasets ZIP Code 20147 is identified as having a moderate level of clustering, whereas in the Loudoun dataset it displays the highest level of clustering, and in the USPS dataset it displays no significant clustering or dispersion. These results clearly illustrate the fact that an analysis with a same variable can obtain different results simply because of a change in the ZIP Code polygon representations. Since there is no authoritative ZIP Code representation it is very difficult to determine which value of spatial autocorrelation should be accepted.

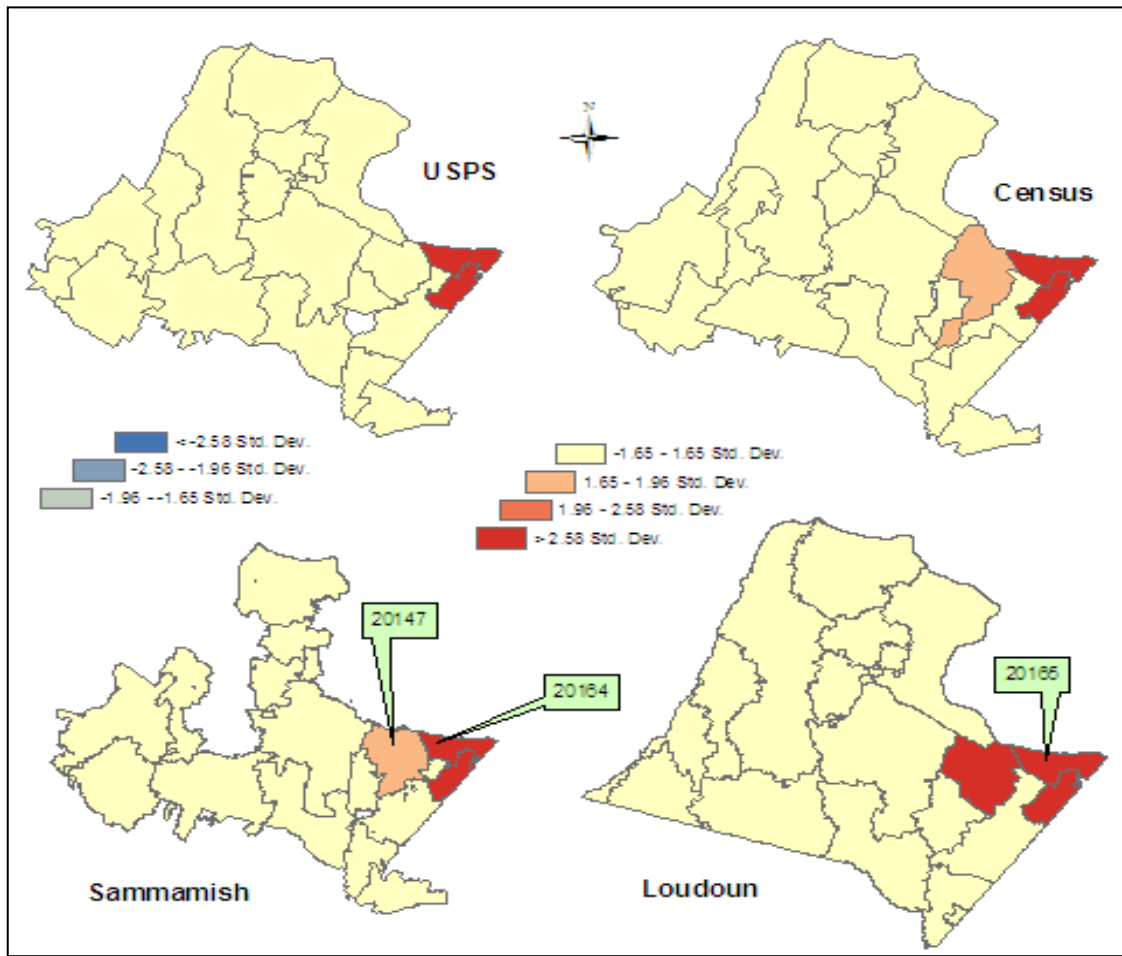


Figure 9: Results of Local Moran's I for population of ZIP Codes across datasets

In this particular case the difference appears to stem from the large difference in the size and shape of ZIP Codes in USPS and Census datasets. Several ZIP Codes in USPS are many times larger than in the Census, and vice versa. This, of course, leads to considerable differences in the populations associated with those ZIP Codes. In conclusion, these outcomes clearly indicate that different values for statistical measures of spatial autocorrelations are possible, perhaps even likely, for a variable depending on the contrasting size, shape and area across different ZIP Code datasets. It suggests that, in

the absence of some strong justification for one ZIP Code dataset over another, it will be difficult to justify the robustness of spatial statistical results when the ZIP code is the unit of analysis.

SECTION 6

THE INFLUENCE OF POLYGON REPRESENTATION ON NETWORK ANALYTICAL RESULTS

The need to estimate the distance between two locations is common in spatial research (Matisziw, Grubestic, and Wei 2008). In telecommunications, health and many other fields of research, network distances are very useful to determine practical separations among locations, to determine the coverage area of a service (Matisziw, Grubestic, and Wei 2008), or to measure accessibility along a network (Curtin, Biba, and Manca 2010; Wan et al. 2012; Ngui and Vanasse 2011).

However, estimated distances are often imprecise due to a lack of specific location data (Krieger et al. 2002; Beyer et al. 2011). This is particularly true in health research where, to protect patients' confidentiality, information is only available on a regional basis, frequently using ZIP Codes or counties as the unit of observation (Wallace 2003; Inagami et al. 2006; Votruba and Cebul 2006). When no further information is available, the geometric center of a geographic area is commonly used as a single point to which all observations are aggregated. For example, the centroid of a polygon may be used to represent the point location of the residence of all individuals known to reside within that area. The use of a ZIP Code area centroid as the representation of a person's location can be problematic as the boundary around a ZIP Code is an arbitrary polygon and can be changed according to the interpolation process used to create the boundary

(Beyer et al. 2011; Krieger et al. 2002). This becomes more complex as multiple ZIP Code maps are available from different data sources creating their own ZIP Code polygon boundaries. This section of the study examines if the variable positions of ZIP Code centroids in different data sources can have an influence on the results of network analyses.

6.1 Network-analytic Literature Survey

The use of network distances of locations is well documented in spatial research. Network distances between locations are important for determining true coverage areas of service locations; for measuring accessibility to some facilities; for assessing and managing emergency response systems; for assessing safety performance within a transportation network; and for many other research (Curtin et.al 2005; Li and Waters 2005; Luo and Qi 2009; Peters and Hall 1999; Grubestic 2008b). Curtin, Biba, and Manca (2010) developed a parcel-network method for determining the walking accessibility of a population living within parcels to nearby transit facilities, utilizing cadastral information and demographic characteristics of parcels as well as using the network distances between the parcels and transit facilities. Foda and Osman (2010) measured the accessibility to some transit facilities and transit access coverage employing pedestrian road networks surrounding the transit facilities. They also developed a set of indices for determining the ratio of actual access coverage (network coverage) to the ideal access coverage (circular coverage) surrounding a transit location. Mishra, Welch, and Jha (2012) determined transit connectivity within a multimodal transportation network where

connectivity is considered as an indicator to quantify and evaluate transit service coverage integrating routes, schedules, socioeconomic, demographic and spatial activity patterns. Patel, Waters, and Ghali (2007) determined the accessibility of areas to cardiac catheterization facilities that are within a 90 minute travel distance from the facilities in the province of Alberta, Canada.

The use of ZIP Code polygon spatial characteristics in research is common. Qureshi, Hwang, and Chin (2002) estimated distances from the ZIP Code of an origin to the ZIP Code of a destination using a great circle distance (GCD) and a network-based model. That study emphasized the importance of network based distance estimation by comparing the distance derived from the two models. Messina et al. (2006) quantified the access to hospitals from ZIP Code centroids in the state of Michigan, considering distance to the nearest hospital and road network density in estimating travel time. Bliss et al. (2012) quantified spatial accessibility of healthcare based on the proximity of a patient's residence within a ZIP Code to health services. Hebert, Chassin, and Howell (2011) examined racial differences in the use of high-quality hospital care influencing neonatal mortality again by using the ZIP Code centroids as the location for each mother.

Additional research has identified the problems of using ZIP Code centroids in network analysis. Govind, Chatterjee, and Mittal (2008) allocate available hospital resources to different types of disease using a network of hospitals within a ZIP Code area, and examine the spatio-temporal pattern of disease incidence within that ZIP Code by incorporating the driving distance to a hospital as well as the types of roads in determining travel time. Grubestic (2008b) found that in delineation of broadband

telecommunication service areas along a street network are overestimated in areas that were demarcated by ZIP Code boundaries. Cudnik et al. (2012) also described how the actual transport distance from a patient's location to health care center changes when using the ZIP Code centroid as a surrogate location. The fact that many research efforts that employ ZIP Code centroid locations have been found in only a review of the recent health care access literature, suggests that this practice is very common, and that the consequences of the practice deserve attention. This research in this section attempts to determine the extent to which network distances between ZIP Code centroids and hospitals change across datasets due to different centroid locations in those datasets.

6.2 Study area and data

To understand the effect of choosing a ZIP Code polygon data layer on the outcome of a network analysis, this study utilizes a road network in Fairfax County, VA (Figure 10). Centroids of ZIP Codes are determined, and the network distances from the ZIP Code centroids to hospitals within the county are examined for different datasets. The results show the variation in the route length that can be generated by different datasets.

In this study, the same ZIP Code interpolated-polygon maps are used as described in Section 3. Specifically, only those that included the ZIP Codes for Fairfax County are employed here (Census Bureau, USPS, Sammamish Data Systems Inc. and Fairfax County). The point layer of 12 hospital locations and the street centerline data of Fairfax

County are collected from the Fairfax County official website

(<http://www.fairfaxcounty.gov>).

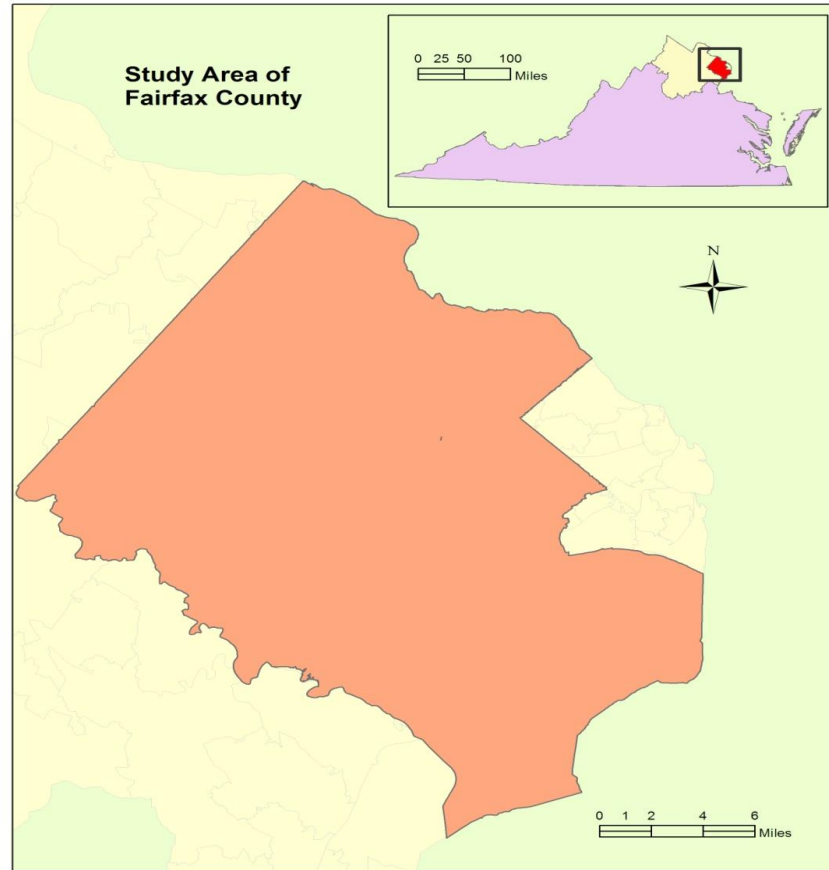


Figure 10: Study area of Fairfax County

A total of 44 common ZIP Codes are used for the network analysis; that is, the ZIP Codes exist in all four of the ZIP Code datasets. 22091 is a unique ZIP Code for the Fairfax County map which, in fact, possesses the same ZIP Code area of 20041 in the

USPS map and a part of 20151 in the Census (Figure 11B). This area is shared by 20151 and 20171 in the Sammamish map (Figure 11A).

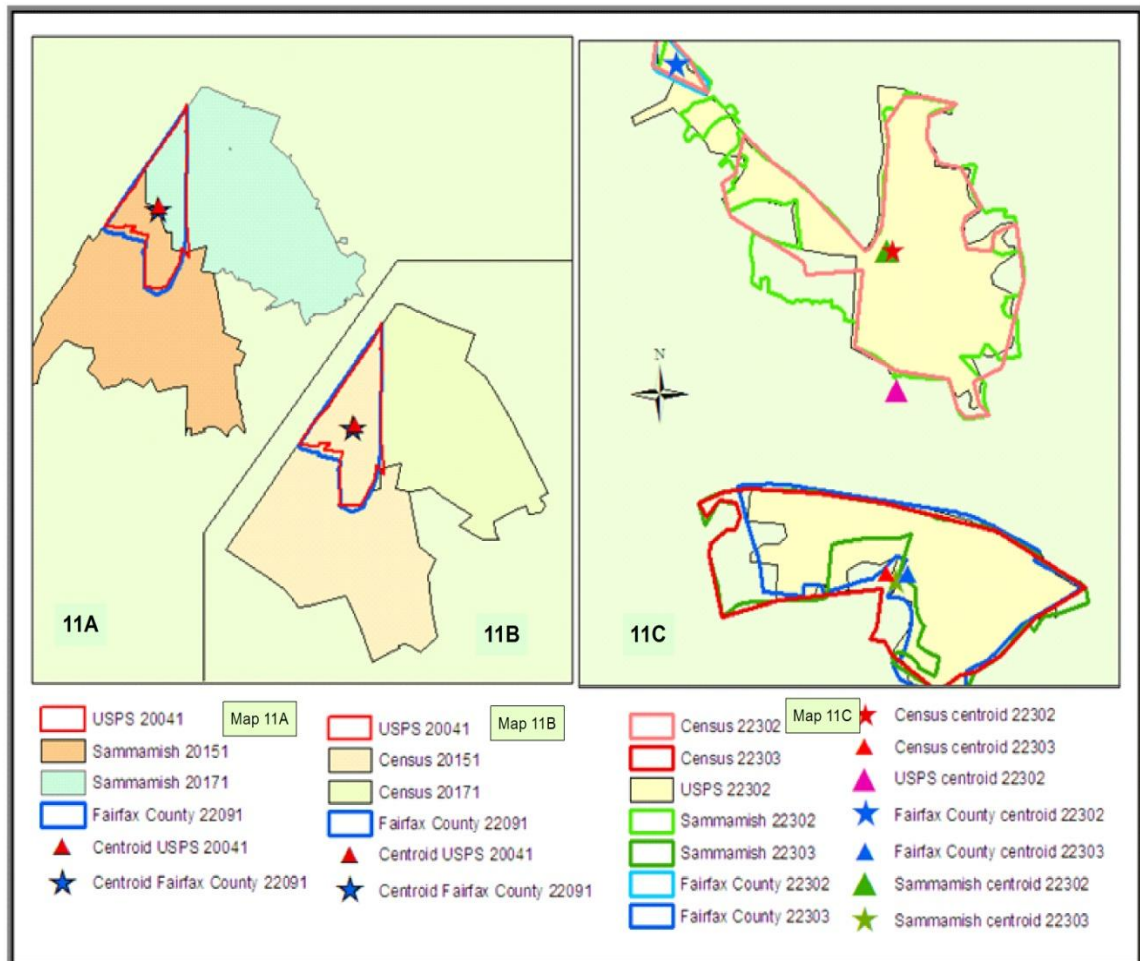


Figure 11: Unmatched ZIP Code that are excluded from the network analysis

Another unmatched ZIP Code is 22303 in the Fairfax County, Census and Sammamish maps. The USPS does not have this ZIP Code and this area is actually a part

of ZIP Code 22302 (Figure 11C). For this reason the USPS map is missing values for ZIP Code 22303 in the network analysis. This is another problem with ZIP Codes as multiple completely detached polygons can have the same ZIP Code number.

6.3 Methods

The analysis is executed using the Network Analyst extension of ArcGIS: first by measuring the network distance between ZIP Code centroids and hospitals and then comparing these distances across the ZIP Code polygon data layers from the Census, Sammamish, Fairfax County and the USPS. All network measurements are conducted along the Fairfax County Road network. To find the shortest distance, an Origin-Destination (OD) matrix is created considering the route length as the network impedance factor. As the centroids do not have any valid address, the network analyst tool in ArcGIS automatically assigns the centroids to the nearest location on a centerline. To avoid problems in this assignment no restrictions (barriers or one way roads) are applied to the network.

The OD matrix creates routes originating from each ZIP Code centroid to every hospital location (destination). This matrix operates faster than the Route solver when it is needed to determine the best route for multiple origins and destinations at once. While the matrix stores the network length in an attribute table, it shows the routes on the map as straight lines. For the purpose of better visualization, the routes in figures are mapped using the Route solver.

For each of the ZIP Codes, 12 routes are generated to hospitals. As the number of ZIP Codes varies over datasets, so does the total number of routes. In this study only the

common routes (528 routes in total) are taken for analysis which have correspondence within all of the datasets. The maximum and minimum distance for a centroid-hospital pair is recorded and compared across the datasets. Since the USPS ZIP Code map has been considered as the standard ZIP Code interpolated-polygon map throughout the studies of this thesis work, the measurements of network distances from ZIP Codes to hospital locations within the Census, Sammamish and Fairfax County datasets are also compared to the USPS dataset.

6.4 Results of network comparisons

6.4.1 Comparison of network distances within and between ZIP Code maps

The cost attribute of the OD matrix reveals that for several ZIP Codes the datasets have negligible difference in network distance but for other ZIP Codes there is a very large disparity among datasets. Some ZIP Codes have very similar centroid positions in multiple datasets. Even though the positions do not exactly match, the differences are negligible and thus essentially identical network distances to hospitals are created for these locations.

For example: ZIP Code 22124 has very similar centroid positions in the Census, Fairfax County, Sammamish and the USPS ZIP Code polygon datasets (Figure 12). A visible distinction is possible only at a scale of 1:20,000 or larger. Figure 12 shows the ZIP Code having slightly different centroid positions in alternative datasets but none of these points is located on a valid edge in the network. The Network Analyst tool assigned the centroids to the same position at the end of the cul-de-sac on Millar Road. Thus the ZIP Code has an identical route from each centroid to a hospital in all polygon data

layers. Only one 8.5 km long route is created for each of the centroids to Inova Fair Oak Hospital.

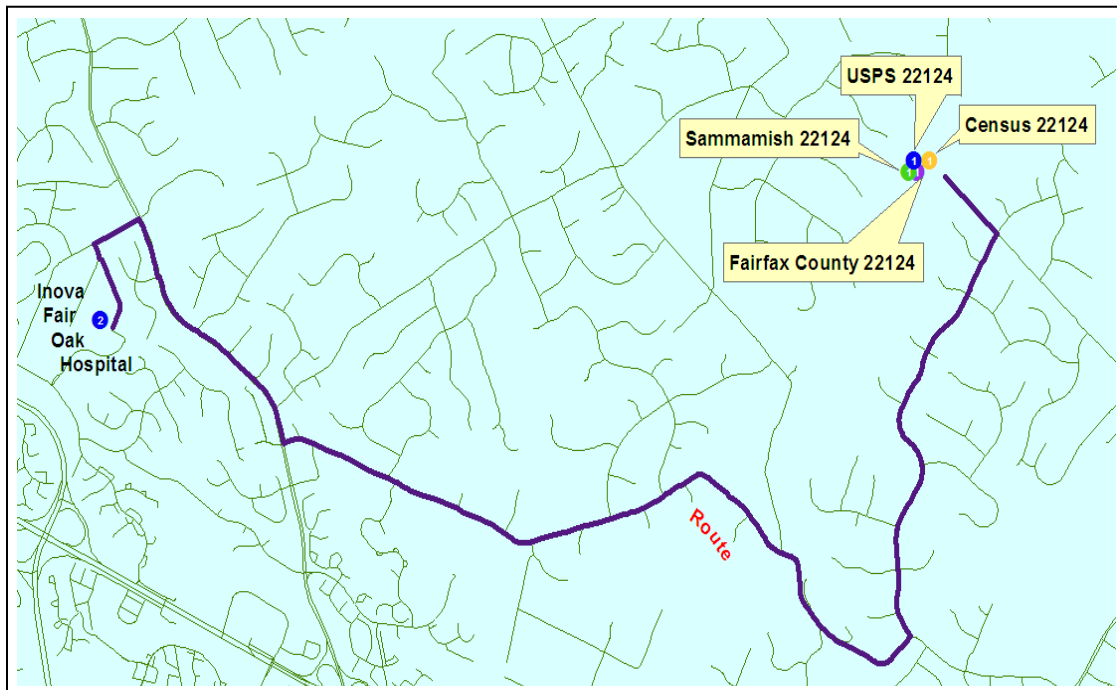


Figure 12: ZIP Code 22124 in different datasets has the same route to the Inova Fair Oak Hospital.

Figure 12 illustrates a different situation in the measurement of network distance using the example of ZIP Code 22066 and the network distance to the Reston Surgery Center. While the network distance is the same for the USPS, Sammamish and Fairfax County map, this is not true for the Census representation. The blue line represents the route (Route 1) for the Census ZCTA 22066 and the red dotted line represents the only route (Route 2) created for the Fairfax County, Sammamish and USPS datasets. In the

inset the routes are drawn to a larger scale to show the detail. On the Census Bureau map, the centroid is located on Aktamar Drive while in other datasets it is assigned to the nearest junction of Haven lane. Route 1 has a length of 12 km which is 0.5 km shorter than Route 2 (length: 12.5 km). This represents a difference in network distance of 4%.

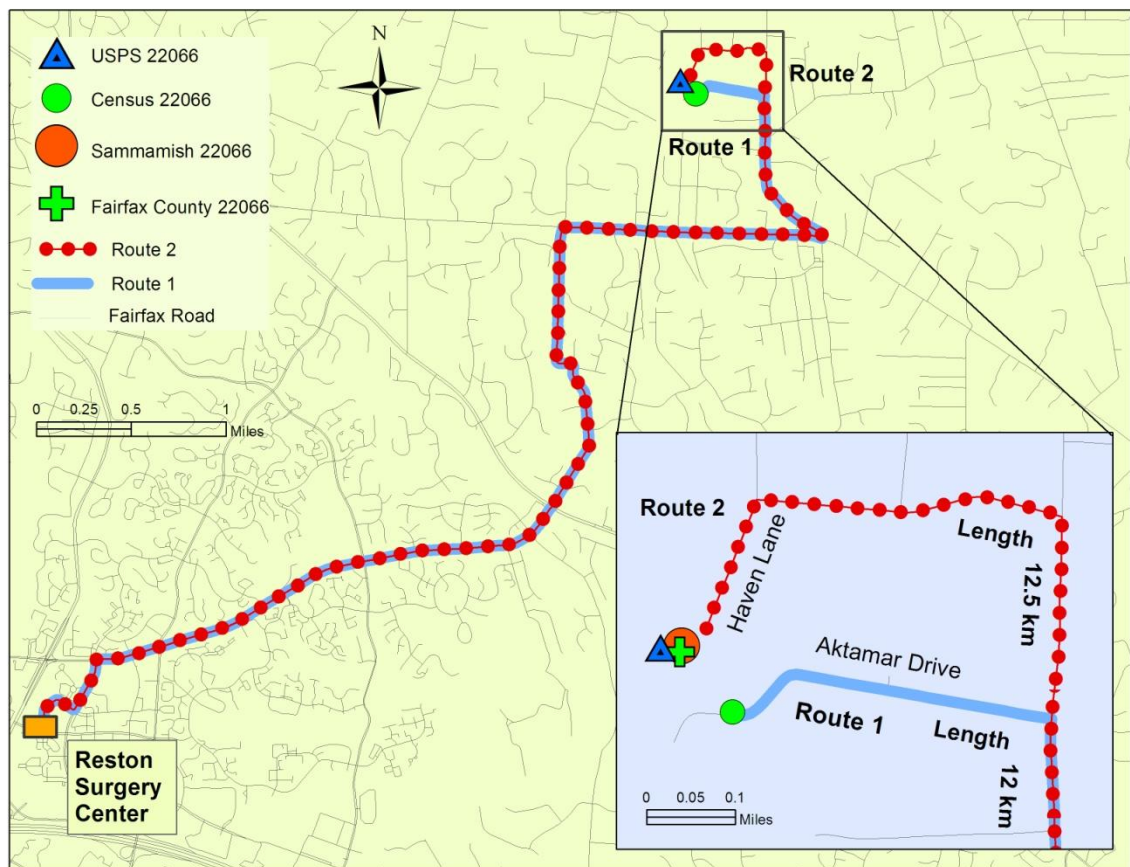


Figure 13: The route between ZIP Code 22066 and Reston Surgery Center is 12 km long in the Census (route 1) and 12.5 km long in the USPS, Sammamish and Fairfax County data (route 2).

While there are a range of differences in network distance, it should be noted that in the worst case scenarios there are some ZIP Codes which show very large differences across the datasets in centroid to hospital distance measurements. For example; the route between ZIP Code 22302 and Dewitt Army Community Hospital has the maximum length of 20 km in the Fairfax County data and a minimum length of 16 km in the USPS data which produces a difference of 4 km (an increase of 25%). The route between the same ZIP Code and Fairfax Surgical Center has a length of 21 km on the Census ZCTA map but 18 km if the Fairfax County ZIP Code map is used for the distance measurement.

The largest variation found in distance measurement is more than 5 km or about 3 miles. Figure 12, mapped at a scale of 1:100,000, shows the route between the Reston Surgery Center and centroid of ZIP Code 20120. The estimated length is 18 Km in the Sammamish dataset (Route 1) but larger than 23 km in the USPS dataset (Route 2). This implies that if a person wants to drive from the ZIP Code to the hospital, there will be a 23 km long driving route if the route is created using the USPS map. But if the driving route is created using the Sammamish map, the driving route will be an 18 km long in a different direction. This could clearly alter the results of any network-analytic research including hospital accessibility. An average, the ZIP Codes within datasets deviate about half a kilometer from each other regarding these route distances.

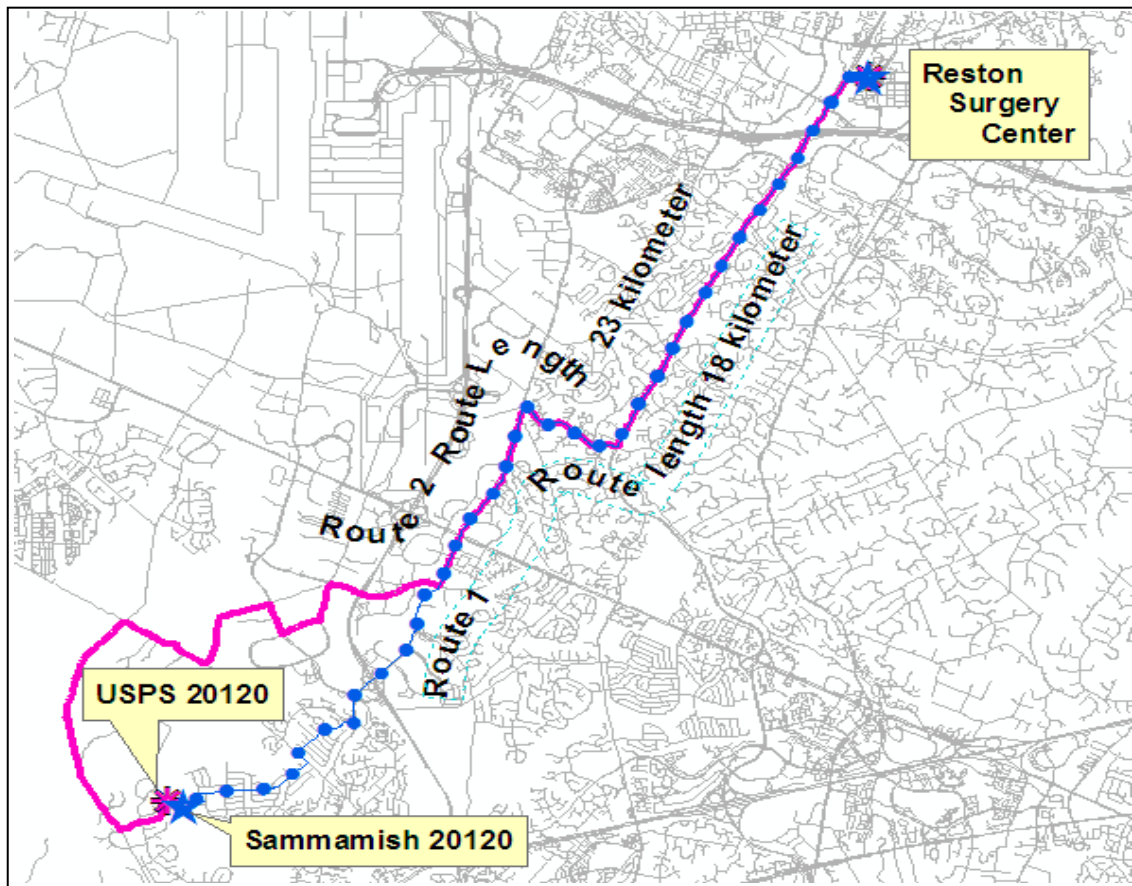


Figure 14: The difference in network distance between ZIP Code 20120 and Reston Surgery Center in Sammamish and USPS datasets

For all of the ZIP Code datasets, a total of 528 routes (12 routes for a single ZIP Code) are created out of which 64 routes (about 12% of the total routes) have a difference of one kilometer or more in route lengths across the datasets.

Figure 15 shows the number of routes for each of the ZIP Codes that have a minimum difference of a kilometer or more across the ZIP Code maps. All 12 routes generated from each of the ZIP Code 20120 and 22079 to the hospitals vary by one km or more within the ZIP Code datasets. ZIP Code 20151, 22015, 22102, 22303, 22310 and

22315 are also contributing to the major route distance variations across datasets. These results suggest that while some ZIP Codes are nearly identical, there are some other ZIP Codes that are very different across datasets. This confirms the results of the descriptive analyses of area, centroid location, and overlap from Section 4.

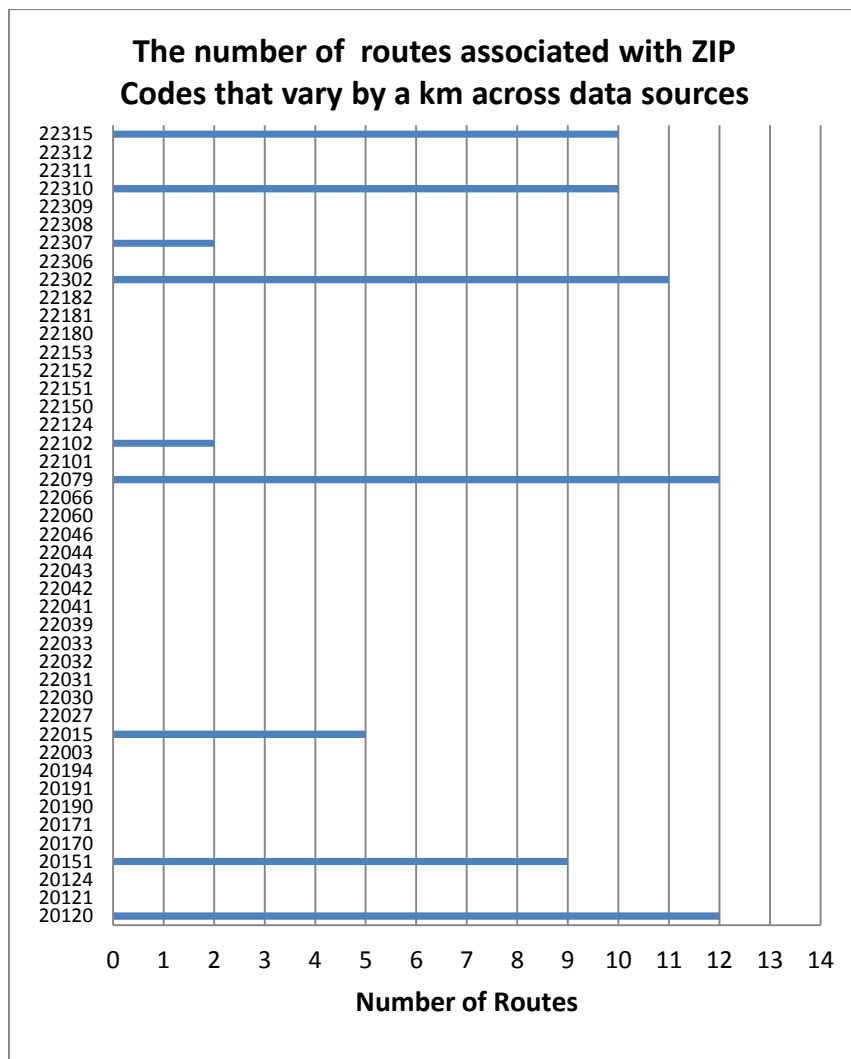


Figure 15: The routes that vary by a km or more across datasets in centroid-hospital network distance calculation

Figure 16 shows the comparison of network distances across pairs of datasets for corresponding routes. Although the average change of route length (0.2-0.3 km) of the total 528 routes may not seem very pronounced across pairs of ZIP Code datasets, some large discrepancies are found for several routes. In two thirds of the cases of pair-wise comparisons, the difference is 5 km or more (across the pairs of Census-Sammamish; Fairfax County-Sammamish; Fairfax County-USPS; and Sammamish-USPS ZIP Code datasets). The largest difference of 3.3 km in route length was found between the Census and Fairfax County datasets. The largest difference in corresponding route lengths between the Census and the USPS datasets is 2.3 km.

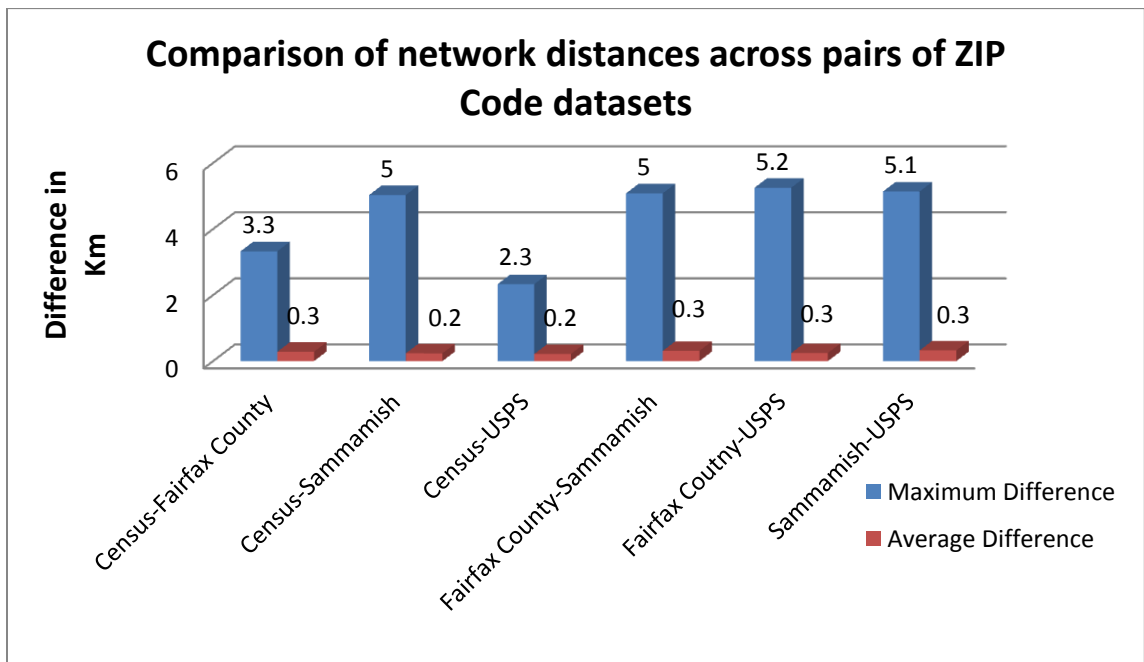


Figure 16: Comparison of network distance between ZIP Code datasets

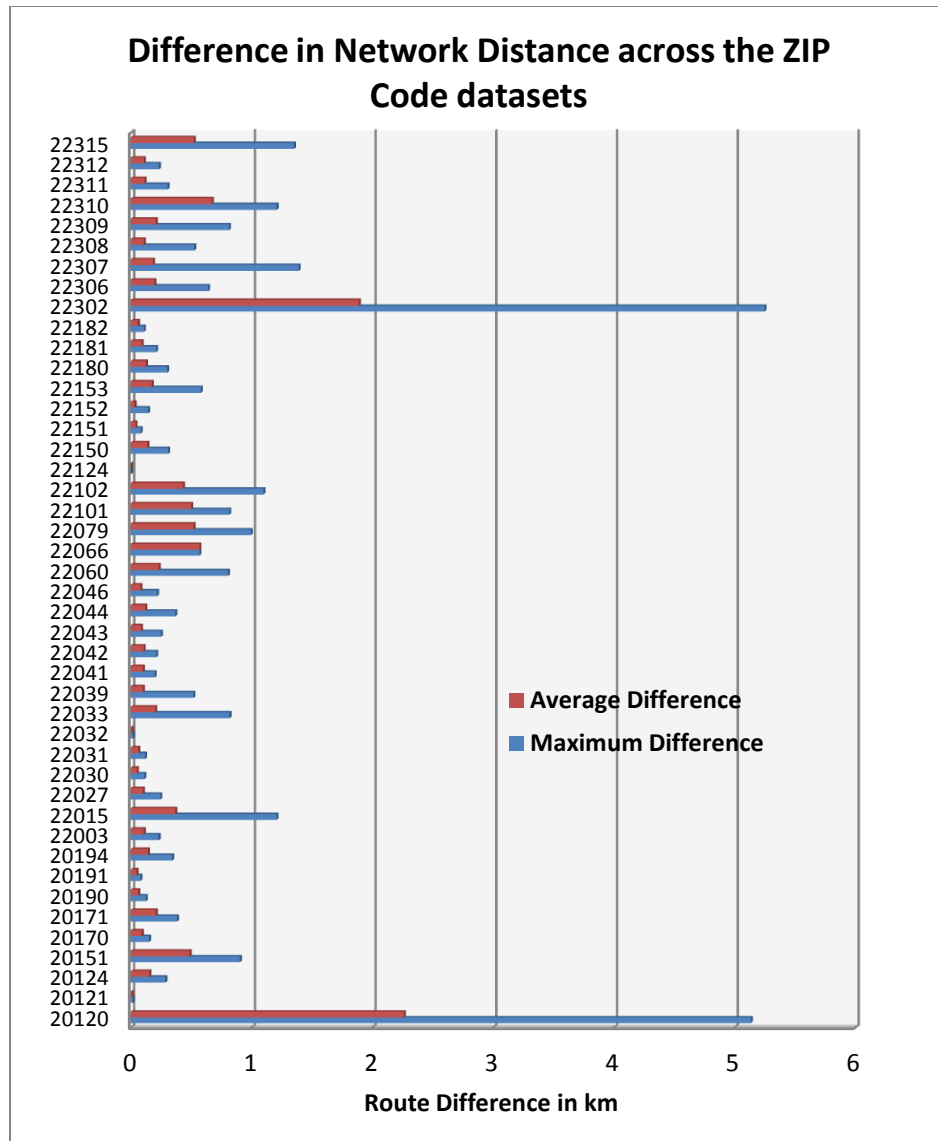


Figure 17: Comparison of network distance within ZIP Code maps

Figure 17 summarizes the differences in route lengths across all the datasets. These are the differences that can occur for a ZIP Code regarding the route lengths across all the datasets. ZIP Code 22302 and ZIP Code 20120 have the highest differences of 5.2

km and 5.1 km respectively. On average, the route lengths of these two particular ZIP Codes vary by 1.9 km and 2.3 km respectively across all 4 datasets. There are several other ZIP Codes for which multiple routes, created from those ZIP Codes to hospitals, vary by about one km or more across the datasets.

6.4.2 Deviations in network distance from the USPS ZIP Code map

Figure 18 summarizes the percent change of the network distances between hospitals and centroids of the ZIP Codes within different datasets compared to that within the USPS dataset. The X-axis plots the percentage changes across a number of bins and the Y-axis shows the frequency of routes within a dataset for which the changes fall in the corresponding bins. These are the percentages of the difference in route lengths within a dataset to the corresponding route lengths within the USPS dataset. The ZIP Code maps do not show any drastic percentage change from USPS on average (1.5% for all hospital locations), but the maps individually provide some of the very large fluctuations. For example; the route between the centroid of 22302 and Inova Mount Vernon Hospital within the Fairfax County map differs by 55% of the corresponding route within the USPS map which is the largest percentage change across the datasets. Several other ZIP Codes within the Fairfax County dataset vary by more than 20% in route lengths from the USPS. The highest percentage change within the Census dataset compared to the USPS dataset is recorded for the route between the ZIP Code 22033 and Inova Fair Oaks Hospital. This change is about 43% of the corresponding route length within the USPS dataset. In the Sammamish map ZIP Code 22315 contributes to the

largest percentage change (about 26%). Most of the ZIP Codes show changes less than 10% in all three datasets. In the Census, Fairfax County and Sammamish, there are 9, 6 and 5 ZIP Codes respectively which differ by 11-20%. For the same order of series, a count of 2, 6 and 4 ZIP Codes change by 21-30%. Only one ZIP Code in Fairfax County changes by 31-40% from the USPS. Each of the Census and Sammamish datasets also has a single ZIP Code for which the network distance varies by 41-50% and 51-60% from the corresponding distance in the USPS.

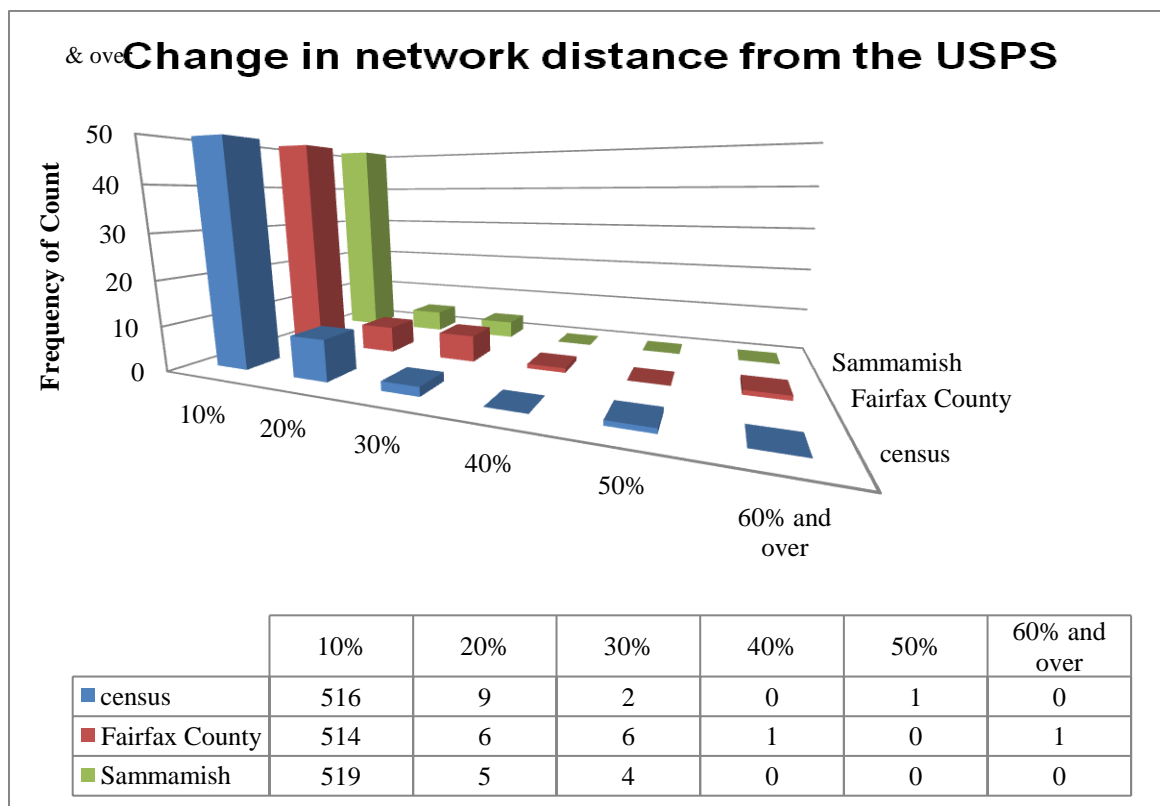


Figure 18: Number of routes that change by different percentages within datasets in centroid-hospital network analysis

These results suggest that using ZIP Code polygon maps can confuse estimates of network distance between two points of interest since the centroid positions vary according to the area and shape of ZIP Code polygons collected from different data sources. Since it is not known which interpolation method these sources use to create ZIP Code polygon maps and since these maps are continually being updated without documentation regarding those changes, the use of centroids of these polygons can produce variable distances for the same ZIP Code. This calls into question the results of the analyses that employ these distances.

SECTION 7

SEGREGATION OF HISPANIC POPULATION

Segregation is an issue that has been studied extensively in geographic research. Often very strong relationships are found regarding segregation by race and different socio-economic variables as well as different health and demographic issues (Wallace 2003; Haas et al. 2008; Orfield and Lee 2005). ZIP Codes are often used as the geographic unit to characterize and analyze the pattern of segregation and its relationships to different variables (Dai 2010; Inagami et al. 2006). This study aims to discover the scope of potential problems that can be generated by measuring segregation of the Hispanic population at ZIP Code level and examine if the outcome can vary according to the data source used for measuring segregation level.

The following segment briefly reviews the literature in the study of segregation using ZIP Codes which is followed by a discussion of the data collection and compilation process. Section 7.3 describes in detail the methodologies used in the study for determining segregation level within ZIP Codes and comparisons of the results across the datasets and section 7.4 discusses results from the study in detail.

7.1 Literature survey

Segregation has not only been linked to race, also, it has also been systematically linked to other forms of segregation. These forms include- segregation among

socioeconomic status, residential location, language, educational institutions, migration patterns, segregation of patient or disease or patterns of disease over time, commercial and industrial markets and many other criteria.

Lankford and Wyckoff (2006) studied the relationship between the racial segregation in elementary and secondary schools and choices of residential locations and schools of white parents. Donato and Garcia (1992) examined the segregation of language in the public schools and how these schools respond to limited English proficient students. Orfield and Lee (2005) discussed the relationship between segregation by race and poverty and teacher quality, test scores and dropout rates in metro Boston. Friedman et al. (2005) examined the residential pattern of immigrant newcomers in Washington D.C. and how the pattern is influenced by their races and ethnicities at the ZIP Code level. Admitting the shortcomings of using ZIP Codes, they went on to use ZIP Codes for measuring segregation as data on new immigrants were available only at the ZIP Code level. Wallace (2003) examined the change in pattern of AIDS incidents over time and influence of race and ethnicity over the disease spreading within the ZIP Codes of New York. Marion (2009) investigated the location of minority owned firms in the highway construction industry in California and how affirmative action may affect the success of firms located in neighborhoods with high segregation of minority residents.

Dai (2010) evaluated the role of black residential segregation on the late-stage diagnosis of breast cancer in metropolitan Detroit and discussed different socio-economic characteristics associated with the segregation. Walton (2009) examined the influence of

segregation on birth weight among Asian, Black, and Latino Americans in an urban environment using two dimensions of segregation: residential isolation and clustering. Inagami et al. (2006) examined the relationship of racial and ethnic segregation with mortality rate within the ZIP Codes of New York City. Rodriguez et al. (2007) studied the relationship among racial composition of ZIP codes in metropolitan areas, the characteristics of dialysis facilities and the outcomes of patients receiving dialysis. Halla et al. (2008) discovers the association of racial segregation with poverty influencing the rate of kidney transplantation among end-stage renal patients within ZIP Codes in the Pacific coast region of the US.

The standard of life of the residents of a segregated region is closely related to the level of segregation of races within that region. Segregation level of a population group also defines the pattern of poverty (Orfield and Lee 2005); health conditions (Dai 2010; Rodriguez et al. 2007; Inagami et al. 2006) educational quality (Orfield and Lee 2005); different socio-economic status (Dai 2010; Haas et al. 2008); and crime occurrence (Shihadeh and Maume 1997). ZIP Codes are very frequently used in research to measure segregation and to evaluate the links between the level of segregation and various factors as mentioned above. The measurement of segregation faces the common challenge of data availability in spatial analyses (as discussed above); particularly when the primary data used in such analyses, are collected and maintained only at the ZIP Code level (Beyer et al. 2011; Bonner et al. 2003; Haas et al. 2008). Based on this review it is clear that ZIP Code spatial representations are frequently used in research regarding spatial segregation. The following sections examine the potential consequences of that reality.

7.2 Data and Study area

This study is centered in Fairfax County (Figure 19, drawn at a scale of 1:250,000) in order to evaluate the discrepancy in the segregation results that can occur when using different data sources. Earlier studies (Friedman et al. 2005) show Fairfax County as a part of the region that received more than 90% of new immigrants in the Washington DC metropolitan area. The percentage of foreign-born residents in the county is more than twice that found nationally (Fairfax County Government, 2011). The County has more Asian and Hispanic immigrants than black immigrants.

The population data at the ZIP Code level is estimated from the 2010 census tract level population data. This study uses population data of 2010 obtained from <http://www.census.gov> for examining the level of segregation of the Hispanic population at ZIP Code level. The ZCTA boundary of 2000 is a little different from the ZCTAs of 2010. Hence, the population data at the census tract level is also used to estimate population in ZCTAs in the Census dataset.

For Census 2010, the minimum threshold for population in a census tract is 1200; therefore any population less than 1200 is not recorded. For example; if a census tract has 1199 people for a population then there will be no population data for that tract. This is why some Census tracts have no population data. For example; the area of ZIP Code 22091 (in the Fairfax County dataset) or 20041 (in the USPS) has some tracts that have no population data available. Much of this area is controlled by the Washington Dulles International Airport (Figure 19). Hispanic population data are also unavailable for some

tracts as the population threshold for race is 100. So the census tracts that have less than 100 people of a certain race will show no population data by race.

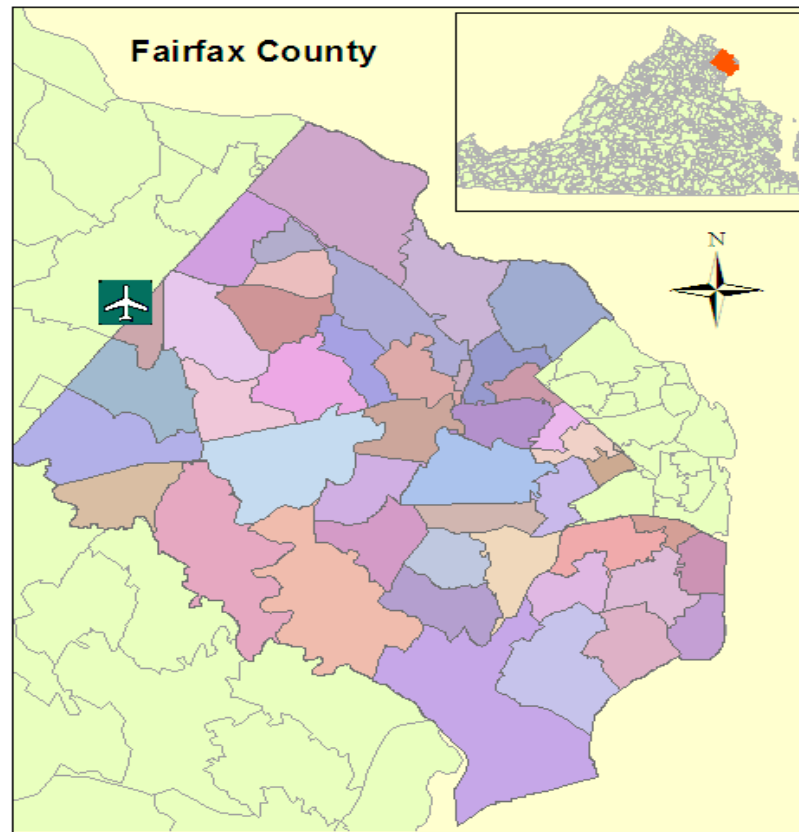


Figure 19: Study area of Fairfax County

7.3 Methods

7.3.1 Segregation index

A variety of indexes have been proposed (Wong 2005; Chang 2006) to efficiently capture multiple dimensions that comprehensively evaluate the level of segregation.

There are five distinct dimensions of segregation explained by Massey and Denton

(1988): unevenness, isolation, centralization, concentration, and clustering. This study employs the isolation dimension to find out if there is any racial or ethnic segregation in Fairfax County and nearby areas. This dimension is particularly efficient for examining racial segregation in areas where that race is predominant. In Fairfax County the white race is dominant (55%) but the number of people from other races has been increasing in recent years (approximately 18% Asian and 16% Hispanic). According to the Fairfax County Community Health Assessment Report (2011), 34% of Fairfax County households speak a language other than English at home. Over 100 different languages are spoken at home by students enrolled in Fairfax County Public Schools (FCPS).

Although the Hispanic population alone is not predominant in this area, together with other racial and ethnic minorities, Fairfax County is becoming a gateway for new immigrants. This study examines the Hispanic population to reveal if any segregation exists for this ethnic group and if so, how the pattern of segregation changes when ZIP Code polygons are used over a range of datasets. The objective of this study is to find out whether ZIP Codes from different datasets can create uncertainty in a spatial analysis for segregation of population groups.

The isolation index refers to the extent to which a member of a minority group comes into contact with members of the same group, compared with residential neighbors in the same unit (Dai 2010; Chang 2006). This index is able to effectively reveal the differential sizes of segregation (Chang 2006). The index not only depends on the percentage of a group of people, but also identifies the extent to which minority members are exposed to each other (Walton 2009). For example, if a region consists of an equal

number of Hispanic and other ethnic groups where all Hispanics live close to each other on one side of that region, the isolation index for that region will be high. A high index value indicates that a Hispanic resident would have a high chance of having other Hispanic residents as neighbors. Thus, this index can reveal the likelihood of these people living close together. Therefore, this study uses the isolation index for segregation measurement. While the outcome of a segregation study can be different based on the index used (Massey and Denton 1988; Chang 2006; Wong 2005), it is not the purpose of this study to perform a comprehensive segregation analysis. The point here is to demonstrate whether or not a specific segregation index will generate variable results based on the choice of ZIP Code representation.

The isolation index (R_j) is solved using Equation 3. Assuming ZIP Code j consists of n census tracts, the segregation index for the Hispanic population within ZIP Code j will be:

$$R_j = \sum_{i=1}^n \frac{H_i}{H_{total}} \times \frac{H_i}{T_i} \quad \text{Equation 3}$$

where i is the i th census tract in ZIP Code j , H_i is the Hispanic population in i , H_{total} is the total Hispanic population in j , and T_i is the total population in i . For a ZIP Code consisting of two or more census tracts, the ZIP Code is overlaid with the tracts and

the population is proportional to the partial census tracts falling within a ZIP code using the areal weighting interpolation method.

Ranging from 0 (no segregation) to 1 (the highest segregation), the isolation index can be interpreted as the chance of having Hispanics as neighbors (Dai 2010; Haas et al. 2008). It evaluates whether the Hispanic population concentrates in a subunit of an area using H_i / H_{total} and how Hispanics and other groups are mixed together in this sub unit using H_i / T_i . Higher numbers of the index indicates higher level of segregation suggesting that the Hispanics would be most likely to have other Hispanics as neighbors.

7.3.2 Detailed implementation of the segregation index

Since the census statistics are unavailable directly for individual ZIP code, this study interpolates population statistics from the census tract to the ZIP Code level using an ‘Areal Weighting Interpolation’ method as described by Goodchild and Lam (1980) and Wang (2006). This overlay method is widely practiced in spatial research to estimate statistics for an area when no direct statistics are found for that area (Wilson and Mansfield 2010; Eicher and Brewer 2001). The detailed methodology of the overlay process of the census tracts and ZIP Codes is described in Figure 20. These methods are implemented in ArcGIS by using a series of “overlay”; “join” and “sum” functions. The ZIP Code map from the Census dataset is used to illustrate the methods. The data of total population (TotPop) and Hispanic population (TotHisPop), collected from the Census Bureau, is joined to the census tract map by ‘Tract’. This map is overlaid on the ZIP Code map by using the analysis tool “Intersect”. The population (IntTotPop) and

Hispanic population (IntHisPop) for the intersected areas are calculated by the proportions of each of the tracts that are inside particular ZIP Codes.

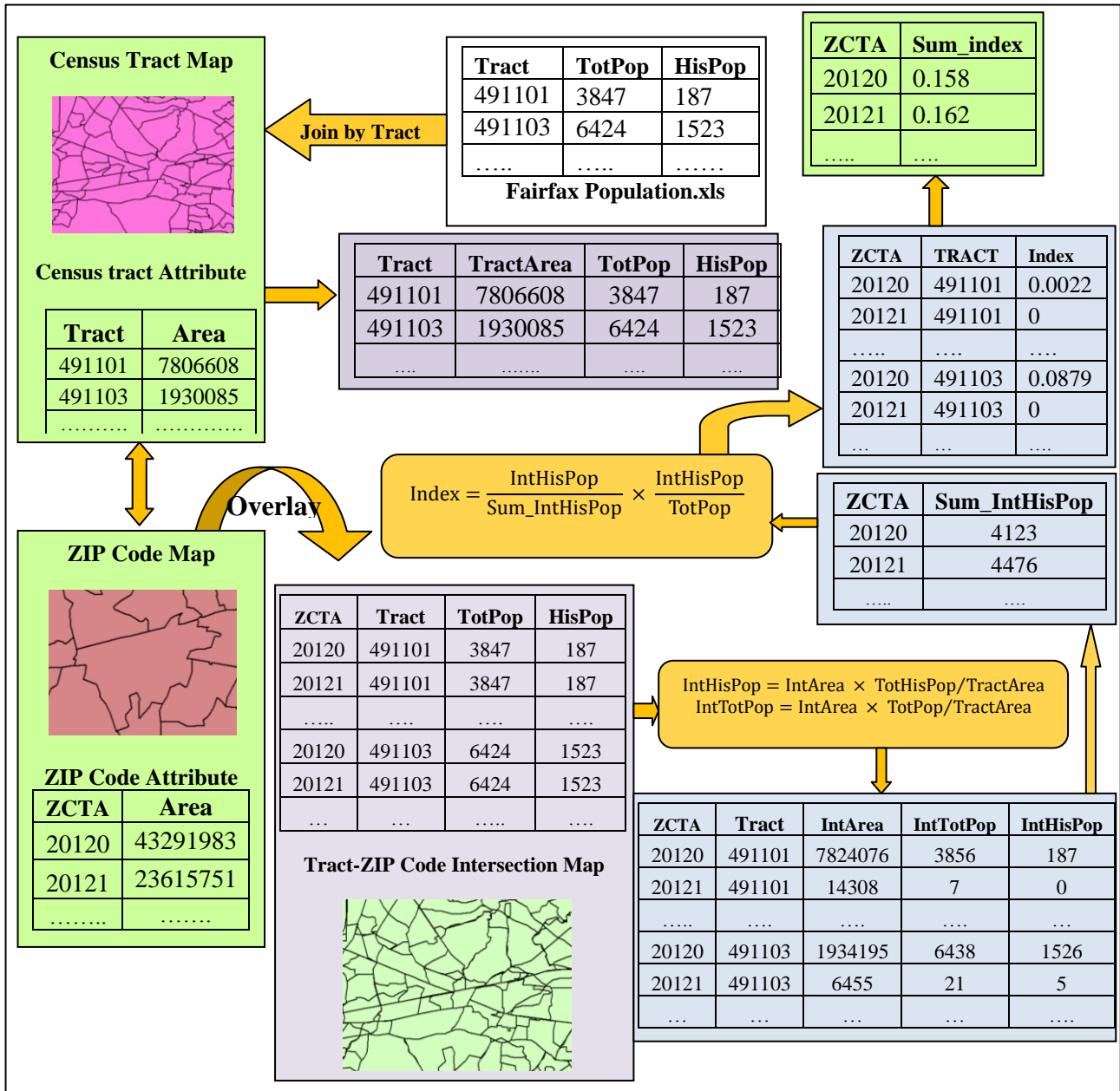


Figure 20: Implementation of Areal Interpolation method and isolation index of segregation.

Total Hispanic population for a ZIP Code is summed to calculate the isolation index of segregation for each of the intersected areas. Finally, for each of the ZIP Codes, the indices of the intersected area with the census tracts are summed to get the total segregation level in ZIP Codes.

7.4 Results

Due to some difficulty in overlaying the census tracts with the ZIP Codes 22180 within the Sammamish dataset, this ZIP Code has been excluded from the analysis. Figure 21 shows the comparison of the level of segregation of the Hispanic population for 43 common ZIP Codes across the datasets. Segregation indices are classified into 5 quantiles which contains ZIP Codes with subsequent values of segregation indices. This method classifies the segregation indices into 5 categories with an equal number of ZIP Codes in each category. The 1st quantile contains the ZIP Codes with the lowest 20% of segregation indices. ZIP Codes within the subsequent quantiles have segregation indices in a lowest to highest succession.

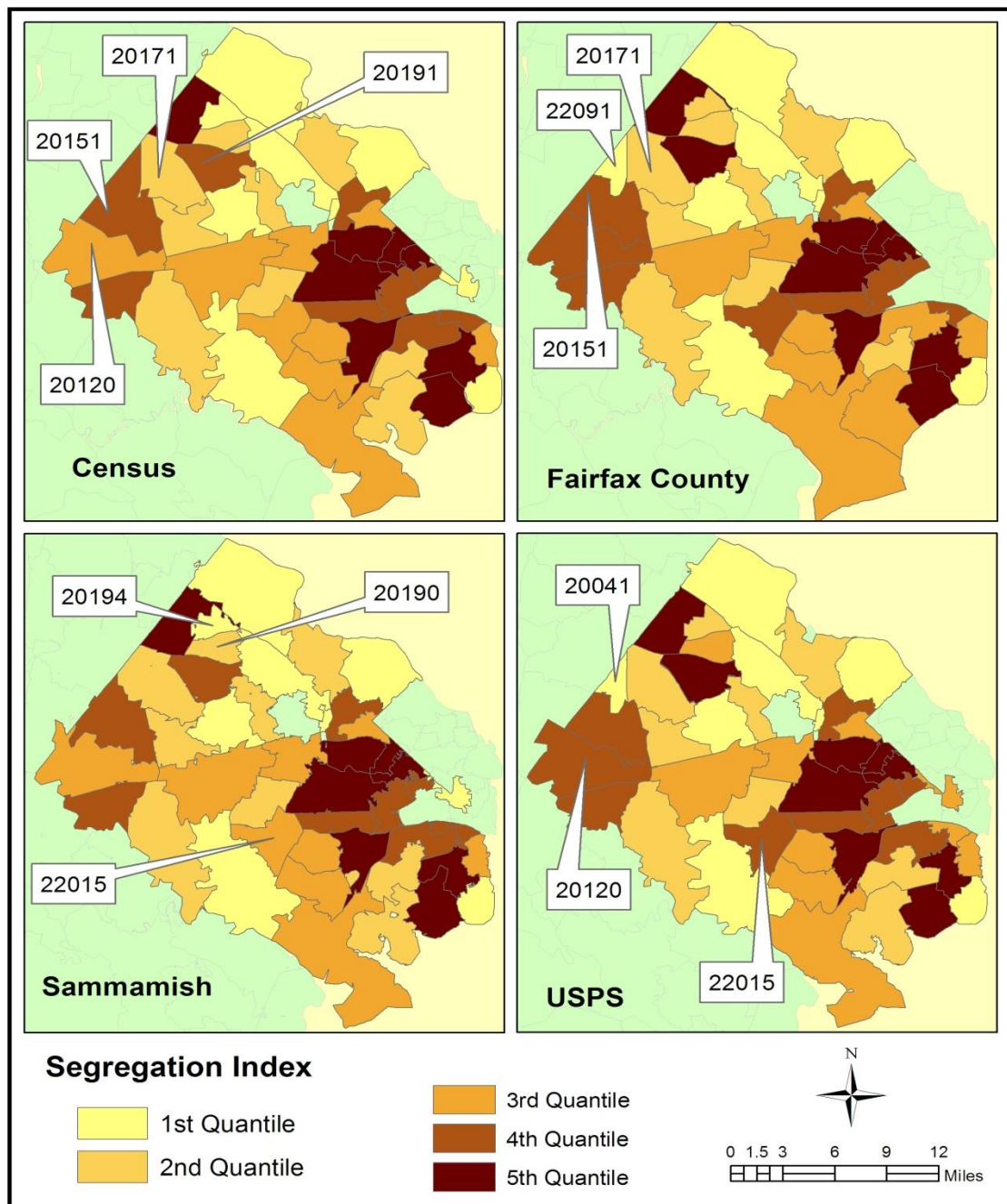


Figure 21: Segregation of Hispanic population in Fairfax County at ZIP Code level within the Census, Fairfax County, Sammamish and USPS datasets.

As noted earlier, together with other nearby counties (e.g. DC, Prince Georges, Montgomery, Arlington, and Alexandria), Fairfax county has been known to have significant segregation of newly migrated populations (Friedman et al. 2005). As a subset of the whole region, Fairfax County may not have very high level of segregation for Hispanic population alone, yet it can give an idea how Hispanic people are likely to have other Hispanic people around them. It is noted that the study area is only to manifest the problems in spatial analysis on ZIP Code polygons. Therefore, the segregation levels themselves are not of primary interest, only the differences in segregation levels across different ZIP Code datasets.

As anticipated, in all of the datasets the highest level of segregation is below 0.6 which can be identified as a medium level of segregation on a 0-1 scale. However, the segregation level varies across different datasets. For example; ZIP Code 20191 is in the 4th quantile within the Census and Sammamish datasets but in the 5th quantile within Fairfax County and USPS datasets. ZIP Codes 20120 and 22015 within the Fairfax County and the USPS dataset fall within the 4th quantile but in the 3rd quantile within the Census and Sammamish datasets. Again, ZIP Code 20191 within the Fairfax County and the USPS dataset is among the top 20% most segregated ZIP Codes being in the 5th quantile but within the other two datasets this ZIP Code is in the 4th quantile. The indices are also different for the ZIP Code 20190 and 20194. This gives a clear understanding that the datasets vary from each other in terms of segregation of Hispanic population.

A direct comparison for some of the ZIP Codes is difficult across the datasets when the ZIP Codes are present in a dataset but not in other datasets. For example; ZIP

Code 22091 in the Fairfax County dataset or ZIP Code 20041 in the USPS are within the 20% least segregated ZIP Codes but the same area falls within the 3rd quantile when part of ZIP Code 20171 in the Sammamish dataset and in the 4th quantile when part of the ZIP Code 20151 within the Census dataset.

Table 5: Percent change in segregation level within the datasets compared to the USPS dataset

	Census	Fairfax County	Sammamish
Max percent change	288	323	286
Min percent change	35	31	35
Average percent change set)	101	102	102

Table 5 summarizes the percentage of the difference in segregation level within individual datasets compared to the USPS dataset. The segregation indices within the datasets may not seem changed from the indices within the USPS dataset on average, however the changes for some of the ZIP Codes are very large. The highest percent changes for ZIP Codes are within the range of 280% to 330% than the corresponding segregation values of the USPS dataset. ZIP Code 22302 within the Fairfax County has a segregation index of 0.04 which is a more than 300% reduction from the segregation index of 0.13 for the corresponding ZIP Code within the USPS dataset.

Table 6: Pair-wise comparison among datasets regarding the level of segregation of Hispanic population.

Pair of Datasets	% of segregation indice within the Census dataset		% of segregation indice within the Fairfax County dataset		% of segregation indice within the Sammamish	
	Max	Average	Max	Average	Max	Average
Census-Fairfax County	111	98	117	102		
Census-Sammamish	124	98			136	102
Fairfax County-Sammamish			102	100	143	100

Table 6 presents the pair-wise comparisons of datasets regarding segregation level of Hispanic population. The segregation indices are relatively similar for the Census-Fairfax County pair. The largest change in segregation index for this pair is about a 111% change from the index within the Census or a 117% change from the index within the Fairfax County dataset. The largest difference for the Census-Sammamish pair is more than 124% from the index within Census dataset or 136% from that within the Sammamish dataset whereas for the Fairfax County-Sammamish pair the difference is about 143% from the index within the Sammamish dataset.

Table 7 summarizes the percentages of the ZIP Codes that switched to a different quantile class when the segregation indices are measured using different datasets rather than the USPS dataset. For example; the USPS dataset has 9 ZIP Codes within the 3rd quantile out of which 2 ZIP Codes have changed their quantile classes across other datasets. In other words, about 22% or one fifth of the ZIP Codes within this quantile has changed their quantile classes in other datasets. More than 16% of the total ZIP Codes (7

out of 43 total ZIP Codes) within the USPS dataset have switched to an another quantile class within the Census dataset. More than 9% and 14% of the total ZIP Codes altered their quantile classes when the segregation level is measured using the Fairfax County and Sammamish datasets respectively.

Table 7: Percentages of the ZIP Codes that switched to a different quantile class across datasets.

% of total ZIP Codes	Census	Fairfax County	Sammamish
1st quantile	17	0	0
2nd quantile	11	11	11
3rd quantile	22	22	22
4th quantile	22	11	22
5th quantile	11	0	11

These results from the measurement of segregation level of the Hispanic population continue to support the hypothesis that a spatial analysis can have variable outcomes according to the data source used for the analysis. Admitting the study area is not highly segregated for Hispanic population alone; the analysis reveals contrasting results for different data sources. This indicates that if this study were to be conducted for all the minority ethnic groups or for all new immigrants in this region or if it would consider some other counties known to have high segregation of minorities (e.g., DC, Prince Georges, Montgomery, Arlington, or Alexandria); the use of ZIP Code polygons

to measure the level of segregation as well as choice of segregation indices would create even more contrast across alternative data sources.

SECTION 8

RANKING OF ZIP CODES

The ZIP Code is a common geographic unit that has been extensively used in spatial analysis. It has been adopted by marketing people and by many other researchers as a standard geographic area, like a city or a county. Analysis has been performed on housing markets and different variables that affect the market of a region (Nagaraja, Brown, and Zhao 2011; Shan 2011; LaCour-Little, Calhoun, and Yu 2011), employing ZIP Codes to describe the characteristics of these relationships. In research on school quality and availability (Horowitz, Keil, and Spector 2009); students' and teachers' characteristics (Fuller and Strath 2001; Lankford and Wyckoff 2006) and many other topics, ZIP Codes are used as the basic geographic entity for analyses.

The quality of public schools is often cited as an important attribute which distinguishes a community (Clark and Herrin 2000; Zahirovic-Herbert and Turnbull 2008). Parents' perceptions of the neighborhoods they live in also have the potential to influence many decisions they make with respect to their children (Carson et al. 2010). Information about property and the school quality within an area is readily available from real estate agents and online. To attract potential buyers, they prominently feature school quality information along with other important house and neighborhood characteristics (Horowitz, Keil, and Spector 2009). Many online real estate websites provide the option to the home buyers to search for new homes based on some variables the buyer chooses.

Even when sellers are not inclined or required to offer information about school quality, this information is widely available to the public. This gives an opportunity for parent home buyers to include school quality measures when determining the value of a particular house.

Many of these websites as well as economic research often use ZIP Codes as the geographic unit to describe the characteristics of a neighborhood (Dan Immergluck 2011; Shan 2011). When searching for new homes in areas that are served by good quality school districts, a prospective home buyer would probably prefer the high ranking ZIP Codes if the ranking is done based on the quality of the available schools within the ZIP Code boundary.

This section studies more examples of spatial analyses that frequently use ZIP Codes as the unit of observation. In order to do so, the average housing prices of Fairfax County are estimated within the ZIP Code polygon areas. The ZIP Codes have also been ranked based on the ranking of schools that serve the ZIP Code area. It is believed that quality schools attract potential new home buyers with a consequent bidding up of the residential property value near the highly ranked schools (Horowitz, Keil, and Spector 2009). Therefore, the ranking of the ZIP Codes have been compared with the market value of properties within this region.

The logic behind the ZIP Code ranking is that the ranking of the schools available within a ZIP Code polygon boundary will affect the desirability of that ZIP Code as a choice for the purchase of a home. If a ZIP Code has higher ranking schools within its boundary area, it is given a higher rank and therefore a higher preference for home

searching. The influence of the school quality on the overall ranking of a ZIP Code will be proportional to the area it serves within that ZIP Code polygon. If a website provides real estate information based on ZIP Codes including the property price and ranking of ZIP Codes for having quality schools, it would let buyers have an understanding as to which ZIP Code has the higher accessibility to high ranking schools and thus would influence their decisions.

8.1 Literature survey

Many public and private organizations rank schools based on examination results and test scores. For example; in the UK, performance of schools are indicated by school rankings published in a 'league table' by the Department for Education and Skill. School rankings are done based on previous test scores over a span of time to predict future school performance and these rankings seriously guide parental choices of schools for their children (Leckie and Goldstein 2009).

Several studies have identified public school quality as a significant determinant of locational choice and property values (Walden 1990; Hayes and Taylor 1996; Clark and Herrin 2000). Studies determining the empirical relationship between school quality and housing prices suggest that parents are willing to pay high value for good schools (Bayer, Ferreira, and McMillan 2007). Even for buyers and owners who don't have school age children, good schools can ensure consistent demand for properties and high return (Max 2010). Areas with good schools tend to be more affluent and vice versa and these areas are less susceptible to mortgage collapse. Even in bad housing markets,

homes in an area associated with great schools generally sell faster than areas with lower ranked schools (Max 2010; Cellini, Ferreira, and Rothstein 2010).

As school quality is an important cause of differences in the prices of residential houses (Horowitz, Keil, and Spector 2009; Haurin and Brasington 1996; Mitchell, Batie, and Mitchell 2010; Carson et al. 2010), failure to consider school quality resulted in a substantial underestimate of the influence on market price (Jud and Watts 1981).

Elementary school test scores are significantly and positively correlated with single-family home prices, controlling for house characteristics, neighborhood effects, and school racial composition (Shan 2011). The location and quality of schools also influence as well as is influenced by the income of the residents living within the school district (Fuller and Strath 2001).

Fuller and Strath (2001) analyze the demographics, earnings, and unequal distribution of the workforce in schools according to race and income and found that inequalities in the supply and quality of early educational organizations and their staffs are related with the economic status of residents within a ZIP Code. Similar work of Fuller and Liang (1996) suggests that distribution and quality of schools are associated with household income, parental education and other demographic characteristics of households within a ZIP Code. Many other studies have been conducted for estimating inequality in school availability and school quality according to socio-economic structures.

However this type of school rankings based on school performance can be misleading. Leckie and Goldstein (2009) discussed that the league tables that measure the

quality of schools based on previous test scores, have no statistical adjustment for the uncertainty arises from predicting future school performances. Using a multilevel model of school effectiveness adjusting for predicting uncertainty, they found that previous school performances cannot predict future school performances. Since most of the schools are statistically not different from the overall mean performance and therefore, these rankings cannot differentiate between future performances of schools.

In much of the research, ZIP Codes are used to analyze the spatial location and characteristics of school districts and the relationship of the housing market with school quality. Shan (2011) analyzed the characteristics of ZIP codes to identify the locations and aspects of reverse mortgage borrowers. Nagaraja, Brown, and Zhao (2011) used ZIP codes to model property sale price over time and location. Kiel and Zabel, 2008 also analyzed the effect of school quality on home price at the ZIP Code level. Hayunga and Pace (2010) discussed the spatial correlation of the location of commercial real estate property with its distances to ZIP Codes. LaCour-Little, Calhoun, and Yu (2011) analyzed the house loan performance and house price appreciation. Pollack et al. (2011) examined the relationship of health and foreclosures while Immergluck (2011) found links between credit score and ZIP Code level characteristics of housing price trends, neighborhood demographics, and other factors. Clearly, ZIP Codes are the spatial representation of choice for a wide range of property and school related research in the United States.

However, while ranking the schools based on previous performances using ZIP Code polygons, an analysis will be done using an invalid geographic unit for predicting

school performance in a way that is statistically not sound. Therefore, it will be compromising two very important criteria: a valid of spatial unit and a valid statistical measurement.

8.2 Study area and data

The study area is the same area of Fairfax County that was used in the previous studies. The ZIP Code polygon maps used in this study (polygon representations of Census ZCTA, Fairfax County ZIP Codes, Sammamish Geocode and USPS ZIP), are collected from the sources mentioned in earlier studies of this thesis paper. The information of school location and school attendance area boundary are collected from the website <http://www.fairfaxcounty.gov/>. The Fairfax County public school board does not rank any school, yet, there are many websites available which rank the schools based on the overall school performance or test scores. The school ranking data has been collected from <http://www.greatschools.org/>.

The attendance areas of 139 elementary schools out of a total 142 schools in the Fairfax County public school website data have been used for this analysis. The attendance boundary is unavailable for three of the elementary schools. This is because the school board assigns a boundary for a school after it has a name that would not change further. For example: the elementary school location data has a school named as ‘Coppermine’ which is also present in the school attendance area boundary layer but in a different name as ‘Coates’. As the school does not have any definite name, it has been given no boundary area and thus is excluded from the analysis.

The Census Bureau primarily created the ZCTA boundaries for the Census 2000 and has updated these boundaries for the Census 2010. However the old ZCTA boundary polygons that were collected for the year of 2000 have been used throughout the thesis work. As the census housing data at ZCTA level changed with the updated ZCTA boundaries, it is problematic to directly use any information from the updated ZCTA boundaries. Fairfax County also creates and maintains the housing information for the ZIP Codes that have also been updated in recent years. Besides, there is no direct information available for the ZIP Code polygons of the USPS and Sammamish. Therefore, property market values on the ZIP Code polygons have been interpolated from the housing information of census 2010 at the tract level, acquired from the website <http://www.census.gov/> rather than using any direct information from the Census Bureau or Fairfax County Website at the ZCTA or ZIP Code level.

There are 44 common ZIP Codes across the datasets out of which 43 ZIP Codes were compared. ZIP Code 22080 in the Sammamish dataset has some problems while overlaid with the census layer and therefore is excluded from analysis.

8.3 Method

8.3.1 Estimating market value of houses

As discussed in earlier chapters, it is problematic to directly associate updated data with the polygon datasets, collected for the year 2000. Therefore, this study uses the new housing data of 2010 for estimating median market value of property in Fairfax County. An areal weighting interpolator method (Wang 2006) has been used here to calculate the average property value from the housing data on census tracts.

The equation used here is as follows. Assuming ZIP Code j consists of n census tracts, the average market value of houses within ZIP Code j will be:

$$V_j = \sum_{i=1}^n \frac{A_i}{A_{total}} \times V_i \quad \text{Equation 4}$$

where i is the i th census tract in ZIP Code j , A_i is the intersected area of j that falls within i , A_{total} is the total area of j , and V_i is the property market value in census tract i . For a ZIP Code consisting of more than one census tract (partly or completely), the market value of property will be proportionate to the area of census tracts falling within that ZIP code using the areal weighting interpolation method. The sum of the values (V_i) within the interpolated areas (A_i / A_{total}) represents the housing prices for the total ZIP Code area.

8.3.2 Estimating ranking of ZIP Codes

The school attendance area or school district boundaries are periodically changed by the School Board (<http://boundary.fcps.edu/>). The boundaries are created based on student capacity of schools; the number of school age children within the surrounding areas of schools and the number of schools within that region. Even though the school districts are determined with no consideration of overlapping ZIP Code areas, it is believed that property values, demographic and economic characteristics of the

neighborhood and locational choice of property are directly influenced by the availability of public schools within a ZIP Code area (Anon. 2010; Clark and Herrin 2000). In this study, ZIP Codes are ranked according to the ranking of the schools that are available within the ZIP Code boundary area.

The areal weighting interpolation method is also used here for ranking the ZIP Codes. This method is more appropriate than spatial joining of ZIP Code polygon layer with school district boundaries. Spatial joining of these data layers includes the ranking of a school within the estimation that falls completely or partially within a ZIP Code. But it does not indicate how much influence the school has on the total ZIP Code area. So, even if a ZIP Code consists of a very tiny portion of a low ranking school district, the school rank can have impact on the total ranking of the ZIP Code. On the contrary, the areal weighing interpolation method ranks a ZIP Code, based on the school ranking within the intersected area.

Assuming ZIP Code j has an area within n school districts, the average ranking of that ZIP Code j will be:

$$R_j = \sum_{i=1}^n \frac{A_i}{A_{total}} \times R_i \quad \text{Equation 5}$$

where i is the i th school district in the ZIP Code j , A_i is the intersected area of j that falls within the attendance area of i , A_{total} is the total area of j , and R_i is the ranking

of the school i . The ranking of ZIP Code j will be the sum of the product of each school rankings(R_i) and the associated intersected areas (A_i / A_{total}).

The Fairfax County public school website only has information for the schools that are located within the county boundary. In this study, no school district outside the county boundary is counted for ranking ZIP Codes. In order to avoid misjudging any information, the ZIP Code polygon maps are clipped with the Fairfax County elementary school boundary map. Due to some editing problems in the Sammamish dataset, ZIP Code polygon of 22180 cannot be overlaid with the school boundary data layer. Therefore, this ZIP Code is excluded from all types of analyses in this study. Figure 22 shows the detailed methodologies employed in ZIP Code ranking based on school rankings.

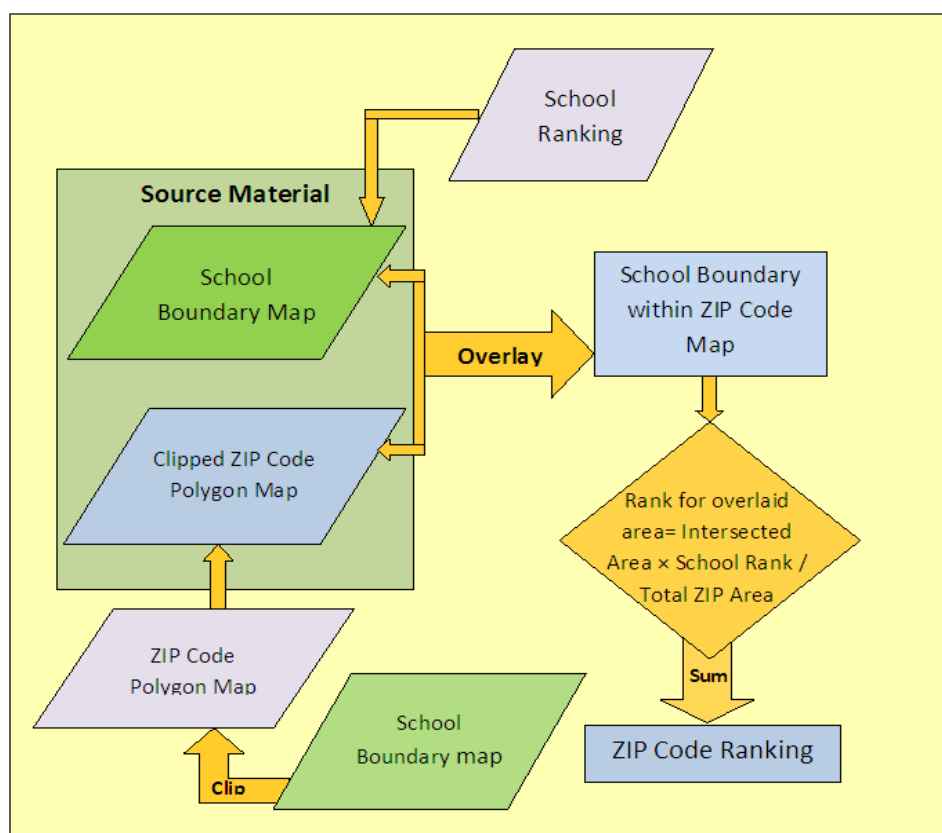


Figure 22: Detail methodology of ZIP Code ranking based on the ranking of available schools within ZIP Code polygon boundary

8.4 Results

Figure 23 shows the ranking of ZIP Codes in 5 defined classes. The ranks vary with the change of the area and shape of ZIP Codes over datasets. Whenever a district boundary of a higher ranking school falls within a ZIP Code polygon area, the rank of that ZIP Code gets higher based on the area served by that school district. In the north-eastern part of the county ZIP Code 22066, 22102, 22101, 22181 and 22182 have high rankings within a range of 9 to 11 within all of the data sources. ZIP Code 22124 has a

high ranking of 9 within the Fairfax County and USPS datasets but a medium-high rank of 7 within the Census and Sammamish datasets respectively.

ZIP Codes 20124, 22152 and 22039, located at the South Western part of the county, show the highest ranking of 9-11 due to having good quality schools within the ZIP Code boundaries. But the ranking of the contiguous ZIP Code 22015 varies due to its differential shapes and areas across the datasets. This ZIP Code has a ranking of 9 within the Sammamish and USPS but a ranking of 8 within the Census and Fairfax County datasets. So, if a home buyer searches for new homes in highly ranked ZIP Codes in the South Western part of the county and relies on an analysis that uses the Sammamish or USPS ZIP Code polygon data layers, the buyer would probably prefer either the ZIP Code 20124, 22039, 22152 or 22015. The buyer may not consider ZIP Code 22015 as his first preference, if the analysis was done on the Census or Fairfax County ZIP Code polygon data layers. ZIP Code 22152 also may not be chosen as it is not the highest ranking ZIP Code in the Census dataset. Other examples of inconsistency in ranking are ZIP Code 20120 20170, 20190, 20191 and 20194 that obtain different ranks in different datasets.

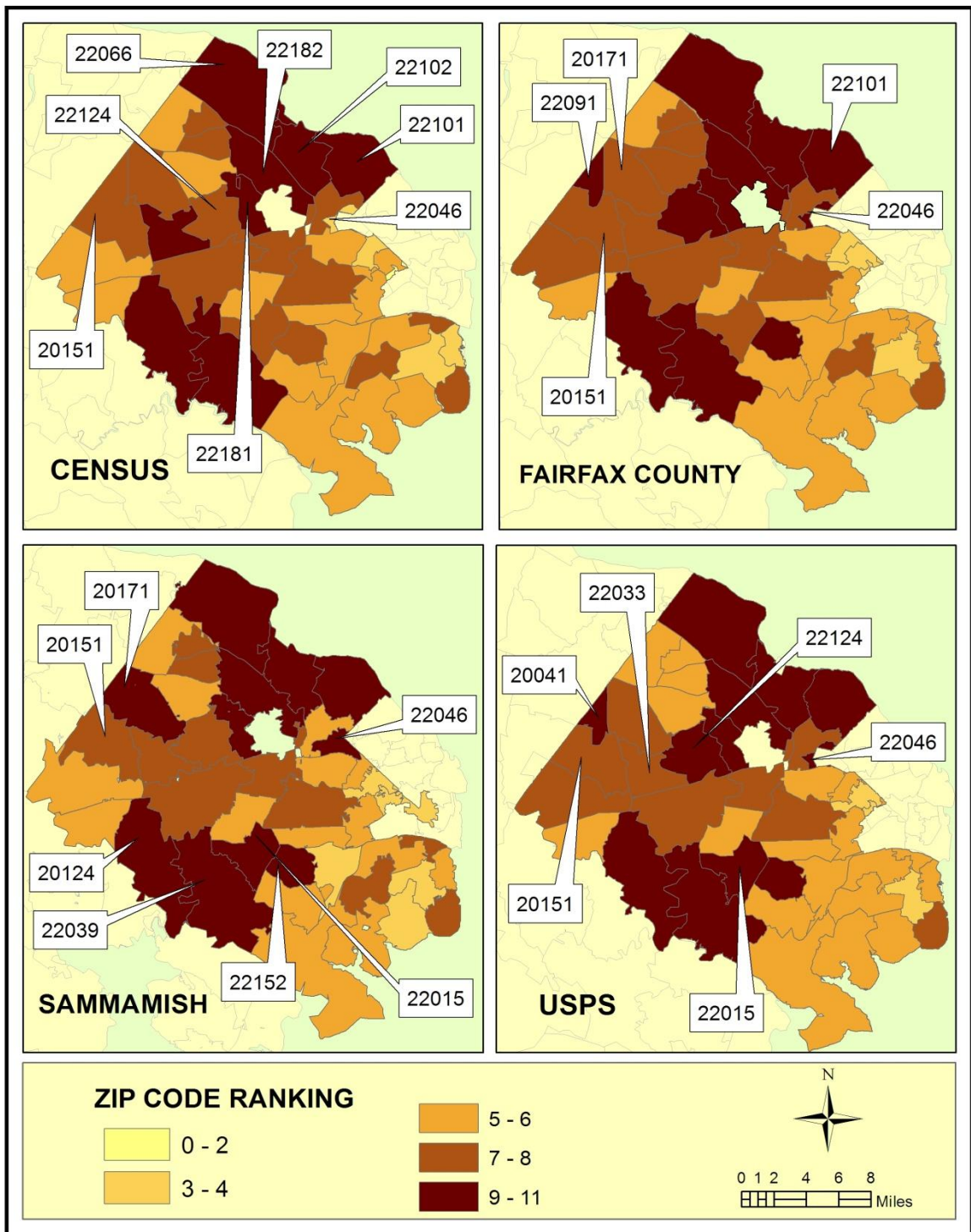


Figure 23: Ranking of ZIP Codes based on school rankings available within ZIP Codes

ZIP Code 22033 (ranking: 9) is one of the highest ranking ZIP Codes within the Census dataset while it has a medium high ranking of 8 within other datasets (

Figure 24, scale: 1:45,000).

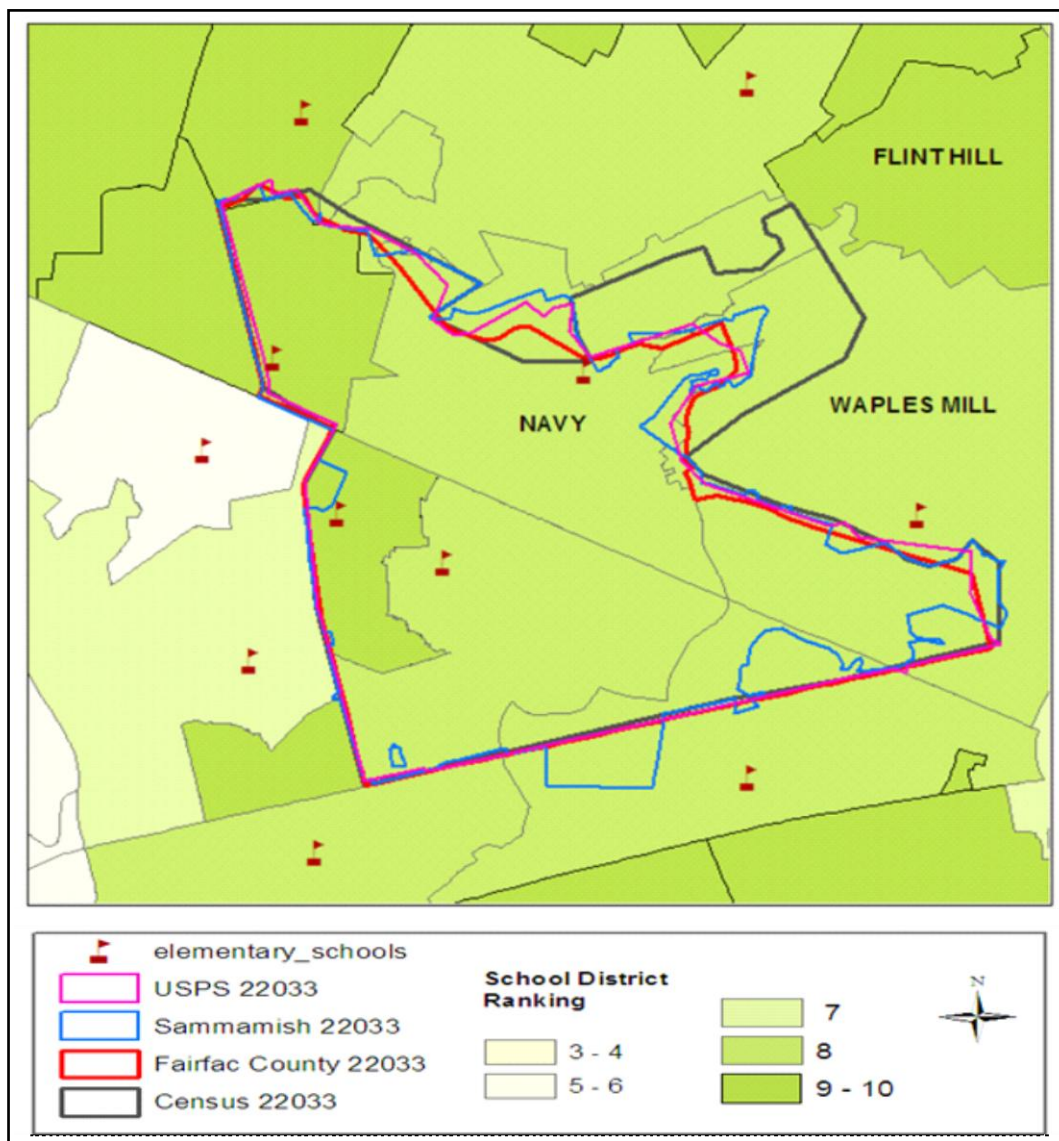


Figure 24: ZIP Code 22033 is ranked as ‘high’ in the Census dataset but ‘medium high’ in Sammamish, Fairfax County and USPS datasets

Two high quality school districts: Navy and Waples Mill school districts serve larger areas of this ZIP Code within the Census than within other datasets. A small portion of another good quality school district (Flint Hill Elementary school, school ranking: 10) falls within the ZIP Code boundary which is unique for the Census dataset and therefore the ranking of this ZIP Code gets higher.

Another example of inconsistency in ranking due to the variation of area and size of ZIP Codes across different data sources is demonstrated in Figure 25 (drawn at a scale of 1:50,000). ZIP Code 20171 has a high ranking of 9 within Sammamish dataset but a rank of 8 within other datasets. The reason for the higher ranking of this ZIP Code is the presence of the highly ranked Floris Elementary school (school rank: 9) that serves a larger portion of this ZIP Code within the Sammamish than within other datasets. As noted, the ranking of schools influences the rank of ZIP Code polygons according to the area that falls within the polygons; this high ranking school makes the rank of ZIP Code 20171 higher within Sammamish than other datasets. The Floris elementary school also covers almost the entire area of 20041 (in USPS) or 22091 (in Fairfax County). So, these ZIP Codes also obtain the highest ranks while estimating ZIP Code ranking.

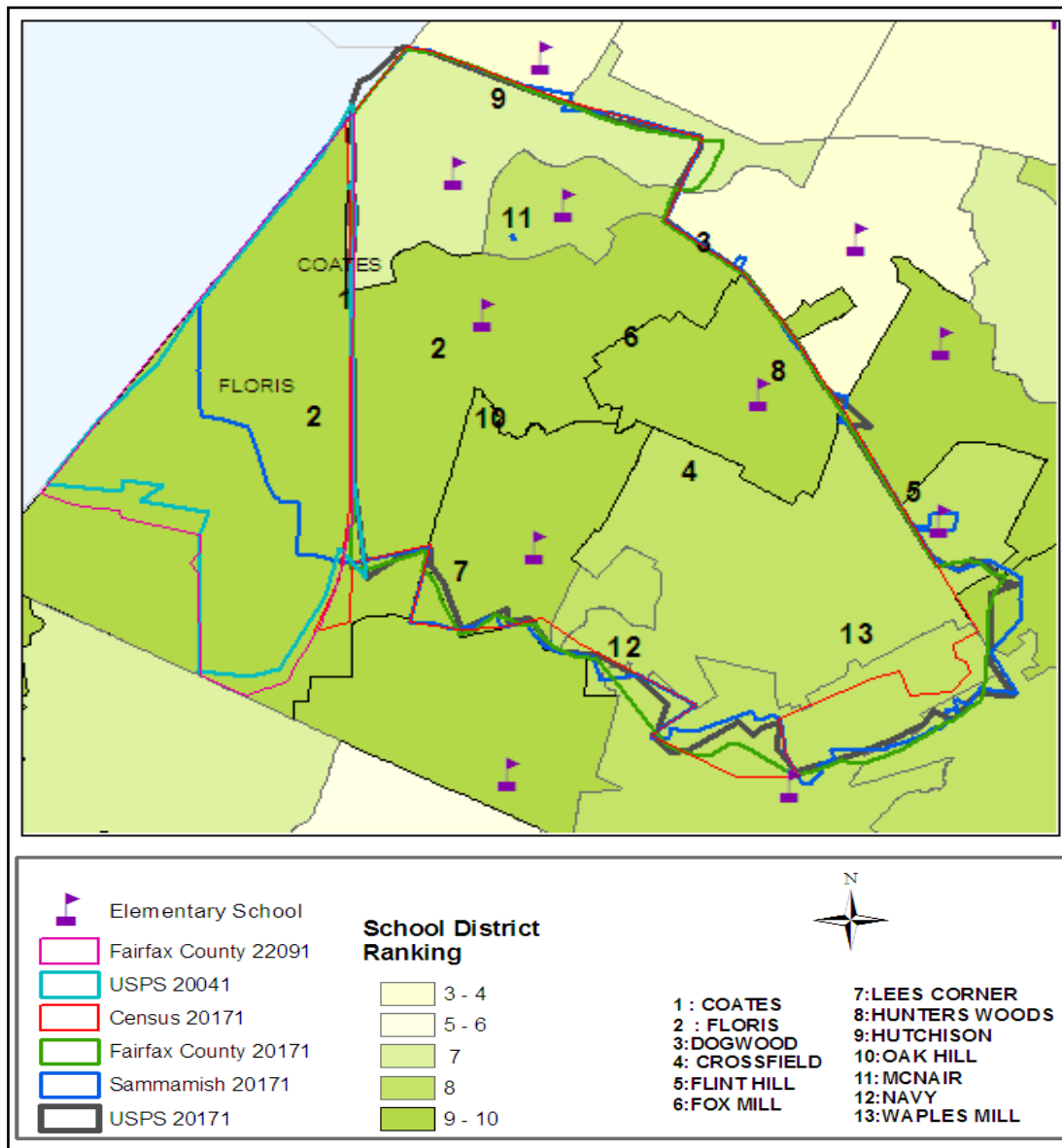


Figure 25: Influence of school ranking on ZIP Code 20041(in USPS) or 22091 (in Fairfax County) and ZIP Code 20171 in different data sources

In Figure 26 the ZIP Codes, that alter the ranking classes across datasets, are shown as percentages of the total ZIP Codes within a specific ranking class within the USPS dataset. Within the class range of 9-11 the USPS dataset has 11 ZIP Codes among

which 7 ZIP Codes are also present in the same class range within the Census dataset and 10 ZIP Codes are present within each of the Fairfax County and the Sammamish datasets. Therefore, 36% of the ZIP Codes within this ranking class range of the USPS dataset have altered the class across the Census dataset and 10% across each of the Fairfax County and Sammamish datasets. 20% and 30% of the total ZIP Codes within the class range of 7-8 switched to another class within the Census, and the Sammamish datasets respectively. 32% of the ZIP Codes within the class range of 5-6 are switched to different class ranges within each of the Census and Fairfax County datasets. 12 out of 19 ZIP Codes or 63% of the total ZIP Codes of this class range within the USPS dataset altered class within the Sammamish dataset. About one third of ZIP Codes within the class range of 3-4 change the class within the Census dataset. ZIP Codes within the class range 0-2 remains within the same class across datasets.

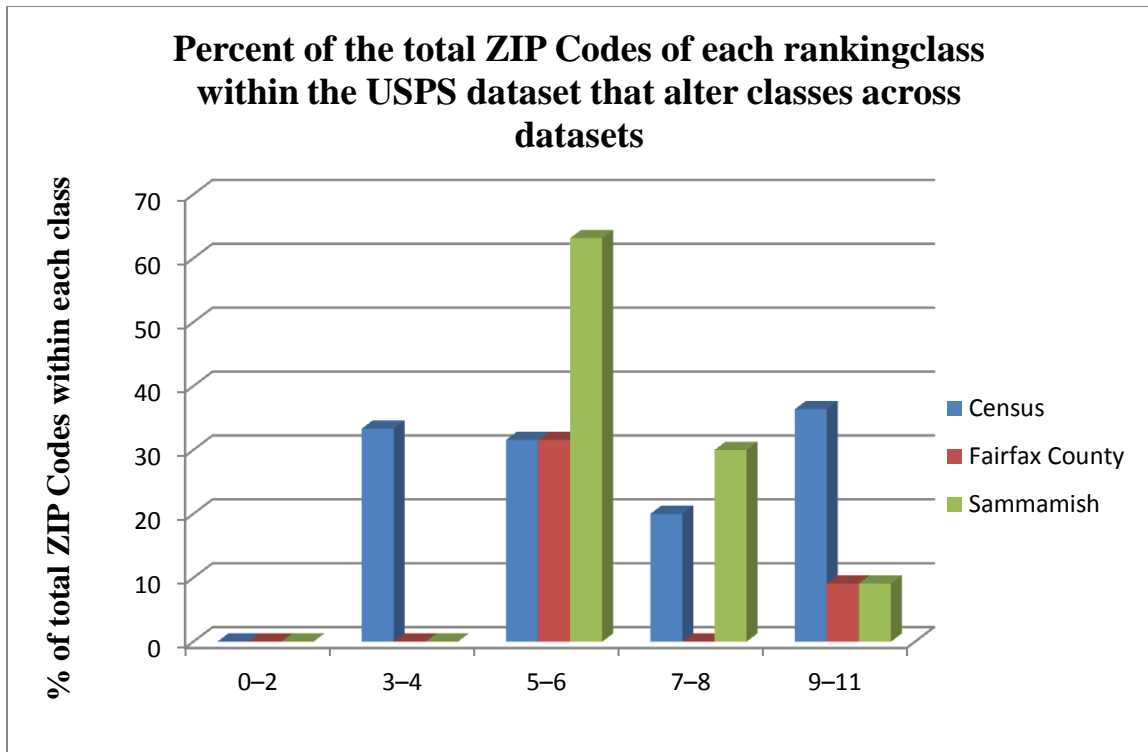


Figure 26: Percent of the total ZIP Codes within the USPS dataset that alter quantile classes across datasets

Table 8 summarizes the pair-wise comparison for the ZIP Codes that switched to different ranking classes across datasets. More than one third of the total ZIP Codes (37%) changed their ranking classes between the USPS and the Sammamish datasets. 30% of the total ZIP Codes switched to another class between each of the pair of USPS-Census and Sammamish-Census datasets. Other pair-wise comparisons also show large percentages of the ZIP Codes that fall within different class ranges between two datasets.

Table 8: Pair-wise comparison of the ZIP Codes that changed ranking classes across the datasets

	ZIP Codes that switched to a different ranking class	
	Number of ZIP Codes	Percent to the total ZIP Codes
USPS - Census	13	30
USPS - Fairfax County	7	16
USPS - Sammamish	16	37
Census - Fairfax County	11	26
Fairfax County - Sammamish	10	23
Sammamish - Census	13	30

The locations of schools also have a huge impact on home values (Max 2010). Property values within a ZIP Code tend to follow the quality of schools available within that ZIP Code and vice versa (Dan Immergluck 2011; Shan 2011). In this study, attempts been made to determine the relationship of housing values to the ZIP Code ranking. The ranking of ZIP Codes is done based on the quality of schools within the ZIP Code. So, it is expected that the housing price will follow the ranking of ZIP Codes. This research also tries to discover the differences of this relationship over multiple data sources. Figure 27 (Scale 1:450,000) shows the median property value within ZIP Codes using Census, Sammamish, Fairfax County and USPS ZIP Code polygon layers. The property values are also classified into 5 defined classes to match with the classification scheme of the ZIP Code ranking.

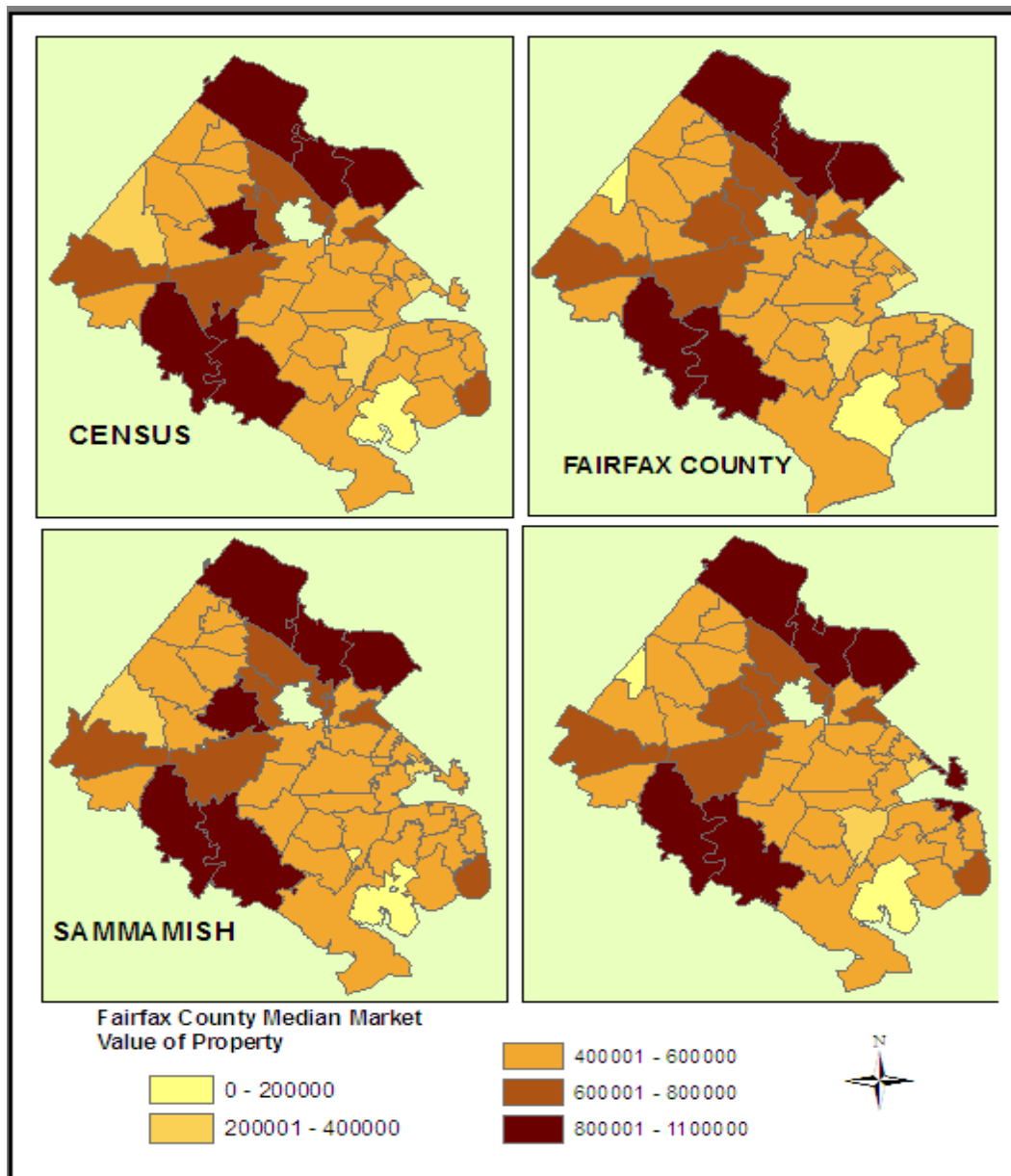


Figure 27: Average Property prices (in Dollar) within ZIP Code polygon boundary in Different data sources.

The average property values are more or less similar for the datasets in the Northern, Southern and middle part of the county. The range of the classes is large enough to hide the difference in property values of a ZIP Code in the data layers. There

are five classes that have been used for the classification of property values in this study to correlate with the result obtained from segregation estimation and ZIP Code ranking. The difference would be more pronounced if the classification includes higher number of price classes. However, there are still some ZIP Codes in the study region that vary in property price with the current classification. Examples of such ZIP Codes are 22124, 22302, 22303, 20041, 22091, 20171, and 20151.

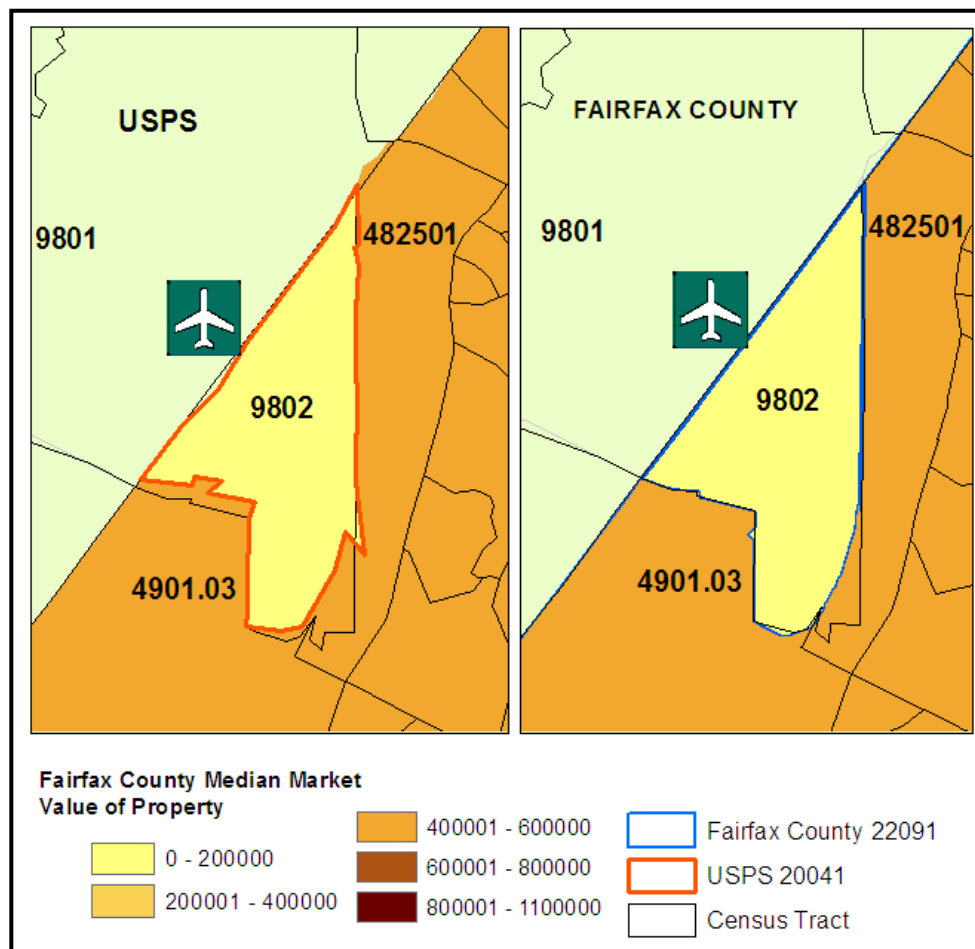


Figure 28: Overlay of ZIP Code 22091 and 20041 with census tracts

There is a unique ZIP Code in the USPS ZIP Code polygon data layer (ZIP Code 20041) which covers the area of another unique ZIP Code 22091 in the Fairfax County dataset. Sharing this area with 20171, ZIP Code 20151 varies in property price over data layers (Figure 28). Most of this area is owned by the Dulles International Airport of Washington DC. As noted, the market prices of property within ZIP Codes are estimated from census tracts based on the areal weight of the tract on the entire ZIP Code area. As the ZIP Codes cover the most of the area of census tract 9802 (that has no housing property within tract boundary) and a very small part of 9801 (also no housing), 4825.01 and 4901.03, the average property value within the ZIP Code is minimal.

There is an inconsistency of the housing price with the ranking of ZIP Codes 20041 and 22091 (Figure 23). It was expected that the housing price would follow the ranking. The ZIP Code is highly ranked based on the available school within its boundary but the results from estimation of property price show low housing price for the ZIP Code area. The explanation of this inconsistency could be the influence of the Floris Elementary school district on the total ranking of the ZIP Code. This school is located in the census tract 4825.01 and almost the entire area of the ZIP Code is served by this school district and ranks the ZIP Code as high. Similarly the disagreement of ZIP Code 20151 and 20171 in these studies also can be explained. However, this could be due to a variety of other factors that influence housing prices.

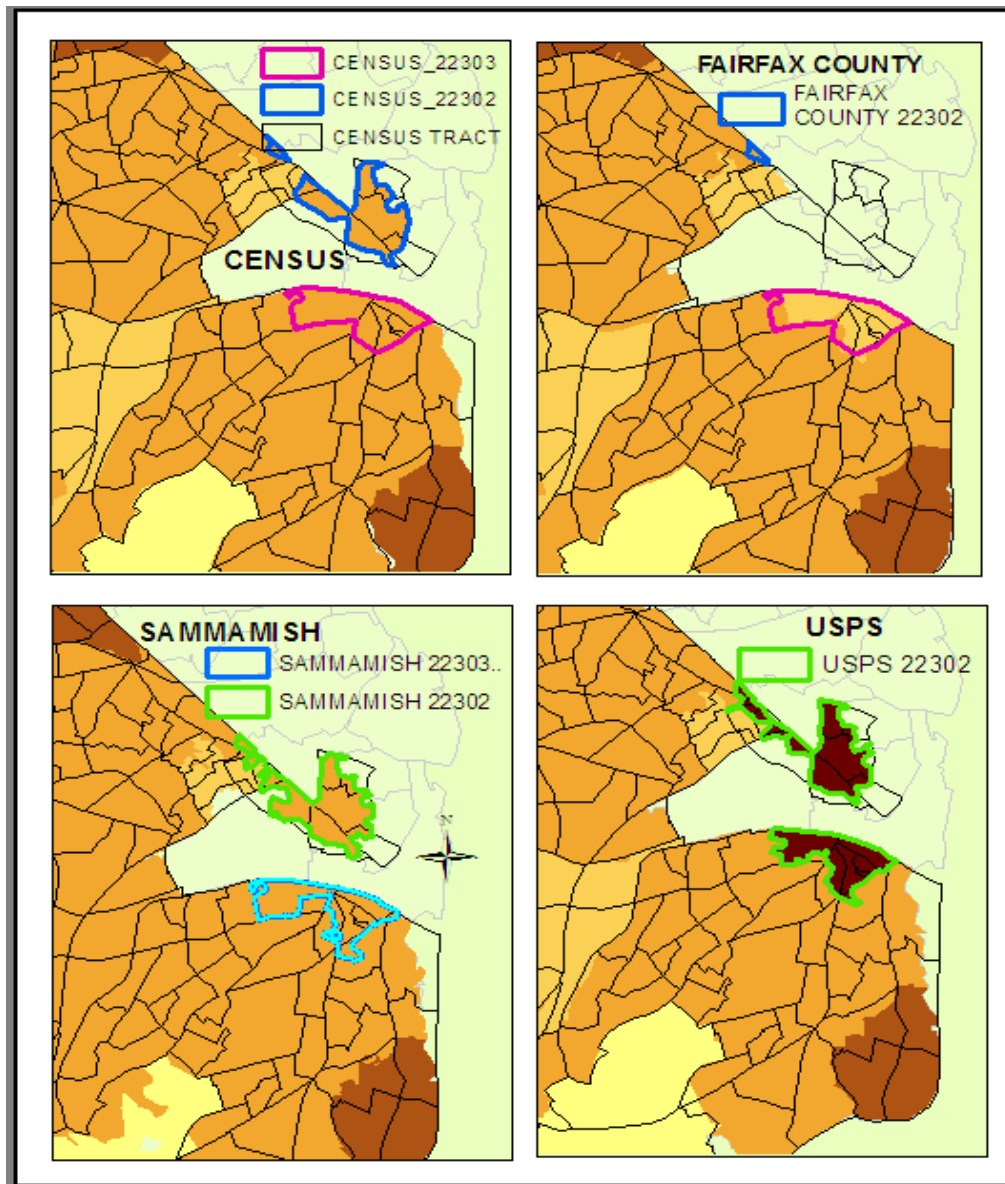


Figure 29: Inconsistency of ZIP Code area creating variable property price

Another example of mismatch within the data sources is ZIP Code 22302 and 22303 (Figure 29). In the USPS, ZIP Code 22302 has two polygons, one of which is known as ZIP Code 22303 in other datasets. In the Fairfax County dataset, a part of

22302 falls within the county boundary. This ZIP Code has medium price range properties (\$400,001-\$600,000) in the datasets but in the USPS, it has the highest property value ranging from \$800,001 to \$11,000,000. The ZIP Code 22303 also varies in property value within datasets.

As expected, the results of the estimation of housing price have similarity with the results of ZIP Code ranking for a large number of ZIP Codes. The ranking of ZIP Codes and housing prices also show consistency with the segregation of Hispanic populations within this region (Section 7, Figure 21). Home values are lower in areas where Hispanic residents are more segregated. The ZIP Codes in the Northern and South Western part of the county have high property values and low segregation. These ZIP Codes are also highly ranked based on the availability of public schools. With over 175,000 students enrolled, the Fairfax County public school is the largest school system in the Baltimore-Washington and Northern Virginia Metropolitan area (www.wikipedia.org). Most of the schools have ranking of 5 or higher (117 out of 139 elementary school used for the analysis). Also, there are very few ZIP Codes in the county that have an average property value less than \$400,000. Therefore, sometimes a high level of segregation can exist in areas with good quality school districts. Although the segregation indices for the ZIP Codes are below 0.6 for all datasets, the ZIP Codes show differences in segregation level of Hispanic residents in alternative datasets. Some of the ZIP Codes with medium range of property value (e.g. ZIP Code 20191) and medium high ranking based on serving school districts (e.g. ZIP Code 22033, ZIP Code ranking range: 7-8) have high levels of segregation. This is possible as the segregation level is classified into 5 classes within the

range of 0 (lowest segregation) to .06 (highest segregation). So, an index value of 0.6 can be shown as highly segregated even if it is actually a medium level of segregation within 0-1 index value of segregation.

Although the studies on segregation, housing price and ranking of ZIP Codes show consistency when the results are compared for a particular dataset, the results are less consistent when multiple datasets are compared. For example the ZIP Codes in the Census dataset have less consistency than the ZIP Codes in the USPS.

Preferences for schools and neighbors shape the way that buyers behave in the housing market, influencing and influenced by the level of residential segregation as well as the quality of schools (Bajari and Benkard 2001). Schools are often one of the most important factors in an area's desirability. Buyers having school age children generally prefer the areas to live within that have accessibility to good ranking schools (Max 2010) and often they rely on real estate agents or websites for information by ZIP Code level. The results obtained from estimating the ranking of ZIP Code, based on quality of school available within the ZIP Code area, suggest that the choice of ZIP Code polygon representation used for an analysis in real estate business can affect the outcome. It can also influence decisions made by home buyers in search for new homes using ZIP Code as the search criterion. A buyer could be confused if housing information is provided by ZIP Code level where information for the same ZIP Code does not match in different searching sites that use ZIP Code polygon maps from different data sources. It is thus suggested not to use ZIP Code as the geographic unit either for simply providing any real

estate related information or in academic and economic research related to housing market.

SECTION 9

ACCESSIBILITY TO EMERGENCY ROOM IN FAIRFAX COUNTY

Accessibility refers to the relative ease by which a location of interest (such as location to work, shopping, recreation, health, law enforcement center and many other service points) can be reached from a given location (Wang 2006). Access to different services or resources are recognized as an important parameter of overall population living standard (Luo, Wei and Qi 2009).

Accessibility is a common issue in many fields at various scales. For example: an economist may want to analyze relationships between jobs and housing or urban commuting patterns within a threshold distance (Peng 1997); a social scientist may be interested in measuring accessibility of rural people to food stores or other facilities (Kaufman 1999) or accessibility to children's playgrounds; a transportation planner may try to explain commuting patterns in an area based on measures of job accessibility (Wang 2000). An epidemiologist may focus on measuring geographic access to health services, especially for high-risk populations and underserved communities or the influence of spatial and aspatial factors on accessibility (Dai 2010; Wing and Reynolds 1988; Messina et al. 2006; Knapp and Hardwick 2000). Spatial access disparities to services is a issue of growing priority for planners and policy makers and thus there is a pressing need to determine the distribution of resources and identify populations who do not enjoy access to various services (Ngui and Apparicio 2011; Wang 2006).

Computing geographic access is complicated as many datasets contain geographic information by region of residence instead of precise address information. In many of the studies ZIP Code centroids are used to represent a demand or a supply location. Often the US Census Bureau population and other information (e.g., demographic, economic, educational) are used to estimate the population and integrate other information at the ZIP Code level. The extensive use of ZIP Codes in accessibility analysis may be due to the reason that ZIP Codes have finer resolution than counties (Parker and Campbell 1998; Knapp and Hardwick 2000) and sometimes data are available mostly at the ZIP code level (Dai 2010; Wang and Luo 2005a; Grubestic 2008a).

For example, the ZIP Code centroid as a point location represents the residence of an individual known to reside within that area, when no further information is available. This often happens when the identity of a group of people need to be protected, such as when medical patients are registered at the ZIP Code level (Franks and Fiscella 2002; Beyer et al. 2011; Cudnik et al. 2012; Fiscella and Franks 2001; Hebert, Chassin, and Howell 2011). Moreover, generally, an individual needs to provide an address with a ZIP Code when accessing a health care center or other facilities. This requires additional geocoding to get census tract or other geographic unit level data from these ZIP Codes (Thomas et al. 2006). ZIP Code centroids are also used as the aggregate location of services that are available within the ZIP Code boundary.

However, the use of a ZIP Code centroid as the representation of a resident's location or a service location within that ZIP Code can be problematic and create a fallacy in the interpretation of the results as well as ecological fallacies. As discussed in

previous sections it is difficult to identify the true boundaries of ZIP Codes and thus to locate the centroids of these ZIP Code polygons.

In this study spatial access is estimated for a population to emergency room care using a two-step floating catchment area method, taking into account both travel time and facility capacity. The travel time is measured using an Origin-Destination matrix along a road network which also identifies the availability of a supply location from a population location within a threshold and vice versa. The facility capacity is measured as the ratio between the number of doctors at the facility (emergency room) and the number of potential patients (population) in its 10 minute catchment area. Due to availability of data at different geographic levels (e.g., cancer patients in Wang and Luo, 2005a are registered to ZIP Codes but population data from the Census Bureau are collected at census tract or block level)- the information at ZIP Code level as well as census tract/block level is incorporated while measuring accessibility. This study assigns the facilities and populations to the ZIP Code centroids and census tract centroids respectively to represent an example that mimics this very common practice of using ZIP Code in accessibility measurement.

9.1 Literature review:

An impressive amount of research has been done on evaluating accessibility from a demand to a service location. Studies carried out by economists, epidemiologists and analysts from many other fields focused on revealing spatial access of a population to a center of service (Knapp and Hardwick 2000; McCarthy and Blow 2004; Messina et al. 2006); evaluating aspatial characteristics of accessibility (Weissman et al. 1991; Hartley,

Quam, and Lurie 1994); some other studies attempt to integrate the spatial and non-spatial factors of accessibility (Wang and Luo 2005a). Patel, Waters, and Ghali (2007) measured the accessibility of populated places to cardiac catheterization facilities and compared the efficiencies of different modes of emergency transportation in terms of travel time within the province of Alberta, Canada.

Many studies use ZIP Code centroids either as supply location (Luo and Wang 2003; Wang and Luo 2005b; Luo, Wang, and Douglass 2004) or demand location (Votruba and Cebul 2006) or both (Messina et al. 2006). Gruenewald, Johnson, and Treno (2002) collected drinkers data from a general-population telephone survey of 1,353 zip code areas in California and examines the relationship of accessibility to alcohol and rate of drinking and driving incidents. Goodman et al. (1997) examines the influence of service-demand distance on hospitalization and mortality rate among population living outside and inside a ZIP Code of the service. Franks and Fiscella (2002) examine the influence of patients' socioeconomic status on physician profiles within a ZIP Code and concluded that adjustments using ZIP Codes yielded comparable effects on the measurement compared to using census tracts. Fiscella and Franks (2001) analyze these effects when patients' addresses are geocoded to patient reported education locations and compared to the census block group level or ZCTA level.

Zhang, Lu, and Holt (2011) employed a population weighted distance to measure potential spatial access to parks from census blocks and to quantify the spatial distribution of neighborhood parks based on residential proximity to parks as well as sizes of parks. To minimize ecological bias in population location within a ZIP Code and

other large spatial units this study used census block as the base geographic unit. Algert, Agrawal, and Lewis (2006) analyzed differences in access to fresh produce between poor ethnic and wealthier non-ethnic neighborhoods. It measures access to food stores from individual addresses within buffer distances around the stores by using Manhattan distance to measure distance of supermarkets from African-American and white neighborhoods.

Govind, Chatterjee, and Mittal (2008) focused on the allocation of available hospital resources to different disease types examining spatio-temporal patterns of disease at the ZIP Code level where the disease incidence values are assumed to be observed at the centroids of census ZCTAs. Hebert, Chassin, and Howell (2011) attempted to discover racial differences in the use of high-quality hospital care by geocoding hospital addresses and assigning mothers to the centroid of the ZIP code of residence for each mother. Many other studies attempt to measure spatial accessibility between sets of demand and supply points (e.g., McCarthy and Blow 2004; Cinnamon, Schuurman, and Crooks 2008; Fu et al. 2009).

Talbot et al. (2000) evaluated the spatial filters in smoothing maps based on filters like fixed geographic size and constant population size. They estimated the population center of a ZIP Code using census block population weight. Schultz, Beyer, and Rushton (2007) also discussed this method for determining population weighted ZCTA centroid to determine the preferred distance an individual has to travel for reaching a service. This method finds a location for which the population weighted blocks, within a ZCTA boundary, would have the least sum of squares distances.

Measurement of potential spatial accessibility depends on spatial factors such as geographic location and distance between the supply or service points and population demand locations (Luo and Qi 2009). It can be measured by proximity of demand location to services, typically in driving distance or driving time (Bliss et al. 2012). There are several methods for assessing accessibility including gravity models, kernel density estimation, and the floating catchment method (Dai 2010; Yang, Goerge, and Mullner 2006; Wang 2006). Each method has some advantages as well some shortcomings. The two-step floating catchment area (2SFCA) is a popular method for measuring accessibility to service providers (Yang, Goerge, and Mullner 2006; Wang and Luo 2005b; Luo and Wang 2003). This study uses the basic 2SFCA which measures accessibility once from demand points and then from the supply or service points within a threshold travel time of 10 minutes along a road network.

9.2 Study area and data

The study area is the same region of Fairfax County that was examined in the earlier studies. The ZIP Code polygon boundaries are taken from the same data sources of the USPS, Fairfax County, Sammamish and Census Bureau. Census tracts and population data are extracted from www.census.gov. As the census tract is the lowest areal unit frequently used in practice for shortage area designation measurement (Dai 2010; Luo and Wang 2003), this study uses census tract as the analysis unit for population. However, incorporation of population data, collected for a year, with tract boundaries from another year creates the potential for temporal mismatches. Yet, it should be noted that this study is dedicated to presenting an example of accessibility

analysis that is very common in practice and thus the issue of incorporation of tract and ZIP Code boundary with population data of 2010 is not a significant concern. Though the emergency room data has the street address and can be geocoded directly to street locations, these addresses are geocoded to ZIP Code areas to simply mimic the practice used in many spatial analyses.

The accessibility from each ZIP code to each health care facility relies on the estimated travel time along a road network extracted from the TIGER/Line files from the www.fairfaxcounty.gov. Fairfax County emergency room data is collected from the website <http://www.yellowpages.com/fairfax-va/emergency-room>. All maps are projected to Equidistant Conic to minimize distortion in distance measurement.

9.3 Method

Accessibility is determined by the distributions of supply and demand and the way they are connected in space (Wang 2006). According to Joseph and Phillips (1984) measures of spatial accessibility include regional availability and regional accessibility. The former is expressed as a population (demand) to provider (supply) ratio within a region. The latter requires more computation and considers complex interaction between supply and demand in different regions often based on a gravity kernel.

In the regional availability approach, interaction across regional boundaries is not adequately accounted for and spatial variability within a region is not completely revealed (Wing and Reynolds 1988; Van Meter et al. 2011). Earlier versions of the floating catchment area (FCA) method (e.g., Peng, 1997), developed for assessing job accessibility, attempted to address these problems (Wang 2006; Wang and Luo 2005b).

The spatial concept behind this method is a circle with the same radius (Wang 2000; Daniel Immergluck 1998) or a fixed travel time range (Wang and Minor 2004; Dai 2010), denoted as the catchment area, floats between the centroids of demand locations and the supply-to-demand ratio within each demand location defines the accessibility for that location (Luo, 2004).

An improved version of this method was developed by Radke and Mu (2000) and later modified by Luo and Wang (2003), referred to as the ‘two-step floating catchment method’ or 2SFCA. It is a special case of gravity model that repeats the process of ‘floating catchment’ twice, once on supply locations and once on demand locations (Luo and Qi 2009; Wang 2006). Subsequently an impressive body of research has been performed, especially in health care research, focusing on the implementation and improvement of the 2SFCA method (Luo and Qi 2009; Ngamini Ngui and Vanasse; Luo and Wang 2003; Wang and Luo 2005b).

A gravity model is a combined indicator of accessibility and availability that counts decreasing accessibility with increasing distance or travel impedance (Guagliardo 2004). It is believed that the frictional coefficient in the distance decay function requires more region specific demand-supply interaction data (Luo and Qi 2009). In an enhanced 2SFCA model, a set of travel time zones around ZIP Code centroids can be used to account for the issue of distance decay but- requires an appropriate number of travel time zones (Luo and Qi 2009).

The 2SFCA method can identify local pockets of poor access in cities compared to the gravity-based model and the kernel density estimation method, yet the method

assumes equal access to health care facilities within a catchment (Ngui and Apparicio 2011; Yang, Goerge, and Mullner 2006). In an enhanced Gaussian 2SFCA, a Gaussian function is used to address the distance decay of accessibility within a catchment area. This method does not require one to determine appropriate travel time zones, but rather accounts for the accessibility loss continuously with increasing distance by a friction-of-distance based Gaussian function (Dai 2010). In this study the basic 2SFCA method is used to avoid complexity in determining any distance decay of accessibility. Only one distance zone is used to estimate accessibility and compare over ZIP Code polygon interpolation maps from different data sources.

The Network Analyst tool in ArcGIS is employed to simulate the shortest travel time between a service (ZIP Code centroid) and demand point (census tract centroid) through the network where speed limits serve as travel impedance. It is admitted that actual travel times may be influenced by actual driving speed (e.g., signal delays or congestions), time finding a parking space, time walking to a facility, time taking public transit and other issues. For example, taking mass transit needs to consider the time driving (or walking) from home to a public transit station, time of taking public transit, and time walking from a station to a facility. So the actual travel time is likely to be longer than the estimated time. Nonetheless, the estimated travel time can effectively capture the variability in geographic access to facilities and is widely used to measure travel impedance in many studies (Luo and Wang 2003; Pedigo and Odoi 2010; Brabyn and Skelly 2002; Thornton, Pearce, and Kavanagh 2011). After calculating the travel time the routes within 10 minute travel distance are included in the analysis.

The 2SFCA method works as follows: first, for each emergency room location j ; search all population locations (k) that are within a threshold travel time of 10 min (d_o) from location j (i.e., catchment area j), and compute the emergency room to population ratio R_j within the catchment area:

$$R_j = \frac{S_j}{\sum_{k \in \{d_{kj} \leq d_o\}} P_k} \quad \text{Equation 6}$$

Where, P_k is the population of census tract k whose centroid falls within the catchment (i.e., $d_{kj} \leq d_o$), S_j is the number of physicians at location j ; and d_{kj} is the travel time between k and j .

Next, for each population location i ; search all emergency room locations (j) that are within the threshold travel time (d_o) from location i (i.e., catchment area i), and sum up the physician to population ratios R_j at these locations:

$$A_i^F = \sum_{j \in \{d_{ij} \leq d_o\}} R_j = \sum_{j \in \{d_{ij} \leq d_o\}} \left(\frac{S_j}{\sum_{k \in \{d_{kj} \leq d_o\}} P_k} \right) \quad \text{Equation 7}$$

Where, A_i^F represents the accessibility at resident location i based on the 2SFCA method, R_j is the physician-to-population ratio at emergency room/physician location

j whose centroid falls within the catchment centered at i (i.e., $d_{ij} \leq d_o$), and d_{ij} is the travel time between i and j .

The first step above assigns an initial ratio to each service area centered at an emergency room location, and thus the travel time between the supply and any demand within the catchment does not exceed the threshold. The second step sums up the initial ratios in the overlapped service areas to measure accessibility for a demand location, where residents have access to multiple emergency room locations. 104 is the ratio of physician to population within the threshold travel time and can be interpreted in the same way. The method can be implemented in GIS by the following procedures (Figure 30):

- 1) Generating centroid of ZIP Codes and census tracts: ZIP Code centroids are determined using the area as the weight field for all the ZIP Code maps. Centroid locations are also determined for the census tracts that fall within the ZIP Code boundaries. This computation is implemented in ArcToolbox by utilizing Spatial Statistics Tools > Measuring Geographic Distribution > Mean Center. Any census tract centroid outside the ZIP Code area in datasets is excluded from analysis.
- 2) Computing distance between tract and ZIP Code: The network distance between a population (census tract centroid) and emergency room (ZIP Code centroid) location is computed using an Origin-Destination Cost Matrix tool in the Network Analyst toolbar. Travel time is used as the impedance factor.

- 3) Extracting distances within a threshold: Based on the distance table (Dist_tract_to_ZIP), records ≤ 10 minutes are selected and exported to a layer Dist10min. The new distance table only includes the distances within the threshold 10 min travel time and thus implements the selection conditions $j \in \{d_{ij} \leq d_o\}$ and $k \in \{d_{kj} \leq d_o\}$ in Equation 7.
- 4) Attaching population and emergency room data to the distance table: The attribute tables of emergency room (ZIPEmerRoom) and population (TractPop) are joined to the distance attribute table Dist10min by corresponding ZIP Code areas and census tracts respectively.
- 5) Summing population around each emergency room location: Based on the updated table of Dist10min, a new table PopbyZIP.dbf is created by summing population by ZIP Code centroid locations. The field Sum_Pop is the total population within the threshold distance from each emergency room location implementing $\sum_{k \in \{d_{kj} \leq d_o\}} P_k$ in Equation 7. It should be noted that a ZIP Code serves only the population whose centroid (census tract centroid) falls within the threshold distance. To avoid complex distances, outside ZIP Code boundaries are removed from the table even if it is located within 10 min travel time.
- 6) Computing initial physician-to-population ratios at each emergency room location: The newly generated PopbyZIP.dbf is joined to the attribute table of Dist10min and a new field DocPopR is added computed as $\text{DocPopR} = 1000 * \text{DocNum} / \text{Sum_Pop}$. This assigns an initial physician-to-population ratio to

each emergency room location, representing the physician availability per 1000 residents. This step implements the term $S_j / \sum_{k \in \{d_{kj} \leq d_o\}} P_k$ in the equation.

- 7) Summing up these ratios by population locations: Based on the updated Dist10min attribute table, the initial ratios (DocPopR)are summed up by population locations (census tracts) to yield a new table RbyTract.dbf. The field sum_ DocPopR sums up the availability of physicians that are available from each population location and thus yields the accessibility A_i^F of Equation 7.
- 8) Mapping accessibility: The table RbyTract.dbf is joined to the census tract shapefile for mapping.

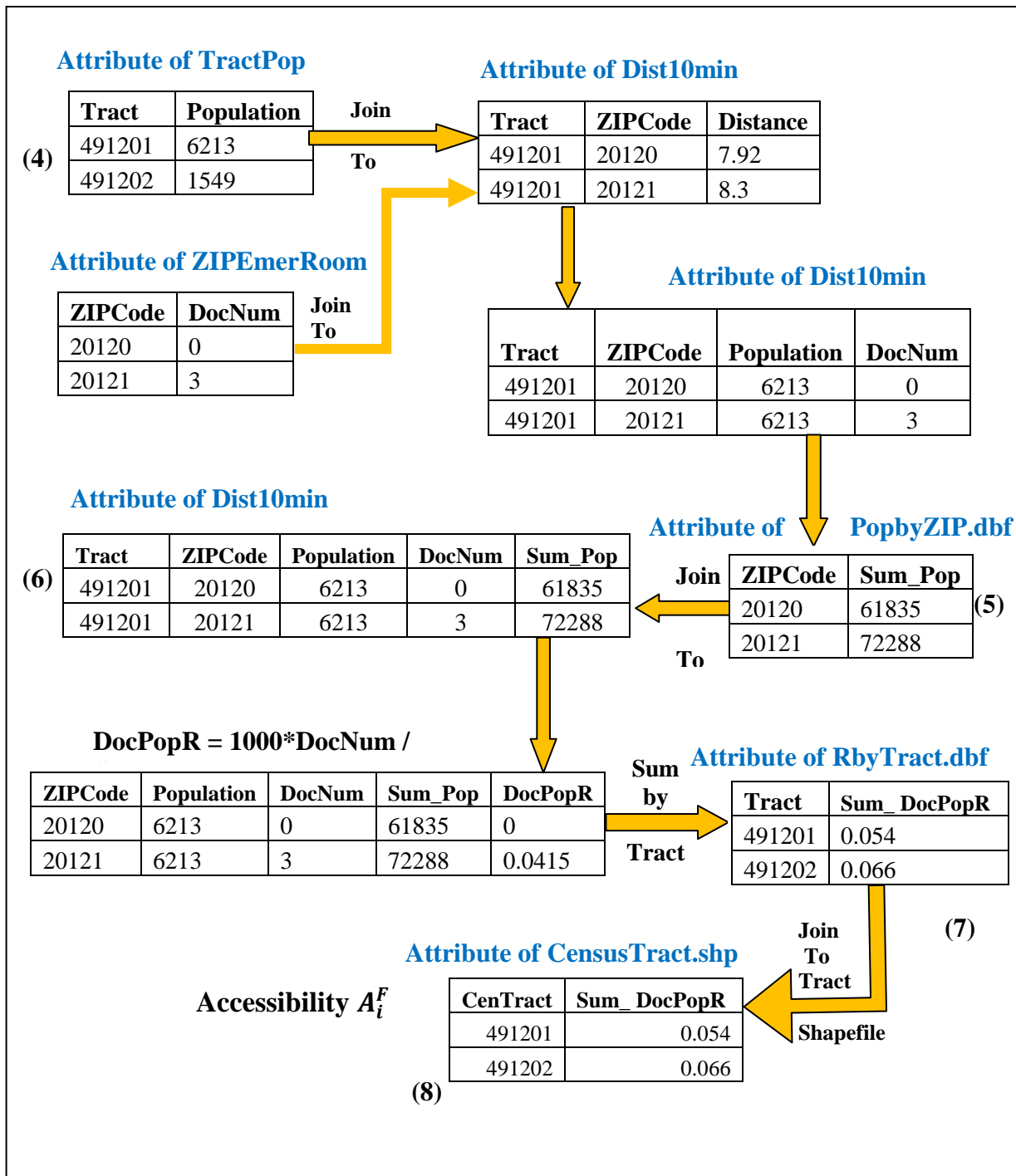


Figure 30: Procedures of implementing 2SFCA in ArcGIS.

9.4 Results

Figure 31 shows the result of the accessibility measurements from population to emergency room within the study area, classified into 5 quantile classes. The datasets are placed side by side in a common defined class for comparing accessibility over these datasets. In general, the results show high accessibility in some parts of the North-East and North-West and moderate to low accessibility in the Middle and Mid-East parts of the county. Census tracts in the upper middle, South-West and South-East regions have the lowest accessibility. But when examining cautiously, the datasets reveal different patterns of accessibility in census tracts.

For example: tract 480100 is within the 5th quantile having a very high accessibility in the Census ZCTA map but this tract has lower accessibility values in other datasets falling within the 4th quantile in the Sammamish and 3rd quantile in the Fairfax County and USPS datasets. Tract 480402 is within the 1st quantile and has a very low accessibility value in the Census dataset, but is within the 5th quantile and has a very high accessibility in other ZIP Code polygon representations. Tract 490103 also is within the 5th quantile in the Sammamish dataset but is within the 3rd quantile in other three datasets. There are several other tracts for which the accessibility values vary and fall within different quantile classes across the datasets.

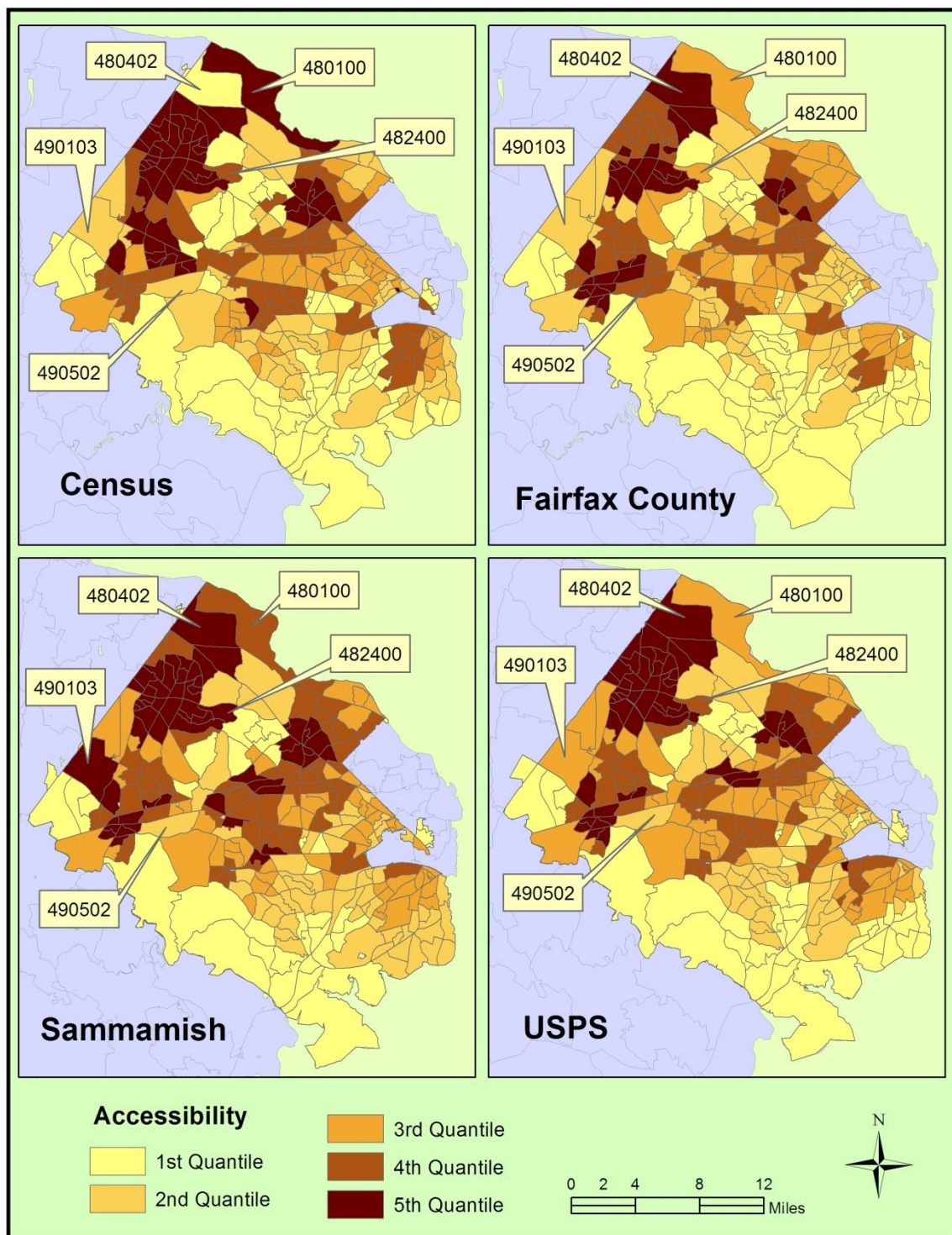


Figure 31: Comparison of accessibility to emergency room across datasets.

The results indicate that the accessibility level of the same ZIP Code varies over different data sources which indicate that an analyst would obtain different measures depending on the data source employed. Although in the literature there are arguments in favor of using ZIP Codes (e.g., Thomas et al. 2006) or Census ZCTAs (e.g. Schultz, Beyer, and Rushton 2007; Krieger et al. 2002), these results challenge that claim and require more research on choosing a reliable data source providing ZIP Code boundaries.

SECTION 10

STATISTICAL SIMILARITY

This section examines a series of statistical similarity measures to determine if the ZIP Code maps truly are significantly different from each other when they are collected from different data sources. The ZIP Code maps have been tested for independence to determine whether there are any significant statistical differences within these maps. A simple linear regression analysis is then done to check whether or not the area values in the datasets have linear correspondence with that of the USPS. All statistics have been done on the IBM SPSS Statistics 19 version and MS Excel 2007.

10.1 Data

All the statistical analyses are done on the area values of the ZIP Codes obtained from previous calculations. There are 166, 159 and 154 valid cases for the USPS, Census and Sammamish datasets respectively. The Alexandria, Arlington, Clarke, Fairfax City, Fairfax County, Frederick, Loudoun and Shenandoah have respective valid records of 8, 11, 6, 3, 44, 11, 22, and 7 cases.

10.2 Methods

10.2.1 Test for normality

A goodness-of-fit test is often employed in order to determine the normality of a distribution. In the test, the observed frequency is compared against a hypothetical

normal distribution. A Chi-square test is a frequently used goodness-of fit test for nominal data whereas the one-sample Kolmogorov-Smirnov (K-S test) test is good for ordinal data (McGrew and Monroe 1999). The Shapiro-Wilk test is another goodness-of-fit test which is used frequently to test normality, most commonly applied to small size datasets (SPSS Inc. IBM 2010). Since the area values on ZIP Codes are ratio data, in this study the K-S test and Shapiro-Wilk tests have been used to examine normality rather than the Chi-square test.

These goodness-of-fit tests can help to determine whether a parametric or a non-parametric test can be used to determine the independence of the ZIP Code maps. If the distribution exhibits normality, a parametric test can be employed. For a non-normal distribution, it is suggested to use a non-parametric test on the distribution (Burt, Barber, and Rigby, 2009). In the Shapiro-Wilk and K-S test, the null and alternative hypotheses are defined as follows

Ho = The population or the dataset fits an expected normal frequency distribution.

H1 = There is a significant difference between the observed and the expected frequencies.

For both tests, the statistical level of significance (p-value) is calculated within a 95% confidence limit. If the p-value comes out to be less than 0.05, the assumption of normality will be violated.

The 'Analyze' menu bar on the SPSS interface has a tool 'Descriptive Statistics' which has an option to 'Explore' different descriptive statistics, box plots and normality plots for a distribution.

The results of the normality tests as well as other descriptive statistics and graphical methods (e.g. Normal Q-Q Plots) are analyzed to determine the normality of the underlying distributions. A Parametric test can be chosen if the normality assumption is met. Otherwise a non-parametric test will be used to examine whether the maps/datasets came from the same distribution; in other words whether they are significantly different from each other.

10.2.2 Choice of parametric or non-parametric test for independence

The Kolmogorov-Smirnov test of normality reveals that the area values of the USPS, Census, Sammamish, and Fairfax County ZIP Code maps are not normally distributed. They have test scores less than 0.05; indeed that are very close to a p-value of zero. Frederick and Loudoun County Datasets have marginal p-values of 0.051 and 0.056 respectively which indicates that they are normally distributed, although very close to a non-normal distribution. The Shapiro-Wilk test for the USPS, Census, Sammamish, Fairfax County, Frederick and Loudoun datasets have test scores less than 0.05 and are not normally distributed. On the other hand Alexandria, Arlington, Clarke, Fairfax City and Shenandoah have a value larger than 0.05 in both the tests indicating their normality. Fairfax City has only three records and therefore it is very difficult to confirm the normality of the distribution.

The descriptive statistics, graphical plots and graphs such as frequency distribution by histogram; variance, skewness and kurtosis of the distributions; box plot; quantile-quantile (Q-Q plot) plot and other exploratory statistics also suggest the non-normal characteristics of the ZIP Code areas within the datasets. Even the datasets that

are found to be normally distributed in normality tests show deviation from expected normal values in Detrended Normal Q-Q plots. In the Q-Q plot for the Census dataset (Figure 32), the frequencies deviate from normal situation especially for low values and some of the high values.

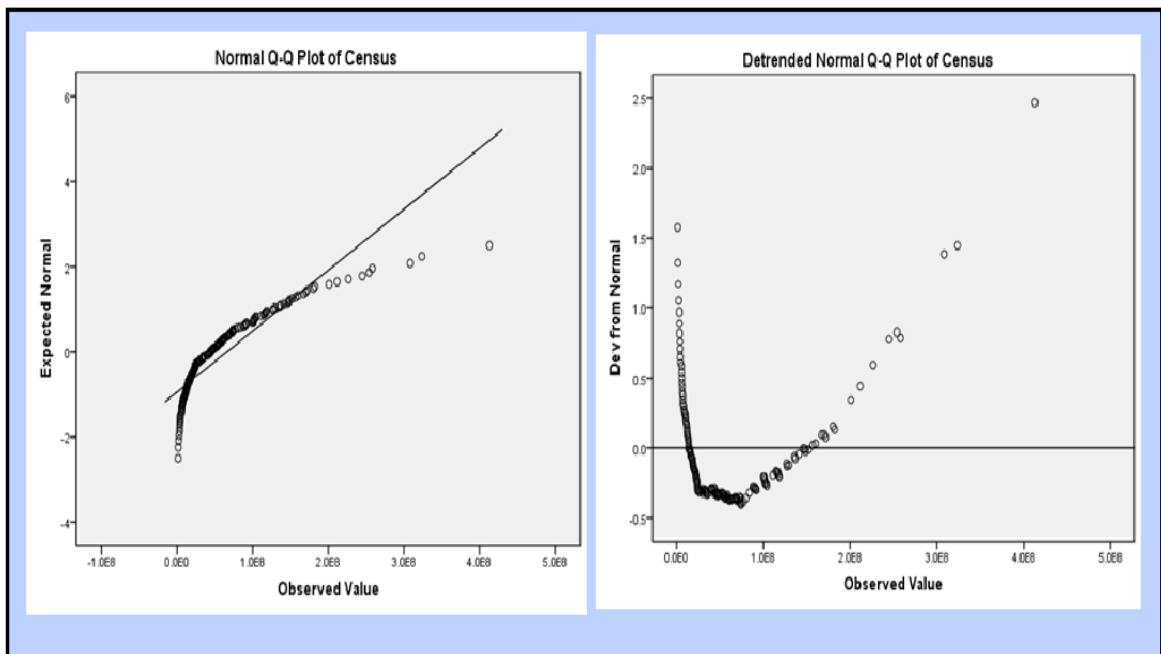


Figure 32: Normality plots of the Census dataset

10.2.3 Test for independence of ZIP Code maps

The fundamental question in this statistical analysis is whether these datasets have any difference between them or whether it can be safely be assumed that they have come from the same or from different population. The inference is examined by comparing the

means or the variances of the area values of ZIP Codes in two or more datasets. In statistics several methods are used for testing independence of two or more distributions. The choice of any test depends on the normality assumption of the distribution. Parametric tests require a distribution to be normal whereas non-parametric tests can be applied regardless of distribution type (Burt, Barber and Rigby 2009; Shaw and Wheeler 1994). A parametric test is more powerful and efficient but can be more easily influenced by any violation of the pre-test assumptions (Burt, Barber and Rigby 2009). A non-parametric test can produce the same result, but loses some statistical detail as it converts the ratio data to nominal or ordinal form (Shaw and Wheeler 1994). It is good to use non-parametric tests if the normality assumption of a distribution is not met.

Because of the non-normal distributions of most of the datasets, a set of non-parametric tests have been chosen for testing independence of the datasets. In all cases the area of the ZIP Code is considered as the testing variable. The null hypothesis and alternative hypothesis are defined as follows:

H_0 = There is no difference between ZIP Code areas across datasets.

H_1 = The datasets are different

The datasets have been tested on two different assumptions: first, they are variables having independent observations across datasets and secondly they have pairs of observations across datasets. In all cases, a 95% confidence interval is applied as it is a very frequently used interval limit in statistical analyses in social science.

The Wilcoxon-Mann-Whitney U test is a 2-sample non-parametric test that assumes the independence of two variables based on the equality of the means. It is the

non-parametric version of an independent sample t-test and does not require the datasets to be normally distributed.

The Kruskal-Wallis test is a k-sample test for ordinal data which compares the mean ranks of multiple groups (George and Mallery 2010). The null hypothesis for this test is that the samples come from the same population such that the probability of a random observation from one group is similar to the measurement of another random observation from another group and thus the probability would be greater than 0.5. In SPSS it is necessary to define object groups to test if mean ranks are the same or not. This study uses eleven groups classifying the ZIP Code datasets considering the area as the variable for which the ranks are going to be tested. The Mann-Whitney U test also has been performed on each pair of groups.

For paired observations, the Wilcoxon signed rank sum test and Friedman test are often used to test the difference between their mean ranks (Burt, Barber and Rigby 2009). The Friedman test is used when there is one independent variable and a normal or ordinal dependent variable (Burt, Barber and Rigby 2009). The Wilcoxon signed rank sum test is the non-parametric version of a paired sample t-test. The difference between the paired values is calculated as follows:

$$d_j = x_{1j} - x_{2j} \quad \text{Equation 8}$$

Where, d_j = difference in mean ranks

x_1 and x_2 = each pair of values

x_{1j} and x_{2j} = j th observations in group 1 and 2.

The underlying methods for determining the mean rank (\bar{d}) and standard deviation (s_d) of the difference are as follows:

$$\bar{d} = \frac{\sum_{j=1}^n dj}{n} \quad \text{Equation 9}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^n (dj - \bar{d})^2}{n-1}} \quad \text{Equation 10}$$

There is another option for performing non-parametric tests without specifying a test by the user. If the user does not specify which non parametric test is to be used, the SPSS software itself chooses some tests for each of the datasets based on the characteristics of area value. This nonparametric (NPTests) test automatically compares an observed dataset to a hypothesized dataset using several non-parametric tests such as McNemar test, Cochran's Q, Wilcoxon matched-pair single rank or Friedman's 2-way ANOVA by ranks. This test option has been applied on each pair of the datasets that have pair observations. Maps are also examined together with the USPS, Census and Sammamish ZIP Code maps. For example the Alexandria map is tested with USPS, Census and Sammamish altogether. It is then compared with each of these three maps

and also with the Fairfax County and Arlington data as these last two datasets have few paired observations with Alexandria.

10.2.4 Linear regression

In regression analysis the correlation between two variables are measured to find out how change in the independent variable influences the dependent variable (Shaw and Wheeler 1994). Although only the area values of maps are being tested, the study uses a simple linear regression considering the USPS as the independent and other datasets as dependent variables. This regression is based on the assumption that: any change of the area values in the USPS does not have any influence on the corresponding area values in other datasets. If the datasets are not independent to each other, the area value of a ZIP Code in the datasets should be correlated with that of the USPS.

The maps are plotted against the USPS ZIP Code map to get an idea of how they are distributed against it. All the maps show moderate to good linear relationship with the USPS. The datasets are then fitted to a curve to find out whether a linear regression can be fitted to them. Many of them do not show a strict linear relationship but the model summaries suggest that a linear regression can be employed on them. Table 9, Table 10, and Figure 33 show the fit for a linear and a non-linear (cubic/quadratic) regression for the Fairfax County and Loudoun County ZIP Code maps.

Table 9: Model Summary and Parameter Estimates for linear regression

Dependent Variable: Fairfax County

Equation	Model Summary				
	R Square	F	df1	df2	Sig.
Linear	.972	1438.372	1	42	.000
Cubic	.980	665.443	3	40	.000

Table 10: Model Summary and Parameter Estimates for linear regression

Dependent Variable: Loudoun

Equation	Model Summary				
	R Square	F	df1	df2	Sig.
Linear	.841	105.641	1	20	.000
Cubic	.870	40.034	3	18	.000
Quadratic	.869	62.998	2	19	.000

For both types, the significance is less than 0.05 which suggests that a cubic or quadratic as well as a linear regression model can be used.

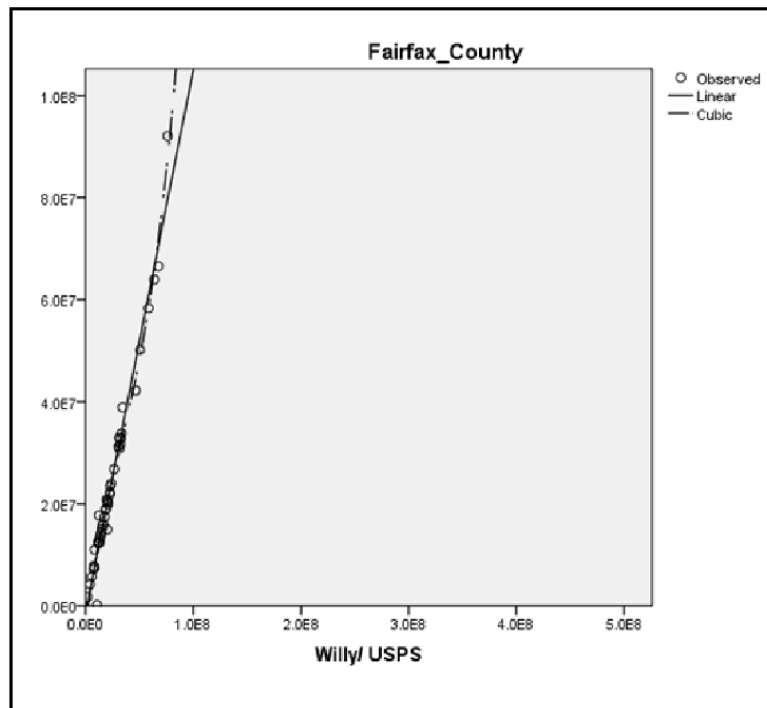


Figure 33: Curve fit for the Fairfax County dataset

10.3 Results

10.3.1 Tests of independence

In the non-parametric test for independence between a pair of datasets (tests automatically chosen by the SPSS) 23 out of 27 cases show a p-value greater than 0.05 which indicates that there is no significant difference between the pair of datasets being compared. In 4 cases or 15% of the total cases (Alexandria-USPS; Clarke -USPS; Frederick-USPS; and Frederick-Census) the null hypothesis of non-different distributions is rejected (Figure 34).

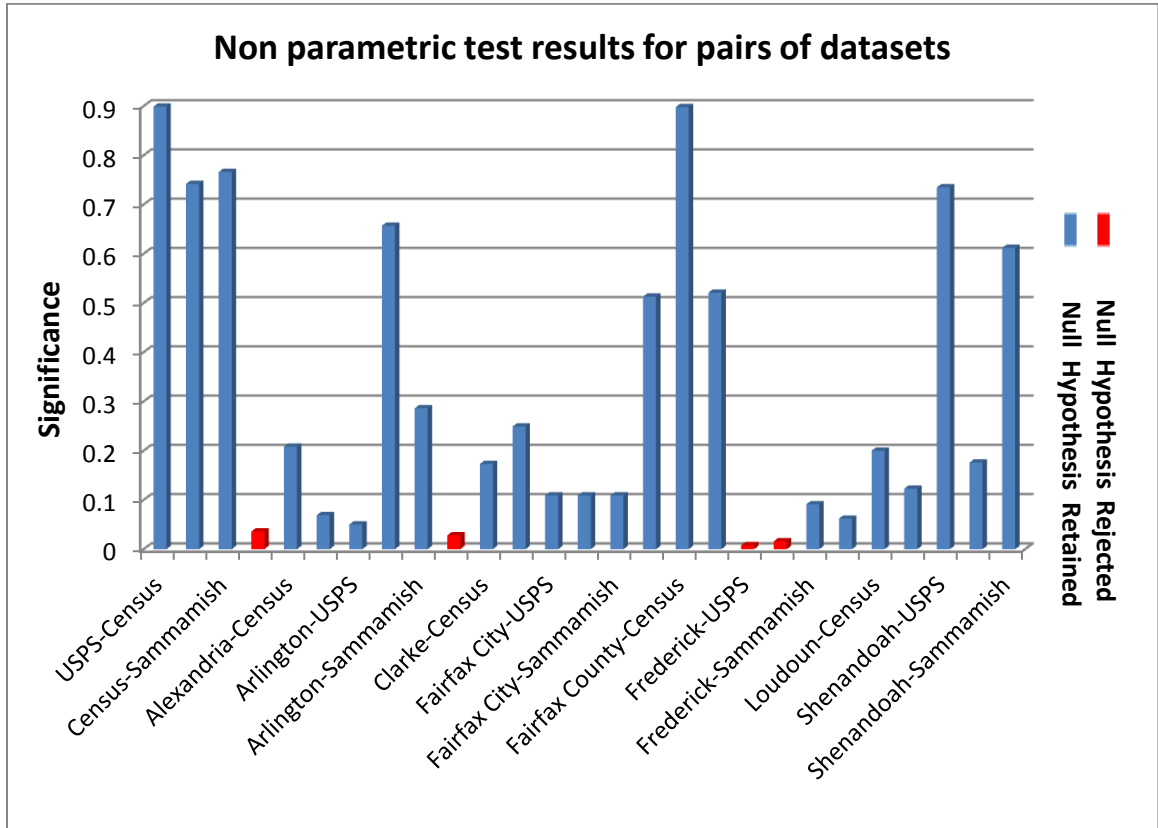


Figure 34: Results from pair-wise non-parametric tests (selected by the software) on the datasets.

The Kruskal-Wallis k-sample test within groups has a Chi-square value of 63.6 at a degree of freedom (df) of 9. The critical value for this df is 16.9190. The test statistic is greater than the critical value and the probability of the null hypothesis being true is zero. This test result indicates that the datasets are significantly different from each other. In the Mann U test the datasets also have significant difference for 50% of total cases. Yet

the results from the Mann U test can be confusing as most of the datasets do not have the similar shape of the distribution.

In the Friedman test of k-related samples (USPS-Census-Sammamish-Clarke and USPS-Census-Sammamish-Frederick) the null hypothesis has been rejected for 20% of the cases. For the remaining 80% of the total cases, the test reveals that there is no difference between the related datasets (Figure 35).

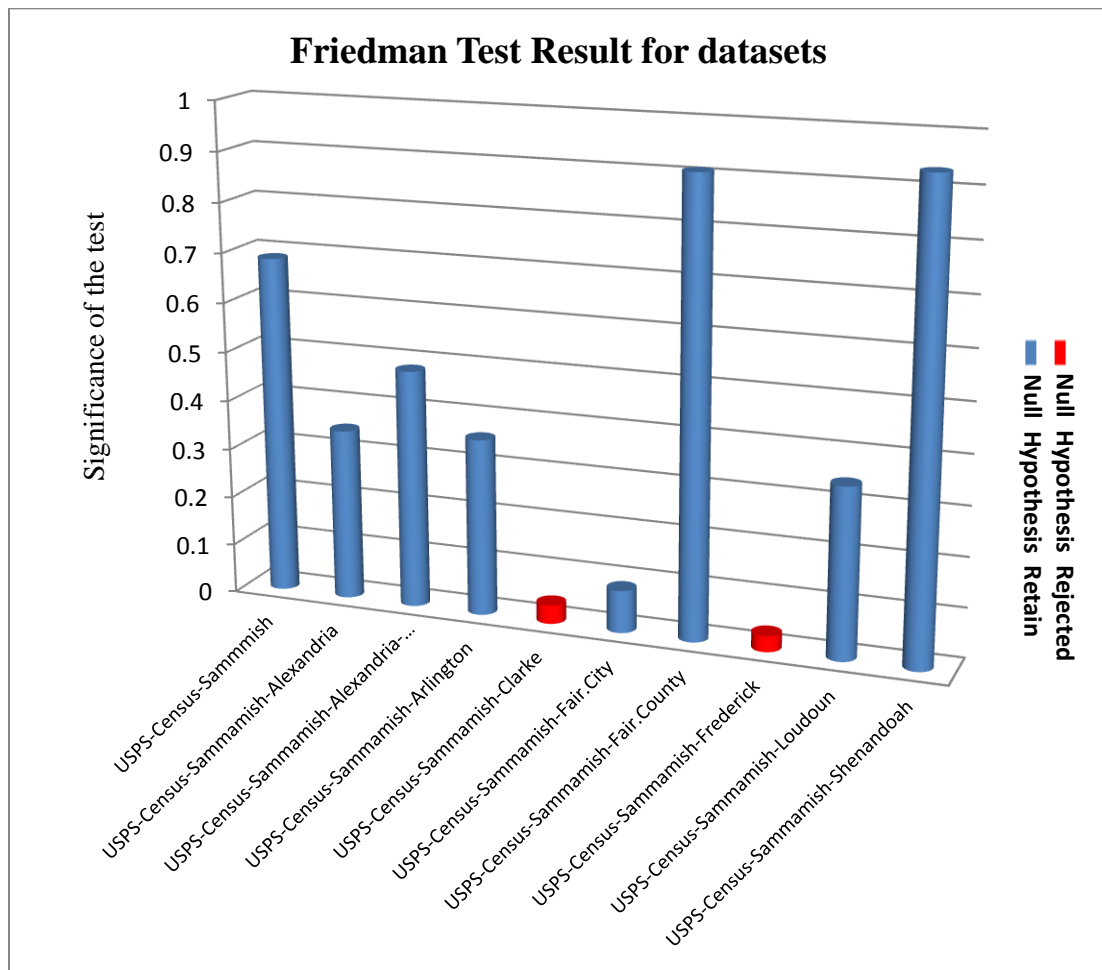


Figure 35: Results from the Friedman non-parametric tests on datasets

Similar results have been found in the Wilcoxon signed rank test. In this 2-related sample test the datasets are similar for 85% of the cases but significantly different for 15% of the cases (Figure 36)

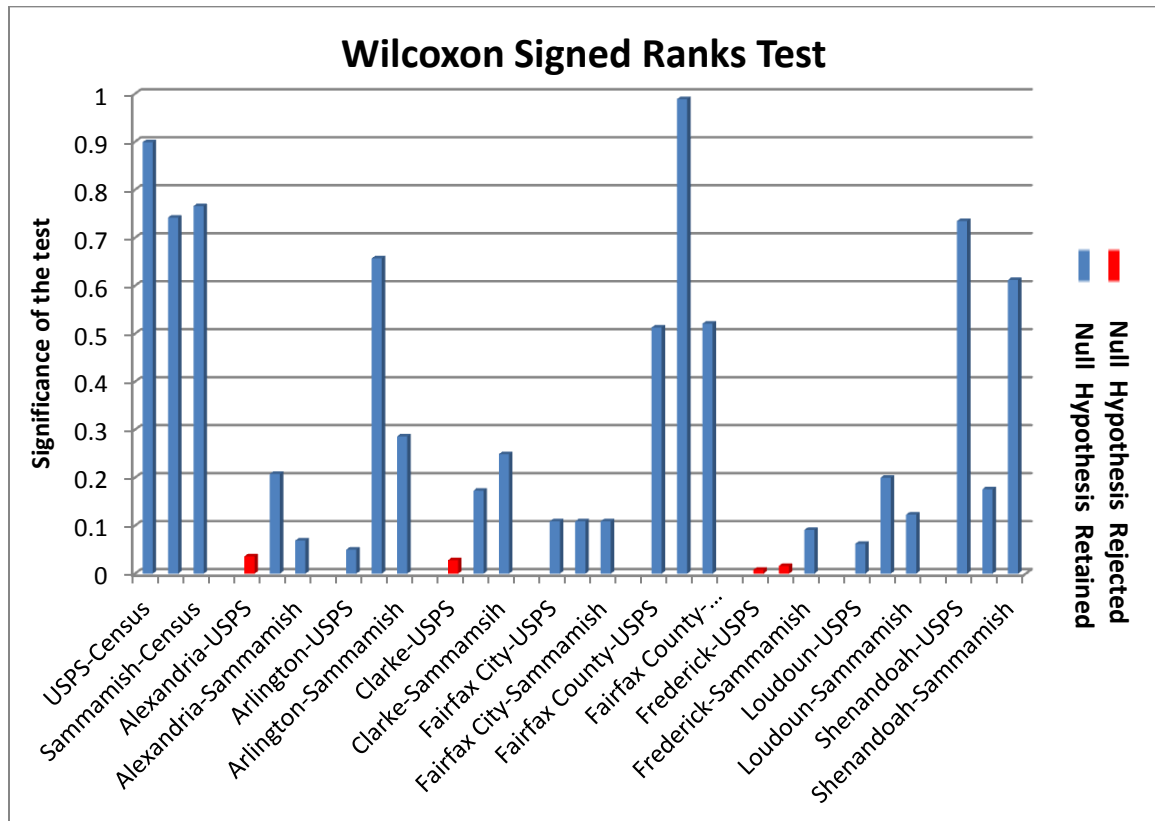


Figure 36: Results from the Wilcoxon Signed Ranks non-parametric tests on datasets.

These non-parametric methods for testing independence between pairs of datasets and across all of the datasets indicate that in most of the cases the datasets do not significantly differ from each other and the area of ZIP Code does not vary within maps. However, for Clarke and Frederick, the probability that the datasets are drawn from the

same population is rejected in all tests. If anyone uses the USPS or Census or Sammamish or Clarke for measuring the area of the ZIP Codes within the Clarke County, the result would be different for each analysis. The same is true if someone uses either of the USPS, Census, Sammamish or Frederick datasets for the ZIP Codes within Frederick County. As the datasets of the ZIP Codes are supposed to be exactly the same, having significant differences for 15-20% cases in all the statistical tests proves the assumption that using different ZIP Code datasets will yield statistically different results in spatial analyses.

10.3.2 Linear regression

For the Census, Frederick, Fairfax County, Loudoun, Sammamish and Shenandoah ZIP Code datasets, the significance of a simple linear regression is smaller than 0.05 (Figure 37). In these cases the null hypothesis of having no influence of USPS dataset on these datasets is rejected. This indicates that an area value within one of these datasets has correspondence within the USPS dataset and is not independent of the USPS dataset. The significance values for the Alexandria, Arlington, Clarke and Fairfax City datasets suggest that the area values within these datasets are not correlated with the USPS area values. Therefore, for 40% of the cases the test reveals no significant correlation between the corresponding area values within the USPS and individual datasets.

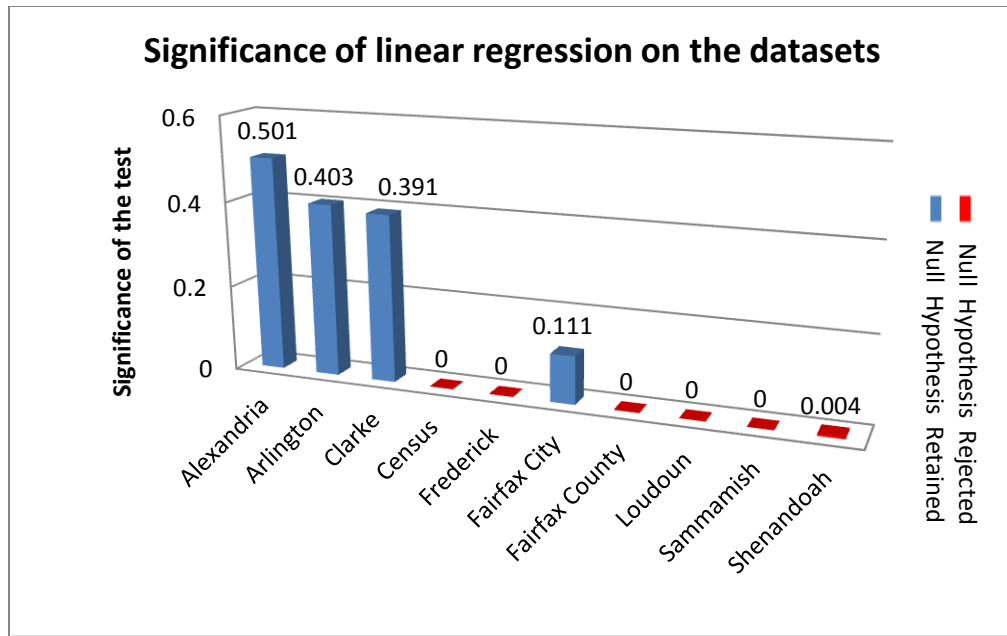


Figure 37: Significance of regression model on the datasets.

However, some of the ZIP Code maps individually have large outliers. For example: in Census there are some outliers that have very large standard deviation (Table 11). ZIP Code 22172 (case no. 77) has a standard residual of 4.43 while for the ZIP Code 22727 (case 149) the standard residual is 5.2.

Table 11: Outliers in a simple linear regression analysis on the USPS and the Census ZIP Code datasets.

Casewise Diagnostics^a

Case Number	Std. Residual	Census	Predicted Value	Residual
18	2.238	141155936	101571908.86	39584027.137
27	2.158	76925640	38752755.23	38172884.772
72	3.582	152243232	88893575.05	63349656.952
77	4.436	118495368	40041368.46	78453999.535
122	-3.394	5719093	65750890.61	-60031797.60
143	-3.584	18792362	82171122.20	-63378760.19
149	-5.185	244886960	336590172.48	-91703212.47
154	-2.344	12845592	54304085.96	-41458493.96
164	-2.248	50100724	89862522.90	-39761798.90

a. Dependent Variable: Census

The results of the independence tests and linear regression reject the probability that the ZIP Code areas are similar across all datasets. Although for many comparisons the areas of ZIP Codes do not show any significant difference across datasets yet for a considerable number of comparisons the datasets are significantly different. These findings support the fact that any spatial analysis employing ZIP Code characteristics will obtain different results with different datasets.

Moreover, it is possible to have similar areas of multiple regions with very different sizes and shapes (Slocum et al. 2008). So even if the area values of the ZIP Code representations show statistical similarity across datasets, it does not eliminate the probability of having significantly different outcome in a spatial analysis using the ZIP

Code maps. Thus, it is needed to do more experiment on the results from spatial analyses conducted in this thesis.

SECTION 11

DISCUSSION

Differences in areas and shapes of ZIP Codes across different data sources can influence the results of data analysis. For example if a researcher tries to analyze crime incidence or population living below poverty level or number of cancer patients and utilizes ZIP Codes as the observation units, the researcher may find 50 patients or 100 crime incidents within a particular ZIP Code when using the USPS dataset whereas for the same ZIP Code there could be only 5 patients or 10 crimes within the Clarke dataset.

Using ZIP Codes as the units of observation can create confusion regarding the outcome of a spatial analysis, as the boundary of ZIP Code polygons collected from a data source often do not match with other data sources (Friedman et al. 2005). Not only may the boundary be mismatched, sometimes ZIP Codes do not even exist in a different data source (Friedman et al. 2005). Therefore, mismatched ZIP Codes cannot be employed in analyses. This is also true for the ZCTA boundaries of the Census Bureau. As the Census Bureau created ZCTAs for tabulating summary statistics from the Census data for the land area covered by each ZIP Code, ZCTAs follow census block boundaries and have little correlation with the USPS ZIP Code boundaries. Geographic analyses of various types of data at the census tract or ZCTA levels may differ from the data at the ZIP code level (Rodriguez et al. 2007). Merging the census-derived and ZIP Code area

data has been shown to leave potential for spatial-temporal mismatch (Inagami et al. 2006). Greater bias can result when using these ZCTA and ZIP Code boundaries than those created by census tracts or block groups (Inagami et al. 2006).

When using ZCTA centroids for spatial analyses, an analyst should be aware of the fact that these may not be the true representation of population location or density. Some areas of ZCTA may have very few people living there or no people at all. Some studies tried to address this problem by not considering ZCTA centroids but the centroid of the largest incorporated area (ZIP Code population center) within the ZCTA. The logic behind this lies in the belief that this measure will provide a better representation of individual's true address.

Data collected for census ZCTAs are also problematic when they are integrated with other information collected at the ZIP Code level from other data sources. Even if a single ZIP Code representation were available it would still be difficult to make a correspondence with the census ZCTAs. For example; the population data of the Census Bureau are available at ZCTA level whereas many health registries; socio-economic; educational; and other organizations collect data at the ZIP Code level. So the population data and health data for the same individual may not assigned to the same area. This introduces potential errors in spatial distribution or accessibility measurement.

ZIP Code centroid distances are frequently used as an approximation of driving distance or driving time between precise geographic locations (Jordan et al. 2004; Goodman et al. 1997). For urban and suburban areas the zip code centroids may be relatively near to an actual residence (when ZIP Code is the demand point) or to a service

(when ZIP Code is the supply point) but the inconsistency increases in rural areas as rural ZIP Codes are larger than urban or suburban ZIP Codes, providing a coarse spatial resolution (Bliss et al. 2012).

Again, distance measures based on ZIP Code centroids are known to overweight locations near boundaries as residences and services located in different ZIP Codes (for example in accessibility measurement) may actually be very close to one another, resulting in estimates that are longer than the true distances (Bliss et al. 2012; Guagliardo 2004). For example; in Votruba and Cebul (2006), in measurement of the effects of patient-hospital distance on the volume of patient to stroke center, more than half of patients were not admitted to the hospital closest to their ZIP Code centroid. This can be true for any accessibility measurement using centroid of a geographic unit as proxy location but it is more troublesome when ZIP Code boundaries are used as they change over time and according to the interpolation method the data sources are using to create the boundaries (Cudnik et al. 2012; Beyer et al. 2011; Grubestic 2008).

Furthermore, the validity of zip codes as proxies for socioeconomic status and other conditions is dependent on the socioeconomic homogeneity within ZIP Codes. Use of zip codes may be less reliable in areas that have greater socioeconomic diversity.

As different sources use different interpolation techniques for creating ZIP Code boundaries and provide no publicly available information about these interpolation techniques, it is difficult to choose and justify a reliable ZIP Code polygon map. In spatial analyses , for example in accessibility measurement, when data are restricted, it is necessary to choose a point location to which population or service data can be attached

and to calculate distances between demand and service locations. This is a very important step in a spatial analysis as any imprecise estimate can compromise the efficacy of policy and programmatic decisions and may misdirect scarce resources.

The choice of a location representing population or service affect any type of measurement and thus the choice should be made carefully. The task would be easier for researchers if some protocols are established for tabulating information at a valid geographic unit and made available those to the public and to assist researchers and practitioners to investigate geographic pattern of accessibility and other spatial measures more effectively and with less uncertainty.

Often data are collected over the span of multiple years. So it is possible to have a case which is assigned to a ZIP Code that is no longer in service. Therefore, it is problematic to preserve information over time using ZIP Codes which compels researchers to eliminate records from the dataset (Krieger et al. 2002; Nancy Krieger et al. 2002). Even though these records are reassigned to current ZIP Codes there remains the potential for significant errors (Schultz, Beyer, and Rushton 2007). It is very difficult to be up to date with the USPS dataset as it changes very frequently. A simple aggregation of exposed data without considering the historical ZIP Code boundaries may introduce potential errors in results (Clary and Ritz 2003). Therefore, any spatial analysis and temporal comparison based on ZIP Code polygons would give flawed outcomes that can be very problematic for scientific research as well as for policy making.

Another shortcoming of using ZIP Codes is the lack of a defined population size as the base for collecting information. The Census Bureau collects information based on

census units for which it maintains an average population size. But there is no definite rule for the population size at ZIP Code level. As a result the ZIP Codes do not have any fixed size for average population across space and data sources. Johnson (2004) discusses a spatial smoothing of population based measurements to address the differing population sizes in ZIP Codes. In spite of the closure of 50 small post offices and addition of 3 new post offices it argues that the ZIP Code polygons are relatively stable service areas. The study combined the ZIP Codes for which the delivery routes changed over a 10 year span of observation.

Some researchers justify the use of ZIP Codes by arguing that they reflect population changes more quickly than census tracts and also updated ZIP Code data are available from private vendors (Carretta and Mick 2003). This particular practice of using updated ZIP Code boundaries has more potential to cause errors in accessibility measurement. The results in this study suggest that even if many commercial products are available they can produce different result for the same analysis and create uncertainty in interpreting these results.

Census tracts are also updated decennially but the temporal changes in tracts are traceable and there are methods to compare census tracts over time (US Bureau of the Census 2011_b). There is no such method for tracing ZIP Code boundaries that are frequently changed by ZIP Code polygon creators creating uncertainty of using ZIP Code boundaries over time. Even if ZIP Code representations from a particular data source were used in all analysis, any reference material should be periodically updated to reflect the newest ZIP Code boundaries updated by that data source. As ZIP Codes are available

from numerous sources, the idea of the updated reference material seems not realistic to maintain with respect to all ZIP Code representations. It is not possible to further use a ZIP Code data if the ZIP Code polygons are updated for any reason, whether it is for operational purpose at the USPS or ZCTA boundary update for removing water body or to match with the USPS ZIP Code boundary. The issue of inconsistency of ZCTA boundaries in spatial analyses at the ZIP Code level has drawn the attention of researchers (Dai 2010; Grubestic and Matisziw 2006; Grubestic 2008; Rodriguez et al. 2007).

Moreover, ZIP codes also do not conform to any other geographic schemes. The census geographic units have some hierarchical system and can recognize other boundaries such as counties or states. There is no such rule for ZIP Codes to follow with respect to other geographies. ZIP Codes can cross state lines, county lines, political jurisdictions (e.g., cities, congressional districts), metro areas, and other geographic boundaries without any correspondence.

Most of the Local Governments maintain their own demographic data from census for only the area and addresses that fall within their own ZIP boundaries. The experiments in this thesis indicate that errors are inevitable when using ZIP Code data for spatial analyses especially when multiple ZIP Code representations are available. No ZIP Code polygon map is entirely correct as they are created bounding around some linear features. Some studies have noticed the challenges with ZIP Codes when the demographic information and other statistics at the ZIP Code level do not match across different data sources due to mismatches in the ZIP Code boundaries (Friedman et al.

2005). Although some studies tried to address this problem by overlaying two different datasets and creating a new boundary, this type of joining may be questionable if either of the datasets lacks permanent boundary layers. To date, no permanent ZIP code boundary layer is available which makes it difficult to rely on the spatial divisions over time (Matisziw, Grubestic, and Wei 2008; Friedman et al. 2005; Grubestic 2008a). For lacking in standardization and very transient nature of ZIP Codes, using these area features does not correspond to real ZIP Code area and causes uncertainty in spatial analyses.

Very often data, available at the ZIP Codes level, are collected from one source and the ZIP Code boundary files are collected from another source. Later these data are joined together for further analyses. For example; if a study tries to evaluate the pattern of a certain type of cancer, it may collect the cancer information from a cancer registry where the cancer records are geocoded to the ZIP Codes of the patient's residence. If the cancer registry does not provide any ZIP Code boundary file, these boundaries must be collected from another source. It is usually not known which sources of ZIP Code boundary a cancer registry uses to register cancer patients or if it creates its own interpolated ZIP Code boundaries. When the study links the cancer data with ZIP Code boundaries collected from the second source (which also employs various types of interpolation methods for creating ZIP Code boundaries), there are possibilities of having large errors in the correspondence between the two datasets. Maybe the worst problem is that most of the time there is no way to measure the errors while compiling two datasets from two different sources. Therefore, the results from these studies remain dubitable.

Despite the limitations ZIP Codes are frequently used in spatial analyses for the advantage of obtaining readily usable information; wide availability; broad geographic coverage; ability to link and compare multiple data sets; and geographic detail of phenomena (Willis et al. 2003; Thomas et al. 2006). ZIP codes are very often used to report an event or phenomenon and thus are included in many datasets of interest. These events or addresses should be address geocoded to obtain their census tract or block groups. Many organizations try to avoid any additional geocoding and thus record information with ZIP Codes. ZIP Codes are also believed to represent greater spatial detail than counties however, Wang (2004) noted that mapping cancer incidents in ZIP Codes may lead to unstable rates. The study combined small ZIP codes based on spatial proximity and by using a population threshold to determine which contiguous ZIP Codes should be combined.

Often ZIP Codes are believed to be a geographic unit that can protect privacy for a group of people. However, Schultz, Beyer, and Rushton (2007) demonstrate that the ZIP Code level of aggregation is unable to sufficiently hide peoples' identity. They refer to the Health Insurance Portability and Accountability Act of 1996 which requires removal of all geographic units smaller than state for protecting patients' information in health related research. However, this article did prefer ZCTA instead of USPS ZIP Code which also suffers from the same problem.

Using census tracts or blocks can reduce the level of problems arising from these ZIP Code issues in the sense that census blocks are more permanent geographic divisions that are a part of nested geographic partitioning system of the Census Bureau (Grubestic

2008a). This ensures existence and repeatability of these divisions because these are used in Census enumerations and even if a new block or finer division is created records for previous blocks would be maintained.

Census blocks or tracts can also be used in research where identity of a person or a group needs to be concealed. Sometime ZIP code polygons are used instead of county boundaries in order to get some detail information. In spite of the advantage of using ZIP Codes or ZCTAs as they offer smaller geographic units than counties in measuring accessibility or other spatial research, Census blocks or tracts can be more efficient for getting more detailed information. It is also convenient for getting readily available information as census blocks are the primary units for tabulating census data.

However, the use of different geographical scales raises scale problems and Modifiable Areal Unit Problems (Fotheringham and Rogerson 2009). For careful analysis, it should be kept in mind that no ZIP Code representation can have correct ZIP Code boundaries as there is no such boundary in reality.

Research that requires geographic precision should avoid using ZIP Codes as the geographic units of observation. The inaccuracies and uncertainties associated with ZIP Codes also can lead to serious mistakes in decision making based on any result that employs ZIP Code polygons or centroids in spatial analysis.

As ZIP code boundaries cannot truly represent any administrative, social or cultural boundaries at national, state or local level; they may not answer fundamental research questions or address key factors for policy decisions. Therefore, a better and

more flexible solution is needed to avoid the use of ZIP Codes for collecting information and utilizing that information in spatial or statistical analyses.

APPENDIX

ZIP Code	USPS	Alexandria	Arlington	Census	Clarke	Fairfax City	Fairfax County	Frederick	Loudoun	Sammamish	Shenandoah
20041	A										
20105	A			D					I	J	
20106	A			D						J	
20107	A									J	
20109	A			D						J	
20110	A			D						J	
20111	A			D						J	
20112	A			D						J	
20115	A			D						J	
20117	A			D					I	J	
20119	A			D						J	
20120	A			D			G		I	J	
20121	A			D			G			J	
20124	A			D			G			J	
20129	A			D					I	J	
20130	A			D	E				I	J	
20132	A			D					I	J	
20135	A			D	E				I	J	
20136	A			D						J	
20137	A			D						J	
20141	A			D					I	J	
20143	A			D						J	
20144	A			D						J	
20147	A			D					I	J	
20148	A			D					I	J	
20151	A			D			G			J	
20152	A			D					I	J	
20155	A			D						J	
20158	A			D					I	J	
20164	A			D					I	J	
20165	A			D					I	J	
20166	A			D					I	J	
20167	A										
20169	A			D						J	
20170	A			D			G		I	J	
20171	A			D			G			J	
20175	A			D					I	J	
20176	A			D					I	J	
20180	A			D					I	J	
20181	A			D						J	
20184	A			D					I	J	
20186	A			D						J	

20187	A			D						J	
20190	A			D			G			J	
20191	A			D			G			J	
20192	A										
20194	A			D			G			J	
20197	A			D					I	J	
20198	A			D						J	
22003	A			D			G			J	
22015	A			D			G			J	
22025	A									J	
22026	A			D						J	
22027	A			D			G			J	
22030	A			D		F	G			J	
22031	A			D		F	G			J	
22032	A			D		F	G			J	
22033	A			D			G			J	
22039	A			D			G			J	
22041	A			D			G			J	
22042	A			D			G			J	
22043	A			D			G			J	
22044	A			D			G			J	
22046	A			D			G			J	
22060	A			D			G			J	
22066	A			D			G		I	J	
22067	A										
22079	A			D			G			J	
22101	A		C	D			G			J	
22102	A			D			G			J	
22124	A			D			G			J	
22134	A			D						J	
22150	A			D			G			J	
22151	A			D			G			J	
22152	A			D			G			J	
22153	A			D			G			J	
22172	A			D						J	
22180	A			D			G			J	
22181	A			D			G			J	
22182	A			D			G			J	
22191	A			D						J	
22192	A			D						J	
22193	A			D						J	
22201	A		C	D						J	
22202	A		C	D						J	
22203	A		C	D						J	
22204	A		C	D						J	
22205	A		C	D						J	
22206	A	B	C	D						J	
22207	A		C	D						J	
22209	A		C	D						J	
22211	A		C	D						J	
22213	A		C	D						J	
22301	A	B		D						J	
22302	A	B		D			G			J	

22304	A	B		D						J	
22305	A	B		D						J	
22306	A			D			G			J	
22307	A			D			G			J	
22308	A			D			G			J	
22309	A			D			G			J	
22310	A			D			G			J	
22311	A	B		D			G			J	
22312	A	B		D			G			J	
22314	A	B		D						J	
22315	A			D			G			J	
22601	A			D				H		J	
22602	A			D				H		J	
22603	A			D				H		J	
22610	A			D						J	
22611	A			D	E					J	
22620	A			D	E					J	
22624	A			D				H		J	
22625	A			D				H		J	
22627	A			D						J	
22630	A			D	E					J	
22637	A			D				H		J	
22639	A			D						J	
22640	A			D						J	
22641	A			D						J	K
22642	A			D						J	
22643	A			D						J	
22644	A			D						J	K
22645	A			D				H		J	
22649	A									J	
22650	A			D						J	
22652	A			D						J	K
22654	A			D				H		J	K
22655	A			D				H		J	
22656	A			D				H		J	
22657	A			D						J	K
22660	A			D						J	K
22663	A			D	E			H		J	
22664	A			D						J	K
22701	A			D						J	
22709	A			D							
22712	A			D						J	
22713	A			D						J	
22714	A			D						J	
22715	A			D						J	
22716	A			D						J	
22718	A			D						J	
22719	A			D							
22720	A			D						J	
22722	A			D							
22724	A			D						J	
22725	A										
22726	A			D						J	

22727	A			D						J	
22728	A			D						J	
22729	A			D						J	
22730	A			D							
22731	A			D							
22732	A			D							
22733	A			D						J	
22734	A			D						J	
22735	A			D						J	
22736	A			D						J	
22737	A			D						J	
22738	A			D						J	
22740	A			D						J	
22741	A			D						J	
22742	A			D						J	
22743	A			D							
22746	A			D							
22747	A			D						J	
22749	A			D						J	

INDEX

REFERENCES

REFERENCES

- Algert, Susan J, Aditya Agrawal, and Douglas S Lewis. 2006. "Disparities in Access to Fresh Produce in Low-income Neighborhoods in Los Angeles." *American Journal of Preventive Medicine* 30 (5) (May): 365–370. doi:10.1016/j.amepre.2006.01.009.
- Bajari, Patrick, and C. Lanier Benkard. 2001. "Demand Estimation With Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach." *National Bureau of Economic Research Technical Working Paper Series* No. 272. <http://www.nber.org/papers/t0272>.
- Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A unified framework for measuring preferences for schools and neighborhoods." *Journal of Political Economy* 115 (4) (August): 588–638. doi:10.1086/522381.
- Beyer, Kirsten MM, Audrey F Saftlas, Anne B Wallis, Corinne Peek-Asa, and Gerard Rushton. 2011. "A Probabilistic Sampling Method (PSM) for Estimating Geographic Distance to Health Services When Only the Region of Residence Is Known." *International Journal of Health Geographics* 10 (1): 4. doi:10.1186/1476-072X-10-4.
- Bliss, Robin L., Jeffrey N. Katz, Elizabeth A. Wright, and Elena Losina. 2012. "Estimating Proximity to Care." *Medical Care* 50 (1) (January): 99–106. doi:10.1097/MLR.0b013e31822944d1.
- Bonner, Matthew R., Daikwon Han, Jing Nie, Peter Rogerson, John E. Vena, and Jo L. Freudenheim. 2003. "Positional Accuracy of Geocoded Addresses in Epidemiologic Research." *Epidemiology* 14 (4) (July 1): 408–412.
- Brabyn, Lars, and Chris Skelly. 2002. "Modeling Population Access to New Zealand Public Hospitals." *International Journal of Health Geographics* 1 (1) (November 12): 3. doi:10.1186/1476-072X-1-3.
- Burt E. James, Barber M. Gerald, and Rigby L. David. 2009. *Elementary Statistics for Geographers, Third Edition*. Third ed. The Guilford Press.
- Carretta, H. Y., and S. S. Mick. 2003. "Geocoding public health data." *American Journal of Public Health* 93 (5) (May): 699–699. doi:10.2105/AJPH.93.5.699.
- Carson, Valerie, Stefan Kuhle, John C Spence, and Paul J Veugelers. 2010. "Parents' Perception of Neighbourhood Environment as a Determinant of Screen Time,

- Physical Activity and Active Transport.” *Canadian Journal of Public Health. Revue Canadienne De Santé Publique* 101 (2) (April): 124–127.
- Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. “The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design.” *Quarterly Journal of Economics* 125 (1) (February): 215–261.
- Chang, Virginia W. 2006. “Racial Residential Segregation and Weight Status Among US Adults.” *Social Science & Medicine* 63 (5) (September): 1289–1303. doi:10.1016/j.socscimed.2006.03.049.
- Cinnamon, Jonathan, Nadine Schuurman, and Valorie A Crooks. 2008. “A Method to Determine Spatial Access to Specialized Palliative Care Services Using GIS.” *BMC Health Services Research* 8: 140. doi:10.1186/1472-6963-8-140.
- Clark, D. E., and W. E. Herrin. 2000. “The impact of public school attributes on home sale prices in California.” *Growth and Change* 31 (3): 385–407. doi:10.1111/0017-4815.00134.
- Clary, Tim, and Beate Ritz. 2003. “Pancreatic Cancer Mortality and Organochlorine Pesticide Exposure in California, 1989-1996.” *American Journal of Industrial Medicine* 43 (3) (March): 306–313. doi:10.1002/ajim.10188.
- Cudnik, Michael T., Jing Yao, Dana Zive, Craig Newgard, and Alan T. Murray. 2012. “Surrogate Markers of Transport Distance for Out-of-Hospital Cardiac Arrest Patients.” *Prehospital Emergency Care* 16 (2) (June): 266–272. doi:10.3109/10903127.2011.615009.
- Curtin, K. M., S. Biba, and G. Manca. 2010. “A New Method for Determining the Population with Walking Access to Transit.” *International Journal of Geographical Information Science* 24 (3) (March): 347–364. doi:10.1080/13658810802646679.
- Dai, Dajun. 2010. “Black Residential Segregation, Disparities in Spatial Access to Health Care Facilities, and Late-stage Breast Cancer Diagnosis in Metropolitan Detroit.” *Health & Place* 16 (5) (September): 1038–1052. doi:10.1016/j.healthplace.2010.06.012.
- Donato, Ruben, and Herman Garcia. 1992. “Language Segregation in Desegregated Schools: A Question of Equity.” *Equity and Excellence* 25 (2): 94–99.
- Eicher, Cory, and Cynthia Brewer. 2001. “Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation.” *Cartography and Geographic Information Science* (April): 125–138.

- Fairfax County Government (2011). *Community Health Status Assessment*. Community Report. Retrieved May 05, 2012 from: <http://www.fairfaxcounty.gov/hd/mapp/pdf/comm-health-assessment.pdf>
- Fiscella, K, and P Franks. 2001. "Impact of Patient Socioeconomic Status on Physician Profiles: a Comparison of Census-derived and Individual Measures." *Medical Care* 39 (1) (January): 8–14.
- Foda, Mohamed, and Ahmed Osman. 2010. "Using GIS for Measuring Transit Stop Accessibility Considering Actual Pedestrian Road Network." *World Transit Research* (January 1). <http://www.worldtransitresearch.info/research/3854>.
- Fotheringham A. Stewart, and Rogerson Peter. 2009. *The SAGE Handbook of Spatial Analysis*. Los Angeles: SAGE Publications.
- Franks, Peter, and Kevin Fiscella. 2002. "Effect of Patient Socioeconomic Status on Physician Profiles for Prevention, Disease Management, and Diagnostic Testing Costs." *Medical Care* 40 (8) (August): 717–724. doi:10.1097/01.MLR.0000020931.02753.72.
- Friedman, S., A. Singer, M. Price, and I. Cheung. 2005. "Race, immigrants, and residence: A new racial geography of Washington, DC." *Geographical Review* 95 (2) (April): 210–230.
- Fu, Linda Y., Nuala Cowan, Rosie McLaren, Ryan Engstrom, and Stephen J. Teach. 2009. "Spatial Accessibility to Providers and Vaccination Compliance Among Children With Medicaid." *Pediatrics* 124 (6) (December 1): 1579–1586. doi:10.1542/peds.2009-0233.
- Fuller, B., and X. Y. Liang. 1996. "Market failure? Estimating inequality in preschool availability." *Educational Evaluation and Policy Analysis* 18 (1): 31–49. doi:10.3102/01623737018001031.
- Fuller, B., and A. Strath. 2001. "The child-care and preschool workforce: Demographics, earnings, and unequal distribution." *Educational Evaluation and Policy Analysis* 23 (1): 37–55. doi:10.3102/01623737023001037.
- George, Darren, and Paul Mallery. 2010. *SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update*. 11th ed. Prentice Hall.
- Goodchild, Mf, and Nsn Lam. 1980. "Areal Interpolation - a Variant of the Traditional Spatial Problem." *Geo-Processing* 1 (3): 297–312.

- Goodman, D C, E Fisher, T A Stukel, and C Chang. 1997. "The Distance to Community Medical Care and the Likelihood of Hospitalization: Is Closer Always Better?" *American Journal of Public Health* 87 (7) (July): 1144–1150.
- Govind, Raw, Rabikar Chatterjee, and Vikas Mittal. 2008. "Timely access to health care: Customer-focused resource allocation in a hospital network." *International Journal of Research in Marketing* 25 (4) (December): 294–300. doi:10.1016/j.ijresmar.2008.07.005.
- Grubestic, Tony H, and Timothy C Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data." *International Journal of Health Geographics* 5 (December 13): 58. doi:10.1186/1476-072X-5-58.
- Grubestic, Tony H. 2008a. "Zip Codes and Spatial Analysis: Problems and Prospects." *Socio-Economic Planning Sciences* 42 (2) (June): 129–149. doi:16/j.seps.2006.09.001.
- . 2008b. "Spatial Data Constraints: Implications for Measuring Broadband." *Telecommunications Policy* 32 (7) (August): 490–502. doi:16/j.telpol.2008.05.002.
- Gruenewald, Paul J., Fred W. Johnson, and Andrew J. Treno. 2002. "Outlets, Drinking and Driving: A Multilevel Analysis of Availability." *Journal of Studies on Alcohol and Drugs* 63 (4) (July 1): 460.
- Guagliardo, Mark F. 2004. "Spatial Accessibility of Primary Care: Concepts, Methods and Challenges." *International Journal of Health Geographics* 3 (1) (February 26): 3. doi:10.1186/1476-072X-3-3.
- Haas, Jennifer, Craig Earle, John Orav, Phyllis Brawarsky, Bridget Neville, and David Williams. 2008. "Racial Segregation and Disparities in Cancer Stage for Seniors." *Journal of General Internal Medicine* 23 (5): 699–705. doi:10.1007/s11606-008-0545-9.
- Halla, Y. N., A. M. O'Harea, B. A. Younga, E. J. Boyko, and G. M. Chertow. 2008. "Neighborhood poverty and kidney transplantation among US Asians and Pacific Islanders with end-stage renal disease." *American Journal of Transplantation* 8 (11) (November): 2402–2409. doi:10.1111/j.1600-6143.2008.02413.x.
- Hartley, D, L Quam, and N Lurie. 1994. "Urban and Rural Differences in Health Insurance and Access to Care." *The Journal of Rural Health: Official Journal of the American Rural Health Association and the National Rural Health Care Association* 10 (2): 98–108.

- Haurin, D. R., and D. Brasington. 1996. "School quality and real house prices: Inter- and intrametropolitan effects." *Journal of Housing Economics* 5 (4) (December): 351–368. doi:10.1006/jhec.1996.0018.
- Hayes, Kathy J., and Lori L. Taylor. 1996. "Neighborhood School Characteristics: What Signals Quality to Homebuyers?" *Economic and Financial Policy Review* (Q IV): 2–9.
- Hayunga, Darren K., and R. Kelley Pace. 2010. "Spatial Statistics Applied to Commercial Real Estate." *Journal of Real Estate Finance and Economics* 41 (2) (August): 103–125. doi:10.1007/s11146-009-9190-2.
- Hebert, Paul L., Mark R. Chassin, and Elizabeth A. Howell. 2011. "The Contribution of Geography to Black/White Differences in the Use of Low Neonatal Mortality Hospitals in New York City." *Medical Care* 49 (2) (February): 200–206. doi:10.1097/MLR.0b013e3182019144.
- Horowitz, John, Stanley Keil, and Lee Spector. 2009. "Do Charter Schools Affect Property Values?" *Review of Regional Studies* 39 (3): 297–316.
- Hurley, Susan E., Theresa M. Saunders, Rachna Nivas, Andrew Hertz, and Peggy Reynolds. 2003. "Post Office Box Addresses: A Challenge for Geographic Information System-Based Studies." *Epidemiology* 14 (4) (July 1): 386–391.
- Immergluck, Dan. 2011. "From Minor to Major Player: The Geography of Fha Lending During the U.s. Mortgage Crisis." *Journal of Urban Affairs* 33 (1): 1–20. doi:10.1111/j.1467-9906.2010.00539.x.
- Immergluck, Daniel. 1998. "Job Proximity and the Urban Employment Problem: Do Suitable Nearby Jobs Improve Neighbourhood Employment Rates?" *Urban Studies* 35 (1) (January 1): 7–23. doi:10.1080/0042098985041.
- Inagami, S., L. N. Borrell, M. D. Wong, J. Fang, M. F. Shapiro, and S. M. Asch. 2006. "Residential segregation and Latino, black and white mortality in New York City." *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 83 (3) (May): 406–420. doi:10.1007/s11524-006-9035-8.
- Johnson, Glen D. 2004. "Small Area Mapping of Prostate Cancer Incidence in New York State (USA) Using Fully Bayesian Hierarchical Modelling." *International Journal of Health Geographics* 3 (1) (December 8): 29. doi:10.1186/1476-072X-3-29.

- Jordan, Hannah, Paul Roderick, David Martin, and Sarah Barnett. 2004. "Distance, Rurality and the Need for Care: Access to Health Services in South West England." *International Journal of Health Geographics* 3 (1) (September 29): 21. doi:10.1186/1476-072X-3-21.
- Joseph, Alun E., and David Phillips. 1984. *Accessibility and Utilization: Geographical Perspectives on Health Care Delivery*. Sage Publications Ltd.
- Jud, Gd, and Jm Watts. 1981. "Schools and Housing Values." *Land Economics* 57 (3): 459–470. doi:10.2307/3146025.
- Kiel, Katherine A., and Jeffrey E. Zabel. 2008. "Location, location, location: The 3L Approach to house price determination." *Journal of Housing Economics* 17 (2) (June): 175–190. doi:10.1016/j.jhe.2007.12.002.
- Knapp, K K, and K Hardwick. 2000. "The Availability and Distribution of Dentists in Rural ZIP Codes and Primary Care Health Professional Shortage Areas (PC-HPSA) ZIP Codes: Comparison with Primary Care Providers." *Journal of Public Health Dentistry* 60 (1): 43–48.
- Krieger, N., P. Waterman, J. T. Chen, M. J. Soobader, S. V. Subramanian, and R. Carson. 2002. "Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas - The public health disparities geocoding project." *American Journal of Public Health* 92 (7) (July): 1100–1102. doi:10.2105/AJPH.92.7.1100.
- Krieger, Nancy, Jarvis T Chen, Pamela D Waterman, Mah-Jabeen Soobader, S V Subramanian, and Rosa Carson. 2002. "Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project." *American Journal of Epidemiology* 156 (5) (September 1): 471–482.
- LaCour-Little, Michael, Charles A. Calhoun, and Wei Yu. 2011. "What role did piggyback lending play in the housing bubble and mortgage collapse?" *Journal of Housing Economics* 20 (2) (June): 81–100. doi:10.1016/j.jhe.2010.11.002.
- Lankford, Hamilton, and James Wyckoff. 2006. "The Effect of School Choice and Residential Location on the Racial Segregation of Students." *Advances in Applied Microeconomics* 14 (December 7): 185–239. doi:10.1016/S0278-0984(06)14008-0.
- Leckie, George, and Harvey Goldstein. 2009. "The Limitations of Using School League Tables to Inform School Choice." *Journal of the Royal Statistical Society: Series*

A (Statistics in Society) 172 (4): 835–851. doi:10.1111/j.1467-985X.2009.00597.x.

- Li, Kaidong, and Nigel M. Waters. “Transportation Networks, Case-Based Reasoning and Traffic Collision Analysis: A Methodology for the 21st Century.” In *Methods and Models in Transport and Telecommunications*, ed. Aura Reggiani and Laurie A. Schintler, 63–92. Berlin/Heidelberg: Springer-Verlag.
http://rd.springer.com/chapter/10.1007/3-540-28550-4_5.
- Luo, Wei, and Yi Qi. 2009. “An Enhanced Two-step Floating Catchment Area (E2SFCA) Method for Measuring Spatial Accessibility to Primary Care Physicians.” *Health & Place* 15 (4) (December): 1100–1107. doi:10.1016/j.healthplace.2009.06.002.
- Luo, Wei, and Fahui Wang. 2003. “Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region.” *Environment and Planning B: Planning and Design* 30 (6): 865 – 884. doi:10.1068/b29120.
- Luo, Wei, Fahui Wang, and Carolinda Douglass. 2004. “Temporal Changes of Access to Primary Health Care in Illinois (1990–2000) and Policy Implications.” *Journal of Medical Systems* 28 (3): 287–299. doi:10.1023/B:JOMS.0000032845.90730.84.
- Marion, Justin. 2009. “Firm racial segregation and affirmative action in the highway construction industry.” *Small Business Economics* 33 (4) (December): 441–453. doi:10.1007/s11187-009-9204-8.
- Massey, Ds, and Na Denton. 1988. “The Dimensions of Residential Segregation.” *Social Forces* 67 (2) (December): 281–315. doi:10.2307/2579183.
- Matisziw, T.C., T.H. Grubestic, and H. Wei. 2008. “Downscaling Spatial Structure for the Analysis of Epidemiological Data.” *Computers, Environment and Urban Systems* 32 (1) (January): 81–93. doi:10.1016/j.compenvurbsys.2007.06.002.
- Max Sarah. 2010. “Good Schools, Bad Real Estate.” *Wall Street Journal*, June 25, sec. Homes.
<http://online.wsj.com/article/SB10001424052748704009804575308951902854896.html?KEYWORDS=Michael+Sklarz>.
- McCarthy, John F, and Frederic C Blow. 2004. “Older Patients with Serious Mental Illness: Sensitivity to Distance Barriers for Outpatient Care.” *Medical Care* 42 (11) (November): 1073–1080.

- McElroy, Jane A., Patrick L. Remington, Amy Trentham-Dietz, Stephanie A. Robert, and Polly A. Newcomb. 2003. "Geocoding Addresses from a Large Population-Based Study: Lessons Learned." *Epidemiology* 14 (4) (July 1): 399–407.
- McGrew, Jr., J. Chapman, and Charles Monroe. 1999. *An Introduction to Statistical Problem Solving in Geography*. 2nd ed. McGraw-Hill Science/Engineering/Math.
- Messina, Joseph P., Ashton M. Shortridge, Richard E. Groop, Pariwate Varnakovida, and Mark J. Finn. 2006. "Evaluating Michigan's Community Hospital Access: Spatial Methods for Decision Support." *International Journal of Health Geographics* 5 (1) (September 22): 42. doi:10.1186/1476-072X-5-42.
- Van Meter, E., A.B. Lawson, N. Colabianchi, M. Nichols, J. Hibbert, D. Porter, and A.D. Liese. 2011. "Spatial Accessibility and Availability Measures and Statistical Properties in the Food Environment." *Spatial and Spatio-temporal Epidemiology* 2 (1) (March): 35–47. doi:10.1016/j.sste.2010.09.009.
- Mishra, Sabyasachee, Timothy F. Welch, and Manoj K. Jha. 2012. "Performance indicators for public transit connectivity in multi-modal transportation networks." *Transportation Research Part a-Policy and Practice* 46 (7) (August): 1066–1085. doi:10.1016/j.tra.2012.04.006.
- Mitchell, Douglas E., Michael Batie, and Ross E. Mitchell. 2010. "The Contributions of School Desegregation to Housing Integration: Case Studies in Two Large Urban Areas." *Urban Education* 45 (2) (March): 166–193. doi:10.1177/0042085908322711.
- Nagaraja, Chaitra H., Lawrence D. Brown, and Linda H. Zhao. 2011. "An Autoregressive Approach to House Price Modeling." *Annals of Applied Statistics* 5 (1) (March): 124–149. doi:10.1214/10-AOAS380.
- Ngamini Ngui, André, and Alain Vanasse. "Assessing Spatial Accessibility to Mental Health Facilities in an Urban Environment." *Spatial and Spatio-temporal Epidemiology* (0). doi:10.1016/j.sste.2011.11.001. <http://www.sciencedirect.com/science/article/pii/S1877584511000566>.
- Ngui, André Ngamini, and Philippe Apparicio. 2011. "Optimizing the Two-step Floating Catchment Area Method for Measuring Spatial Accessibility to Medical Clinics in Montreal." *BMC Health Services Research* 11: 166. doi:10.1186/1472-6963-11-166.
- Orfield G. & Lee C. 2005. *Why segregation matters: Poverty and educational inequality*. Cambridge, MA: Harvard University, The Civil Rights Project. Retrieved May 05, 2012 from: <http://www.civilrightsproject.harvard.edu>

- Parker, E B, and J L Campbell. 1998. "Measuring Access to Primary Medical Care: Some Examples of the Use of Geographical Information Systems." *Health & Place* 4 (2) (June): 183–193.
- Patel, Alka B, Nigel M Waters, and William A Ghali. 2007. "Determining Geographic Areas and Populations with Timely Access to Cardiac Catheterization Facilities for Acute Myocardial Infarction Care in Alberta, Canada." *International Journal of Health Geographics* 6 (October 16): 47. doi:10.1186/1476-072X-6-47.
- Peck Emily. 2008. "Buying a new home: How important is the school district?", The Wall Street Journal, Feb 20, 2008. Retrieved May 10, 2012 from: <http://blogs.wsj.com/developments/>.
- Pedigo, Ashley S., and Agricola Odoi. 2010. "Investigation of Disparities in Geographic Accessibility to Emergency Stroke and Myocardial Infarction Care in East Tennessee Using Geographic Information Systems and Network Analysis." *Annals of Epidemiology* 20 (12) (December): 924–930. doi:10.1016/j.annepidem.2010.06.013.
- Peng, Zhong-Ren. 1997. "The Jobs-Housing Balance and Urban Commuting." *Urban Studies* 34 (8) (July 1): 1215–1235. doi:10.1080/0042098975600.
- Peters, J, and G B Hall. 1999. "Assessment of Ambulance Response Performance Using a Geographic Information System." *Social Science & Medicine* (1982) 49 (11) (December): 1551–1566.
- Pollack, Craig Evan, Shanu K. Kurd, Alice Livshits, Mark Weiner, and Julia Lynch. 2011. "A Case-Control Study of Home Foreclosure, Health Conditions, and Health Care Utilization." *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 88 (3) (June): 469–478. doi:10.1007/s11524-011-9564-7.
- Qureshi, M. A., H. L. Hwang, and S. M. Chin. 2002. "Comparison of distance estimates for commodity flow survey - Great circle distances versus network-based distances." In *Transportation Data and Information Technology Research: Planning and Administration*, 212–216. Washington: Transportation Research Board Natl Research Council.
- Radke, John, and Lan Mu. 2000. "Spatial Decompositions, Modeling and Mapping Service Regions to Predict Access to Social Programs." *Annals of GIS* 6 (2): 105–112. doi:10.1080/10824000009480538.
- Roberts, Sam. 2007. "An Elite ZIP Code Becomes Harder to Crack." *[[The New York Times]]* (March 21): p. C15.

- Rodriguez, Rudolph A., Saunak Sen, Kala Mehta, Sandra Moody-Ayers, Peter Bacchetti, and Ann M. O'Hare. 2007. "Geography matters: Relationships among urban residential segregation, dialysis facilities, and patient outcomes." *Annals of Internal Medicine* 146 (7) (April 3): 493–501.
- Schultz, Alan, Kirsten Beyer, and Gerard Rushton. 2007. "Using ZIP-Æ Codes as Geocodes in Cancer Research." In *Geocoding Health Data*, ed. Barry Greene, Michele West, Gerard Rushton, Josephine Gittler, Marc Armstrong, Claire Pavlik, and Dale Zimmerman, 37–67. CRC Press.
<http://www.crcnetbase.com/doi/abs/10.1201/9780849384332.ch3>.
- Shan, Hui. 2011. "Reversing the Trend: The Recent Expansion of the Reverse Mortgage Market." *Real Estate Economics* 39 (4): 743–768. doi:10.1111/j.1540-6229.2011.00310.x.
- Shaw, Gareth, and Dennis Wheeler. 1994. *Statistical Techniques in Geographical Analysis, 2nd Edition*. 2nd ed. Wiley.
- Shi, X. 2007. "Evaluating the Uncertainty Caused by Post Office Box Addresses in Environmental Health Studies: A Restricted Monte Carlo Approach." *International Journal of Geographical Information Science* 21 (3): 325. doi:10.1080/13658810600924211.
- Shihadeh, Edward S., and Michael O. Maume. 1997. "Segregation and Crime The Relationship Between Black Centralization and Urban Black Homicide." *Homicide Studies* 1 (3) (August 1): 254–280. doi:10.1177/1088767997001003004.
- Talbot, T O, M Kulldorff, S P Forand, and V B Haley. 2000. "Evaluation of Spatial Filters to Create Smoothed Maps of Health Data." *Statistics in Medicine* 19 (17-18) (September 15): 2399–2408.
- Thomas, Avis J., Lynn E. Eberly, George Davey Smith, and James D. Neaton. 2006. "ZIP-Code-Based Versus Tract-Based Income Measures as Long-Term Risk-Adjusted Mortality Predictors." *American Journal of Epidemiology* 164 (6) (September 15): 586–590. doi:10.1093/aje/kwj234.
- Thornton, Lukar E., Jamie R. Pearce, and Anne M. Kavanagh. 2011. "Using Geographic Information Systems (GIS) to Assess the Role of the Built Environment in Influencing Obesity: a Glossary." *International Journal of Behavioral Nutrition and Physical Activity* 8 (1) (July 1): 71. doi:10.1186/1479-5868-8-71.
- US Bureau of the Census. 2000. "Census 2000 ZCTAs ZIP Code tabulation areas technical documentation." Available at http://www.census.gov/geo/ZCTA/zcta_tech_doc.pdf. Accessed May 30. 2012.

- . 2001. Census 2000 ZIP Code tabulation area (ZCTA) frequently asked questions. Available at: <http://www.census.gov/geo/ZCTA/zctafaq.html>. Accessed May 14, 2012.
- . 2011. "ZIP Code tabulation areas (ZCTAs)." Available at: <http://www.census.gov/geo/ZCTA/zcta.html>. Accessed Dec 30, 2011
- . 2011b. "Census 2010 Participant Statistical Areas Program (PSAP) and Census Bureau Input on Census Tracts and Block Groups". Available at: http://www.census.gov/geo/www/psap2010/tract_criteria.pdf. Accessed May 3, 2011
- Votruba, Mark E., and Randall D. Cebul. 2006. "Redirecting Patients to Improve Stroke Outcomes." *Medical Care* 44 (12) (December): 1129–1136. doi:10.1097/01.mlr.0000237424.15716.47.
- Walden, Michael L. 1990. "Magnet Schools and the Differential Impact of School Quality on Residential Property Values." *Journal of Real Estate Research* 5 (2): 221–230.
- Wallace, R. G. 2003. "AIDS in the HAART era: New York's heterogeneous geography." *Social Science & Medicine* 56 (6) (March): 1155–1171. doi:10.1016/S0277-9536(02)00121-1.
- Walton, Emily. 2009. "Residential Segregation and Birth Weight Among Racial and Ethnic Minorities in the United States." *Journal of Health and Social Behavior* 50 (4) (December 1): 427–442. doi:10.1177/002214650905000404.
- Wan, Neng, F. Benjamin Zhan, Bin Zou, and Edwin Chow. 2012. "A Relative Spatial Access Assessment Approach for Analyzing Potential Spatial Access to Colorectal Cancer Services in Texas." *Applied Geography* 32 (2) (March): 291–299. doi:10.1016/j.apgeog.2011.05.001.
- Wang, Fahui. 2000. "Modeling Commuting Patterns in Chicago in a GIS Environment: A Job Accessibility Perspective." *The Professional Geographer* 52 (1): 120–133. doi:10.1111/0033-0124.00210.
- . 2004a. "Spatial Clusters of Cancers in Illinois 1986-2000." *Journal of Medical Systems* 28 (3): 237–56.
- . 2006. *Quantitative Methods And Applications in Gis*. CRC Press.
- Wang, Fahui, and Wei Luo. 2005a. "Assessing Spatial and Nonspatial Factors for Healthcare Access: Towards an Integrated Approach to Defining Health Professional Shortage Areas." *Health & Place* 11 (2) (June): 131–146.

- . 2005b. “Assessing Spatial and Nonspatial Factors for Healthcare Access: Towards an Integrated Approach to Defining Health Professional Shortage Areas.” *Health & Place* 11 (2) (June): 131–146.
- Wang, Fahui, and W. William Minor. 2004. “Where the Jobs Are: Employment Access and Crime Patterns in Cleveland.” *Annals of the Association of American Geographers* 92 (3) (November 5): 435–450. doi:10.1111/1467-8306.00298.
- Wang, Fahui, ed. 2004b. *Geographic Information Systems and Crime Analysis*. IGI Global. <http://www.igi-global.com/chapter/integrating-gis-maximal-covering-models/18826>.
- Weissman, J S, R Stern, S L Fielding, and A M Epstein. 1991. “Delayed Access to Health Care: Risk Factors, Reasons, and Consequences.” *Annals of Internal Medicine* 114 (4) (February 15): 325–331.
- White, E, and T E Aldrich. 1999. “Geographic Studies of Pediatric Cancer Near Hazardous Waste Sites.” *Archives of Environmental Health* 54 (6) (December): 390–397. doi:10.1080/00039899909603370.
- Willis, Alette, Daniel Krewski, Michael Jerrett, Mark S Goldberg, and Richard T Burnett. 2003. “Selection of Ecologic Covariates in the American Cancer Society Study.” *Journal of Toxicology and Environmental Health. Part A* 66 (16-19) (October 22): 1563–1589. doi:10.1080/15287390306425.
- Wilson, James L., and Christopher J. Mansfield. 2010. “Disease, Death, and the Body Politic.” *International Journal of Applied Geospatial Research* 1 (3): 49–68. doi:10.4018/jagr.2010070104.
- Wing, P, and C Reynolds. 1988. “The Availability of Physician Services: a Geographic Analysis.” *Health Services Research* 23 (5) (December): 649–667.
- Wong, D. W. 2005. “Formulating a general spatial segregation measure.” *Professional Geographer* 57 (2) (May): 285–294. doi:10.1111/j.0033-0124.2005.00478.x.
- Yang, Duck-Hye, Robert Goerge, and Ross Mullner. 2006. “Comparing GIS-Based Methods of Measuring Spatial Accessibility to Health Services.” *Journal of Medical Systems* 30 (1): 23–32. doi:10.1007/s10916-006-7400-5.
- Zahirovic-Herbert, Velma, and Geoffrey K. Turnbull. 2008. “School quality, house prices and liquidity.” *Journal of Real Estate Finance and Economics* 37 (2) (August): 113–130. doi:10.1007/s11146-007-9081-3.

Zhang, Xingyou, Hua Lu, and James B. Holt. 2011. "Modeling Spatial Accessibility to Parks: a National Study." *International Journal of Health Geographics* 10 (1) (May 9): 31. doi:10.1186/1476-072X-10-31.

CURRICULUM VITAE

Tunaggina Subrina Khan graduated from Viqarunnisa Noon School and College, Dhaka, Bangladesh in 2001. She received her Bachelor of Sciences from University of Dhaka, Bangladesh in 2006.