ACCESS TO HOSPITALS IN A REGULATED HEALTH CARE SYSTEM: IMPLICATIONS FOR UTILIZATION

By

Paul L. Delamater

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Geography

2012

ABSTRACT

ACCESS TO HOSPITALS IN A REGULATED HEALTH CARE SYSTEM:
IMPLICATIONS FOR UTILIZATION

By

Paul L. Delamater

Hospital use varies among populations due to access, socio-demographic characteristics, and overall health care needs. Further, the interaction between populations and health care providers is often mediated by governing bodies, such as Certificate of Need programs, that regulate the supply of health care resources. The intersection of the spatial and aspatial components of access and utilization within a regulated health care market provide the backdrop for this research. The outcomes provide insights that inform future health services research and offer guidance for public policy initiatives. The research approach adopted in this dissertation addresses both methodological and theoretical issues related to *the study* of access and utilization and the nature of the relationship between them. The project is comprised by three sequential studies tied together within the framework of assessing access and utilization in a regulated health system.

The first study examines methods used to measure distance among locations. Specifically, the work addresses the theoretical and applied implications of using raster and network data models for identifying areas with limited geographic accessibility. The findings suggest that the network data model provides a more accurate framework for estimating vehicular travel time along roadways, while the raster data model offers advantages in scenarios where roadways are not the primary route of travel. The second study offers a methodology for clustering spatial observations having multiple attribute values. The specific focus of the work is the formation of Hospital Groups, the allocation units used in a state-level methodology for predicting future hospital bed demand. The main outcome of the research is the methodology itself, which provides a substantial advance over the previous methodologies used in health services research by way of its ability to cluster observations based on overall

patterns of health care utilization and geographic location, simultaneously. Using knowledge gained from the first two studies, the final portion of the dissertation explores the relationship between the availability of hospital beds and the utilization of hospital services. The focus of the study is Roemer's Law, which states that *a hospital bed built is a bed filled*. The findings of this study provide strong support for the concept that greater levels of hospital bed availability lead to higher hospital utilization rates. This relationship is confirmed at various levels of data aggregation, demonstrating that the observed impact of availability on utilization is stable across geographic scales of analysis.

The main outcomes of this research can be separated into those relating to advancement in health services research and those relating to public policy. From a public policy perspective, this dissertation offers updated methodologies for identifying areas with limited geographic accessibility and grouping health-based observations. In addition, the final study finds strong evidence of the effects of Roemer's Law, thus providing support for the continued regulation of hospital bed availability. This dissertation also contributes significant new knowledge to the field of health services research. The specific salient outcomes include: detailing both the theoretical and applied differences between the raster and network data models for estimating travel time among locations, offering a methodology that simultaneously clusters observations based on comprehensive patterns of utilization and geographic location, and producing compelling, robust evidence that hospital availability has a positive, significant relationship with hospital utilization rates.

ACKNOWLEDGEMENTS

There are a number of people that have provided me with the support that was invaluable in completing this dissertation. Most importantly, I would like to thank Dr. Joseph Messina for his willingness to continue as my academic advisor after my time away from the program and for his seemingly unwavering belief that I would finish this work. A special amount of gratitude goes to my committee members for helping me to expand my understanding of health and medical geography and pushing me to fully realize the goals of my research. I commend them for the amount of time they spent fielding hastily written emails, reading and commenting on rough manuscript drafts, and listening intently while I fired numerous questions and ideas toward them. I would like to thank my parents for their encouragement and support. Finally, I want to thank Jeni Lee for her strength and understanding throughout the process of completing this research. I definitely could not have finished without her support and I may not have even attempted it without her encouragement.

TABLE OF CONTENTS

List of Tables **ix**

List of Figures **x**

Introduction **1**

Study #1

Measuring geographic access to health care: raster and network-based methods **12**

2.1 Abstract . . . 12
2.2 Background . . . 13
2.2.1 Access and geographic accessibility . . . 15
2.2.2 Data models . . . 17
2.3 Case study . . . 21
2.4 Data and methods . . . 22
2.4.1 Roads data . . . 22
2.4.1.1 Speed limit classification . . . 23
2.4.1.2 Road hierarchy . . . 24
2.4.1.3 Network comparison . . . 26
2.4.2 Population and hospital data . . . 28
2.4.3 Raster-based method . . . 30
2.4.4 Network-based method . . . 31
2.4.5 Sensitivity . . . 32
2.5 Results . . . 33
2.5.1 Underserved areas . . . 33
2.5.2 Limited Access Areas . . . 34
2.5.3 Sensitivity . . . 37
2.5.3.1 Speed limits . . . 37
2.5.3.2 Population representation . . . 37
2.6 Discussion . . . 38
2.7 Conclusions . . . 47

Study #2

Regional health care planning: a methodology to cluster facilities using community utilization patterns **50**

3.1 Abstract . . . 50
3.2 Background . . . 51
3.2.1 Clustering . . . 54
3.2.2 Case study . . . 55
3.2.2.1 Unreliable measure of hospital similarity . . . 57
3.2.2.2 Poorly defined home areal units . . . 59
3.2.2.3 Subjective modification by expert panel . . . 61

3.2.2.4 New methodology to cluster hospitals . . . 62
3.3 Methods . . . 63
3.3.1 Overview . . . 63
3.3.2 Input data . . . 63
3.3.3 Clustering algorithm . . . 65
3.3.4 Determining the number of Hospital Groups . . . 66
3.3.5 New hospital assignment . . . 69
3.4 Results . . . 70
3.5 Discussion . . . 75
3.6 Conclusions . . . 78

Study #3

Do more hospital beds lead to higher hospitalization rates? A spatial examination of Roemer's Law 80

4.1 Abstract . . . 80
4.2 Introduction . . . 81
4.3 Materials and Methods . . . 83
4.3.1 Research design . . . 83
4.3.2 Case study . . . 86
4.3.3 Population data . . . 88
4.3.4 Travel time . . . 88
4.3.5 Ethics statement . . . 89
4.3.6 Hospital utilization . . . 89
4.3.7 Spatial accessibility . . . 91
4.3.8 Clustering methodology . . . 94
4.3.9 Methods to remove multicollinearity . . . 97
4.3.9.1 Principal components analysis . . . 97
4.3.9.2 Bivariate regressions . . . 101
4.3.9.3 Test variance inflation factor . . . 101
4.3.10 Regression models . . . 103
4.4 Results . . . 104
4.5 Discussion . . . 107
4.6 Limitations . . . 110
4.7 Conclusions . . . 111

Conclusions 113

5.1 Overall contributions . . . 113
5.2 Future research . . . 116
5.2.1 Geographic accessibility . . . 116
5.2.2 Clustering health care observations . . . 118
5.2.3 Roemer's Law . . . 120
5.2.4 Spatial structure . . . 123
5.2.5 Health insurance . . . 124
5.2.6 Spatial accessibility . . . 124
5.2.7 Access and utilization in a regulated health care system . . . 125

Appendices **127**
A: R code to implement the Thomas Methodology 128
B: R code to implement the new clustering methodology 141
C: Testimony– Blue Cross Blue Shield of Michigan/Blue Care Network 154
D: Additional Figures and Tables 157
E: Additional R Code 174

References **236**

LIST OF TABLES

1 Travel speeds (miles per hour, mph) used in custom-built network datasets . 27

2 Turn delays (seconds) used in custom-built network datasets 27

3 Mean difference in travel time and road distance between Google Maps and custom-built networks in shortest path analysis 28

4 Comparison of underserved areas 34

5 Comparison of Limited Access Areas 35

6 Comparison of underserved areas and LAAs identified with speed limits assigned to roads 38

7 Comparison of results from block centroid population assignment method with original travel speed settings 38

8 Michigan roads by travel speed 41

9 Initial candidate solutions 73

10 Attribute variable set 102

11 Coefficient statistics 106

D.1 Cluster solutions and *incF* scores 162

D.2 Number of components and % of the total variance explained for each functional set of variables 165

LIST OF FIGURES

1 The triangle of human ecology 6

2 Research design 8

3 A) Network data model and B) Cost example 18

4 A) Raster data model and B) Cost example 20

5 Conversion of vector road data to raster cells 21

6 Hierarchical classification system for speed limits 25

7 Travel time estimates from custom-built networks compared with travel time estimated from Google Maps 29

8 Example of raster filter 31

9 Service areas (and resulting underserved areas) produced by network-based method 32

10 Underserved areas 35

11 Example of the similarities and differences between network and raster-based underserved areas 36

12 Limited Access Areas 37

13 Limited Access Areas with block population assignment method 39

14 Conversion of vector roads data to raster data format with slowest route rule 42

15 Service area delineation in areas where no roads are present 44

16 Michigan's current Subareas 53

17 Subareas produced by the Thomas Methodology using current data 58

18 *RI* and *CI* values for two hospitals of different sizes 60

19 Local minima and random starting locations with the K-means algorithm . . 67

20 Initial candidate solutions for Hospital Groups 71

21 Hospital Groups created using new clustering methodology 72

22 Population distribution and hospital locations in Michigan 87

23 Age adjusted hospital utilization (U_{STD}) and bed distribution (Av, E2SFCA) in Michigan 89

24 Distance decay of hospital utilization in Michigan 93

25 Zip Code clusters 96

26 Standardized coefficients for weighted SAR models 105

27 Example of rezoned region 111

D.1 *incF* scores for cluster solutions in set S 157

D.2 Moran's I of regression residuals for weighted OLS regression model 158

D.3 Moran's I of regression residuals for weighted SAR and CAR models 159

D.4 Levene Test of regression residuals for SAR and CAR models 160

D.5 Levene Test of regression residuals for weighted SAR and CAR models . . . 161

Introduction

The United States health care system is decentralized and fragmented, while also growing increasingly expensive over the last 40 years (Kaiser Family Foundation, 2009). Although some states mediate the availability of health care services through Certificate of Need (CON) programs, the US health care system has generally followed "market" forces in its evolution, resulting in an inequitable distribution of resources (Angell, 2008; Kuttner, 2008) and disparities in access to health care services. Concurrently, the costs of health care services have risen dramatically in recent years including an increased burden of out-of-pocket costs being placed on consumers (Cunningham, 2010; Wennberg, 2005). Despite spending more money on health care than any country in the world, the US lags far behind the leaders in numerous measures of public health outcomes (Murray and Frenk, 2010). Furthermore, the increased commercialization and profit-maximizing behavior of health care providers has resulted in distorted resource allocation of services and escalating costs (Kuttner, 2008).

Access to health care services can be defined as the ability to secure appropriate and effective health care services in a timely manner. It is well understood that access arises from a combination of both spatial and aspatial factors. In addition, utilization of health care services varies among populations and is dependent upon both access *and* factors unrelated to access such as overall health needs, socio-demographic characteristics, and perceptions of the health care system (Andersen and Newman, 1973). Understanding how access to services affects service utilization and health outcomes has been identified as being of great importance in health services research (Higgs, 2009).

In the US, both researchers and the popular media have largely placed an emphasis on exposing the financial barriers in access to health care due to the prohibitive costs of health care services and significant uninsured or underinsured populations. Past research has shown that a lack of health insurance is associated with less service utilization and worse health outcomes (Freeman et al., 2008). Other studies have explored the spatial aspects of access to health care services, indicating that a large number of people in the US have limited

geographic access to services such as emergency departments (Carr et al., 2009), specialty physicians (Rosenthal et al., 2005), and cancer care (Onega et al., 2008). Another large body of research, most notably by Wennberg and colleagues in the Dartmouth Atlas of Health Care, has explored small area variations in health care spending (Fisher et al., 2003), utilization (Wennberg, 2005), and outcomes (Welch et al., 2011), exposing disparities that exist throughout the US.

Although health care delivery has shifted increasingly towards profitability (Kuttner, 2008), health care planning and regulation in the US generally attempted to achieve two broad goals: 1) promote public health by ensuring that the supply of services meets the population's needs and 2) contain health care costs by regulating the supply of services to a level congruent with the need of the population. Regulation is often enforced through state-based Certificate of Need (CON) programs. The primary goals of CON programs are to contain health care costs by limiting the supply of health care services to only those *needed* by the population and to achieve equal access to health care (McGinley, 1995). Passage of the National Health Planning Act of 1974 required states to implement CON programs to receive federal funding for certain programs such as Medicare and Medicaid. However, this act was repealed in 1986 under concerns that it had failed to achieve its goal of reducing overall health care costs (Finn, 2007). Although their merits have been questioned over the past 40 years (see US Federal Trade Commission, 2004; Rivers et al., 2007; Ferrier et al., 2010) and they are no longer federally mandated, 35 states currently employ some form of CON program (National Conference of State Legislatures, 2011).

A number of states continue to regulate the supply of acute care hospitals, inpatient hospital beds, and hospital services through CON programs (Langley et al., 2010). Given that the plurality of overall health care expenditure in the US is for inpatient hospital care (Kaiser Family Foundation, 2009), hospitalizations, and thus hospitals, are logical candidates for cost control measures. The high costs of inpatient hospitalizations, in conjunction with the generally accepted implications of Roemer's Law (Shain and Roemer, 1959; Roemer,

1961), *a bed built is a bed filled*, serve as the current justification for continued regulation of hospital-based resources through CON programs.

Theoretically, access and utilization should have a direct relationship with each other considering that access measures the "potential" to utilize services (Aday and Andersen, 1974). The study of this relationship has a long history in health and medical geography and health services research. As noted by Hunter et al. (1986), Jarvis' study from the mid 19th century considered the effects of distance on admissions to mental health hospitals. Jarvis noted that the number of people from a given area admitted to a mental hospital declined with increasing distance from the hospital, postulating that this effect was not due to an abundance of people with mental health problems near the facilities. These ideas gave rise to Jarvis' Law, that health care utilization decreases with increasing distance from the location of the service. Additionally, the previously mentioned Roemer's Law was delivered in the late 1950s, defining the relationship between hospital bed availability and hospital utilization. Although only two are mentioned here, each demonstrate historical attempts by researchers to understand how access-related factors affect health services use.

More recent research has provided contrary or inconclusive findings in regards to the direct relationship between access and utilization (e.g., Goodman et al., 1997; Wright and Ricketts III, 2010). As the understanding of spatial structure and spatial processes in health services research has progressed, shortcomings of previous research are exposed. However, in spite of improved knowledge and methodological capabilities, the intertwined spatial and aspatial components of access and utilization make characterizations of this relationship extremely difficult. In addition, factors such as clinical practice variation among areas (Wennberg, 2005) and supply-induced demand complicate research efforts. Hence, few studies have linked access and utilization together in a comprehensive and coherent framework acknowledging the spatial and aspatial components of each. As a result, simply stated, current health services research lacks a clear understanding of how access affects the volume of utilization nor how it affects where people seek care.

The primary goal of this research is to provide a more complete understanding of how access to hospitals impacts hospital utilization. However, given the complexity of this issue, the research approach adopted in this dissertation addresses not only the nature of the relationship between access and utilization, but also methodological and theoretical issues related to *the study* of this relationship. I explore access and utilization of hospitals in Michigan, a health care system that has been under CON regulation for 40 years. Michigan serves as an excellent case study for this work due to 1) a physical landscape with two separate peninsulas that complicate traditional distance measurements, 2) a large variation in regional population density (both urban and rural areas) and hospital availability, allowing for access and utilization to be examined over a wide range of settings, and 3) an overall system of hospitals that have been, historically, relatively stable due to CON regulation.

Michigan implemented a CON program in 1972, thus it is one of the longest-tenured, currently active programs in the US (Finn, 2007). As part of its overall CON program, Michigan regulates the availability of hospital beds such that any hospital wanting to add licensed beds to their facility, relocate their existing facility (more than 2 miles from the existing facility), or construct a new facility must file a CON application and demonstrate a population need for the additional beds (Michigan Department of Community Health, 2009). The state implements a bed need methodology to predict future population demand for acute care hospital beds (Langley et al., 2010), thus providing hospitals the necessary information for CON applications. In addition, Michigan identifies regions in the state with limited geographic access to acute care hospitals (Messina et al., 2006), providing little resistance for hospitals or hospital systems expanding into these regions (Michigan Department of Community Health, 2009).

Many of the ideas that ultimately led to this research were formed, in part, while attending various meetings with members of Michigan's Department of Community Health; the Michigan CON Commission; and academics formerly involved with Michigan's CON program. However, the experience of working in a scientific advisory role for the most recent

hospital bed CON Committee proved to be the most influential. This committee included various stakeholders from Michigan's hospitals, hospital systems, and insurance companies who were assembled to review the state's hospital bed standards in the spring of 2011. Throughout six months of meetings, the need for information that would inform not only health services researchers, but also health care providers and policy makers became apparent. The most obvious need was for a better understanding of the spatial aspects of health care access and utilization, especially as they relate to health care policy. Paul-Shaheen and Carpenter (1982) noted: *there are no purely technical answers in health policy*; hence, this dissertation not only explores issues related to health services, but also acknowledges the intertwined nature of health research, policy, and regulation, and has aimed to provide original, robust, and *useable* findings.

The triangle of human ecology provides a useful conceptual framework for the study of overall population health (see Figure 1(A), Meade and Emch, 2010). The state of health, found in the center of the triangle, is influenced by population, habitat, and behavior and interactions among these characteristics. Each of these broad characteristics comprise three sub-characteristics. In Figure 1(B), the portion of the triangle explored in this research is illustrated. In this framework, hospital utilization is considered a behavior, not a health outcome or description of the state of health. Because hospitalization is used in an attempt to restore health in cases of illness or injury, a state of comprimised health can be assumed. However, because the health outcomes associated with hospitalization are not assessed within this work, considering hospital utilization as a state of health is not justified. To explore the spatial aspects of meta-relationship between access and hospital utilization, characteristics of the population and their built habitat are considered, most notably the interaction among population location, the transportation infrastructure, and hospital location.

The overall research project includes three sequential studies in which I explore spatial accessibility characterization, health care utilization patterns, and the relationship between the access and utilization. The specific outcomes of these studies are 1) the development

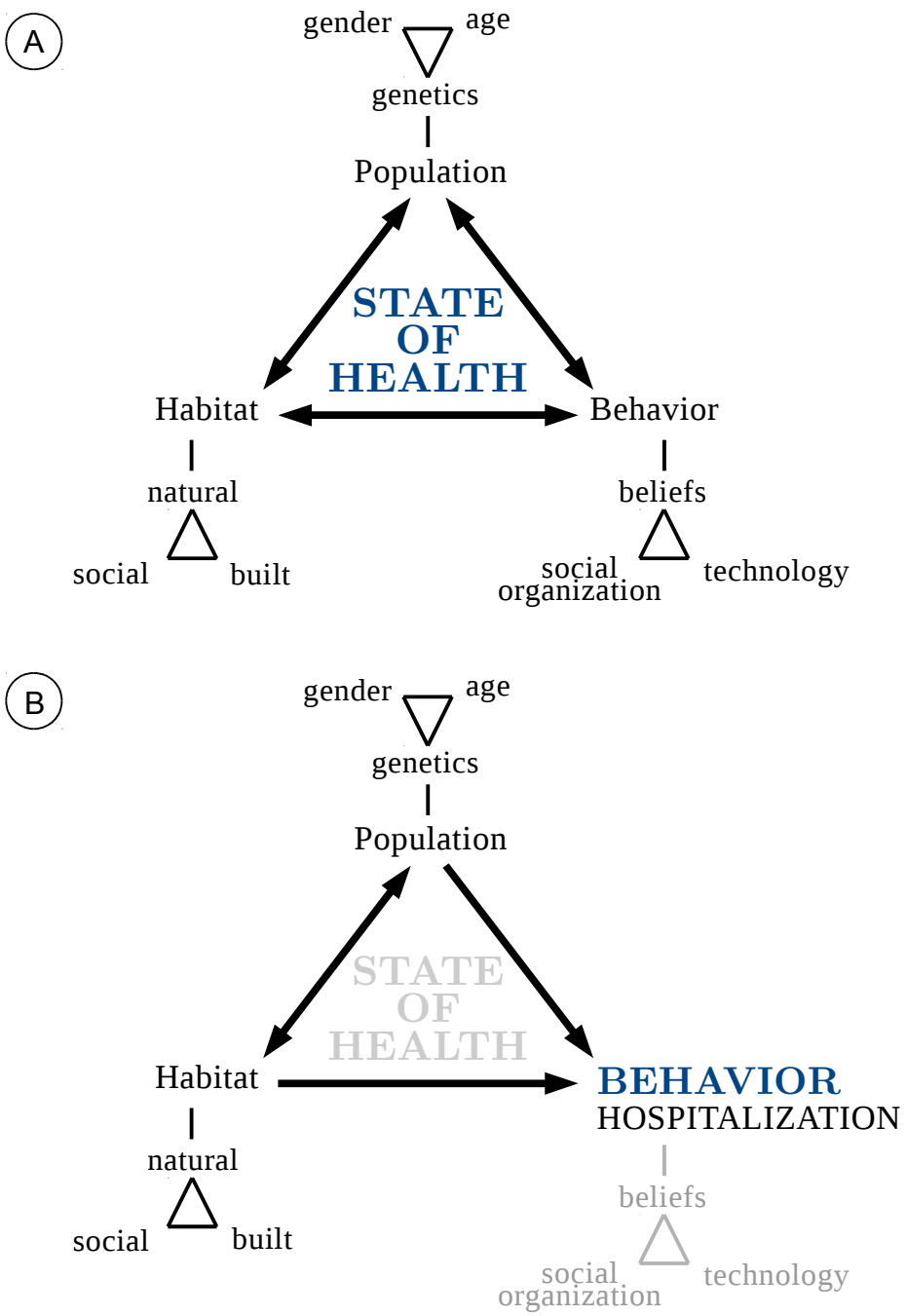

Figure 1: **The triangle of human ecology.** A) the original (Meade and Emch, 2010, redrafted by Paul L. Delamater) and B) the portion of the overall triangle that is explored in this research. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

of a system for measuring distance and travel time among locations, 2) a methodology for comparing and clustering observations based on utilization patterns and location, and 3) an assessment of the relationship between access and utilization of hospital services. Because the measurement of both access and utilization patterns are heavily dependent upon the ability to accurately characterize and measure distance and spatial relationships, studies 1 and 2 focus on methodological problems regarding the characterization of location, distance, and spatial patterns in health services research. The final study uses this knowledge to explore the relationship between access to hospitals and hospital utilization in Michigan. Figure 2 shows the interconnected nature of these studies as they pertain to the overall research goal.

Because each study was submitted independently for publication, they all contain the sections (e.g., Abstract, Introduction, Conclusions, etc.) required for submission. In this dissertation, each study is presented as an individual chapter with its original formatting intact. These chapters are written in the first-person plural point of view. While I conceived and performed the research and drafted the resulting manuscripts, my committee provided helpful guidance throughout the research process and manuscript preparation, thus they were included as co-authors of the submitted versions.

In the following paragraphs, I provide a short summary of the background and aims of each individual study. Then, each study is presented as a stand-alone chapter. In the Conclusions, the work is synthesized within the framework of assessing access and utilization within a health care system regulated by a CON program. The outcomes provide insights that inform future health services research, while also offering guidance for future public policy initiatives.

Study 1: Measuring geographic access to health care: raster and network-based methods. Traditional measures of geographic accessibility to services have been replaced with more elaborate gravity-based metrics that incorporate the distance, supply, and potential demand (Ngui and Apparicio, 2011), thus integrating accessibility and availability into one comprehensive measure (e.g., the 2 Step Floating Catchment Area (2SFCA),

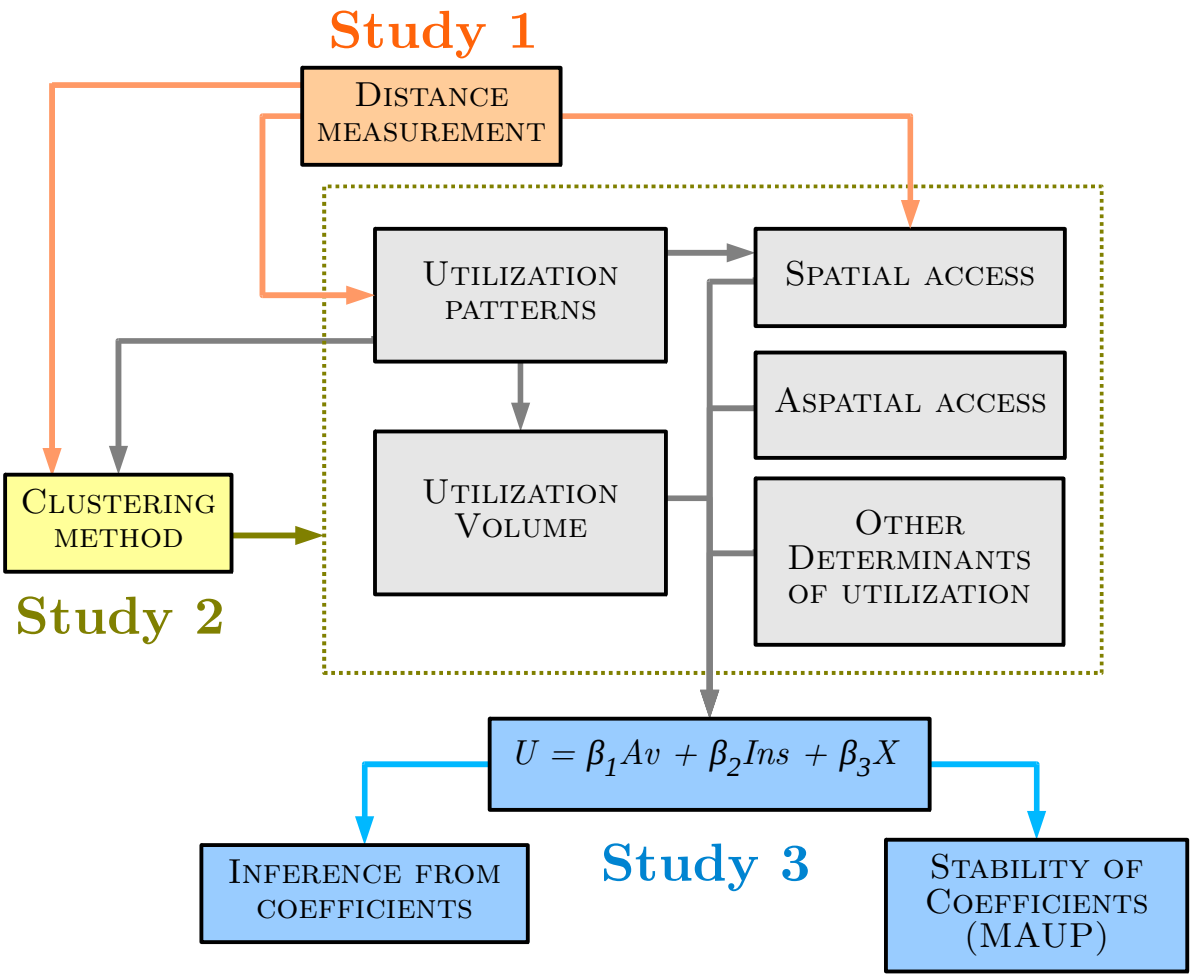

Figure 2: **Research design**

Luo and Wang, 2003). Because these metrics are calculated with limited or no actual utilization data, they rely heavily on accurate measures of distance among locations. Past studies have regularly measured distance using a straight line (Euclidean), not accounting for true connectivity or travel impedances between locations (Jones et al., 2010; Martin et al., 2002). More recently, advances in data availability and processing capability have provided researchers the ability to incorporate both road connectivity (road distance) and travel speed (travel time) in their characterizations of distance. Although these measures are generally accepted to be more accurate representations of the *friction of distance* or *travel burden* between locations (see Pedigo and Odoi, 2010; Shahid et al., 2009), there remains uncertainty regarding their implementation. Most importantly, because speed limits are often not included in roads data, travel speed must be estimated based on available road attribute data. Concurrently, both raster and vector-based (network) methods are available to for researchers to calculate travel time and distance along a road network.

This study compares raster and network-based methods of calculating travel time between locations. The specific case study focuses on the identification of regions with limited geographic accessibility to Michigan's hospitals. I develop a speed limit classification system based on road attribute data and explore its robustness by comparing network travel time estimates with those from Google Maps. Thirty minute travel time buffers (service areas) are constructed around each Michigan hospital using both raster and network-based methods. Underserved areas are identified as those falling outside of the travel time buffers. To understand each method's sensitivity to speed limit settings, the speed limit classifications are modified and changes in the resulting underserved areas are compared.

Study 2: Regional health care planning: a methodology to cluster facilities using community utilization patterns. In health services research, the methods used to create small (geographic) areas has been given little attention (Shwartz et al., 2001). Small areas are constructed by combining disaggregated population units based on some level of shared similarity among units. Wennberg and Gittelsohn (1973) offered a method

to create hospital service areas by aggregating population units to a single facility (or two near-by facilities) based on a simple plurality rule of utilization, noting that over 85% of care was delivered by hospitals within the service area boundaries. However, this method often requires *manual* adjustment to ensure contiguity (e.g., Klauss et al., 2005) and is problematic in urban areas where service use is distributed similarly to many near-by facilities (Thomas et al., 1981). Another set of methods, based on distance from each facility (e.g., Garnick et al., 1987; Schuurman et al., 2006), rely heavily upon the assumption that bypass of the nearest facility will be minimal. Although these approaches have scientific merit, they assume that each population unit should be *tied* to a specific hospital (or set of hospitals). Additionally, by ignoring where *the rest* of the population seeks care, they do not account for the overall patient utilization patterns and are, thus, incomplete in their comparison.

I provide a methodology to cluster observations based on their overall utilization patterns and geographic location. Specifically, this methodology is used to group Michigan's acute care hospitals into "Hospital Groups." The limitations in Michigan's current method to group hospitals (Thomas et al., 1981) are identified, emphasizing the importance for a methodology that emphasizes overall patterns of utilization, not hospital-based competition. Along with the clustering methodology, I supply a heuristic that assists in determining the appropriate number of clusters in the data, a common difficulty in clustering applications (Jain, 2010). Although the specific case study proposed focuses on grouping hospitals, the theoretical underpinnings are such that the methodology can be used to group any set of spatial observations with multiple attributes. Importantly, it can be used to create health service areas while addressing limitations found in the plurality and distance-based methods.

Study 3: Do more hospital beds lead to higher hospitalization rates? A spatial examination of Roemer's Law. Roemer's Law defines a positive relationship between the availability of hospital beds and the use of hospital services. Past research has provided support for Roemer's Law (e.g., Ginsburg and Koretz, 1983; Harris, 1975; Kroneman and Siegers, 2004; Pasley et al., 1995; Shwartz et al., 2011; Wennberg, 2005); however,

other studies have found conflicting results (e.g., Alexander et al., 1999; Rohrer, 1990; van Doorslaer and van Vliet, 1989) or inconclusive results (e.g., Clark, 1990). The intertwined relationships among population health, access, use of health care services, and outcomes provide a number of research dilemmas, both theoretically and methodologically. Perhaps, the most difficult is defining and characterizing the availability of hospital beds. Although counting the number of beds in a hospital is trivial, measuring the overall availability of those beds *to a population* is a much more complex task related to distance, demand, and access-related factors. Unsophisticated measures of hospital bed availability, such as container-based methods or simple distance measures (Joseph and Phillips, 1984; Guagliardo, 2004), ignore the multifaceted nature of access and the spatial and geographic nature of health care service use. Others have noted that the observed effects of Roemer's Law may be due to oversimplified methods used to assign hospital beds to regions (Folland and Stano, 1990). In addition, statistical methods that do not incorporate spatial structure in the relationship between availability and utilization are at risk of being biased due to the effects of spatial autocorrelation.

This study explores the relationship between access and utilization of hospital services using an ecological research design that integrates individual behavioral models of health care utilization in an explicitly spatial context. I characterize both the spatial and aspatial components of access while also controlling for other determinants of hospitalization throughout my theoretical and applied models. As a result, the effects of Roemer's Law can be identified and quantified. The ecological study design implemented in this study necessitates that the relationship between access and utilization be explored at varying scales of analysis to examine the effects of the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984; Fotheringham and Wong, 1991). Therefore, I explore the stability of the relationship by performing the analysis at numerous levels of data aggregation.

Measuring geographic access to health care: raster and network-based methods

2.1 Abstract

Background: Inequalities in geographic access to health care result from the configuration of facilities, population distribution, and the transportation infrastructure. In recent accessibility studies, the traditional distance measure (Euclidean) has been replaced with more plausible measures such as travel distance or time. Both network and raster-based methods are often utilized for estimating travel time in a Geographic Information System. Therefore, exploring the differences in the underlying data models and associated methods and their impact on geographic accessibility estimates is warranted. **Methods:** We examine the assumptions present in population-based travel time models. Conceptual and practical differences between raster and network data models are reviewed, along with methodological implications for service area estimates. Our case study investigates Limited Access Areas defined by Michigan's Certificate of Need (CON) Program. Geographic accessibility is calculated by identifying the number of people residing more than 30 minutes from an acute care hospital. Both network and raster-based methods are implemented and their results are compared. We also examine sensitivity to changes in travel speed settings and population assignment. **Results:** In both methods, the areas identified as having limited accessibility were similar in their location, configuration, and shape. However, the number of people identified as having limited accessibility varied substantially between methods. Over all permutations, the raster-based method identified more area and people with limited accessibility. The raster-based method was more sensitive to travel speed settings, while the network-based method was more sensitive to the specific population assignment method employed in Michigan. **Conclusions:** Differences between the underlying data models help

Submission information: Submitted to the International Journal of Health Geographics on January 13, 2012. Accepted on April 10, 2012. Published on May 15, 2012: Volume 11, Issue 15. Authors on manuscript: Paul L. Delamater, Joseph P. Messina, Ashton M. Shortridge, Sue C Grady.

to explain the variation in results between raster and network-based methods. Considering that the choice of data model/method may substantially alter the outcomes of a geographic accessibility analysis, we advise researchers to use caution in model selection. For policy, we recommend that Michigan adopt the network-based method or reevaluate the travel speed assignment rule in the raster-based method. Additionally, we recommend that the state revisit the population assignment method. **Keywords:** Health care access, geographic accessibility, limited access areas, underserved populations, health services.

2.2 Background

Disparities in the geographic accessibility of health care services arise due to the manner in which people and facilities are arranged spatially. Specifically, health care services are provided at a finite number of fixed locations, yet they serve populations that are continuously and unevenly distributed throughout a region (Joseph and Phillips, 1984). Although inequalities in accessibility are inevitable due to this configuration, the extent to which they manifest is a product of the unique spatial arrangement of the health care delivery system, the location and distribution of the population within a region, and the characteristics of the transportation infrastructure. Of particular concern are scenarios that result in large distances between people and health care facilities. These populations experience greater difficulty in gaining access due to increased travel times, often coupled with poor transportation infrastructure and a lack of public transportation options (Arcury et al., 2005).

The spatial or geographic dimensions of access have received considerable attention from planners and researchers for many years (Cromley and McLafferty, 2002). Referred to as spatial accessibility (Guagliardo, 2004), the spatial dimensions of access include accessibility and availability of services. Accessibility (or geographic accessibility) is a measure of the "friction of distance" or "burden of travel" between locations, whereas availability generally measures the number of services in comparison to the number of potential users of the service. Identifying areas with limited spatial accessibility of health care services allows planners to

understand the effects of opening, closing, or relocating health care facilities or modifying the services offered by existing facilities (McGuirk and Porell, 1984). Thus, accurate and detailed representations of spatial accessibility are imperative to describe and understand the overall access picture.

Changing technology and the availability of detailed spatial data have allowed for the representation of geographic accessibility in a GIS to more closely resemble the real-world phenomena of travel. Early studies acknowledged that the travel costs among locations were more complex than those provided by straight-line (Euclidean) distance measures (see Shannon et al., 1973), yet this particular representation of geographic accessibility has been the most widely used in past health services research (McLafferty, 2003). Although Euclidean distance has shown to be correlated with travel time (Apparicio et al., 2008; Haynes et al., 2006; Phibbs and Luft, 1995), it does not incorporate topological structures or the transportation infrastructure (Jones et al., 2010), both of which are likely to influence travel travel time. As computational power and data collection/storage capabilities have improved, more detailed representations of geographic accessibility have emerged, incorporating the transportation infrastructure (e.g., roads $\rightarrow$ travel distance), travel impedance (e.g., speed limits $\rightarrow$ travel time), and various modes of travel (public transportation $\rightarrow$ travel time).

The flexibility provided by GIS allows for multiple data representations of the same real-world phenomena. Specifically, travel costs can be represented using a field-based model (raster) or an object-based model (vector). The vector data model can also be extended to incorporate network or graph features and is referred to as a "network" data model. Whereas a raster vs. vector debate in regards to spatial data representation and analysis in GIS has been present for many years in the GIS and Geography literature (see Couclelis, 1992; van Bemmelen et al., 1993; Goodchild et al., 2007), the issues have not been fully explored in health services research. Considering the importance placed on the role of distance and travel in health care accessibility studies, we believe that an examination of the data models and methods is warranted. Thus, the purpose of this paper is to compare geographic accessibility

measured as travel time using both raster and network (vector) based models of spatial data representation. We aim to illuminate both the conceptual and practical differences between models and their methodological implications in measuring geographic accessibility. Specifically, we address the following questions over the course of this manuscript:

- What are the basic assumptions when constructing a conceptual model of travel?
- What are the specific abstractions in the raster and network representational models of travel in a GIS?
- What are the similarities and differences in results between data models?
- How do the underlying differences in data models affect the results?

The manuscript is organized as follows. First, we offer a short review of access and geographic accessibility. Next, the spatial data models and methods used to calculate travel costs are summarized. In the following section, we describe our case study and report on the specific data and methods used in analysis. Next, we report our results and discuss the similarities and differences between methods. Lastly, we discuss the implications of our findings for measuring geographic accessibility.

2.2.1 Access and geographic accessibility

Access to health care is a multifaceted and complex concept, dependent upon the characteristics of both the population in need of services and the health care delivery system (Aday and Andersen, 1974). Penchansky and Thomas (1981) identified five distinct dimensions of access which were classified by Khan (1992) into spatial components (accessibility and availability) and aspatial components (affordability, accommodation, and acceptability). Access to health care can also be classified into potential and realized delivery of services (Aday and Andersen, 1974; Joseph and Phillips, 1984) based on whether actual utilization data of the services is incorporated (realized) or based solely on the characteristics of the services offered (potential).

In recent health service research, distance is commonly measured as vehicular travel time over a road network calculated in GIS (Higgs, 2004). However, other measures such as travel distance or Euclidean distance are also regularly used (Higgs, 2009; McLafferty, 2003). By incorporating real-world connectivity provided by the road infrastructure, travel distance offers a more accurate characterization of the distance among locations compared to Euclidean distance. Yet, travel distance does not recognize the variations in travel impedance (speed limits or travel speeds) often found between rural and urban environments. Although Euclidean and travel distance are computationally less expensive and require fewer inputs, respectively, recent improvements in spatial data processing capabilities and drive distance analysis allow for vehicular travel time to be modeled more easily in a GIS (Jones et al., 2010). We acknowledge that travel time estimates offer the most accurate representation of the cost of travel for measuring geographic accessibility based on a number of recent studies in health services research discussing the subject (see Apparicio et al., 2008; Martin et al., 2002; Pedigo and Odoi, 2010; Shahid et al., 2009).

A number of assumptions regarding real world phenomena are required prior to spatial representation and modeling. In the case of forming a conceptual for model travel time, the initial assumption is that the unique and personal experience of travel among locations can be sufficiently characterized and estimated using spatial data and models. Rather than attempting to isolate and discuss all the factors influencing travel time, we instead point out the general assumptions present in many geographic accessibility models constructed for population-based studies. First, the models assume that each person in the population has similar driving characteristics and comparable vehicles. Another assumption is that each person experiences the same travel conditions, therefore variation in factors influencing travel time such as the day, time of day, local traffic patterns, and weather are held constant. The models also assume that all people possess knowledge of and choose to travel along the shortest path between locations. Increased availability of desktop and internet-based trip planners has likely diminished the overall impact of this assumption, yet it remains salient in

travel time models. Finally, due to limitations in data availability and data processing capabilities, the location of a population is often assigned to a single point location. Therefore, the travel time estimates originating from this location are assumed to be a reliable proxy for the travel time experienced by each member of the population. Although these assumptions hide significant variability, they are necessary when conducting population-based studies due to the unpredictability of potential factors influencing travel (Witlox, 2007) and the lack of individually georeferenced data. Hence, GIS-based travel time estimates should aim only to capture the average situation encountered, a suitable metric for most accessibility studies (Haynes et al., 2006).

2.2.2 Data models

The differences between raster and network data models have been extensively documented in many GIS textbooks and research papers (e.g, Longley et al., 2010). Although the conceptual models of space, input data formats, and computational algorithms employed in processing these data differ, the basic premise behind the calculation of travel time is quite similar for both. Travel time is modeled as a function of distance and travel speed and can be conceptualized as the *cost of movement*. A number of data products based on cost of movement can be calculated using a GIS. However, due to their importance in assessing geographic access, we focus our discussion on a minimum cost path between locations and a catchment or service area corresponding to a point location. In the following paragraphs, the data formats and corresponding cost of movement concepts are summarized for both the network and raster models.

The basic network data model comprises a series of nodes (points) that are connected by edges (lines). Because the nodes and edges are the sole geometric features defined in the data model, any place not falling on the network is essentially "undefined" or empty space. Therefore, location and movement within the network data model are confined solely to the edges and nodes (see Figure 3(A)).

In the representational model of travel time, the cost to traverse an edge is defined by

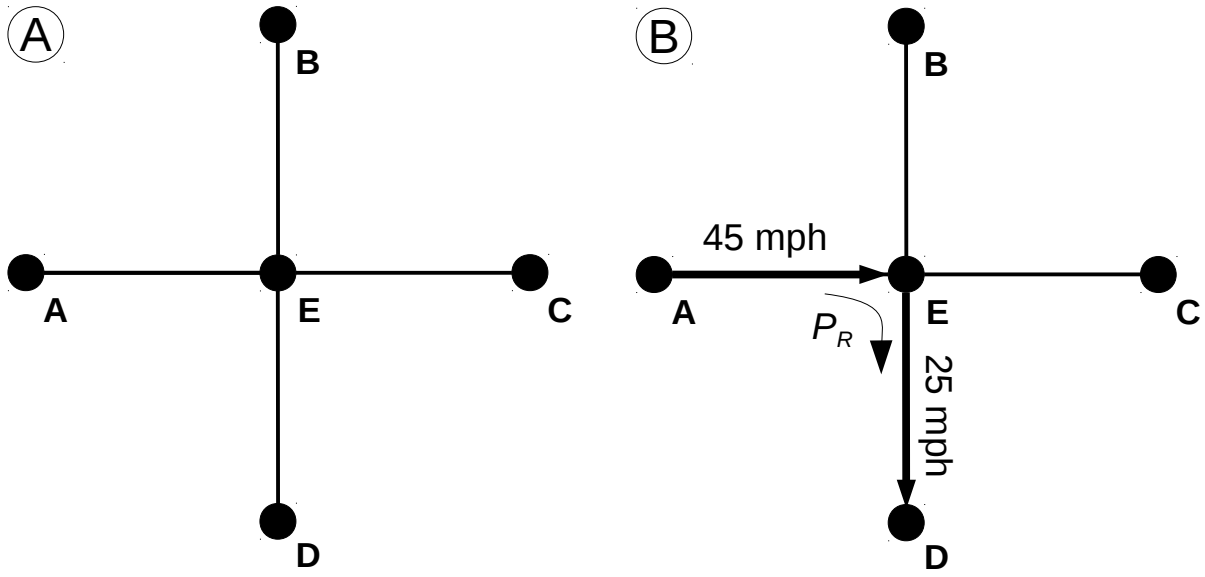

Figure 3: **A) Network data model and B) Cost example**

the edge length and its associated travel speed. Additionally, the network data model can be augmented to include a penalty for a directional change at a node (i.e., a time penalty or turn delay when making a turn at an intersection). In this case, movement through a node is assigned an angular direction, relative to the original direction of travel, and the corresponding delay for that directional change is applied. An example of travel within a network model is detailed in Figure 3(B), showing travel from Node A to Node D in a simple network. The travel time (T_{AD}) for the trip can be calculated such that

$$T_{AD} = \frac{d_{AE}}{S_{AE}} + \frac{d_{ED}}{S_{ED}} + P_R \tag{1}$$

using edge distance A-E (d_{AE}), edge distance E-D (d_{ED}), travel speed of edge A-E (S_{AE}), travel speed of edge E-D (S_{ED}), and the turn delay for making a 90° right hand turn at Node E (P_R).

Many recent studies of health service accessibility have utilized the network data model for calculating travel time estimates (Dai, 2010; Pedigo and Odoi, 2010; Schuurman et al.,

2010; Wan et al., 2011). The network data model is appealing for representing vehicular travel time or distance considering that road segments (edges) are connected at road intersections (nodes), upholding real-world connectivity among locations. Results of path calculations are likely to be very similar to those experienced in the real world due to the similarities between the data model structure and the true travel environment (Kwan and Hong, 1998). Because areal features are not defined in the network data model, service area calculation requires that edges (lines) must be converted to a polygon representation. The polygon represents the areal extent of the edges within the service area, but requires an approximation of undefined space in the original data model.

The raster data model is composed of a series of regularly sized and spaced cells (or pixels). Cells are arranged in a lattice with explicit spatial boundaries, thus all locations within the boundaries of the lattice are represented by their 2 dimensional coordinate location. In this data model, travel occurs through cell to cell movement wherein a specific cost is designated for each cell, representing the time required to traverse the cell.

In most GIS software packages, movement occurs in only cardinal directions (*Rook's case*) or in both cardinal and diagonal directions (*Queen's case*, see Figure 4(A)). However, other software packages offer more flexible options such as Knight's case movement (Lopez-Quilez and Munoz, 2009). Travel time is calculated using the cell dimensions and travel speed assigned to the cell. Unlike the network model, the length of individual steps in a route is based on the cell resolution of the data and thus, constant throughout the entire raster grid. Figure 4(B) contains a graphic representation of possible travel routes between cell A and cell D in the raster model. In this case, the journey can be accomplished by taking a similar route as shown in Figure 3(B) whereas the route goes from cell A to cell E to cell D. Travel time (T_{AD}) for this route would be calculated such that

$$T_{AD} = \left(\frac{\frac{d}{2}}{S_A} + \frac{\frac{d}{2}}{S_E}\right) + \left(\frac{\frac{d}{2}}{S_E} + \frac{\frac{d}{2}}{S_D}\right) \tag{2}$$

where d is the distance between cell centers, which is equal to cell resolution, and travel speed

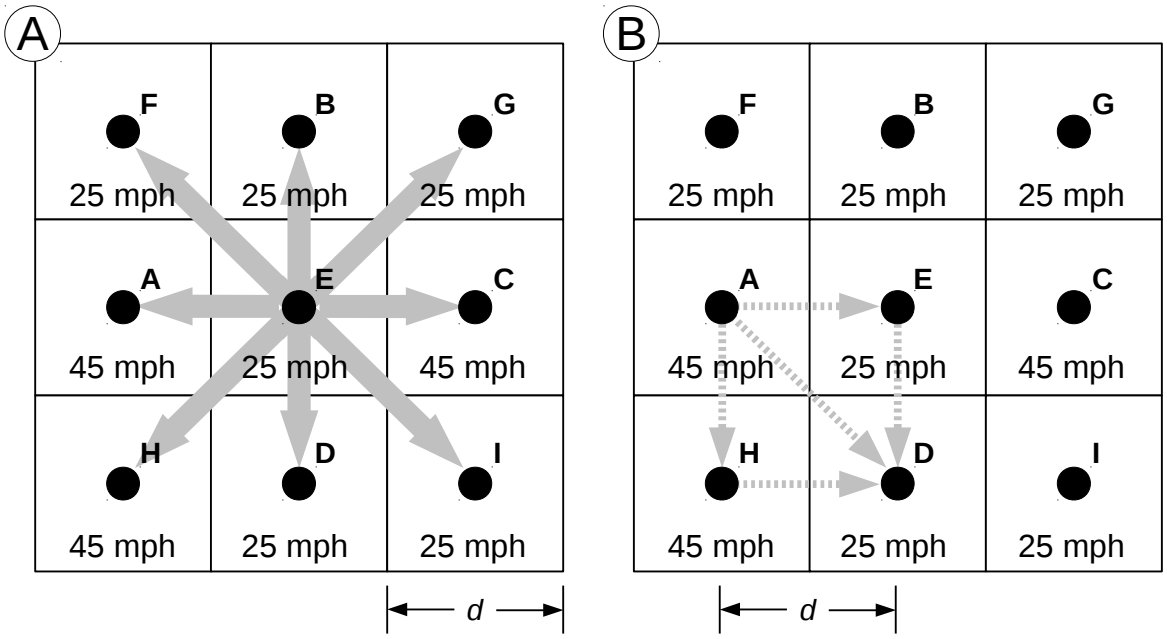

Figure 4: **A) Raster data model and B) Cost example**

(S_i) is defined for each cell. Division by 2 occurs for each step in the movement because half of each cell is traversed with each step. In this case, to travel from Point A to Point E, half of d is traversed at 45 mph and half is at 25 mph. The journey can also be completed by taking the diagonal, direct route between the two points such that

$$T_{AD} = \frac{\frac{\sqrt{2}}{2} * d}{S_A} + \frac{\frac{\sqrt{2}}{2} * d}{S_D} \tag{3}$$

where the increase in distance traveled for the step is accounted for by using the Pythagorean theorem to adjust the distance term.

The raster data model has been used to calculate travel time in health service accessibility studies (see Martin et al., 2002; Messina et al., 2006; Ray and Ebener, 2008; Tanser et al., 2006). Because all locations are explicitly defined in the raster data model, it is attractive for creating service areas, especially in regions without an all-encompassing transportation network (Tanser et al., 2006).

Roads data are generally available as vector features and must be converted to a raster

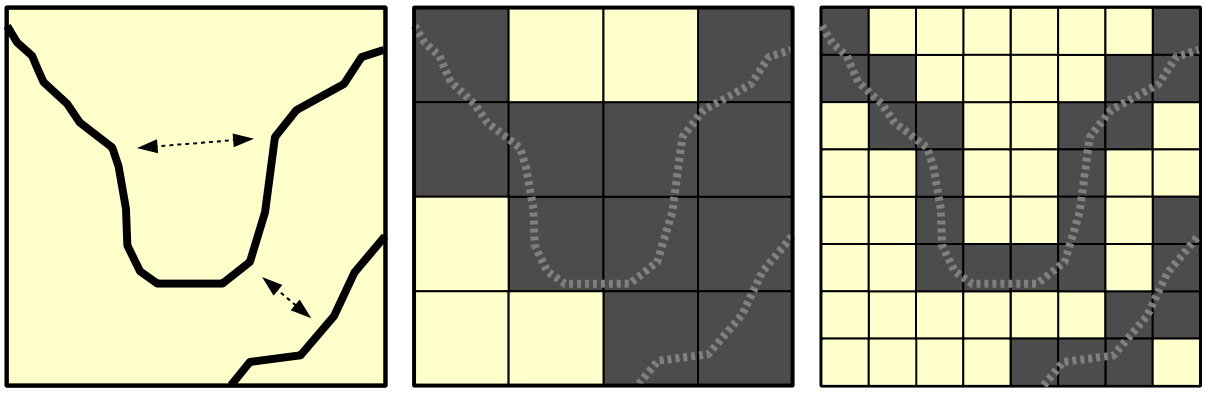

Figure 5: **Conversion of vector road data to raster cells.** The original roads (black lines on left) are converted to a cell-based representation with large cell sizes (middle), resulting in an overconnected travel grid. Smaller cells (right) improve the topological structure of the travel grid. However, the two roads are still erroneously connected in this scenario.

representation. This process requires specification of a cell resolution. The abstraction process necessitates decision rules for assigning a travel speed to cells in which multiple roads (with varying speed limits) fall inside the cell bounds and/or cells in which no roads are present. When the vector roads data are converted to cells, the roads cease to exist as unique and individual entities (e.g., highways, surface streets, ramps, etc.) and become a surface of travel speeds (see Figure 5). In the raster data model, the strict topology that governs real world travel along roads is replaced by predefined directional movement among cells. Thus, in routing applications, the raster data model has the potential to produce unexpected results (Sander et al., 2010; Upchurch et al., 2004). Furthermore, travel time estimates may be either overestimated or underestimated depending upon the geometric complexity of the road network and the cell resolution.

2.3 Case study

Our case study explores the geographic accessibility of hospitals in Michigan. The Michigan Department of Community Health (MDCH) identifies Limited Access Areas (LAA) as a part of the state's Certificate of Need (CON) program, thus offering a formal definition of

areas with limited geographic accessibility with which to compare methods. The state also serves as an excellent study area to conduct a travel time analysis due to a unique physical geography (two separate peninsulas with irregular shorelines) and highly variable mix of urban and rural regions (Martin et al., 2002).

As defined by statute (Michigan Department of Community Health, 2009), an LAA is any geographic area containing a population of 50,000 that is more than a 30 minute drive time (utilizing the slowest route available) to the nearest acute care hospital offering 24 hours/day 7 days/week emergency room services. LAA maps are used by the MDCH and Michigan's CON Commission to evaluate applications to construct new hospitals or branch locations and requests to add or modify existing hospital services.

In Messina et al. (2006), the authors presented a raster-based GIS methodology used to measure travel time to hospitals and identify underserved areas and LAAs in Michigan. This methodology is re-implemented using updated population and health service facility data from 2010. Underserved areas and LAAs are also identified using a network-based travel time analysis. Both methods are tested for sensitivity to travel speed settings and changes in the population assignment method. The results of the raster and network-based methods are compared and implications for measuring geographic accessibility are explored.

2.4 Data and methods

2.4.1 Roads data

Both the network and raster-based methods of calculating travel time among locations are heavily dependent upon a detailed and accurate representation of both road location (length) and travel speed (impedance). The 2009 road network database (Michigan Geographic Framework Version 10a) was acquired from the Michigan Center for Geographic Information (MCGI, http://www.michigan.gov/cgi). The location of each road segment is provided along with attributes including, but not limited to: length, road name, data source, National Functional Classification (NFC) code, Framework Classification Code (FCC), and

legal ownership.

2.4.1.1 Speed limit classification The estimation of travel speed for each road segment, in the absence of measured travel speed data, can be accomplished most accurately using the posted speed limit and surface material of the road segment. Speed limits define the maximum legal travel speed, whereas surface material helps to determine realistic travel speeds (n.b., reasonably lowered speeds on unpaved roads in rural areas). Because neither speed limit nor road surface type are included as attributes in the MCGI roads database, we developed a hierarchical classification system to assign estimated travel speed to each road segment. Traditional methods of assigning travel speeds or speed limits are generally simple classifications using *only* the FCC or the NFC of each road segment (see Birkmeyer et al., 2003; Nallamothu et al., 2006; Berke and Shi, 2009). Our classification system for assigning travel speed offers a significant advantage over traditional methods by incorporating NFC, FCC, and road ownership into in a hierarchical decision tree, rather than relying on a single road attribute class.

The actual speed limits of Michigan roads are based upon road classification, landuse of surrounding areas, or average travel speed. Statutory speed limits are those set throughout the state for a certain set of roads (i.e., 70 mph for expressways, 55 mph for state and county roadways, and 25 mph for roads in business or residential areas), whereas modified speed limits are assigned when roads require a speed limit below 55 mph, but above 25 mph. National guidelines state that modified speed limits be based upon the 85^{th} percentile speed of all travelers during free flowing traffic and ideal weather conditions. The length of a speed zone should be at least one half of a mile and the number of speed limit changes along a given route should be kept minimal (Michigan Office of Highway Safety Planning).

In preliminary investigations, we found that the NFC system provided valuable information for speed limit assignment, but should be superseded or supplemented with FCC or road ownership. For instance, in small rural communities, road ownership better char-

acterized observed speed limits than the NFC system, where the cutoff value for an urban population is 5,000 people. Using only the NFC attribute, the speed limits for streets in many small communities (rural villages and towns with populations less than 5,000) would be mis-assigned as they are not distinguished from other rural roads. Each of the many scenarios encountered will not be discussed in detail; however, a graphic depiction of the complete hierarchical classification system is found in Figure 6. Development and preliminary evaluation of the classification system included personally traveling road networks in southeast and mid-Michigan, documenting the actual speed limits.

2.4.1.2 Road hierarchy Each road was assigned a "hierarchy" value in an effort to control traffic flow within the network data model. The MCGI roads data did not contain attribute information describing real-world connectivity at road intersections (e.g., overpasses and underpasses). All intersections are presumed traversable if no connectivity rules are established, leading to an over-connected network and likely underestimation of travel times if not accounted for. True connectivity could not be established for all roads in the state due to the large number of intersections in the roads dataset (n > 500,000) along with a lack of reference data. Therefore, our efforts were directed towards establishing realistic connectivity between expressways and surface streets.

We utilized the hierarchy attribute in conjunction with a turn delay to account for the absence of connectivity information at expressway intersections in the MCGI data. In ArcGIS^{TM}, turn delays in a network dataset can be assigned not only by the direction of the turn, but also by the hierarchy values of the intersecting roads. Using the FCC attribute in the roads data, all expressways were assigned a hierarchy value of 1, all ramps (leading onto and off of expressways) were assigned a value of 2, and all remaining roads (surface streets) were assigned a value of 3. Considering that real-world traffic flow between expressways and surface streets is restricted to only entrance and exit ramps connecting the two road types, we assigned an artificially high turn delay (20 minutes) to any direct turn between

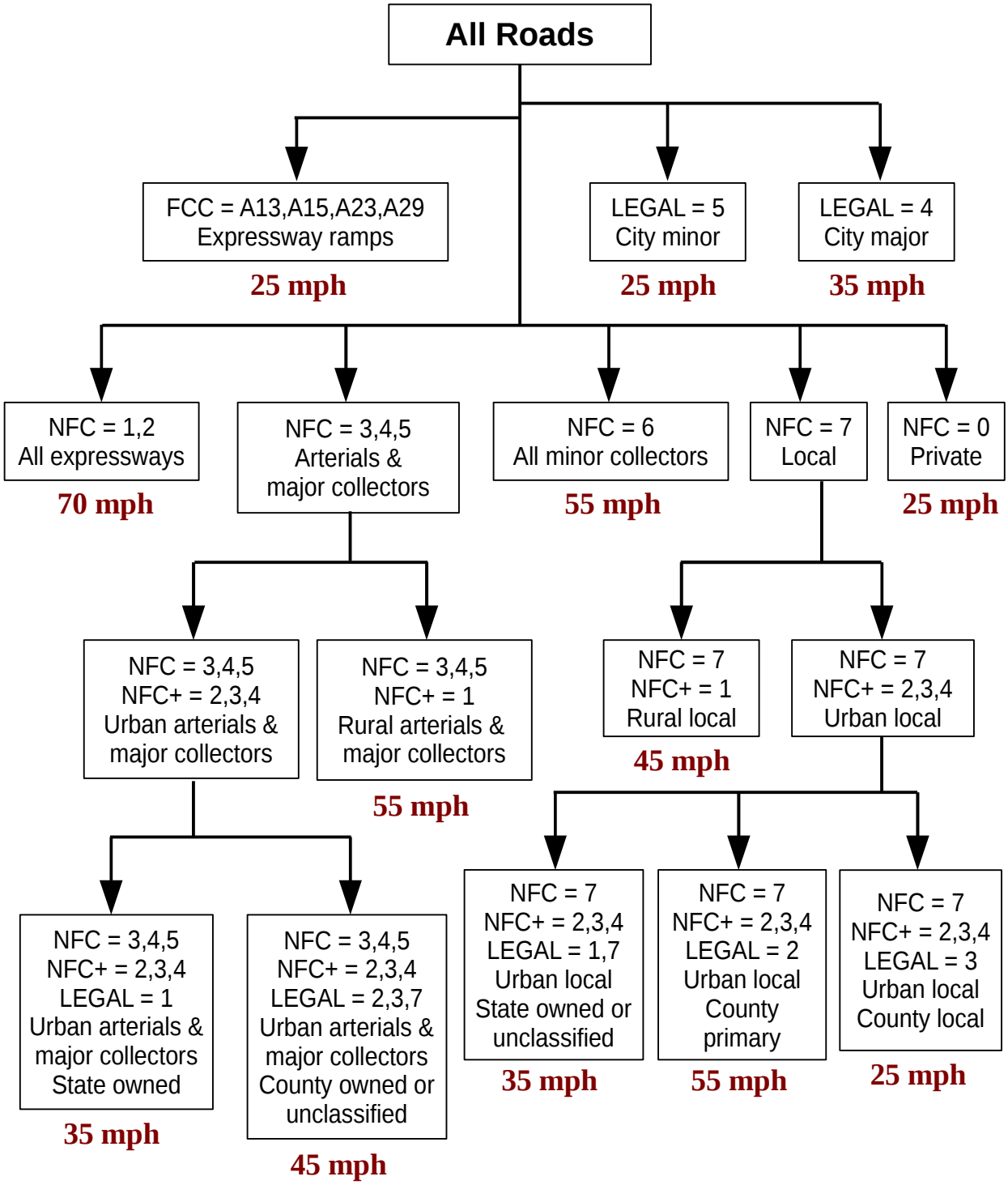

Figure 6: **Hierarchical classification system for speed limits**

expressways and surface roads (hierarchy values 1 and 3). This prevented the network solver from choosing to make a "non-existent" turn between surface streets and expressways due to the unrealistically high turn delay between road hierarchy values. Essentially, expressway connectivity within the network was restricted to match actual driving conditions, thus improving the accuracy of travel time estimates.

2.4.1.3 Network comparison Five network datasets were created and explored to better understand how changes to the speed limit classification system (see Table 1) and the penalties assigned for turn delays (see Table 2) affected the estimated travel times. Although the Michigan Office of Highway Safety Planning offers guidelines for assigning road speed limits (Michigan Office of Highway Safety Planning), we were unable to locate reference data for comparative purposes. Furthermore, collecting enough actual travel time data to allow for formal statistical testing was not feasible. Given these limitations, we compared travel time estimates to results obtained from Google Maps^{TM}. The results from Google Maps were not considered true travel times due to the lack of methodological documentation available and a substantial number of speed limit errors that were manually identified in their roads data. However, because the Google Maps travel time estimates are derived from independent source data, the comparison allowed us to assess whether the travel speeds and turn delays of our custom built networks provided *reasonable* travel time estimates[1] (see Wang and Xu, 2011).

A "shortest path" analysis was completed for 1618 routes covering a broad range of travel distances (range = 0.5 - 647 miles, mean = 185.41 miles) and route types (e.g., rural, urban,

[1]The dominance of Google Maps in web-based mapping applications (BuiltWith Trends, 2012) does not guarantee that their roads data, travel speed data, or travel time estimates are, in fact, accurate. However, given the large and growing number of users, we believe that there is a low likelihood that the Google Maps source data contain a substantial amount of significant errors.

Table 1: **Travel speeds (miles per hour, mph) used in custom-built network datasets**

Road Type	**N1**	**N2**	**N3**	**N4**	**N5**
Expressways	70	60	60	62	65
Ramps	25	25	25	25	20
City owned, major	35	30	30	35	30
City owned, minor	25	20	20	25	20
Private	25	25	25	25	20
Minor collectors	55	55	55	45	50
Rural arterials and major collectors	55	55	55	45	50
Rural local	45	45	45	45	40
Urban, state owned arterials and major collectors	35	35	35	35	30
Urban, county owned arterials and major collectors	45	45	45	45	40
Urban, state owned local	35	35	35	35	30
Urban, county primary local	55	55	55	45	50
Urban, county local	25	25	25	25	20

Table 2: **Turn delays (seconds) used in custom-built network datasets**

Turn Type	**N1**	**N2**	**N3**	**N4**	**N5**
Non-existent expressway turn	1,200	1,200	1,200	1,200	1,200
Reverse (non U-turn)	8	8	10	45	20
Left	4	5	8	30	8
Right	2	3	5	15	5
Straight (with crossroad)	1	0	2	1	1
Straight (no crossroad)	0	0	0	0	0

Table 3: **Mean difference in travel time and road distance between Google Maps and custom-built networks in shortest path analysis**

	Time (minutes)	**Distance (miles)**
Network 1	18.39	2.84
Network 2	8.29	6.41
Network 3	1.54	4.42
Network 4	2.33	3.04
Network 5	0.87	2.55

suburban)[2]. All networks provided reasonable travel time estimates compared to Google Maps (see Figure 7 and Table 3). Network 5 was considered the most suitable for estimating travel time in this application. The travel speeds specified in Network 5 are a simple 5 mph reduction of the initial speed limit values from our hierarchical classification system, offering an objective method to account for sub-optimal driving and traffic conditions and the presence of stop signs, traffic lights, and other mechanisms for traffic control not present in the roads database. Additionally, the turn delays (outside of the expressway turn delay) in Network 5 are conservative, but conventional, estimates for normal surface street turns (Price, 2008, 2009).

2.4.2 Population and hospital data

2010 block population data and boundary files were acquired from the US Census Bureau (http://www2.census.gov/census_2010/, http://www.census.gov/geo/www/tiger/). Michigan statute requires that LAAs be identified using zip code population data, therefore the block population data were aggregated to their corresponding Zip Code Tabulation Area (ZCTA) boundaries (n = 978), herein referred to as zip codes. Because the census blocks nest perfectly inside the zip code boundaries, the block population polygons were converted to geographic centroids and spatially joined to the zip code boundary file. The population

[2]A custom-written automated query function was implemented in R^{TM}. The function sent origin and destination locations to the Google Maps API and returned the resulting travel times and distances.

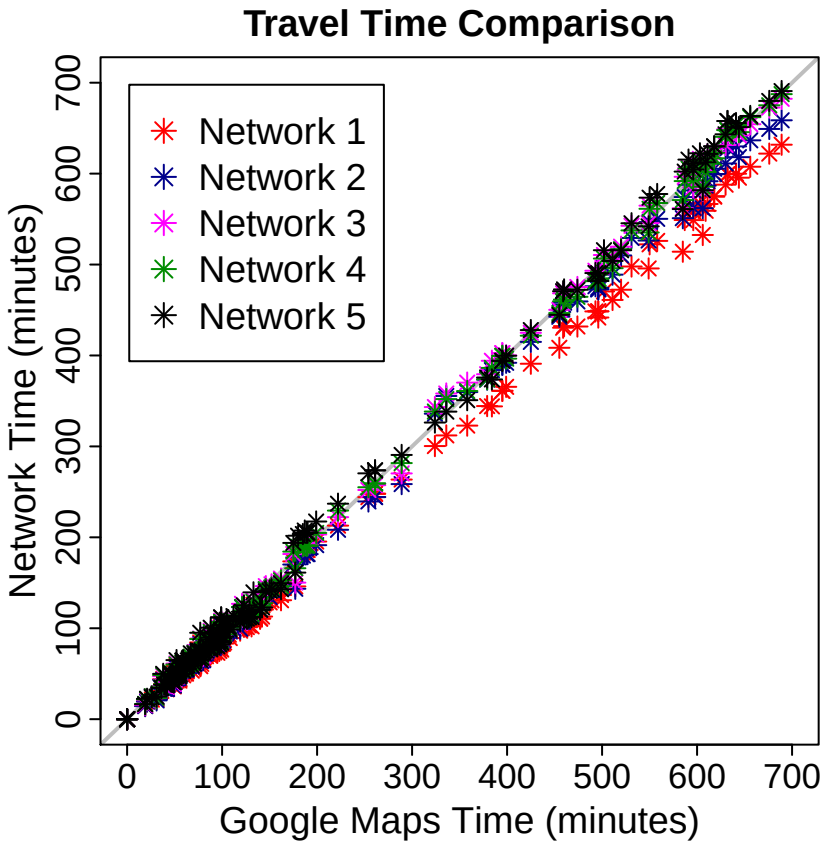

Figure 7: **Travel time estimates from custom-built networks compared with travel time estimated from Google Maps**

of each zip code was calculated by summing the population of all the block centroids falling within its boundaries. Michigan's total population was 9,883,640 in 2010.

Location and attribute data for 169 hospitals in Michigan were acquired from the MDCH. The hospital addresses were geocoded in ArcGIS and converted to point features. Hospital attribute data were used to identify and subset those hospitals offering acute care and 24/7 emergency room services, resulting in 137 hospitals.

2.4.3 Raster-based method

The raster-based method used to identify LAAs is documented extensively by Messina et al. (2006) and MDCH (Michigan Department of Community Health, 2009). Thus, it will only be summarized here. First, roads data were converted to a raster grid of 1 km cells wherein the travel speed for each cell was defined as speed of the *slowest* road falling inside the bounds of the cell. Because each cell required a specific travel speed, cells containing no roads were assigned 3 mph as an estimate of non-vehicular travel speed. Travel time or cost for traversing each cell was calculated using the cell length and specific travel speed. An accumulated cost surface was created wherein cell values represented the total travel time from the cell to the nearest hospital location (i.e., least cost path for each cell). To identify underserved areas, the accumulated travel time surface was reclassified into a Boolean surface based on whether the cell was greater than 30 minutes from a hospital location. The grid representing underserved areas was then filtered to remove any groups of less than three contiguous cells (using Queen's case connectivity). The filtering process was conducted in an effort to remove single cells and very small areas where no roads were present, but were generally "inside" the 30 minute travel bounds. Using a connectivity filter in lieu of a "count-only" filter ensured that areas near the edges of the actual underserved areas were not trimmed. Figure 8 shows an example of the filtering process near an underserved area in southern Michigan. After the filtering process, the underserved areas were converted from a raster grid to a vector data format (polygons) wherein a unique ID was assigned to each contiguous underserved area.

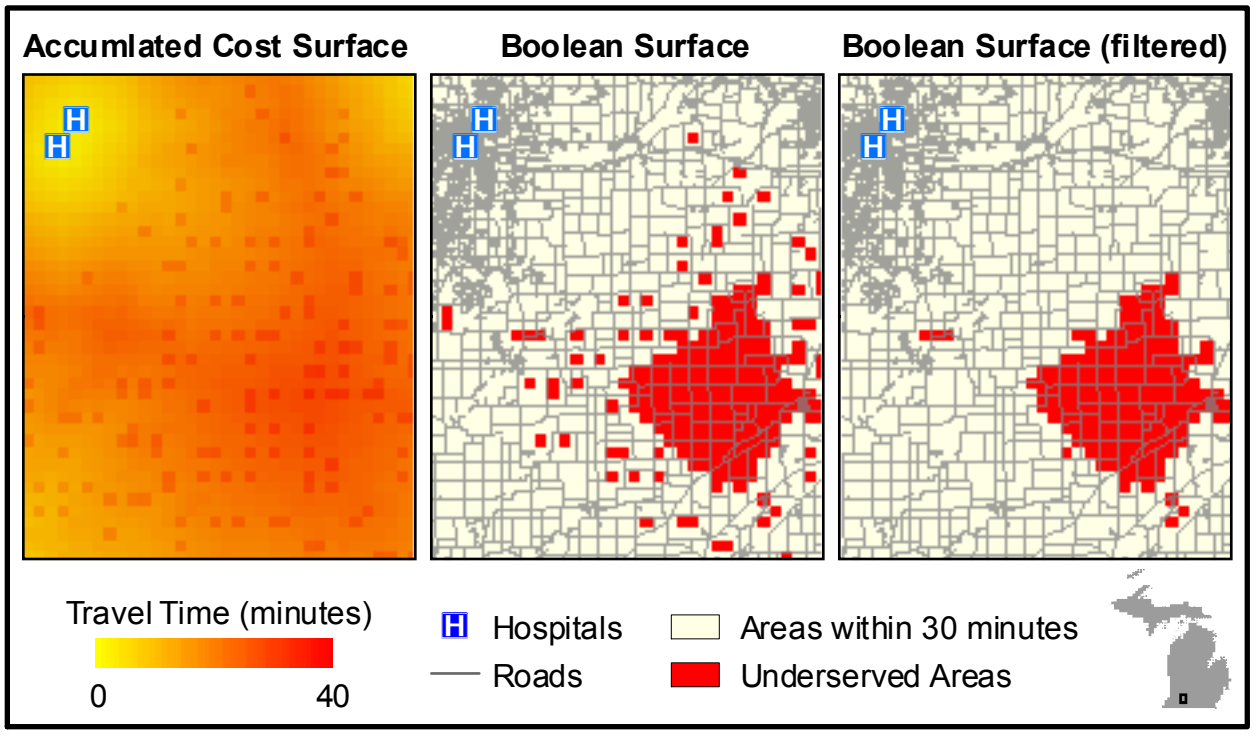

Figure 8: **Example of raster filter**

The population assignment method, according to Michigan's guidelines for identifying LAAs, requires that the *entire* population of a zip code be assigned to the underserved area if *any* portion of the zip code polygon falls inside of the underserved area. Thus, the underserved area polygons and zip code polygons were spatially joined in the GIS such that each underserved area polygon was assigned the summed population of all intersecting zip code polygons. Underserved areas with a total population of 50,000 or greater were then classified as Limited Access Areas.

2.4.4 Network-based method

ArcGIS Network Analyst was employed for all network-based analysis. Prior to converting the vector roads database to a network data format, each line segment was assigned a travel time value calculated using the line segment's length and estimated travel speed. Upon the conversion to the network data format, travel time was specified as the cost value for edges.

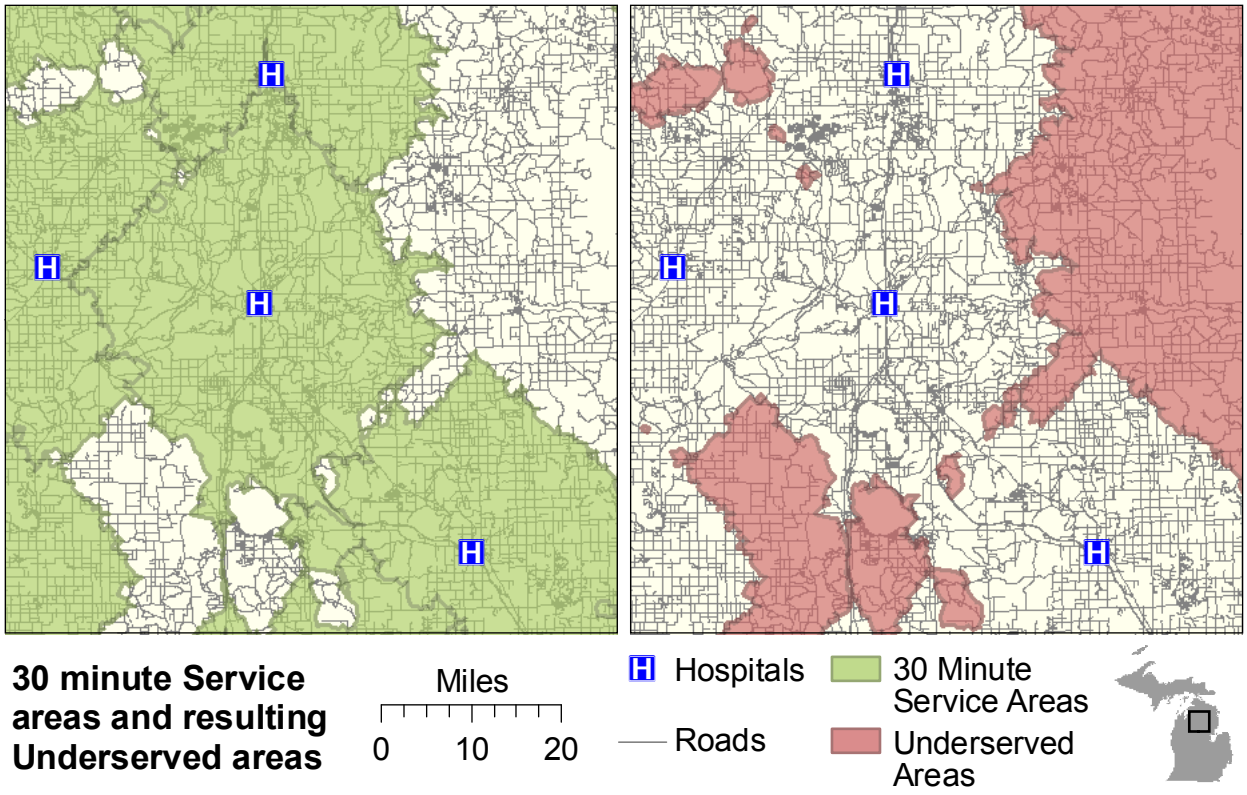

Figure 9: **Service areas (and resulting underserved areas) produced by network-based method**

Turn delays were defined to both control traffic flow and to model expected slowdowns in travel speed accompanying directional changes as detailed previously.

After the network was built, we created 30 minute travel time polygons for each of the hospital locations using the "Service Area" function. Underserved areas were identified by clipping the service area polygons from a state base map, essentially finding the inverse of the 30 minute travel areas throughout the state (see Figure 9). Population data were assigned to each underserved polygon and the LAAs were subset using the methods detailed in the previous section.

2.4.5 Sensitivity

To assess each method's sensitivity to the input roads data, the preceding steps for the raster and network methods were carried out a second time using the original speed limits of the

roads as opposed to the travel speeds in Network 5. In the raster-based analysis, the speed limit of cells with no roads present were raised to 10 mph. This test was conducted in an effort to uncover the variability in the results associated with small changes in the travel speed settings. Although this was not a comprehensive sensitivity analysis, exploring the difference in results due to the changes in the travel speed settings allowed us to estimate the relative importance of the settings for each method and the overall robustness of each data model.

We also evaluated each method for sensitivity to the scale of the data used to assign population to underserved areas. Instead of assigning the population using the zip code polygons, we assigned population using the US Census block centroids. In this method, a block's population was assigned to an underserved area only when the centroid fell within the bounds of underserved area polygon. Then, the population of all block centroids were summed and new LAAs were then identified using the updated population totals within the underserved areas. The results of the population assignment by census block were compared to the original results for both the raster and network-based methods. Considering that the block estimates of population are closer to the "true" number of people within the underserved areas (Apparicio et al., 2008), this comparison allowed us to evaluate which method is more sensitive to the population assignment method specified in Michigan's statute.

2.5 Results

2.5.1 Underserved areas

The underserved areas identified using both the raster and network-based methods are found in Figure 10 and Table 4. Overall, the raster-based method identified more total area, zip codes, and population as being underserved than the network method. The raster method produced fewer unique contiguous areas than the network method. Examination of Figure 10 reveals that this result was due to larger and more contiguous areas in the raster output. The most notable difference between methods is the total population identified as being

Table 4: **Comparison of underserved areas (Percent figures reflect proportion of state totals)**

Underserved Areas	**Raster**	**%**	**Network**	**%**
Area (km^2)	52,971	*35*	40,043	*26*
Number of unique areas	223		386	
Number of zip codes	410	*42*	316	*32*
Total population (zip code)	2,258,452	*23*	1,280,257	*13*

underserved. Whereas the raster method reports that 23% of Michigan's population (≈2.26 million) lives in underserved areas, the network method identified only 13% (≈1.28 million), a difference of nearly one million people.

As Figure 10 illustrates, the underserved areas identified by both methods share similar shapes resulting in a general agreement in the overall configuration of underserved places throughout the state. We compared the spatial configuration of the underserved areas by conducting an overlay analysis. The total overlapping area (the areas identified by *both* methods) was 38,667 km^2, comprising 71% of the total area identified by *either* method (54,347 km^2). The network-based results are a nearly perfect subset of the raster-based results; only 1,376 km^2 were identified uniquely by the network method. Figure 11 shows a detailed example where each method produced both overlapping and unique results.

2.5.2 Limited Access Areas

The results of the LAA identification are found in Figure 12 and Table 5. Again, the raster method produced more total area, zip codes, and total population identified in LAAs. Similar to the results of the underserved areas, the most notable difference between methods is the total population identified. The raster-based method identified over 1.8 million people in LAAs, whereas the network-based method identified just over 650,000, a difference of over one million residents. Because the LAAs are a subset of the underserved areas, the spatial configuration produced by each method are similar.

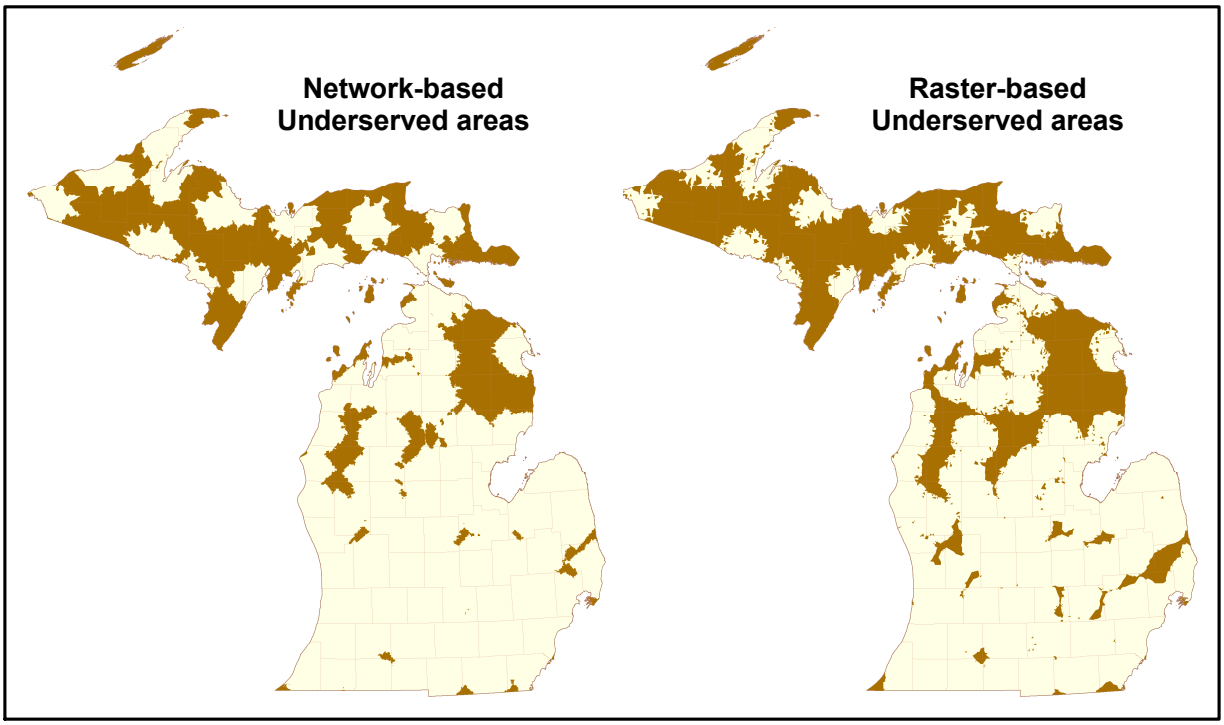

Figure 10: **Underserved areas**

Table 5: **Comparison of Limited Access Areas.** % figures reflect proportion of state totals

Limited Access Areas	**Raster**	**%**	**Network**	**%**
Area (km^2)	49,080	*32*	34,634	*23*
Number of unique areas	15		6	
Number of zip codes	328	*33*	199	*20*
Total population (zip code)	1,830,028	*19*	654,755	*7*

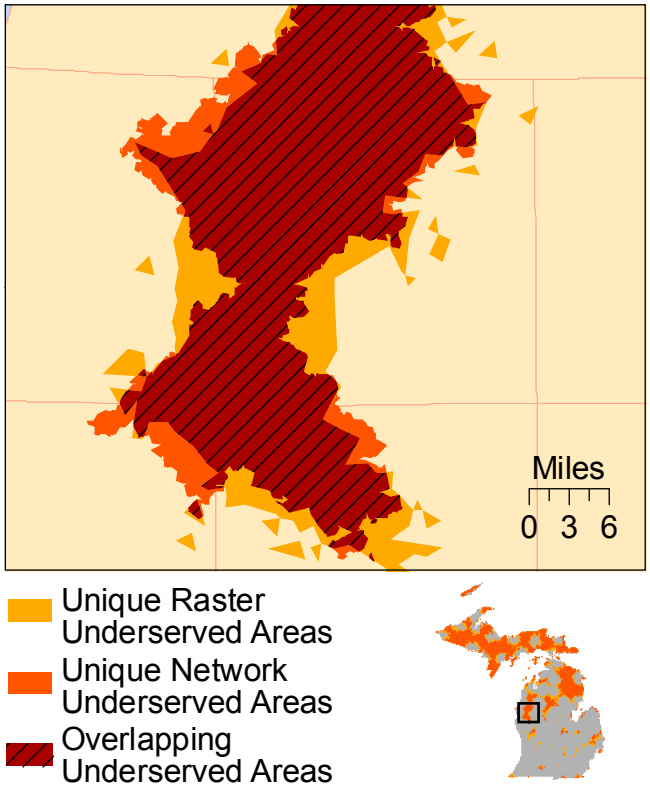

Figure 11: **Example of the similarities and differences between network and raster-based underserved areas**

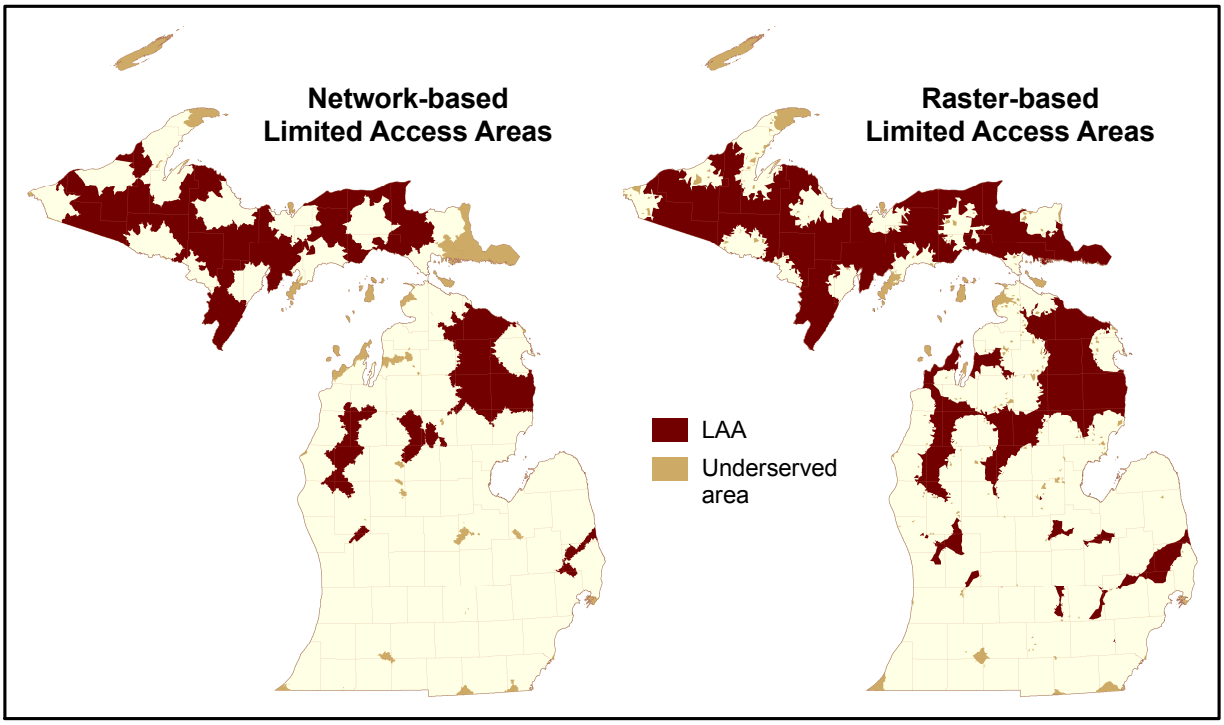

Figure 12: **Limited Access Areas**

2.5.3 Sensitivity

2.5.3.1 Speed limits The results for underserved areas and LAAs, using both the network and raster-based methods, are presented in Table 6. The table contains the initial areas identified and the areas identified using the actual speed limit values of the input roads data (+5 mph). Interestingly, the network-based method identified more people as being underserved, whereas the raster-based method identified more once the LAA criteria of 50,000 people was applied to the underserved areas.

2.5.3.2 Population representation Table 7 displays the number of people in underserved areas and LAAs when the population is assigned using the US Census block centroids. In both the raster and network-based methods, the use of a less aggregated population data source identifies far fewer people as being underserved within the state. A new set of LAAs were identified using the original 50,000 population criteria, but with population assigned

Table 6: **Comparison of underserved areas and LAAs identified with speed limits assigned to roads.** % change reflects change compared to initial travel speed settings.

Underserved Areas	**Raster**	**% change**	**Network**	**% change**
Area (km^2)	37,945	*-28*	31,815	*-21*
Number of unique areas	61	*-73*	390	*1*
Number of zip codes	238	*-42*	255	*-19*
Total population (zip code)	856,150	*-62*	1,000,612	*-22*

Limited Access Areas	**Raster**	**% change**	**Network**	**% change**
Area (km^2)	35,404	*-28*	19,343	*-44*
Number of unique areas	6	*-60*	3	*-50*
Number of zip codes	194	*-41*	117	*-41*
Total population (zip code)	694,562	*-62*	333,290	*-49*

Table 7: **Comparison of results from block centroid population assignment method with original travel speed settings.** % change reflects change compared to zip code intersection method.

Block centroid	**Raster**	**% change**	**Network**	**% change**
Underserved population	489,588	*-78*	191,420	*-85*
Limited access population	288,118	*-84*	0	*-100*

using the block population in lieu of the zip code populations. Figure 13 shows the resulting LAAs. Only three LAAs were identified using the raster-based method and no underserved area met the population criteria using the network-based method, although two areas nearly met the criteria with populations of 45,786 and 47,849.

2.6 Discussion

The results of the analysis show that large areas in Michigan are outside of a 30 minute travel time from an acute care hospital and thus have limited geographic accessibility, regardless of which data model is employed. Using the state's current methods, we found that over 2.2 million residents would be considered underserved and over 1.8 million residents would be classified as having limited access. The network-based method identifies fewer total residents

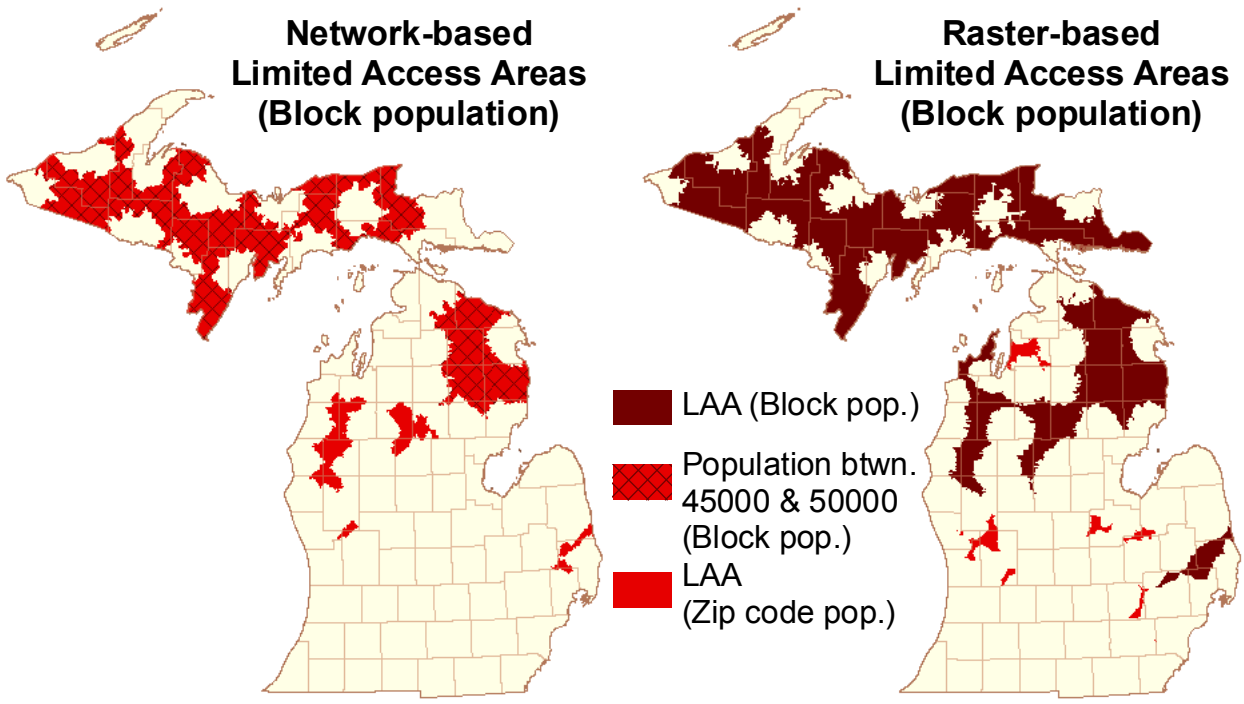

Figure 13: **Limited Access Areas with block population assignment method**

as underserved (≈1.28 million) and as having limited access (≈650,000). The results are less dramatic after "raising" the speed limits of the input roads data by 5 mph. However, both the raster and network-based methods identified large numbers of underserved and limited access populations in this scenario. Modifying the population assignment method resulted in far fewer people as both underserved and having limited access using both methods. Notably, the network-based method in conjunction with the block population assignment did not identify any official LAAs, although nearly 200,000 would be considered underserved in this scenario and two underserved areas nearly meet the 50,000 person LAA threshold.

The general location of the underserved areas and LAAs are similar between raster and network-based methods. Much of the underserved area is found in sparsely populated regions in Michigan's Upper Peninsula and northern Lower Peninsula. However, both methods identified small areas in the more populated central and southern Lower Peninsula. These smaller underserved areas are located in rural regions between urban centers. The raster-

based method identified larger, more contiguous underserved areas, thus more were classified as being LAAs.

In both the network and raster data models, the cost to travel among locations is based on the distance separating places and travel speed. Given these meta-parameters, the 71% agreement in total area identified as underserved is not completely surprising. However, in all of the tests performed in this analysis, the raster-based method identified more total area as underserved and as LAAs in comparison to the network-based method, warranting further examination. Figures 10 and 12 show that both methods identified similar patterns of underserved areas and LAAs throughout the state, however the raster method's results are universally larger. These results appear to be due to the underlying difference in the data models and the abstraction process occurring when converting the vector road data to a raster representation. The differences in the data models' characterization of space are worth reinforcing such that they directly influence geographic accessibility measurement. The raster data model defines space as a continuous surface where each cell within the data extent has a specific location and attribute value. The network data model defines space as an empty container that is populated only by features having specific locations and attributes. In the following paragraphs, we explore these differences and their implications for conducting geographic accessibility studies.

Given the structural constraints of the raster data model, accessibility calculation necessitates converting the vector road data to a cell-based representation. The conversion process requires a decision rule for assigning the speed limit to a cell when multiple roads are present within the cell bounds. Although a number of decision rules exist (e.g., the highest travel speed or the mean travel speed of roads within the cell), each increases the uncertainty of travel time estimates in the raster method. In the case study, because Michigan statute requires that the speed limit of the cell be determined by the slowest route available, only a small percentage of cells are assigned to the higher speed categories (i.e., highways and expressways) due to the presence of nearby slower roads. This results in a general overesti-

Table 8: **Michigan roads by travel speed**

Travel Speed (mph)	Network %	Raster %	Difference
20	30.78	38.92	8.14
30	5.99	0.36	-5.63
40	40.75	49.33	8.58
50	19.73	11.20	-8.53
65	2.76	0.19	-2.57

mation of the time required to travel among locations. Figure 14 contains an example that illustrates the dilemma produced by the abstraction process. In the example, an expressway traversing a medium-sized town nearly disappears after the conversion to the raster data format. Although Figure 14 shows a very specific example, the impact of this decision rule in the conversion process is not trivial when summed over the entire state. Table 8 contains the proportions of the roads in each travel speed class in the original vector format (based on road length) and after conversion to the raster format (based on cell counts). Notably, the raster format contains a higher proportion of roads in the 20 and 40 mph classes and less in the rest of the travel speed classes. As Figure 14 illustrates, this clearly inhibits high-speed travel. The result of slower travel speeds is an overestimation of travel time among locations and an increased amount of area identified as being underserved. As Table 4 shows, the raster-based method identified nearly 13,000 km^2 more total area as being underserved than the network-based method. In addition, the raster-based underserved areas were larger on average than the network-based areas (237.54 km^2 vs. 103.74 km^2). Larger contiguous underserved areas increase the probability that the 50,000 population threshold will be reached for LAA classification. Hence, the raster-based method identified nearly 1.2 million more people in LAAs than the network-based method.

All areas of the state should be accounted for in the LAA identification process (Messina et al., 2006). This creates a conundrum- LAAs are conceptually based upon vehicular travel time, yet some places in the state do not have any roads present. In the raster data model, all

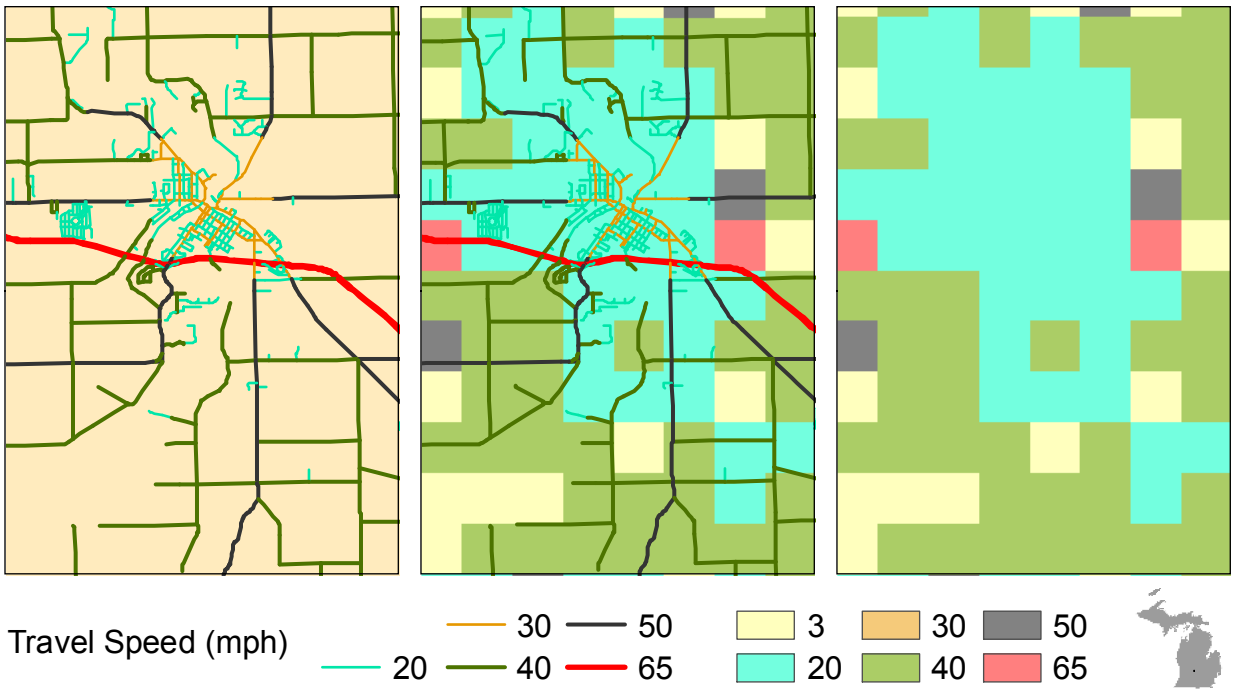

Figure 14: **Conversion of vector roads data to raster data format with slowest route rule**

locations within the data extent are explicitly defined and measurable. Hence, to be included in the service area estimation, each cell *must* be assigned a specific travel speed even if no roads are present within the cell. The network model does not define "space" outside of the network features (i.e., places not located on a node or edge feature). Therefore, non-road areas are undefined and not directly measured in service area calculation. Because the two data models diverge greatly in their characterization of space without roads, each method requires specific techniques to account for the presence of non-road areas when identifying geographic service areas based on vehicular travel time estimates.

In the raster method, non-road cells are not distinguished from cells with roads. Therefore, by assigning an artificially low travel speed value to non-road cells (e.g., walking speed), vehicular-based travel time estimates originating at these cells will be artificially high. Regions near the origin of the service area will be less affected than those located towards the periphery of the serve area extent. For example, the travel time to exit a 1 km non-road cell

with a travel speed of 3 mph is 6.21 minutes. When a specific threshold value for a service area is implemented, the higher travel time estimates for non-road cells result in regions or cells identified as "non-served" areas even though they fall *within* the extent of the larger service area (see Figure 8). When combined with the conservative population assignment method employed by Michigan, the non-road cells have the potential to significantly bias the results of the analysis. Therefore, we implemented the filter process to limit the number of non-road cells identified as underserved. As observed in the results of the speed limit sensitivity analysis, the raster-based method is much more sensitive to changes in the input speed limits. The 5 mph increase in travel speeds led to a 28% reduction in the total area (15,000 km^2) and 62% reduction in the population (1.4 million) identified as underserved, far outpacing the changes observed in the network-based method. Whereas some of the raster-based method's sensitivity can be attributed to the cell-based representation of roads and the predefined directional movement (considering that travel occurs in large 1km steps between cells), we believe that much of it is due to the change in speed for the non-road cells (from 3 mph to 10 mph).

"Non-road" areas are also accounted for in the network-based method; however, this process is not as apparent due to the output format of the data produced using ArcGIS Network Analyst. The "Service Area" function produces polygon features which are in turn used to clip a state base map to find non-served areas. Albeit indirectly, all areas in the state are measured when implementing the network-based method to identify service areas. Although this technique appears straight-forward, it is not without uncertainty. Service area polygons constructed from the network-based data model are actually areal approximations of the network edges (roads) within a specified travel time from the origin location. In Network Analyst, the network edges are converted to a triangulated irregular network (TIN) data structure with travel time estimates along the edges as the "height" value. Service area polygons are then formed by subsetting the TIN to only those areas falling within the specified travel time (ESRI, 2010). Figure 15 shows a service area where large regions, both

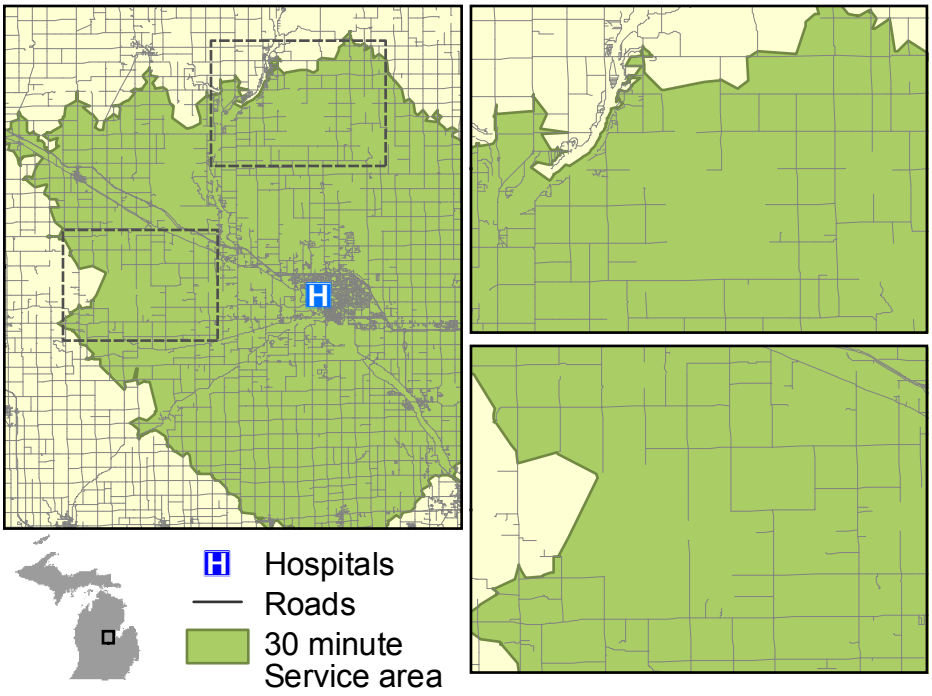

Figure 15: **Service area delineation in areas where no roads are present**

inside and near the bounds, have no roads. The figure includes two detailed examples of non-road areas to help illustrate the abstraction process of generating a polygon from a set of lines. In the upper right example, the non-road area is nearly completely enclosed by roads within 30 minutes, thus the entirety of the non-road area is considered "served". In the lower right example, the non-road area is bisected by the boundary of the service area. Specifically, the "cut out" region in the service area appears to be a remnant of the TIN conversion and subsetting technique. In theory, this particular boundary could be located anywhere within the non-road area; therefore, its true location is uncertain. The uncertainty associated with the polygon generation process raises questions regarding the validity of the service area boundaries produced by Network Analyst. However, we did not find any evidence that this led to a large amount of over or under-representation of underserved areas (and hence, LAAs) in our case study.

Because the conceptual models of space differ significantly between data models, topo-

logical relationships governing movement among locations are also highly dissimilar. In the raster model, connectivity is defined solely by cell proximity- movement only occurs in single step increments in predefined directions from the cell. The network data model, on the other hand, enforces strict connectivity rules within the data structure itself; travel only occurs along the edges of the network and directional changes can only be accomplished at nodes. Because the actual cost of travel between locations is highly dependent upon the connectivity provided by the transportation network linking the locations, the models' differences in defining connectivity lead to dissimilar travel time estimates. Specifically, real-world connectivity is not accounted for in the raster data model. Therefore, travel routes among locations may be geographically warped, resulting in inaccurate travel time estimates. For example, in Figure 14, all cells surrounding the 65 mph cell (on the right side of the map) have the potential to "route" through this cell. However, in the original vector road data, no ramp connects the surface streets to the expressway within this cell. Only the cell to the left and bottom of the 65 mph cell are actually connected to this cell. Therefore, movement is less restricted in the raster model than in the real-world and travel time estimates will generally be underestimated. In our case study, we believe that the underestimation of travel speeds was offset by the previously discussed overestimation of travel time due to the "slowest route" assignment rule.

Reducing the cell size of the input data used in the raster-based method would result in improved travel time estimates. Specifically, smaller cells will increase the probability of a single road falling within each cell, negating the impact of the decision rule to assign travel speeds to multi-road cells. In addition, as cell size is reduced, the topological similarity between the raster travel speed surface and the original roads data increases (see Figure 5). As a result, travel time estimates would be more accurate for cells falling on or near the road network, providing improved results in simple distance measurements and routing applications. However, for service area identification, reducing the cell size would also lead to an increase the number of non-road cells in the raster data. This would likely require a more

sophisticated method to create the travel speed surface, a more elaborate filtering process to remove these cells, or a polygon generating algorithm similar to the one employed in the network-based method. Additionally, reducing cell size may lead to substantial increases in processing time and data storage requirements (Upchurch et al., 2004; Schuurman et al., 2006).

By design, the zip code population assignment rule used in Michigan is conservative (Messina et al., 2006) in that it attempts to minimize the likelihood of source A errors (Current and Schilling, 1990). Hence, by assigning the entire zip code population regardless of the amount of area overlapping an underserved area, the true population with limited geographic accessibility is almost certainly overestimated. The results from the block population assignment method illustrate the magnitude of the overestimation. The percent change values in Table 7 show that the network-based method was more sensitive to the block population assignment method, overall. This is likely a result of the differences in the size and shape of the underserved areas produced by each method. On average, the raster-based method produced larger contiguous underserved areas. Due to the abstraction and filtering processes (see Figure 8) in the raster-based method, the *minimum* size of an underserved area is 3 cells ($3km^2$). The network-based method has no such size restriction. This difference has three main implications in relation to population assignment. First, larger areas increase the likelihood that an individual area will intersect multiple zip codes when assigning population using the zip code intersection method, resulting in more underserved areas meeting the LAA population criteria (See Tables 4, 5, 6, and 7). Second, unequally sized underserved areas can be assigned the same population. For example, using the intersection method, a very small area that falls on the border of two zip codes would be assigned the same population as a larger area completely covering the two zip codes. However, third, larger areas increase the likelihood that an underserved area will contain a block centroid when the population assignment method is modified. Considering that the average size of the raster-based underserved areas were generally larger than their network counterparts, the raster-based method

was less affected by the change in the population assignment method.

2.7 Conclusions

We have presented a comparison of raster and network-based methods for measuring geographic access to health care facilities. Specifically, we have explored how both conceptual and practical differences in the underlying data models have the potential to influence travel time estimates. In Michigan, each data model and method produced underserved areas and LAAs with similar configuration and shape, but of varying size. Specifically, the raster-based method identified 132% more land area as underserved than the network-based method. After assigning population to the underserved areas, the results clearly indicate that these spatial differences resulted in substantial variation in the number of people with limited geographic accessibility to acute care hospitals. In fact, the raster-based method identified 176% more *people* than the network-based method, a difference of nearly one million state-wide. Using the 50,000 population minimum for an underserved area to be deemed an LAA, the differences were even greater with the raster-based method identifying 142% more land area and 279% more people in LAAs.

Because speed limit data were not available for Michigan roads, travel speeds were estimated using the available road attribute data. Although we presented a detailed hierarchical speed limit classification system, the unavailability of the true speed limits, the variability in road surface types, and the large number of roads throughout the state make a perfect characterization of travel speeds impossible. Therefore, we tested each data model for sensitivity to changes in the travel speed settings. The method using the raster data model was more sensitive to the input speed limits of the roads data. Specifically, a small increase in travel speed settings produced greater changes in the resulting underserved areas and population identified when compared to the network-based method.

Messina et al. selected the raster-based method to fulfill the requirement that all areas of the state be measured directly while assessing geographic access in Michigan (Messina et al.,

2006). However, we have illustrated that converting the roads data to a 1 km cell resolution leads to a substantial loss of topological relationships due to the abstraction process. In addition, the coarse resolution requires a decision rule to assign travel speeds to cells with multiple roads present, resulting in a lower precision travel speed dataset. A reduction in cell size would provide a travel speed surface more similar to the original roads data along with better travel time estimates and more accurate routing results. Uncertainty associated with travel speed classification systems is always present in these kinds of large, unconstrained travel models. Future application of raster data modeled geographic access should explore alternatives to the methods described here for assigning travel speeds to cells with multiple roads and cells where no roads are present. Furthermore, an examination of the effects of cell size is also warranted in future research efforts as it was not considered here.

As noted earlier, the conservative population assignment method currently employed in Michigan likely overestimates the number of people in underserved areas (and thus in LAAs). We implemented an alternative population assignment method using higher spatial resolution data. Our findings suggest that the network-based method was more sensitive to the block population data assignment method. This sensitivity is likely due to the overall smaller underserved areas produced by the network-based method and its lack of a minimum size filter as was employed in the raster-based method. However, this finding speaks more to the population assignment method used by Michigan rather than the results of the travel time analysis. Thus, we believe that the overestimation of the population with limited geographic accessibility, regardless of whether the network or raster-based method is employed, warrants further evaluation.

Both the network and raster data models provide a valid structure for constructing travel time models. A definitive conclusion regarding the superiority of one or the other is unjust, however, due to the lack of true reference data to compare each against. Therefore, we recommend that, when measuring geographic access for health-related applications, researchers consider how the data models and associated methods employed may potentially influence

their results. Because the raster data model defines all areas as traversable, the raster-based method appears more suitable when estimating travel time service areas for non-vehicular travel modes or in regions where travel is not restricted to roads. For estimating vehicular-based travel time, we contend that the network data model provides a more accurate characterization of the topology governing vehicular travel. Therefore, for this travel mode, we believe that the network-based method is the appropriate choice to identify areas with limited geographic access to health care services.

Acknowledgements: This work was funded by the Michigan Department of Community Health, Certificate of Need Program. The authors would like to thank the two anonymous reviewers for their helpful comments and Larry Horvath for his assistance in determining the final set of hospitals used in the analysis.

Regional health care planning: a methodology to cluster facilities using community utilization patterns

3.1 Abstract

Background: Community-based health care planning and regulation necessitates grouping facilities and areal units into regions of similar health care use. Limited research has explored the methodologies used in creating these regions. We offer a new methodology that clusters facilities based on community utilization patterns and geographic location. **Case study:** Our case study focuses on Hospital Groups in Michigan, the allocation units used for predicting future inpatient hospital bed demand in the state's Bed Need Methodology. We detail the scientific, practical, and political concerns that were considered throughout the formulation and development of the methodology. **Methods:** The clustering methodology employs a 2-step K-means + Ward's clustering algorithm to group hospitals. The final number of clusters is selected using a heuristic that integrates both a statistical-based measure of cluster fit and characteristics of the resulting Hospital Groups. **Results:** Using recent hospital utilization data, the clustering methodology identifies 35 Hospital Groups in Michigan. After extensive research, review, and discussion, the new clustering methodology was approved by Michigan's Certificate of Need Commission to replace the state's previous methodology. **Conclusions:** Despite being developed within the politically charged climate of Certificate of Need regulation, we provide an objective, replicable, and sustainable methodology to create Hospital Groups. Because the methodology is built upon theoretically sound principles of clustering analysis and health care service utilization, it is suitable for grouping either facilities or areal units. **Keywords:** Health care utilization, hospital planning, certificate of need, clustering, K-means, Ward's.

Submission information: Submitted to Source Code for Biology and Medicine on February 22, 2012. Submitted to BMC Health Services Research on April 1, 2012. Authors on manuscript: Paul L. Delamater, Ashton M. Shortridge, Joseph P. Messina.

3.2 Background

Health care planning and regulation in the United States has generally attempted to achieve two broad goals: 1) promote public health by ensuring that the supply of services meets the population's need and 2) contain health care costs by regulating the supply of services to a level congruent with the need of the population. Regulation is often enforced through state-level Certificate of Need (CON) programs, which attempt to enable a sufficient supply of service to meet the population's health care needs without providing a large oversupply or duplication of services (Ferrier et al., 2010). CON programs require that proposals for additional health care services or facilities demonstrate an unmet need prior to approval. Although their merits have been questioned over the past 40 years (see US Federal Trade Commission, 2004; Rivers et al., 2007; Ferrier et al., 2010) and they are no longer federally mandated, CON programs persist throughout the US.

A number of states implement CON programs to regulate the supply of acute care hospitals, inpatient hospital beds, and hospital services (Langley et al., 2010). Considering that the costs of hospital-based care make up a plurality of overall health care spending (Kaiser Family Foundation, 2009), hospitals are a logical target for cost containment measures. Additionally, Roemer's Law (Roemer, 1961) states that *a bed built is a bed filled*, implying that an oversupply of hospital beds results in more and possibly unnecessary hospitalizations and costs.

Health care services are used by people, but are supplied by health care professionals who deliver these services at hospitals, clinics, and other facilities. Although the demand for hospital services can be considered an attribute of people or populations, the supply only exists at hospitals. In addition, the areal units used to aggregate populations rarely, if ever, contain residents who use a single health care facility (Bay and Nestman, 1984). To enable community-based planning of health care resources, communities and/or hospitals are grouped to form regions of similar health care use. Thus, planning occurs at a regional level wherein the supply of health care resources available to the community are measured against

community need. CON programs predict or evaluate the relationship between supply and demand of hospital beds, necessitating methods or techniques for grouping both population units and hospitals (e.g., Illinois General Assembly, 2012; New York State Department of Health, 2012; North Carolina Department of Health and Human Services, 2012).

Very limited research emphasis has been placed on grouping or clustering hospitals based on similarity in community utilization. Methods for clustering hospitals using multivariate data received attention from health services researchers in the 1970s and 1980s. These studies, however, were more focused on identifying hierarchical structure in the overall system of hospitals or identifying similar hospitals for determining reimbursement levels (Berry Jr., 1973; Elayat et al., 1978; Klastorin and Watts, 1981, 1982; Vertrees and Manton, 1986). More recently, this research topic has been revived in response to changes in health care delivery and organization (Dubbs et al., 2004; Luke, 2006; Zwanziger and Khan, 2008).

John Griffith, J. William Thomas, and colleagues explored the subject of service communities over 30 years ago (Griffith, 1972; Thomas, 1979; Thomas et al., 1981; Griffith et al., 1981), providing a clustering methodology that groups communities and hospitals simultaneously. The State of Michigan adopted the Thomas methodology (Thomas et al., 1981) for the creation of the state's Subareas (see Figure 16), the allocation units used in Michigan's Bed Need Methodology. In 2011, Michigan's CON Commission recommended a review of the Thomas Methodology. When implemented with current data, the methodology did not produce an acceptable Subarea configuration. Additionally, a number of theoretical and practical issues were identified, raising concerns that the methodology was no longer suitable to identify Subareas in light of the changes in hospital use and utilization patterns since its adoption over 30 years ago.

A review of the literature provides little guidance toward alternative or improved methods to group health care facilities. The branch of research most related to this particular problem is the creation of small areas. Yet, the methods used to create small areas have received little attention (Shwartz et al., 2001). Although multiple methods have been proposed

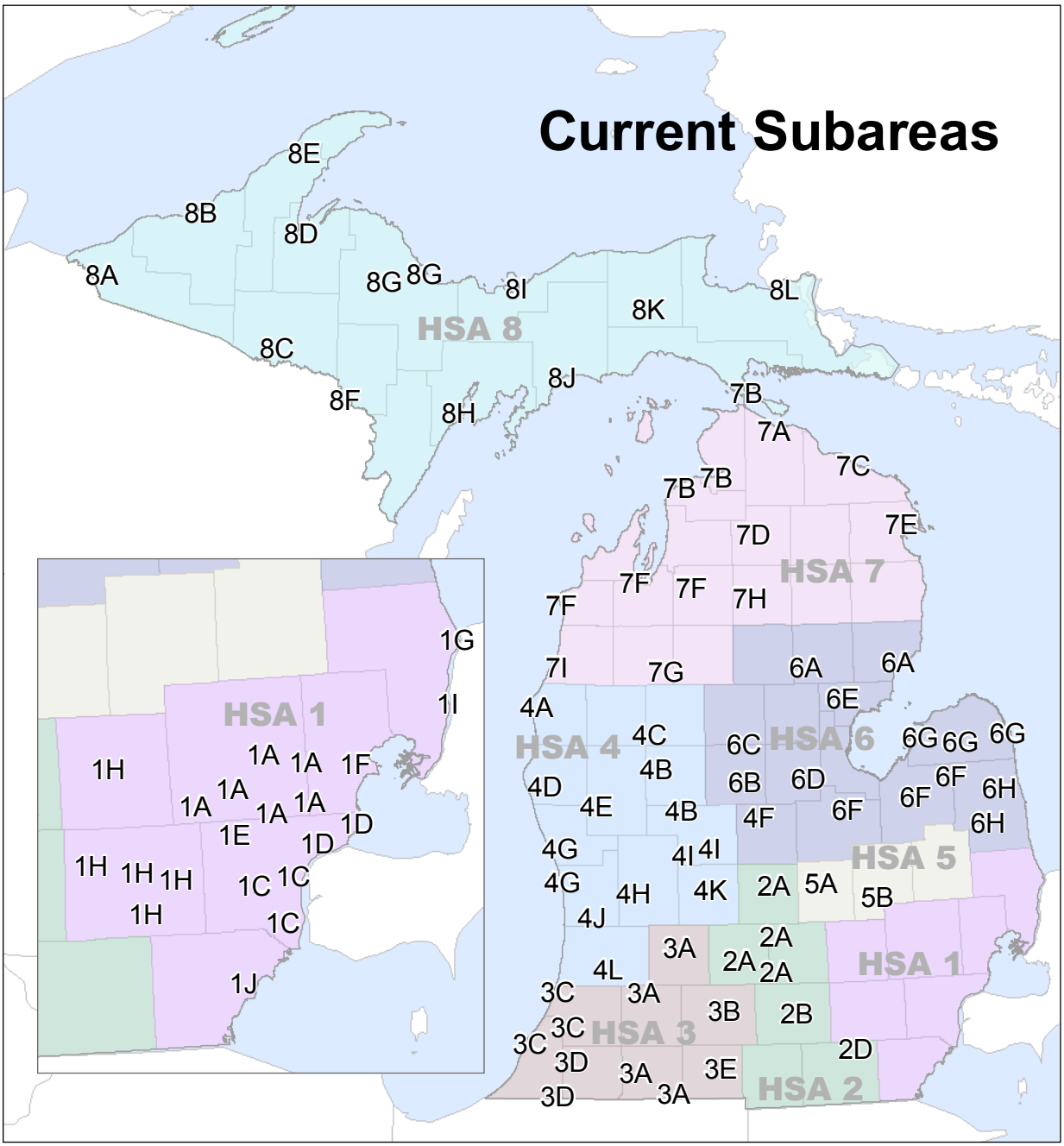

Figure 16: **Michigan's current Subareas.** Labels indicate hospital location and Subarea membership. Underlying colors represent Michigan's Health Service Areas (HSA). The data have been thinned for display purposes.

more recently to group communities into health service regions (e.g., Goodman et al., 2003; Shortt et al., 2005; Klauss et al., 2005), they are extensions of the straightforward, yet unsophisticated, plurality method employed by Wennberg and the Dartmouth Atlas group (Wennberg and Gittelsohn, 1973).

Here, we present a new clustering methodology that groups hospitals based on overall community utilization patterns and geographic location. The methodology is objective, replicable, and sustainable, offering a substantial improvement over the previous methodology. Furthermore, the methodology uses generally accepted clustering techniques and can be easily transferred to create small areas for health service studies. The source code necessary to replicate our clustering methodology is provided to ensure that the specific techniques we employ are unambiguous (See Appendix B).

Our manuscript is organized as follows: First, we offer a brief overview of clustering analysis and methods. The overview provides an introduction to a number of topics that were considered during the development of the clustering methodology. Next, we detail our case study and discuss the scientific, practical, and political concerns that were encountered while reviewing the Thomas Methodology and developing the new methodology. The clustering methodology is then provided in detail. We present the resulting hospital clusters and discuss the implications for adopting the methodology. Finally, we explore pathways in which our methodology can be extended for use in other health service applications.

3.2.1 Clustering

The overall objective in most clustering analyses is to assign individual observations into natural groups or clusters. Jain (2010, p. 652) states that the operational definition of clustering is:

> Given a *representation* of n objects, find K groups based on a measure of *similarity* such that the similarities between objects in the same group are high while the similarities between objects in different groups are low.

A large majority of clustering algorithms can be described as either hierarchical or partitional in nature. Hierarchical algorithms use an n x n similarity matrix to recursively form nested clusters over all possible values of K. Partitional algorithms divide observations into a user-defined number of clusters and utilize an n x n similarity matrix or an n x m matrix of observations, where n observations have m attributes or data dimensions.

Applied cluster analysis requires the analyst to make a number of subjective decisions. Prior to clustering, the attributes (or variables) used to describe similarity among observations must be determined, a potentially subjective process (Klastorin and Watts, 1981). Additionally, a large number of clustering techniques exist, creating a "user's dilemma" in the technique selection process (Dubes and Jain, 1976). Finally, determining the number of clusters or groups, K, is one of the most difficult problems in cluster analysis (Steinley, 2006; Jain, 2010). Milligan and Cooper (1987) provide a comprehensive review of clustering and cluster analysis, offering a seven-step structure to guide the clustering process.

3.2.2 Case study

In 2011, the State of Michigan CON Commission formed a Hospital Bed Standard Advisory Committee (HBSAC) to investigate issues related to the state's Hospital Bed Standards (see Michigan Department of Community Health, 2009). One charge of particular concern was to explore the methodology used to calculate the necessary supply of hospital beds needed to meet the state's future population demand (Bed Need Methodology). As part of this charge, the HBSAC formed a working group that focused on the specific methodology employed to create Subareas, the allocation units used in the state's Bed Need Methodology. The HBSAC working group was composed of various stakeholders in Michigan's health care industry including representatives from hospitals, hospital systems, and health insurance providers. The authors were commissioned by Michigan's Department of Community Health (MDCH) to provide the HBSAC with technical and scientific support throughout the process of reviewing the Thomas Methodology and to offer alternative approaches to create Subareas or other modifications of the Thomas methodology.

The HBSAC working group's initial concerns revolved around: 1) the legitimacy of the current Subarea configuration, 2) the high frequency of single hospital Subareas in the current configuration (32 out of 64), 3) the plausibility of re-implementing the Thomas methodology, which includes an initial automated clustering method and a secondary step where the results are reviewed and modified by an expert panel, and 4) the suitability of the Thomas methodology itself, given changes in health care delivery in the 30 years since its adoption. Despite efforts to trace the history of Michigan's Subareas, we were unable to locate detailed records or accounts of previous configurations. Outside of minor changes in 2002, we believe that the Subarea configuration had not undergone significant modification since the original formulation in the late-1970s.

Although a detailed description of the methodology is offered in Thomas et al. (1981) and the Hospital Bed Standards Michigan Department of Community Health (2009), portions of the methodology remain cryptic. A similar problem was experienced by researchers at Michigan State University when tasked with implementing Michigan's Bed Need Methodology and detailed by Langley et al. (2010). Therefore, the initial action required explicitly defining the Thomas Methodology (see Appendix A) and running the methodology with up-to-date population and hospital utilization data. We used the R programming language and environment (R Development Core Team, 2011) to complete this task. We selected R because of the statistical, graphical, and data processing capabilities it provides. Other benefits of using R are that it is a multi-platform open source language, is highly customizable, and can be augmented with a number of additional packages.

As Figure 17 illustrates, the Thomas Methodology does not provide a solution resembling

the current Subarea configuration when implemented with recent hospitalization data[3]. Most notably, only 21 Subareas were identified. The dissimilarity can most likely attributed to changes in hospital utilization patterns that have occurred since the last time the methodology was run. However, because the original Thomas Methodology results have been modified by an expert panel, we cannot state this with complete certainty. In addition, we identified theoretical and methodological issues in the current methodology that provided concern. These included an unreliable measure of hospital similarity; poorly defined home areal units; and subjective modification by an expert panel.

3.2.2.1 Unreliable measure of hospital similarity The Thomas Methodology clusters hospitals based on overlapping home areal units, defined by patient utilization patterns expressed using Relevance Index (RI) values. For a hospital, h_i, RI values are defined at each population unit (set of j units) such that

$$RI_{i,j} = \frac{Pd_{i,j}}{\sum Pd_j} \tag{4}$$

where $Pd_{i,j}$ is the number patient days used by residents of areal unit j at hospital i and $\sum Pd_j$ is the total number of patient days used by residents of areal unit j. Although RI is calculated for each hospital, the measure actually provides more information about communities rather than the hospitals. Because the patient days are summed for the areal unit in the denominator, the RI value describes the importance of the hospital to the community. Thus, using RI values to compare hospitals provides little information about the similarity

[3]Interpreting the definition of the "home areal unit" of each hospital or cluster of hospitals in the Thomas Methodology was especially problematic. The original manuscript is quite vague in its discussion of home areal units. Unfortunately, the definition in the Hospital Bed Standards does not offer clarification. Therefore, we implemented multiple versions of the Thomas Methodology, each with a slightly different interpretation of the home areal unit. Although each produced unique results, none provided Subareas that were similar to the current configuration. The results presented in Figure 17 defined the home areal unit as the zip code in which the hospital is located. This implementation also allowed the algorithm run until clustering was completed.

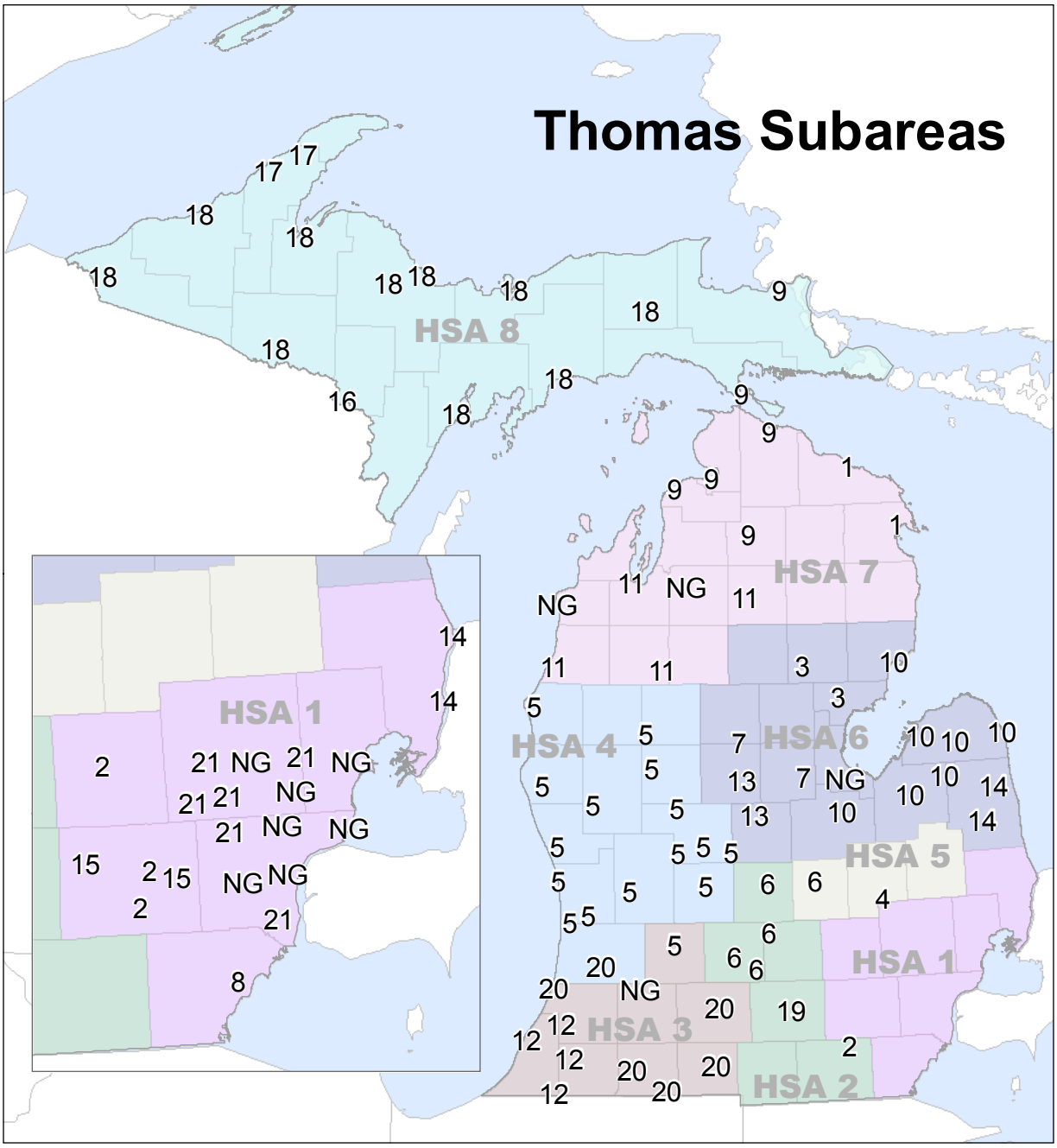

Figure 17: **Subareas produced by the Thomas Methodology using current data.** 34 hospitals did not possess the required minimum home area to be included in the Thomas Methodology. Because no details are provided by the methodology with regards to handling these cases, they were removed from the clustering process. They have been assigned NG (non-groupable) for display purposes. The data have been thinned for display purposes.

of the overall hospital utilization patterns.

An alternative measure of utilization patterns, the Commitment Index (CI), is a hospital-based representation of patient utilization patterns. CI is defined for a hospital, h_i, at each population unit (set of j units) such that

$$CI_{i,j} = \frac{Pd_{i,j}}{\sum Pd_i} \tag{5}$$

where $\sum Pd_i$ is the total number of patient days at hospital h_i. CI values measure the importance of each population unit to the hospital. Unlike RI, CI values are not directly influenced by the size of the hospital (as measured by number of inpatient beds). For example, two hospitals located near each other, one small and one large, may have very similar patterns of utilization when expressed as CI values (e.g., in Figure 18).

Although Griffith contends that RI is "more useful" than CI (Griffith, 1972), we find that only to be true for defining service populations or exploring market penetration for a single hospital. It has little utility for comparative purposes. Conversely, the CI values provide a suitable measure of similarity among community utilization patterns. As Figure 18 illustrates, the two hospitals have very similar patterns of patient utilization, drawing comparable percentages of their total patients from the surrounding areal units. The Pearson's correlation coefficient (r) of the two hospitals' CI values confirms the similarity with near perfect correlation ($r = 0.975$). Although correlation is also high between the hospitals' RI values ($r = 0.855$), the similarity in community utilization patterns is not nearly as apparent due to the differences in magnitude of the RI values. Furthermore, the Euclidean distance between the hosptials' CI values is 0.034, whereas the distance between RI values is 0.437.

3.2.2.2 Poorly defined home areal units In the Thomas Methodology, hospitals are clustered iteratively based on RI values in home areal units. However, these home areal units are poorly defined once hospitals have been clustered. Specifically, the home areal unit of the entire cluster is assigned as the home areal unit of a single cluster member hospital.

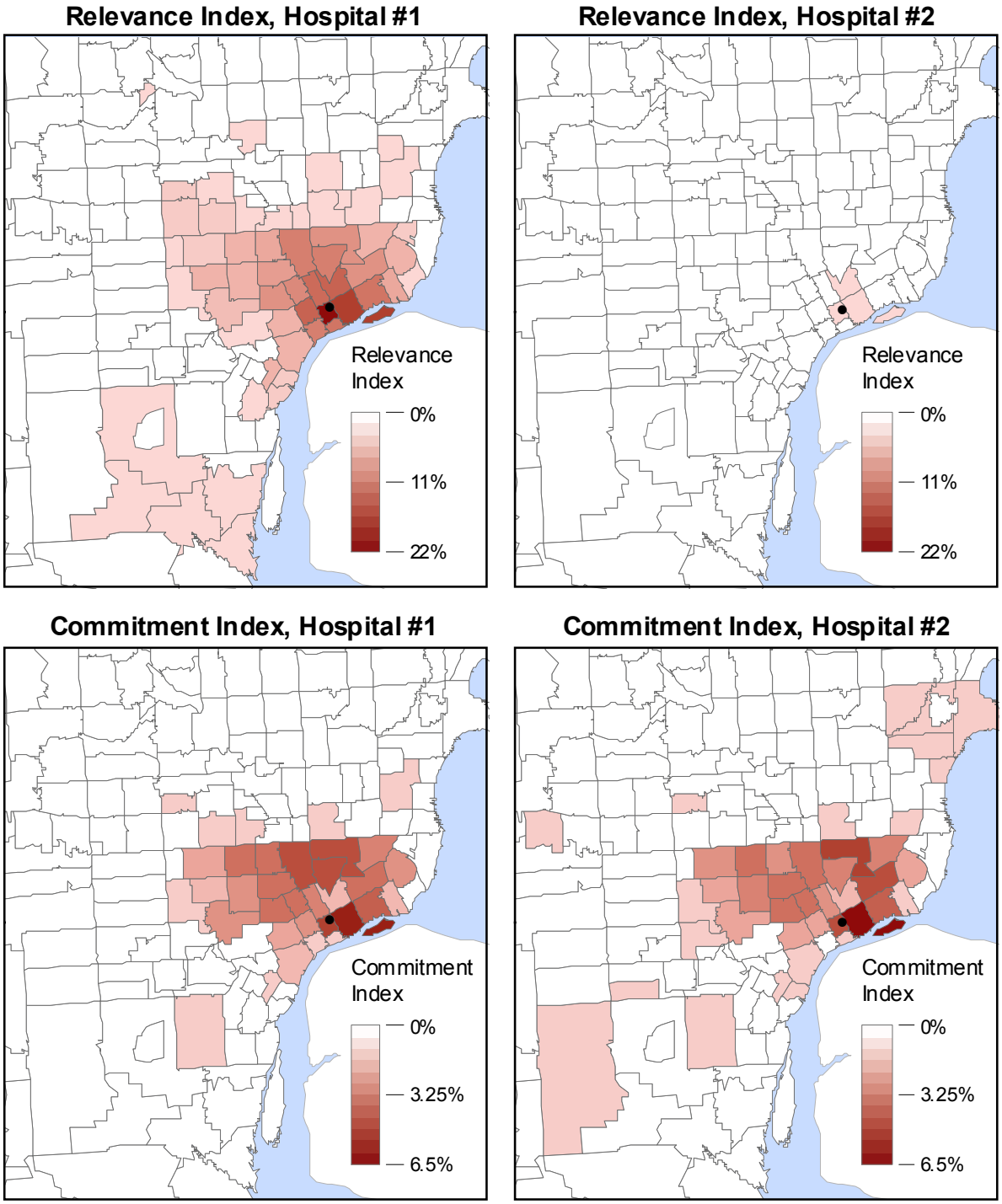

Figure 18: ***RI* and *CI* values for two hospitals of different sizes.** Hospital #1 has 470 licensed inpatient beds and Hospital #2 has 94.

Because the methodology further clusters these groups based on overlap within the single home areal unit, it does not acknowledge that multiple hospitals compose the cluster. This results in scenarios where hospitals grouped into the same Subarea may share little to no similarity. For example:

- Hospital A is clustered with Hospital B based on Hospital B's RI in Hospital A's home areal unit.
- Once the hospitals are clustered to form Cluster AB, the home areal unit is assigned as Hospital B's home areal unit.
- When Cluster AB is further clustered with Hospital C, the criteria for clustering is based on Hospital C's RI value in Cluster AB's home areal unit. Because Cluster AB's home areal unit was defined as Hospital B's home areal unit alone, overlap between Hospital C and Hospital A's home areal units is not considered.

In this scenario, Hospital C and Hospital A may share little or no similarity in the newly formed Cluster ABC. Because the Thomas Methodology iterates until there is little overlap among home areal units, this can lead to very large clusters (see Clusters #5 and #18 in Figure 17) or geographically distorted clusters (see Cluster #10 in Figure 17).

3.2.2.3 Subjective modification by expert panel In the Thomas Methodology, the Subareas results provided by the clustering algorithm are passed along to an expert panel for modification. Thomas et al. (1981, p. 46) state:

> Based on members' knowledge of hospital relationships and other factors influencing the reasonableness of proposed groupings, the committee is asked to decide whether the objectively determined clusters are in fact appropriate. ... Thus the committee makes the final determination, using the patient origin data analysis as one important source of information.

Although this step offers the potential to incorporate useful qualitative or local knowledge into Subarea formulation, it is also raises practical concerns with regards to implementation. The Hospital Bed Standards do not provide guidance regarding the composition of the expert panel or the scope of their charge. Additionally, by modifying the Subareas *post hoc*, the original results of the Thomas Methodology are lost, leaving no record that would allow for the utility of the automated method itself to be examined.

3.2.2.4 New methodology to cluster hospitals After discussing the theoretical concerns and application-oriented limitations present in the Thomas Methodology, the HBSAC working group opted to explore alternate approaches to creating Subareas, rather than choosing to modify the parameters of the Thomas Methodology in such a way that the methodology would provide reasonable results. In addition, the group decided to replace the term, *Subarea*, with *Hospital Group* to better reflect the nature and specific use of these units within the context of the overall Bed Need Methodology. For the remainder of this manuscript Subareas will be referred to as Hospital Groups.

Our overall goal in creating the new Hospital Group methodology was for the method itself to be as ***objective***, ***replicable***, and ***sustainable*** as possible. Considering the subjectivity present in clustering applications and the vast number of possible clustering methods, we placed emphasis on the higher-level theoretical issues, rather than specific application-oriented concerns. Preliminary discussions with the HBSAC working group focused on the identification of measurable hospital characteristics that could be used to compare and cluster similar hospitals. From this discussion, two characteristics were deemed as the most important, 1) that hospitals drew their patients from similar communities and 2) that hospitals were geographically proximate. Given these meta-parameters, we presented the HBSAC working group with a variety of suitable clustering methods. Because CON-related proceedings have the potential to become highly political affairs, we initially presented only the clustering methods themselves, rather than offering "results" of the methods. This left the

HBSAC to form their opinions based on the merits and appropriateness of the clustering methods themselves, not the Hospital Groups they produced.

3.3 Methods

3.3.1 Overview

The new clustering methodology employs a 2-step K-means + Ward's algorithm to create Michigan's Hospital Groups. This algorithm compares observations across multiple attribute values, allowing for both community utilization patterns and hospital location to be evaluated simultaneously in cluster formation. In this, specific patient hospitalization data and travel distance measurements among hospitals are required. The methodology includes a heuristic to determine the number of Hospital Groups, K, based on statistical measures of cluster fit and characteristics of the Hospital Group solution. We also include a set of techniques to assign a *new* or *proposed* hospital to the existing Hospital Group solution in case this scenario arises.

The source code used to implement the overall methodology[4] can be found in Appendix B. We utilize the R programming language using only base package functions to allow for portability across operating systems. The code in Appendix B has also been modified slightly from the actual code presented to the HBSAC in an effort to make it more generalizable. In the following sections, we provide a detailed description of each step in the clustering methodology.

3.3.2 Input data

The methodology requires georeferenced hospital utilization data. We employ data from the Michigan Inpatient Database (MIDB), a nearly exhaustive record of the state's inpatient hospitalizations. Each patient record includes the discharging hospital, the zip code of the

[4]Although not discussed in the manuscript, the CON approved source code contains additional steps to assign a numeric identifier to the resulting Hospital Groups based on their geographic location and bed inventory.

patient's residence, patient demographic information, and diagnostic codes. Using the most recent three years of MIDB data, the number of patient days used at each hospital by residents of each Michigan zip code are arranged in an n x z origin-destination (OD) matrix. Three years of data are included to ensure that recent patterns of state-wide hospital utilization are captured without the fluctuations possible in a single year. All existing hospitals that reported their inpatient data to the MIDB for any portion of the three year period are included. In this, reporting is essentially universal throughout the state's hospitals. The n x z matrix of patient days is converted to a CI matrix (for each hospital in n) using Eq. 5.

The geographic location of each hospital is represented as an 1 x n vector of the travel distances to the other hospitals in the state. When consolidated, this results in an n x n OD distance matrix. The use of an n-dimensional representation of location, in lieu of traditional 2-D locational attributes such as x,y geographic coordinates, is necessary to account for Michigan's particular physical characteristics and transportation infrastructure. Most notably, Euclidean distance measurements may lead to misrepresentations of true distances among locations near shorelines. For example, using only x,y coordinates to define location, hospitals in Michigan's "thumb" region (HSA 6) in Figure 16 would be considered *near* hospitals to their northwest, not accounting for the true magnitude of their separation due to the Saginaw Bay. Distances among hospitals are calculated as travel distances on Michigan roads using a custom-built network model (Delamater et al., 2012). After the n x n matrix is assembled, the distance entries are rescaled from 0 to 1 by dividing each by the maximum distance between any two hospitals. The rescaling process ensures that the range of values in the hospital utilization matrix and distance data matrix are similar(Milligan and Cooper, 1988; Steinley et al., 2004).

The utilization matrix and distance matrix are joined to form a final data matrix containing n rows or observations with m $(z + n)$ attribute values per observation.

3.3.3 Clustering algorithm

The K-means clustering algorithm is employed as the primary method to create Hospital Groups. The specific algorithm employed is that of Hartigan and Wong (1979), the default option in R's base package `kmeans()` function. Given a set of n observations with m associated attribute values to be partitioned into K clusters, K-means attempts to find the cluster solution (C) that minimizes the sum of the squared errors ($J(C)$) between cluster members (x_i) and their associated cluster center (c_k) over all clusters.

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{6}$$

Although their origins are closely related, two distinct characteristics of the basic K-means algorithm provided concern for identifying Hospital Groups. First, solving Eq. 6 is an NP-hard problem (Drineas et al., 2004), essentially rendering it non-computable in any acceptable amount of time. Thus, K-means relies on an search algorithm to approximate the solution and likely finds a locally optimal solution, rather than the globally optimal solution (Steinley, 2003; Jain, 2010). Second, the basic K-means method employs a random initialization procedure for the search algorithm. Given that the input data were of high dimensionality, the resulting Hospital Group solution identified by the randomly initiated K-means algorithm would likely vary between model runs. Therefore, the results would not be reproducible.

To examine the variability associated with the random initialization of K-means and for the presence of local optima, we initially grouped the hospitals into 50 clusters using 5,000 random starting locations. Although there were roughly 9 x 10^{203} possible solutions[5], the observed variability in the output cluster solutions was much higher than initially expected; each random start provided a unique 50 cluster solution.

To stabilize the clusters provided by the K-means algorithm, we "seed" it with rational

[5]Based on $K^N/K!$ (Kaufman and Rousseeuw, 2005) where $K = 50$ and $N = 158$.

starting locations in lieu using of the random start method (Milligan, 1980). Ward's hierarchical clustering algorithm (Ward, 1963) was employed to initially cluster the hospitals and provide the seed locations. The cluster centers produced by Ward's algorithm are a K x m set of locations that define the central location of each cluster in m-dimensional space. They are used as initial locations in the K-means search algorithm, creating a 2-step K-means + Ward's clustering algorithm. Because Ward's algorithm provides deterministic results, this effectively and efficiently removed the stochastic element present in K-means initialization. In addition, for $K = 50$, the cluster solution identified by K-means + Ward's provided a superior fit to solutions from *all* 5,000 model runs using K-means with random starts (see Figure 19). Although we cannot confirm that the K-means + Ward's algorithm provided the globally optimal solution, we are encouraged that a single model run produced such a large improvement in the fit of the cluster solution.

3.3.4 Determining the number of Hospital Groups

As was discussed earlier, one of the more difficult problems facing any applied cluster analysis is determining the number of clusters in which to group the data. Researchers have noted that the selection of K is largely subjective (Elayat et al., 1978), may be politically influenced (Klastorin and Watts, 1981), or completed by an analyst with expert domain knowledge (Jain, 2010). The members of the HBSAC working group were steadfast that the number of Hospital Groups (K) should be derived from the data itself, not explicitly predetermined prior to the clustering process nor modified after clustering is completed. However, no method or measure exists to definitively answer the question, "how many clusters should the data be grouped into?". Therefore, in conjunction with the HBSAC working group, we developed a heuristic to determine the number of Hospital Groups, incorporating a statistical approach along with a set of decision rules.

We define k as the set of integer values from 2 to n-1. A Hospital Group solution is created for each value in k using the K-means + Ward's clustering algorithm, allowing all possible values of K to be evaluated. The first step in the heuristic to determine the final

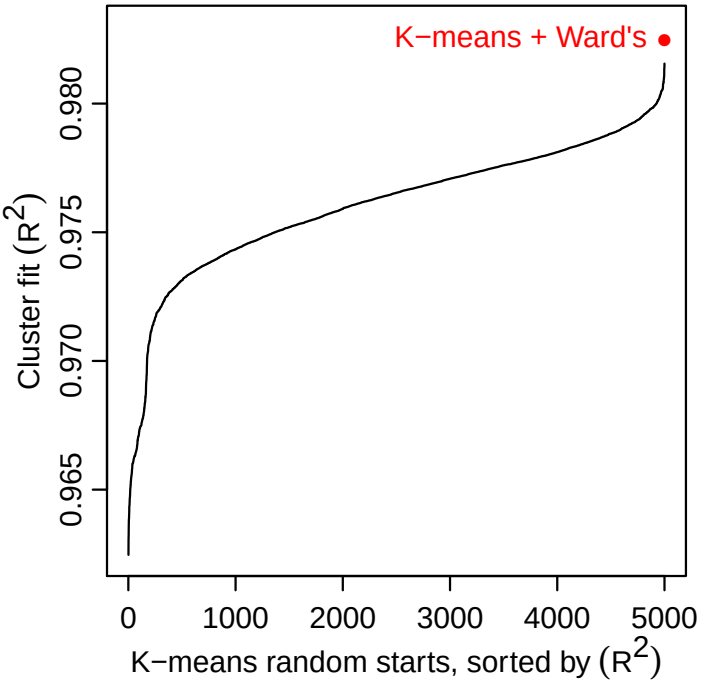

Figure 19: **Local minima and random starting locations with the K-means algorithm.** 5,000 K-means model runs for $K = 50$ produced 5,000 unique cluster solutions (black line). A single model run of the K-means + Ward's algorithm provided another unique cluster solution with a better fit (red point) than any of the 5,000 stand-alone K-means solutions.

value of K is to calculate the incremental F statistic ($incF$) (Gujarati, 1988) for each solution in k,

$$incF_i = \frac{\frac{R_i^2 - R_{i-1}^2}{k_i - k_{i-1}}}{\frac{1 - R_i^2}{n - (k_i - 1)}} \tag{7}$$

where

$$R^2 = 1 - (RSS/TSS). \tag{8}$$

RSS and TSS are the residual sum of squared error and total sum of squared errors, respectively, calculated for each cluster solution in k ($J(C)$ from Eq. 6 is equal to RSS). R^2 is an overall measure of the "fit" of the cluster solution to the original data. The incremental F statistic measures only the amount of fit gained from allowing an additional cluster in the solution, while also penalizing for adding this additional cluster. Because increasing K will almost certainly improve the R^2 of the cluster solution, $incF$ offers a measure that incorporates both fit and K. Initial candidate solutions are selected by identifying those with local maxima in $incF$ (all solutions where $incF_k > incF_{k-1}$ and $incF_k > incF_{k+1}$).

After the initial candidate solutions are identified, a set of decision rules is employed to select the final value of K. The HBSAC working group offered two qualifications for a suitable Hospital Group configuration, 1) that no individual Hospital Group contains more than 20 hospitals and 2) that the number of "single hospital" Hospital Groups is minimized. First, all initial candidate solutions where any single Hospital Group contains more than 20 hospitals are removed. Next, for each of the remaining solutions, the number of single hospital Hospital Groups is noted. The solution(s) having the minimum number of single hospital Hospital Groups is/are retained. If multiple solutions meet these criteria, the final solution is selected by choosing the candidate with the maximum K from the remaining solutions.

3.3.5 New hospital assignment

The HBSAC working group requested that the new clustering methodology include steps to assign a new or prospective hospital to the existing set of Hospital Groups. In the Thomas Methodology, this task was accomplished re-running the entire methodology with market survey data (projected RI values for the new hospital) added as a new observation. The HBSAC working group members doubted the veracity of these survey data and requested a simplified approach that did not require a market survey or rerunning the clustering methodology. We designed a method wherein a new hospital is assigned to an existing Hospital Group using geographic location.

A geocoded location of the new hospital is required to calculate the travel distance from the new hospital to each existing hospital. These distances are placed in a 1 x n vector, which is rescaled using the maximum distance between any two hospitals in Michigan (see Input data) and arranged such that the entries are in the same order as the entries in the original travel distance matrix.

Like the Ward's algorithm, the 2-step K-means + Ward's algorithm produces a K x m matrix of cluster centers. The cluster centers from the Hospital Group solution are subset to only those columns corresponding to the travel distance attributes (column numbers z+1 to m), resulting in a K x n matrix. This subset represents the geographic location of the existing Hospital Group centers in n-dimensional space.

The Euclidean distance (d) from the new hospital to an existing Hospital Group center is calculated

$$d = \sqrt{\sum_{i=1}^{n} (c_i - h_i)^2} \tag{9}$$

where c_i is the cluster center for the Hospital Group and h_i is the rescaled distance vector for the new hospital. A d value is calculated from the new location to each existing Hospital Group. The new hospital is assigned to the Hospital Group having the minimum d value.

3.4 Results

We implemented the new Hospital Groups clustering methodology using inpatient hospitalization data from 2007 to 2009, which included 169 acute care hospitals. A small number of hospitals reported their inpatient data to the MIDB in tandem with another hospital or set of hospitals. The hospitals reporting together are owned by the same health care system and are located very near each other geographically. Therefore, these were treated as a single observation for the purposes of clustering[6]. Two hospitals did not report any patient records to the MIDB and were removed prior to clustering. The final data matrix consisted of 158 observations with 1065 attributes (CI values for 905 zip codes and rescaled travel distance to 160 hospital locations).

A Hospital Group solution was created using the 2-step K-means + Ward's algorithm for each value of K from 2 to 157. We implemented the heuristic to select the number of Hospital Groups for the final solution. 49 initial candidate solutions were identified using $incF$ values (see Figure 20 and Table 9). Next, candidate solutions of less than 29 clusters were removed due to the maximum number of hospitals in a single Hospital Group. From the remaining candidate solutions, the minimum number of single hospital Hospital Groups was 1. Therefore, all solutions greater than 35 clusters were removed from consideration. From the remaining candidate solutions, 35 was the maximum value of K and selected as the final Hospital Group solution (see Figure 21).

[6]Because these hospitals were each associated with a unique geographic location, their travel distance measurements were slightly dissimilar. To calculate the travel distances for the grouped set, we took the mean of the hospitals comprising the group. However, when calculating the number of "single hospital" Hospital Groups during the clustering methodology, the grouped set was not considered a single facility.

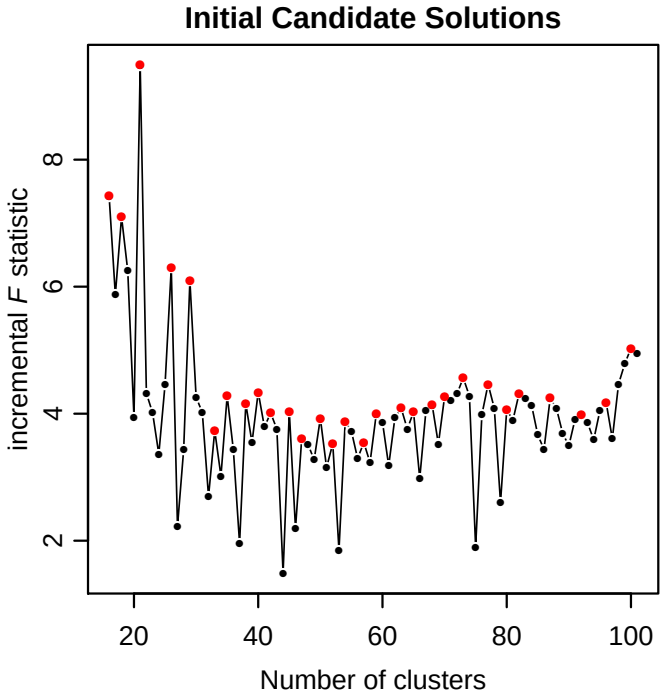

Figure 20: **Initial candidate solutions for Hospital Groups.** Data are truncated for display purposes. Red points represent local maxima in $incF$ values.

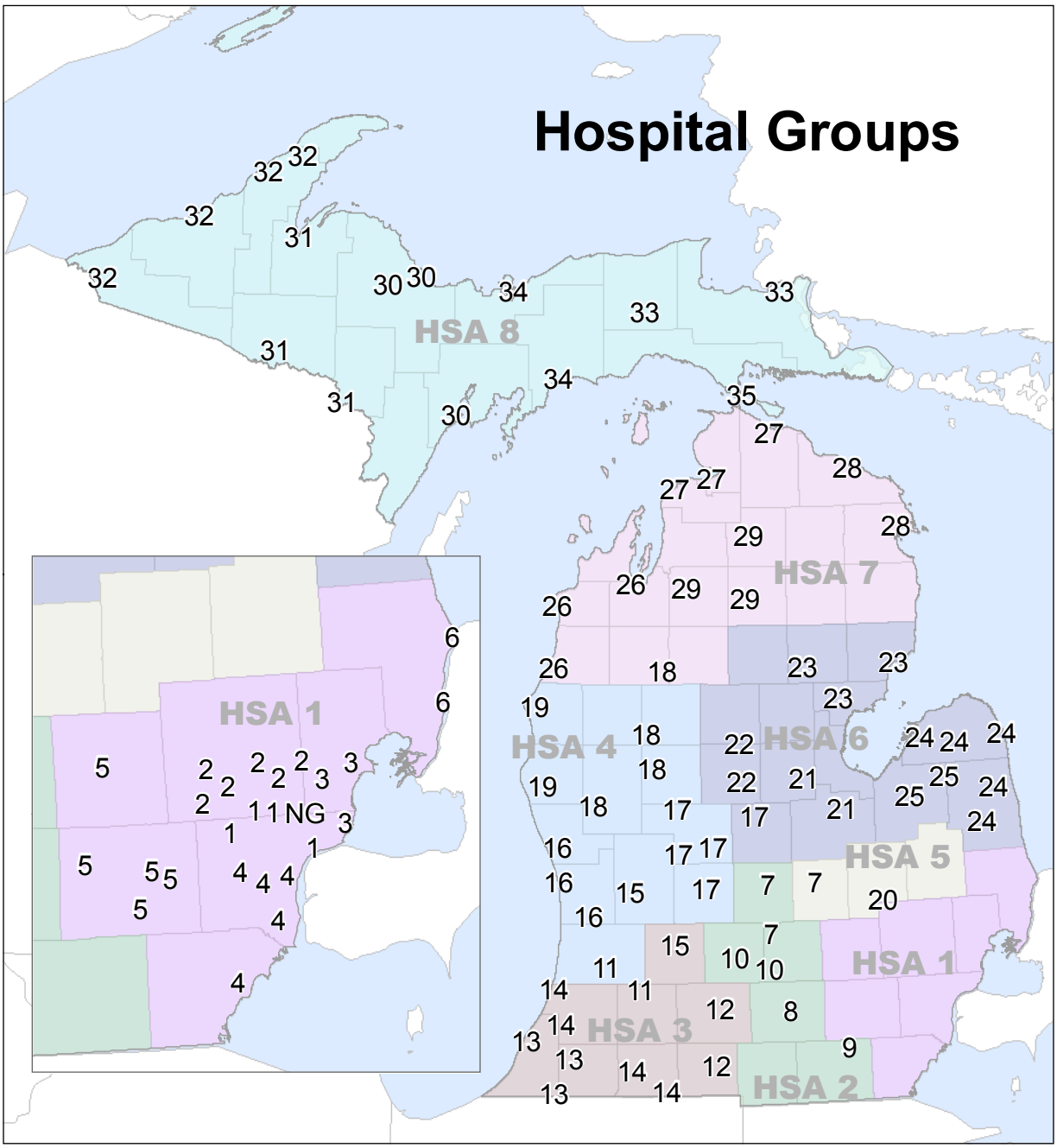

Figure 21: **Hospital Groups created using new clustering methodology.** The data have been thinned for display purposes.

Table 9: **Initial candidate solutions.** SH are the number of single hospital clusters in the overall solution and Max is the maximum number of hospitals in any cluster in the Hospital Group solution. Solutions with less than 29 clusters have Max > 20 and were removed from consideration. From the remaining solutions, the minimum SH value was 1. Therefore, solutions with SH > 1 were removed from consideration. From the remaining 3 solutions (29, 33, 35), the 35 cluster solution was the maximum K and selected as the final Hospital Group solution.

Clusters	***incF***	**SH**	**Max**
3	94.78	0	91
8	25.39	0	58
11	14.58	0	48
16	7.43	0	48
18	7.10	0	45
21	9.49	0	36
26	6.30	1	24
29	6.10	1	17
33	3.73	1	17
35	4.28	1	17
38	4.16	2	17
40	4.34	4	17
42	4.01	7	17
45	4.03	8	16
47	3.61	8	16
50	3.93	10	16
52	3.53	12	16
54	3.88	14	13
57	3.55	17	12
59	3.99	20	12
63	4.09	25	12
65	4.03	28	12

Cont. on next page

Table 9 – *Cont. from previous page*

Clusters	***incF***	**SH**	**Max**
68	4.14	32	12
70	4.27	35	12
73	4.56	38	12
77	4.46	43	9
80	4.06	45	8
82	4.32	47	7
87	4.24	54	7
92	3.98	63	7
96	4.18	70	7
100	5.03	75	7
104	5.43	80	7
107	5.06	83	7
109	3.94	86	7
112	3.61	89	7
114	3.76	91	7
117	2.93	96	7
121	3.06	101	7
124	2.83	103	5
126	2.62	107	5
128	2.46	109	5
130	2.41	112	5
135	2.17	119	5
137	1.95	120	5
141	2.12	126	5
149	2.10	136	3
154	2.38	145	3
156	2.64	149	3

To explore the stability of the Hospital Groups provided by the methodology, we re-

created a 35 Hospital Group solution using data from 2004 to 2006. This allowed us to test the resulting Hospital Groups with data from an independent time period with no overlapping years. Because a small number of hospitals closed and opened during this time frame, after clustering, the hospitals were normalized such that only hospitals open during both time periods were compared. The normalization step was completed post-clustering as to not influence the results of the 2004-2006 Hospital Group solution. Overall, the two 35 Hospital Group solutions were in agreement on 93.37% of hospitals (155 of 166 hospitals). 30 of the 35 Hospital Groups produced using the 2004-2006 data were an exact match (both group size and hospital membership) with their counterparts from the 2007-2009 data.

3.5 Discussion

Following extensive review, the HBSAC recommended that the new clustering methodology for Hospital Groups be adopted into Michigan's Hospital Bed Standards by a unanimous vote. The recommendation was presented to the state's CON Commission, who approved unanimously to move the methodology forward to the public comment stage. After allowing for public comments, the new methodology was again approved by the CON Commission to be adopted into Michigan's Hospital Bed Standards[7].

Using the 2007-2009 utilization data, the new clustering methodology reduced the number of Hospital Groups in Michigan from 64 to 35. During development of the methodology, the HBSAC strongly believed that the number of single hospital Hospital Groups in the state should be decreased. Therefore, an emphasis was placed on this characteristic in the heuristic to select the final number of Hospital Groups. In the 35 Hospital Group solution, only one Hospital Group contained a single hospital (2.86% of the groups). This result was substantially different than the current configuration wherein 50% of the 64 Subareas contain a single hospital.

The overall fit of the original 64 Subarea configuration is slightly better than that of the

[7]The CON Commission is scheduled to meet and approve the methodology in June, 2012.

35 Hospital Group solution ($R^2 = 0.984$ vs. $R^2 = 0.973$). However, this is not entirely surprising, given that model fit is influenced heavily by the number of clusters. Using the F statistic, which incorporates both R^2 and the number of clusters, we found that the 35 Hospital Group solution ($F = 130.21$) outperformed the original configuration ($F = 92.86$). It is important to note that a direct comparison of the 64 Subarea and the 35 Hospital Group solutions can be somewhat misleading given that they were created with very disparate methods *and* do not have a similar number of clusters. Methods and procedures to evaluate clustering methods or algorithms generally compare cluster solutions with the same number of clusters or compare the cluster to solutions to a random clustering of observations. Therefore, a more appropriate statistical test of the methods would require that the number of output clusters be similar. For example, the fit of a K-means + Ward's 64 cluster solution ($R^2 = 0.989$ and $F = 135.71$) is better than the current 64 Subarea configuration. Yet, caution is also warranted in interpreting these results considering the uncertainty surrounding the modification performed on the original output of the Thomas methodology in the 64 Subarea configuration.

Comparing the statistical fit of the 64 Subareas to the 35 Hospital Groups (or a 64 K-means + Ward's cluster solution) does not provide a suitable evaluation of the two methods in light of the final purpose for allocating beds in the Bed Need Methodology. Rather, the small number of clusters produced by the Thomas Methodology, when implemented with recent hospitalization data, speaks more to the overall utility of the Thomas Methodology itself. Therefore, the most basic advantage provided by the new clustering methodology is that it produces an usable and actionable number of Hospital Groups.

While the new methodology was generally lauded by members of the HBSAC and CON Commission, there are potential implications for hospital bed distribution within the state. The reduction of the number of single hospital Hospital Groups assumes a more regional view of community-based need than the previous configuration. While the initial move toward more regional-level planning and regulation units is consistent with other states'

CON programs, the actual consequences for inpatient hospital bed distribution and access in Michigan remain to be seen. We are encouraged, however, by our preliminary tests showing that the 35 Hospital Group configuration did not substantially alter predictions of the state's future bed demand.

Another issue to consider is the use of alternative data for clustering hospitals. Because the focus of this application is to define Hospital Groups for inpatient hospital bed planning, we chose only to include inpatient hospitalization data. However, other measures such as the American Hospital Association's case-mix adjusted discharges may be explored in the future. Adjusted discharges incorporate both inpatient and outpatient hospital visits, possibly offering a more complete characterization of community health care utilization. Additionally, raw inpatient days do not provide insight into the efficacy of the hospitalizations or their overall contribution to public health (Thomas et al., 1983). For our specific application, we do not consider the use of inpatient hospitalization data as a limitation. However, we do acknowledge the limitations of these data and future research would benefit from exploring alternative data sources for clustering hospitals.

While the clustering methodology was designed specifically to create groups of hospitals, the concepts are transferable to the creation of health service areas or small areas. One of the most notable topics in health services research over the past 30 years has been the exploration of small area variation in health care utilization (Wennberg and Gittelsohn, 1973; Wennberg, 2005), spending (Fisher et al., 2003), and outcomes (Welch et al., 2011) in the US. These studies often rely on an aggregation method wherein *small areas* are formed by grouping disaggregated population units into larger regions based on similarity in health services use. The method implemented by Wennberg and colleagues at Dartmouth employs a simple plurality rule, grouping areal units based on a single CI value, not their overall patterns of utilization (Wennberg and Gittelsohn, 1973). In rural communities, this process is generally straightforward considering that much of the population's health care needs are provided by a single facility. Because urban areas often contain a greater number of facilities,

service use by any given community is often distributed similarly among facilities (Thomas et al., 1981), complicating small area creation and/or service area definition. Using our clustering methodology, community utilization patterns can be expressed as the CI values from areal units to hospitals. The areal units could then be clustered into regions of similar hospital use, where the overall utilization patterns and location are considered. However, we note that an additional step would be required to link the clustered areas to specific hospitals or groups of hospitals using this methodology.

3.6 Conclusions

The goal of our new clustering methodology to create Hospital Groups was for it to be as ***objective***, ***replicable***, and ***sustainable*** as possible. Given the politically and economically charged climate surrounding CON regulation in Michigan, a full recasting of the theoretical approach to cluster hospitals was no small undertaking. A number of possible clustering methods were presented to the HBSAC working group and each could be considered "objective" given that they are data-driven. However, we believe that placing our focus on the concepts of hospital similarity and the theoretical underpinnings of the methods, rather than results, allowed for a politically objective overall methodology to emerge. In addition, we implement a heuristic that selects the final number of Hospital Groups based on desirable characteristics of the solutions instead of relying on a predefined number. The use of a heuristic does not completely remove all subjectivity from our methodology; the HBSAC working group members determined which characteristics were acceptable for selecting the final number of clusters. However, by including the decision rules in the methodology, the new clustering methodology provides a level of transparency that was not present in the post-clustering modification step of the previous methodology.

Two distinct interpretations of "replicable" are fulfilled by the clustering methodology. First, by integrating the K-means and Ward's clustering algorithms, we have effectively removed the unconstrained stochastic element associated with random starting locations in

K-means. Each time the methodology is run with the same data, it will produce the same final Hospital Group solution (both the configuration of the Hospital Groups and the number of Hospital Groups). By supplying the source code necessary to implement the methodology, we have provided it in an unambiguous format. Additionally, the methodology is built upon well-known clustering algorithms allowing it be transferable in other statistical packages.

We examined the sustainability of the clustering methodology by creating a 35 Hospital Group solution using hospitalization data from 2004-2006. The high level of agreement in the composition and size of the resulting Hospital Groups suggests that the methodology captures long-term community hospital utilization patterns in Michigan. Therefore, when the clustering methodology is run in the future, Hospital Group configuration will not change dramatically unless community utilization patterns have significantly changed.

We believe that the appropriate levels of consideration were given to the scientific, practical, and political concerns encountered during the developmental process. The new clustering methodology offers substantial improvement over the previous methodology, as it is unambiguously actionable and produces superior results. Furthermore, the methodology is generalizable such that it is suitable for clustering both facilities or areal units within a variety of health care service applications.

Acknowledgements: This research was supported by the Michigan Department of Community Health, Certificate of Need Program. Additionally, the authors would like to thank Dr. Bruce Pigozzi for his insightful suggestions regarding clustering methodologies and evaluation procedures.

See Appendix C for Blue Cross Blue Shield's testimony regarding the proposed clustering methodology.

Do more hospital beds lead to higher hospitalization rates? A spatial examination of Roemer's Law

4.1 Abstract

Background: Roemer's Law, a widely cited principle in health care policy, states that hospital beds that are built tend to be used. This simple but powerful expression has been invoked to justify Certificate of Need regulation of hospital beds in an effort to contain health care costs. Despite its influence, a surprisingly small body of empirical evidence supports its content. Furthermore, known geographic factors influencing health services use and the spatial structure of the relationship between hospital bed availability and hospitalization rates have not been sufficiently explored in past examinations of Roemer's Law. We pose the question, "Accounting for space in health care access and use, is there an observable association between the availability of hospital beds and hospital utilization?" **Methods:** We employ an ecological research design based upon the Anderson behavioral model of health care utilization. This conceptual model is implemented in an explicitly spatial context. The effect of hospital bed availability on the utilization of hospital services is evaluated, accounting for spatial structure and controlling for other known determinants of hospital utilization. The stability of this relationship is explored by testing across numerous geographic scales of analysis. The case study comprises an entire state system of hospitals and population, evaluating over one million inpatient admissions. **Results:** We find compelling evidence that a positive, statistically significant relationship exists between hospital bed availability and inpatient hospitalization rates. Additionally, the observed relationship is invariant with changes in the geographic scale of analysis. **Conclusions:** This study provides evidence for the effects of Roemer's Law, thus suggesting that variations in hospitalization rates have origins in the availability of hospital beds. This relationship is found to be robust across geo-

Submission information: Submitted to PLoS Medicine on May 3, 2012. Submitted to PLoS ONE on June 12, 2012. Authors on manuscript: Paul L. Delamater, Joseph P. Messina, Sue C. Grady, Vince WinklerPrins, and Ashton M. Shortridge.

graphic scales of analysis. These findings suggest continued regulation of hospital bed supply to assist in controlling hospital utilization is justified. **Keywords:** Roemer's Law; hospital utilization; supplier-induced demand; access; spatial accessibility; Certificate of Need

4.2 Introduction

Roemer's Law famously and simply states, *hospital beds that are built tend to be used* (Shain and Roemer, 1959, p.71). Although the authors' original intent behind the statement is debatable, the most common interpretation is that as the supply of hospital beds increases the use of hospital services also increases. Roemer's Law has fostered the belief that excess hospital beds leads to an *over*utilization of hospital services, when the observed demand outpaces the population's actual need for services (Mulley, 2009). Hospital utilization rates rise, therefore, due to higher levels of inpatient admissions which may or may not lead to longer stays, contributing to higher costs. Wennberg (2005) suggests that Roemer's Law may be due to physicians being influenced by a subliminal knowledge regarding the availability of hospital beds.

In the USA, the high costs of inpatient hospitalizations, in conjunction with the generally accepted implications of Roemer's Law, serve as the justification for state-based Certificate Of Need (CON) programs. CON programs are independent entities that are responsible for regulation of the supply of health care services such that the supply meets the population's health care needs without an oversupply or duplication of services. Given that the plurality of overall health care expenditure in the USA is for inpatient hospital care (Kaiser Family Foundation, 2009), hospitalizations, and thus hospitals, are logical candidates for cost control measures. Supply is regulated by CON programs (Ferrier et al., 2010) wherein an unmet demand for services must be demonstrated prior to CON approval of new expenditures for hospital construction or expansion. Currently in the USA, 35 states have some form of CON program with 28 states specifically regulating the supply of acute care hospital beds (National Conference of State Legislatures, 2011).

Roemer's Law defines a positive relationship between the availability of hospital beds and the use of hospital services. Past research has provided support for the effects of Roemer's Law (e.g., Ginsburg and Koretz, 1983; Harris, 1975; Kroneman and Siegers, 2004; Pasley et al., 1995; Shwartz et al., 2011; Wennberg, 2005), while other research has found conflicting (e.g., Alexander et al., 1999; Rohrer, 1990; van Doorslaer and van Vliet, 1989) or inconclusive results (e.g., Clark, 1990). The intertwined relationships among population health, access, use of health care services, and outcomes provide a number of research dilemmas, both theoretically and methodologically. Perhaps, the most difficult dilemma is defining and characterizing the availability of hospital beds. Although counting the number of beds in a hospital is trivial, measuring the overall availability of those beds *to a population* is a much more complex and influenced by distance, demand, and access-related factors. Measures of hospital bed availability such as container-based metrics or simple distance (Joseph and Phillips, 1984; Guagliardo, 2004) ignore the multifaceted nature of access and the spatial and geographic nature of health care service use. Others have noted that the observed effects of Roemer's Law may be due to oversimplified methods used to assign hospital beds to regions (Folland and Stano, 1990). In addition, statistical methods that do not incorporate spatial structure in the relationship between access and utilization are at risk of being misestimated due to the effects of spatial autocorrelation.

As Wennberg and colleagues (1999, p.2) have noted, *in American health care, geography is destiny.* The important role of spatial factors in health care services use have not been been given full consideration when exploring Roemer's Law. Hence, we believe a substantive re-examination is warranted.

So, the critical question remains, "does the availability of hospital beds affect hospital utilization?". Whereas Roemer's natural experiment (Roemer, 1961) was based on a regional study when a single hospital added a substantial number of inpatient beds, we approach this issue by examining an entire hospital system, comprising the hospitals, populations, and transportation infrastructure that connects populations to hospitals. We employ an ecolog-

ical research design that integrates individual behavioral models of health care utilization in an explicitly spatial context. Thus, the research question is reframed to ask, "Accounting for space in health care access and use, is there an observable association between the availability of hospital beds and hospital utilization?".

We characterize both the spatial and aspatial components of access such that their individual and combined contributions can be subsequently identified. Furthermore, by controlling for other determinants of hospital utilization, we isolate the effects of hospital bed availability on the utilization of hospital services, thus allowing us to statistically examine the effects of Roemer's Law on hospitalization rates. In addition, we explore the stability of the relationship between hospital bed availability and hospital utilization by constructing models at varying scales of geographic analysis.

4.3 Materials and Methods

4.3.1 Research design

The Andersen model of health service utilization serves as the underlying theoretical framework in our research: utilization of health services results from a predisposing component, an enabling component, and illness level or "need" (Andersen and Newman, 1973). This framework is appealing because characteristics of both the population and the health care delivery system are integrated into a single model:

$$U = f(n, P, E, N) \tag{10}$$

where n is the number of people, P is the predisposing component, E is the enabling component, and N is need for services. The enabling component in the Anderson model roughly equates to access, but does not provide a detailed characterization. We extend the Andersen model using the theoretical framework offered by Penchansky and Thomas (1981) that defines access as the "fit" between the population in need of services and services offered. In this framework, access results from a combination of five separate dimensions. Khan

(1992) classified the dimensions into spatial components: accessibility (Ac) and availability (Av) and aspatial components: affordability (Af), acceptability (Ap), and accommodation (Am). In addition to the five access components proposed, we add a mobility component (M) to capture differences in the ability to overcome distance (Paez et al., 2010). Portions of the extended access framework cross over through P and E from the Andersen model. Therefore, we define:

$$P = f(Ag, G) \tag{11}$$

$$A = f(Ac, Av, Af, Ap, Am, M) \tag{12}$$

$$N = f(H, \varepsilon_{\mathrm{h}}) \tag{13}$$

where

$$H = f(In, Ed, Et). \tag{14}$$

Ag and G are the age and gender structure of the population, A is access, and H is the health status of the population. It is important to highlight the distinction between need (N) and demand (U) for services in this framework. Although a certain amount of U is predictable based on known demographic characteristics of the population, N arises from the general health status of the population and, for hospitalizations, includes a stochastic element triggered by unpredictable instances of ill-health (Feldstein, 1966). Measuring N is problematic in health services research given that patients and health professionals often evaluate the need for services differently (Donabedian, 1972), resulting in cases of both unmet need and unnecessary utilization. Therefore, in Eq. 13, H is a measure of the health of the population and ε_{h} is a random variable representing occurrence of ill-health. Oleske (2009) report six approaches to measuring health care need, yet all are essentially proxies for estimation of H. Thus, we employ socio-economic status (SES) measures, income (In), education (Ed), and ethnicity/race (Et), as proxy measures of population health (see Young (2005), pp.153-154 for a discussion of inclusion of ethnicity/race in health models). Although there may be questions regarding causality between SES and health, SES has shown to be

significantly correlated with both morbidity and self-assessed health status (Norris et al., 2003) in the US and internationally (Young, 2005).

Our theoretical model is supplemented by accounting for variations in hospital utilization among populations that may not be fully captured in Eq. 14. We use the number of Low Variation (LV) hospitalizations (U_{LV}) to help capture this variability. LV hospitalizations are those with little clinical-based doubt regarding the need for hospitalization (Wennberg, 2005); therefore, variations in LV hospitalization rates can be considered as arising from the actual health care needs of the population. We also consider hospitalizations for Ambulatory Care Sensitive (ACS) conditions (U_{ACS}) in our theoretical model. This class of hospitalizations (also known as preventable hospitalizations) are those where inpatient hospitalization may be avoided if primary care is available (Bindman et al., 1995) and accessible (Ricketts et al., 2001). Hence, we control for variation in hospital utilization due to inadequate access to primary care. In combination, we label these variables as the *case mix* of a population's hospital usage, offering proxy measures of health variation not captured in P or H.

Given dissimilar population sizes among areal units or zones, we normalize all variables by population size producing rate-based (e.g., beds / person) or proportional (e.g., % of population with insurance) measures where applicable. Therefore, we remove n from the theoretical model when moving to an applied model. In addition, due to the differences in age structure among populations, we age-standardize the hospitalization rates. Hence, we remove Age from the theoretical model and specify a full model of hospital utilization,

$$U_{std} = f(G, Ac, Av, Af, M, In, Ed, Et, LV_{std}, ACS_{std}, \varepsilon_{\mathrm{h}}) \tag{15}$$

which allows for examination of the relationship between U and hospital bed availability while controlling for differences in demographic characteristics and health status among

populations [8].

The proposed framework is implemented in an explicitly spatial context, acknowledging the role of geography in interactions among populations and hospitals. First, because all populations do not have equivalent geographic access to the same hospital services, we incorporate the spatial character of hospital utilization by limiting our analysis to only those hospitalizations where services were demanded locally. Second, we overcome container-based measures of hospital bed availability by calculating a metric that captures the interaction between distance, hospital bed supply, and demand. Third, we employ spatial regression models which incorporate the spatial structure of the proposed framework, thus counteracting the problems associated with spatial autocorrelation.

The ecological study design requires that we address issues stemming from the Modifiable Areal Unit Problem (MAUP, Openshaw, 1984; Fotheringham and Wong, 1991). The MAUP arises when correlation or regression-based analysis is influenced by the particular resolution or zoning scheme of the data. In extreme cases, regression coefficients may flip from positive to negative or statistical significance may be greatly altered when an alternate scale of analysis or zoning methodology is implemented (Chi and Zhu, 2008; Mobley et al., 2008; Wright and Ricketts III, 2010). Therefore, we explore the stability of Roemer's law by evaluating the relationship between hospital bed availability and hospital utilization over varying levels of data aggregation.

4.3.2 Case study

Our case study explores the relationship between hospital bed availability and utilization for the state of Michigan. As of 2010, Michigan had a population of 9,883,640 residents

[8]The other access-related variables, Ap and Am, have been removed from the theoretical model for the following reasons: 1) Acceptability was defined by Penchansky and Thomas (1981) as capturing the religious or racial/ethnic fit between a person and the health care facilities, thus is very likely outdated. 2) Accommodation attempts to account for waiting times, hours of operation, telephone appointment systems, and other non-supply related factors of the health care facility. These factors should be quite constant among modern hospitals.

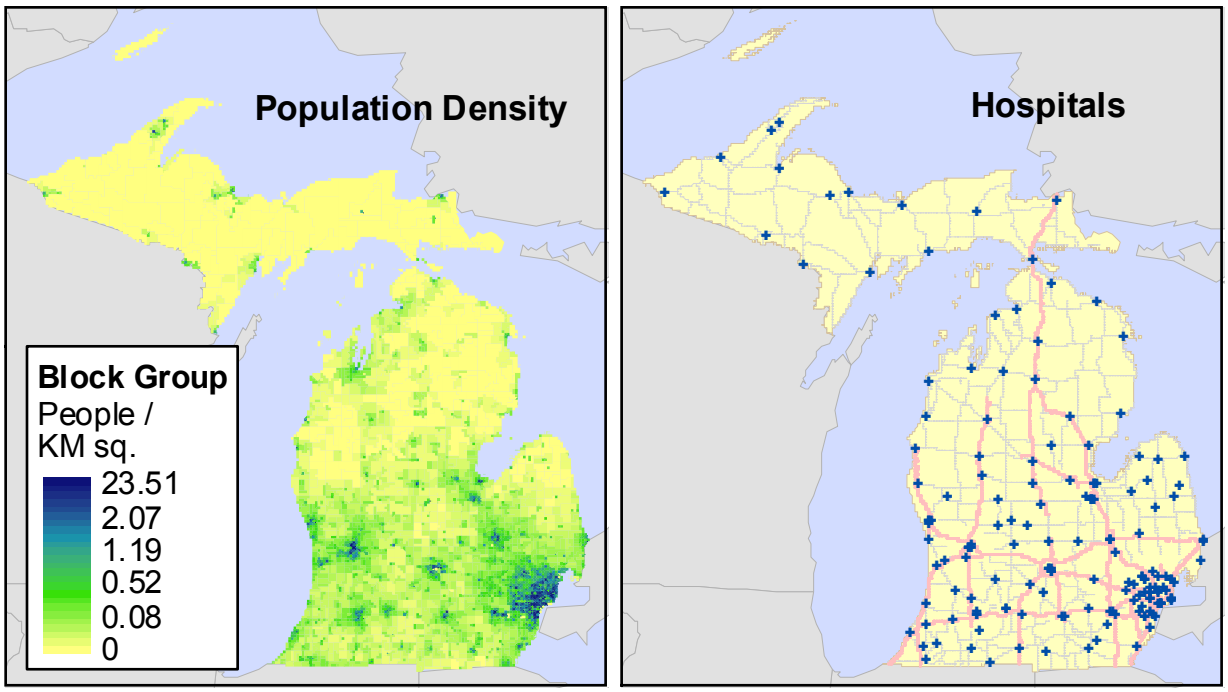

Figure 22: **Population distribution and hospital locations in Michigan.**

served by 169 acute care hospitals with 26,180 total licensed inpatient beds. In 2010, there were 1,127,576 hospital admissions of Michigan residents to Michigan hospitals and a total of 5,313,149 days spent in hospitals, resulting in an overall patient day usage rate of 0.537 patient days per person. For every 1000 people, there were 9.51 hospital admissions per month, which is slightly higher than the national averages of 8/1000 found by Green et al. (2001) and 9/1000 as reported by White et al. (1961).

Michigan employs a CON program to regulate the availability of inpatient hospital beds (Messina et al., 2006). To assess the needs of the population, a bed need methodology is implemented to predict the future demand for hospital beds, which is compared with current levels of supply (Langley et al., 2010). Michigan serves as a satisfactory study area due to the large number of hospitalizations and population, the state's relatively stable system of acute care hospitals, and a diverse collection of rural and urban areas with varying population densities, health care services distributions, and demographic characteristics (see Figure 22)

by which to examine Roemer's Law.

4.3.3 Population data

The Zip Code boundary data used for Michigan were acquired from the ESRI ArcGIS^{TM} v10 data CD[9]. The 2010 population and demographic attribute data were acquired from the US Census Bureau (http://2010.census.gov). Block-level data for age, gender, race/ethnicity were aggregated to their respective Zip Code boundaries. The age-specific data were aggregated into 5 year categories for 0 to 84 years of age with an additional category for 85 and older. Income, education, and mobility attributes were culled from the 2006-2010 American Community Survey 5-year estimates (http://www.census.gov/acs/www/). These data are available at the block group level and were aggregated to the Zip Code boundaries. A small number of block groups were not reported (48 blocks with a population of 52,593, roughly 0.5% of the total state population). Values for the missing block group data were estimated using a weighted average of first-order (queen's case) neighboring values (Bivand et al., 2008). First-order neighbors are defined as areas sharing a common boundary. 2009 Small Area Health Insurance Estimates (SAHIE, http://www.census.gov/did/www/sahie/) data were used for health insurance rates. For this analysis, we only considered the health insurance status of people under 65 years of age. Because SAHIE data are only available at the county level, Zip Code-level data were estimated using the age-specific rates found in the SAHIE data and age-specific population distribution of the Zip Codes.

4.3.4 Travel time

Travel time data were derived using a custom-built network model. The most recently available roads database (2009 version 10a, http://www.michigan.gov/cgi) was downloaded from the Michigan Center for Geographic Information and used to construct the network travel model. Travel speeds for each road were assigned using the road attribute data and a

[9]Prior to the analysis, the 908 unique Zip Codes were aggregated into 895 Zip Codes due to mismatches between the spatial data and the hospital utilization data

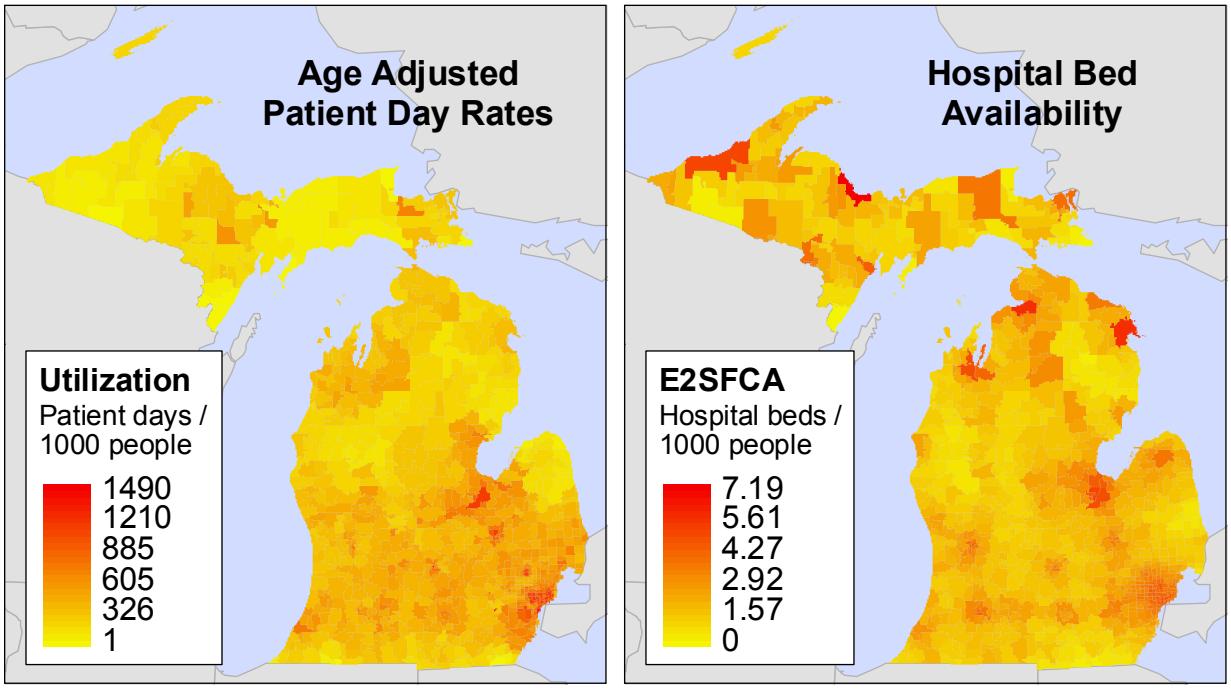

Figure 23: **Age adjusted hospital utilization (U_{STD}) and bed distribution (Av, E2SFCA) in Michigan.**

hierarchical speed limit classification system (Delamater et al., 2012).

4.3.5 Ethics statement

The Michigan Hospital Inpatient Database (MIDB) consists of routinely collected information on patient's hospital discharge for billing purposes. The patients provided consent for their information to be stored in the hospital database but that information is protected under HIPPA rules. All identifiable patient information was removed from the MIDB prior to use in this research. The Michigan State University Internal Review Board determined the use of this de-identifiable data exempt (IRB #07-362).

4.3.6 Hospital utilization

Inpatient hospitalization data were gathered from the 2010 Michigan Inpatient Database (MIDB), a comprehensive record of the state's inpatient hospitalizations. For each non-psychiatric hospital admission excluding normal newborns, the age, principal discharge di-

agnosis (ICD-9-CM), length of stay in days (LOS), Zip Code of residence, and admitting hospital were collected. Travel time was attached to each discharge, calculated from the population-weighted centroid of the Zip Code of residence and the location of the admitting hospital (Berke and Shi, 2009). Hospitalizations occuring more than 60 minutes from the patient's residence were removed from the analysis. This geographic constraint accounts for two scenarios in which hospitalizations would not be affected by the hospital bed availability of nearby hospitals, thus confounding the analysis. First, it removes hospitalizations where patients traveled a long distance due to the availability of hospital-specific *services*, not hospital bed availability. Second, the constraint removes hospitalizations that occured when the patient was a significant distance away from their residence (e.g., while on vacation) and not affected by local hospital bed availability. While the 60 minute cutoff value is arbitrary, it is based on previous research exploring spatial accessibility in regions having highly rural populations (McGrail and Humphreys, 2009). Of the total patient days in 2010, 93.2% were served by a hospital within 60 minutes of the patient's residence.

The LV hospitalization data used in this analysis included discharges for Myocardial Infarction, Ischemic Stroke, and Hip Fracture (Fisher et al., 1994)[10]. ICD-9-CM codes for the ACS hospitalizations were culled from the Dartmouth Atlas of Healthcare (Wennberg et al., 1999). In 2010, there were 659,997 patient days for ACS conditions and 229,834 for LV conditions.

Because the age distribution of populations is not homogeneous among areal units, the hospitalization data were standardized via the direct method of standardization (Meade and Emch, 2010). Michigan's 2010 population was used as the standard population. Age standardization was accomplished in a two step process. Some of the state's Zip Codes contain small populations in each age-specific category and thus violate the 20/50 rule for calculating health-related incidence rates (Klein et al., 2002). In addition, as previously

[10]ICD-9-CM codes: Myocardial Infarction (410), Ischemic Stroke (431, 434-438), and Hip Fracture (808)

mentioned, inpatient hospitalizations are also subject to random fluctuations of ill-health events. Therefore, the first step in the age standardization process was to calculate each areal unit's age-specific patient day usage rates using an local Empirical Bayes (EB) smoothing method (Marshall, 1991). This smoothing method assumes that the patient day count data follow a Poisson distribution, while also borrowing strength from the patient days and populations of neighboring regions (Bivand et al., 2008; Odoi et al., 2003). The neighborhood structure for the EB smoothing process was defined via first-order neighbors. Once the age-specific rates were smoothed, each areal unit's age-specific patient day rates were multiplied by the age-specific distribution of Michigan's population. To calculate the overall patient day rate, the age-specific data were summed and divided by the total state population (see Figure 23).

Following the age-standardization process, the hospital utilization rate data were converted to a Standardized Rate Difference (SRD) by subtracting the average utilization rate of the entire state from the age-adjusted utilization rate of each observation. This simple scalar transformation did not affect the magnitude of the data; however, it did allow for easier interpretation of the results such that observations with rates greater than 0 are higher than the state average and those less than 0 are lower.

4.3.7 Spatial accessibility

Recently, a set of gravity-based GIS measures of spatial accessibility have been proposed that allow both availability and accessibility to be integrated by including measures of supply, demand, and distance simultaneously (Ngui and Apparicio, 2011). The general form of the gravity-based models can be represented as

$$A_i^G = \sum_{j=1}^{n} \frac{S_j f(d_{ij})}{\sum_{k=1}^{m} P_k f(d_{jk})} \tag{16}$$

where A_i^G is the spatial accessibility for population zone i, S_j is the attractiveness of a facility at location j, $f(d_{ij})$ is an impedance (decay) function based on the distance (d) from

zone i to location j, $f(d_{ik})$ is an impedance function based on the distance from location j to zone k, and P_k is the population in zone k. The total number of zones and facilities are n and m, respectively.

We employed the enhanced two-step floating catchment area (E2SFCA), a gravity-based metric proposed by Luo and Qi (2009), to measure the availability of hospital beds. One drawback in using gravity-based measures is that the unit (A^G) is often difficult to interpret. The E2SFCA overcomes this limitation by providing availability values in easy to understand, container like units (hospital beds per person). The E2SFCA improves on its predecessor, the two-step floating catchment area (2SFCA, Radke and Mu, 2000; Luo and Wang, 2003), by replacing a dichotomous distance characterization with distance or service area "bands" radiating from each service location. The FCA measures overcome the theoretical limitations of container-based measures by allowing the catchment areas for supply and demand locations to "float" based on travel distance or travel time in lieu of adherence to administrative boundaries. To accomplish this, the potential demand is calculated for each facility, which is in turn used to calculate the supply available at each areal unit.

The E2SFCA requires weight values to allocate demand and supply to the distance bands using the theory of distance decay. The three functions most oftenly used to model distance decay in gravity-based measures are the Inverse power, Exponential, and Gaussian (Kwan and Hong, 1998). Gravity-based models are generally limited by the arbitrary selection of a distance decay function and the associated β parameter that describes the magnitude of decay (Schuurman et al., 2010). However, because the actual travel patterns of Michigan residents are known, our study is not limited by this arbitrary selection process. Using the actual utilization patterns of state residents, the distance decay function and associated parameter values were empirically estimated using a non-linear regression model.

Initial investigations showed that the oft-used distance decay functions did not adequately fit the utilization patterns. However, the downward log-logistic decay function (de Vries et al.,

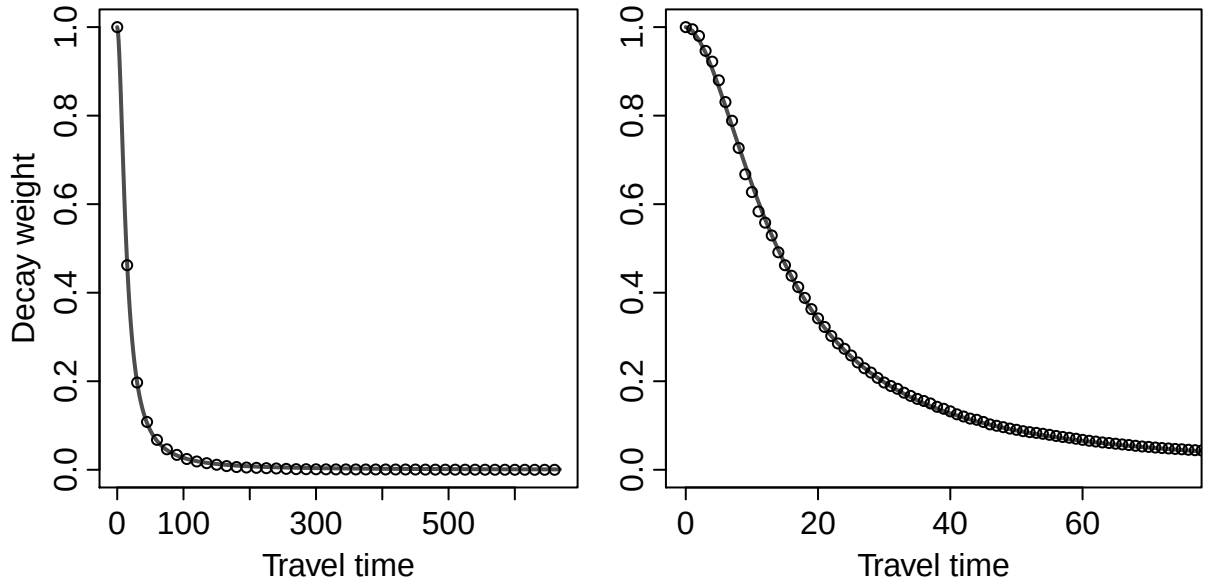

Figure 24: **Distance decay of hospital utilization in Michigan.** Left) over the entire range of the inpatient travel data. Right) a subset of the travel data. The circles are the data points (thinned for display purposes) and the line is the downward log-logistic function fit to the data.

2009),

$$W = \frac{\gamma}{1 + (\frac{d}{\beta_0})^{\beta_1}} \tag{17}$$

provided a superior characterization of the observed decay pattern and thus was employed to estimate the weights (W) for the E2SFCA calculation (See Figure 24). In Eq. 17, the γ parameter controls W at $d = 0$. Therefore, because W must equal 1 at $d = 0$, we were able to simplify the parameter estimation process by setting γ equal to 1. We estimated the two remaining decay parameters (β_0 and β_1) using the non-linear least squares estimator available in **R** (R Development Core Team, 2011). The resulting parameter values were β_0 = 13.89 and β_1 = 1.82. Both parameters were statistically significant ($p < 2$ x 10^{-16}) and the model produced a low residual standard error (RSE = 0.003) with an excellent curve fit (see Figure 24).

In the first step in the E2SFCA, the supply is calculated at each facility. Using the network dataset, travel time rings were created for each hospital at 5 minute intervals to a

maximum of 45 minutes and a final ring was created from 45 to 60 minutes to incorporate travel in the rural regions in the state (McGrail and Humphreys, 2009). A W value was assigned to each travel ring using the downward log-likelihood function of each travel time value comprising the ring (e.g., the 5-10 minute ring W value is the mean of the W values for 5-10 minutes). The population data were spatially joined to the travel time rings. The supply (R_j, beds / person) is calculated at each facility as follows:

$$R_j = \frac{S_j}{\sum_{k \in [D_r < 60]} P_k W_r} \tag{18}$$

where S_j is the number of licensed hospital beds at hospital j, P_k is the set of population of units falling within the set of travel time rings (D_r), and W_r is the set of associated weight values for the travel time rings. Census block centroid points were used in this step as they offered the most accurate representation of population location.

The second step of the E2SFCA calculates the availability of hospital beds (Av) as moderated by distance as follows:

$$Av_i = \sum_{k \in [d_{i,j} < 60]} R_k W_k \tag{19}$$

where Av_i is the availability of hospital beds at population unit i, R_k is the set of hospitals within 60 minutes of population unit i, and W_k is the set of weights based on the travel time from unit i to hospital j for all hospitals in k using Eq. 17. We completed this step using the travel time from the population weighted Zip Code centroids to the hospitals, thus calculating the availability of hospital beds at the Zip Code level (see Figure 23).

4.3.8 Clustering methodology

Much of the available literature regarding data aggregation in health services research pertains to the creation of *small-areas* for investigating health disparities among regions (e.g., Wennberg and Gittelsohn, 1973). Generally speaking, these methods use geodemographic

characteristics of the initial areas to create clusters of homogeneous, contiguous regions (Rey et al., 2011). Although a number of methods have been proposed for creating small-areas, these were deemed inappropriate for our study. Specifically, we believe that implementing a method that clusters the areal units by the same attributes that were being used to explore Roemer's Law would essentially be optimizing the aggregation process to achieve a stronger statistical outcome (Openshaw, 1984). Hence, the level of objectivity in our test of the MAUP would be diminished (Swift et al., 2008).

Given this problem, we implemented a clustering methodology that incorporates hospital utilization patterns and geographic location, identifying geographically promixal areal units whose populations use a similar set of hospitals (Delamater et al., Under review). The resulting clusters are based on similarities in hospital use; however, they are not explicitly optimized based on the same geodemographic attributes used to construct the regression models. Essentially, the clustering methodology is based on principles garnered from small-area studies, but does not produce the statistical bias likely present when using the same set of attributes for the purpose of grouping the data *and* constructing the regression models.

The initial observation units (Zip Codes) were grouped into clusters using the K-means clustering algorithm with rational starting locations provided by Ward's Hierarchical clustering (Milligan, 1980). We clustered the original Zip Code data based on their hospital utilization patterns and geographic location simultaneously. The utilization pattern data were an n x m matrix containing the proportion of each Zip Code's total inpatient hospital days (1:n) spent at each hospital (1:m), otherwise known as the Commitment Index (CI, Griffith, 1972). The location of each observation is defined by the travel time from each Zip Code (population weighted centroid) to each hospital, thus comprising another n x m matrix. Representing location as a set of travel distances, rather than coordinates from a traditional planar coordinate system (e.g., latitude and longitude), allows for factors influencing the true separation among places (i.e., road infrastructure, travel speeds, or the physical landscape) to be more accurately characterized (Jones et al., 2010). The travel time data were rescaled

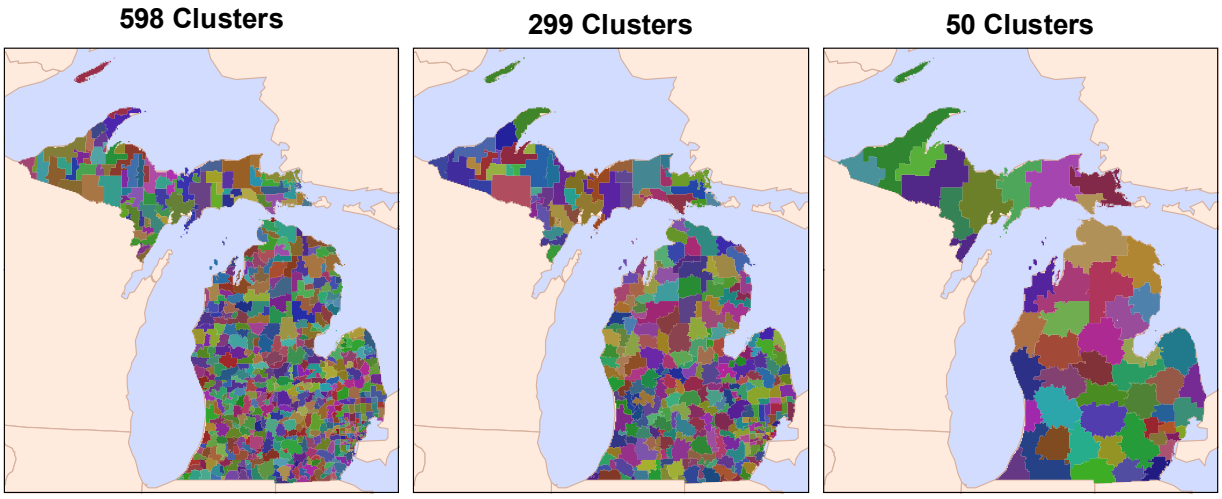

Figure 25: **Zip Code clusters.**

to match that of the CI data (0-1) by dividing by the maximum travel time between any Zip Code and hospital pair. The two n x m matrices were appended to create the final data matrix input to the clustering methodology.

The clustering methodology was run iteratively such that it provided a cluster solution for the set of all possible clusters from 2 to 894 (the set, S). We subset the resulting set S by implementing a selection method based on the incremental F score ($incF$) of each cluster solution (Delamater et al., Under review; Gujarati, 1988). $IncF$ measures only the amount of "fit" gained from allowing an additional cluster within the solution, while also penalizing for adding this additional cluster. Therefore, local maxima in $incF$ scores represent cluster solutions that provide an substantial improvement in the fit when compared with its immediate neighbors. From the initial set S, 276 cluster solutions had local maxima in $incF$ scores, thus they were selected as the levels of aggregation for the regression analysis (see Appendix Figure D.1 and Table D.1). Figure 25 provides three example maps from the final set of cluster solutions. The attribute data for each Zip Code were aggregated based on cluster membership. In addition, we added the non-clustered data (with the 895 Zip Code observations) for a final set of 277 levels of aggregation.

4.3.9 Methods to remove multicollinearity

Considering that the ultimate goal of the analysis was inference on the coefficient values from a regression analysis, multicollinearity in the independent variable set would invalidate the observed coefficient values. Substantial correlation (Pearson's Correlation Coefficient, $r > 0.5$) was observed among the independent variables. We addressed the multicollinearity using a suite of methods as described in the following sections.

4.3.9.1 Principal components analysis We performed a Principal Components Analysis (PCA) on functional "sets" of variables: income/education, ethnicity/race, transportation, mobility, and case mix. By producing uncorrelated component variables, PCA reduces the number of independent variables without a large reduction in the explanatory power of the independent variable set (Jolliffe, 2002). For example, at most scales of data aggregation, the seven variables within the income/education variable set yielded only a single component. Rather than attempting to identify which of the seven variables would be included in the regression analysis, we were able to include a single income/education component that sufficiently described the entire suite of variables (Graham, 2003; Vyas and Kumaranayake, 2006). Because the data were not standardized, we used the correlation matrix for the PCA (Jolliffe, 2002). We also employed a varimax rotation of the results to assist in interpretation of the component structure (Luginaah et al., 2001).

General methods to determine the number of components to extract include manual interpretation of the results or "rules of thumb" (Rogerson, 2006), thus were not applicable for our study given the large number of PCA runs that were necessary to complete the multi-scale analysis. Therefore, we implemented a heuristic that allowed for automation of the process to select the number of components extracted. We added a randomly generated variable to each of the variable sets included in the PCA analysis and generated components. Because PCA provides the loadings on each component for each input variable, the component most heavily influenced by randomness was identified. The PCA was then reim-

plemented without the random variable, extracting only those components describing more variation in the data than randomness.

The functional sets of variables, the input data, and the interpreted output of the PCA are as follows (see Appendix Table D.2 for detailed information including the number of components extracted and the amount of variation captured by the extracted components for each functional variable set at each level of data aggregation):

- **Income/education**
  - *Input variables*
    1. Median household income
    2. Median earnings (16+)
    3. % less than high school education (25+)
    4. % with high school eduction (25+)
    5. % with associates degree (25+)
    6. % with bachelors degree (25+)
    7. % with graduate degree (25+)
  - *Components*
    1. Income and education (SES): High scores reflect populations with higher education, income, and earnings[11]
- **Ethnicity/race**
  - *Input variables*
    1. % White
    2. % African American

[11]In 19 of the 277 levels of aggregation, 2 components were identified: one with high scores on education and another with high scores on income and earnings. The impacts of this split are noted in the Results section.

3. % Hispanic
4. % Asian
5. % American Indian or Alaskan Native (AIAN)
6. % Hawaiian or Pacific Islander (HWPI)

– *Components*

1. Race (BLACK): High scores reflect populations with higher proportions of African Americans and lower proportions of Whites

2–5. Minority population components: High scores reflect observations with higher proportions of Hispanic (HISP), Asian (ASIAN), AIAN, and HWPI populations[12]

- **Means of Transportation to Work (Transportation)**

  – *Input variables*

  1. % Automobile (16+)
  2. % Car pool (16+)
  3. % Public transportation (16+)
  4. % Motorcycle (16+)
  5. % Walk, Bicycle, other (16+)

  – *Components*

  1. Transportation (TRAN1): High scores reflect populations that are less reliant on automobiles as the means for their journey to work

[12]The number of components for ethnicity/race were highly variable across the levels of aggregation. The breakdown was as follows: 1 component (32), 2 components (127), 3 components (64), 4 components (52), 5 components (2). The component interpretations are noted in the Results section.

2. Shared transportation (TRAN2): High scores reflect populations with a larger number of people using car pools for their journey to work[13]

- **Average Travel Time to Work (Mobility)**
  - *Input variables*
    1. % 0-9 minutes (16+)
    2. % 10-19 minutes (16+)
    3. % 20-29 minutes (16+)
    4. % 30-39 minutes (16+)
    5. % 40-59 minutes (16+)
    6. % 60-89 minutes (16+)
  - *Components*
    1. High mobility (MOB1): High scores reflect populations that have a higher proportion of long distance (greater than 40 minute) commuters
    2. Medium mobility (MOB2): High scores reflect population that have a higher proportion of medium distance (20-40 minute) commuters and a lower proportion of short distance (less than 10 minutes) commuters
- **Hospitalizations (Case Mix)**
  - *Input variables*
    1. Age-adjusted rate of LV hospitalizations
    2. Age-adjusted rate of ACS hospitalizations

[13]In 37 of the 277 levels of aggregation, only a single component was identified: one with high scores on non-automobile means of transportation. The component TRAN2 was not included in the final regression analysis as we did not believe that a sufficient theoretical relationship existed between populations with a higher proportion of carpoolers and hospital utilization.

– *Components*

1. Case mix (CASE): High scores reflect populations that have higher rates of both LV and ACS hospitalizations

4.3.9.2 Bivariate regressions Because we were interested in the individual impacts of Av and Af on hospital utilization, these variables (E2SFCA and INS) were held out of the PCA analysis. However, we found that E2SFCA was moderately correlated with the African American population component (BLACK) and INS was moderately correlated with the SES component. In addition, the case mix component (CASE) was also correlated with the African American population component (BLACK). Although the moderate correlation would not invalidate the regression results, we wanted to identify the isolated effects of these variables. Therefore, we adopted the strategy of regressing the variable of interest on its associated correlated variable and using the residuals for further analysis (Graham, 2003). In this, the residuals function as the "unexplained" portion of the variable of interest, allowing *both* variables to be included in the final model. For example, the variable E2SFCA becomes the availability of hospital beds *not* associated with BLACK and is thus recast as E2SFCA-resid. This process was completed independently at all levels of aggregation when r was greater than 0.4. The F scores of the overall model and coefficients were tested to ensure the linear models provided significant (p value < 0.05) results.

4.3.9.3 Test variance inflation factor We calculated the variance inflation factor (VIF) for the set of independent variables (see Table 10), removing those with a VIF > 2 (Graham, 2003). The variables were removed in a reverse step-wise fashion starting with those considered the least established predictors of hospital utilization toward the most (from bottom to top in Table 10). For example, if TRAN1 and SES both had a VIF > 2, then TRAN1 would be removed first in the stepwise process. As the level of aggregation increased and the number of observations became smaller, correlation among the independent variables increased substantially. As a result, we did not perform any subsequent analysis at scales of

Table 10: **Attribute variable set.** TM is the variable label in the modified Andersen model from Eqs. 11-13 and TM2 is the label from the full model specified in Eq. 15.

TM1	TM2	Abbr.	Name	Description
Dependent variable:				
U	U_{std}	SRD	St. Rate Difference	Difference between the age standardized hospitalization rate and the state's age standardized rate
Independent variables:				
N	ACS_{std}, LV_{std}	CASE-resid	Case mix	ACS and LV component not explained by BLACK
A	Ac, Av	E2SFCA-resid	Hospital Bed Availability	E2SFCA not explained by BLACK
A	Af	INS-resid	Health Insurance	INS not explained by SES
N	In, Ed	SES	Income/education	High income and education component
P	G	FEMALE	Gender	Female population
N	Et	BLACK	Ethnicity/race	African American component
N	Et	HISP	Ethnicity/race	Hispanic component
N	Et	ASIAN	Ethnicity/race	Asian component
N	Et	AIAN	Ethnicity/race	American Indian or Alaskan Native component
N	Et	HWPI	Ethnicity/race	Hawaiian or Pacific Islander component
A	M	TRAN1	Transportation	Non-automobile reliant component
A	M	MOB1	High mobility	Long commutes to work component
A	M	MOB2	Medium mobility	Medium commutes to work component

aggregation with fewer than 37 clusters/observations.

4.3.10 Regression models

As noted earlier, previous studies of the effects of Roemer's Law have not incorporated spatial structure. The main implication of this particular model misspecification is that regression coefficients may have contained artificially low standard errors, leading to the rejection of the null hypothesis when it should have been accepted. Initial tests of non-spatial linear models showed high spatial autocorrelation in the residuals with first-order neighboring values (see Appendix Figure D.2). To account for this phenomena, we used two sets of spatial error models (Anselin, 1988), Simultaneous and Conditional Autoregressive Regression models (SAR and CAR, respectively). Both models use the general form,

$$Y = \beta X + \mu \tag{20}$$

where

$$\mu = \lambda W \mu + \epsilon. \tag{21}$$

In the spatial error model, Y is a vector of SRD observations; X and B are matrices of independent variables and coefficients, respectively; μ is a vector of autocorrelated residuals; λ is the autoregressive coefficient; W is a neighborhood weight matrix; and ϵ is a vector of non-autocorrelated residuals.

SAR and CAR models differ in their treatment of the spatial pattern in the dependent variable (Anselin, 2003; Chi and Zhu, 2008). In the SAR model, the spatial pattern is explained only by the independent variables, simultaneously over all observations. The CAR model uses the independent variables to explain the spatial pattern of the dependent variable, but also conditions the value of the dependent variable on its neighboring values (Anselin, 2003). For all regression models, we defined W as first-order neighbors. No prior information in our data suggested whether the SAR or the CAR model were more appropriate for this analysis. Additionally, we were unable to locate past research that provided compelling

justification for the use of one over the other.

A Levene test confirmed heteroscedasticity in the models' residuals due to differing population sizes among areal units (Rogerson, 2006, see Appendix Figure D.4). Therefore, we implemented weighted SAR and CAR models (Bivand et al., 2008; Sparks and Sparks, 2010) using the inverse of the square root of the population size as the weights. This specification led to a substantial alleviation of the heteroscedasticity in the residuals (see Appendix Figure D.5).

We constructed the SAR and CAR regression models at each level of data aggregation produced by the clustering methodology. An automated stepwise-like process was employed to remove independent variables that were insignificant predictors of hospital utilization rate. The initial regression model was constructed and the independent variables were tested for significance (p value < 0.05). If all variables were significant, the process terminated. If any were insignificant, the variable having the highest p value in the model was removed and a new model was constructed. This process continued iteratively until only statistically significant independent variables remained in the final model.

4.4 Results

In total, the SAR and CAR models were constructed at 268 levels of aggregation. In 12 and 31 models for the weighted SAR and CAR models, the spatial parameter (λ) was insignificant and the model considered invalid. The overall coefficient values of the independent variables were very similar among the SAR and CAR models over all levels of aggregation; however, the results of the CAR model contained latent spatial autocorrelation in the residuals at higher levels of aggregation (see Appendix Figure D.3). Considering these findings, we believe the CAR model was misspecified at these scales of analysis and report only the results of the SAR model. Selected standardized coefficient values for the SAR model are found in Figure 26 and Table 11 contains an overview of all coefficient values.

In general, the magnitude of the statistical relationship among the independent variables

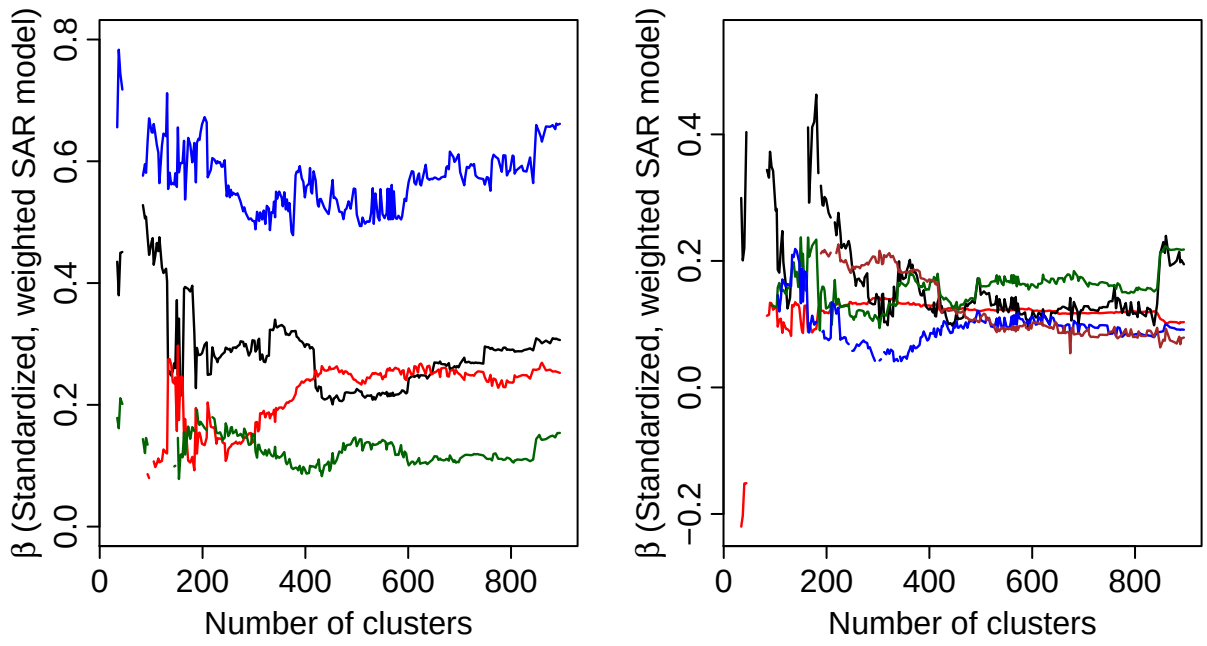

Figure 26: **Standardized coefficients for weighted SAR models.** LEFT: E2SFCA-resid (red), CASE-resid (black), INS-resid (green), BLACK (blue), RIGHT: SES (black), TRAN1 (brown), MOB1 (green), MOB2 (blue), λ (red). All coefficients are significant at a p value < 0.05.

Table 11: **Coefficient statistics.** Total is the number of times the variable is present; Model is the number of times that the variable was included in the initial model ($VIF < 2$); positive is the number of time the variable's coefficient was significant (p value < 0.05) and positive in the final model; negative is the number of time the variable's coefficient was significant (p value < 0.05) and negative in the final model; and insig is the number of times the variable was insignificant and removed from the model.

	weighted SAR model				
Variable	**Total**	**Model**	**positive**	**negative**	**insig**
CASE-resid	268	268	268	0	0
E2SFCA-resid	268	268	254	0	14
INS-resid	268	268	252	0	16
FEMALE	268	254	19	0	235
SES1	268	256	256	0	0
SES2	17	17	0	13	4
BLACK	268	268	268	0	0
ASIAN	66	63	44	4	15
AIAN	99	98	0	95	3
HISP	106	103	102	0	1
HWPI	137	137	41	34	62
TRAN1	268	248	219	0	29
TRAN2	238	0	0	0	0
MOB1	268	260	254	0	6
MOB2	268	252	237	0	15
λ	268	268	252	4	12

and hospital utilization was quite stable across levels of aggregation. In particular, hospital bed availability (E2SFCA), LV and ACS hospitalization rates (CASE), health insurance coverage (INS), proportion of African Americans (BLACK), high income and education (SES), and higher mobility (MOB1 and MOB2) had consistent, positive relationships with hospital utilization rates across levels of aggregation.

4.5 Discussion

Although Roemer initially seemed somewhat surprised that his statement had been bestowed the status of a *law* (Roemer, 1961), our findings provide compelling evidence to support this claim. We found that a positive, significant relationship exists between hospital bed availability and hospital utilization rates while controlling for numerous other determinants of hospital utilization. Additionally, this relationship was consistent across levels of data aggregation providing support that the origin of the observed effect is not a product of the scale of analysis.

In previous studies, Alexander et al. (1999) and Clark (1990) found that hospital beds per capita was not a significant predictor of hospital use rates in Michigan. In Alexander et al., SES variables were the most significant predictors of hospital utilization, whereas board certified physicians and registered nurses per hospital bed were significant predictors in Clark's study. In contrast, our results illustrate that *both* SES and bed availability have significant impacts on hospital utilization rates; however, we did not consider measures of physicians or nursing as variables in our models. A number of factors cause concern in the results of these previous studies. First, although Alexander et al. controlled for temporal autocorrelation in their regression models, neither study acknowledged the spatial structure of their observations, thus likely misspecifying their regression models. Second, in both studies, hospital bed availability was calculated using a summation of the beds and population within the administrative unit boundaries, not incorporating the travel behavior of patients. Third, both studies were limited to regional-level observation units (58 over

Michigan's lower peninsula for Alexander et al. and 53 over Michigan's lower peninsula excluding Detroit for Clark) and a single scale of analysis.

As Figure 26 illustrates, in the weighted SAR model, the coefficient for E2SFCA decreases slightly as the data are aggregated to a regional-level scale. The most similar level of aggregation used in our analysis to those employed by Clark and Alexander et al. is 70 clusters (58 observations in the lower peninsula). At this level of aggregation, the weighted SAR model provides a positive, significant coefficient for hospital bed availability; however, the λ parameter is insignificant in this model. In a non-spatial weighted OLS regression with 70 clusters, we find that hospital bed availability is again not a significant predictor of utilization rates. These results likely stem from the homogenization of the data that occurs as the level of aggregation moves towards this regional scale of analysis. Interestingly, the level of aggregation used by Alexander et al. and Clark is very near an observed threshold where λ and E2SFCA become insignificant in the set of SAR models. In fact, at 88 clusters, E2SFCA is a positive and significant predictor and the λ parameter is also significant, suggesting that both Alexander et al. and Clark's studies may have produced different findings had they used less aggregated data. As a result, the effects of hospital bed availability on utilization rates may go undetected at regional-level scales. More specifically, our results provide empirical evidence of a threshold level in the ability to observe the effects of Roemer's Law in small area studies.

Recent research has shown the danger in statistical inference garnered from ecological-based relationships at a single geographic scale of analysis. Wright and Ricketts III (2010), in a review of Kravet et al. (2008), showed that coefficient values related to the supply of health care resources may change in significance and even direction as the scale of analysis changes by way of data aggregation. Their work highlights the problems associated with the MAUP in health-based research. In our study, the stability of the coefficients across levels of aggregation suggest that the observed relationships are not highly susceptible to variation due to the scale in which the data are aggregated to. Although levels of aggregation smaller

than Zip Codes could not be tested (due to the privatization of the hospitalization data), the overall statistical strength and invariant nature of the relationship between hospital bed availability and hospital utilization provide strong evidence that our findings are not a product of the MAUP.

With support of Roemer's Law demonstrated, we turn our attention toward the implications of our research with regards to CON programs. Past research has suggested that over the past 40 years CON programs have not been successful in controlling health care costs (Ferrier et al., 2010; Rivers et al., 2007; US Federal Trade Commission, 2004). A recent study by Conover and Sloan (2003) reported that Michigan's CON program had not effectively contained hospital costs and recommended that the state abandon regulation of acute care hospital beds. Whereas the effects of hospital bed availability on health care costs were not considered, the findings do suggest that efforts to control hospital bed availability will affect hospital utilization rates. Furthermore, the significant, stable, and positive nature of the observed relationship indicates that CON-based regulation of hospital bed supply to levels consistent with the needs of the population is justified.

Although it was not the focus of the analysis, our results also showed a strong, positive association between a higher proportion of Black and Hispanic populations and higher rates of hospital utilization. Given that other possible determinants of hospital utilization, SES and access to primary care (ACS hospitalizations), often associated with contributing to poorer health in disadvantaged populations were controlled for in our models, these findings are troubling from a social justice perspective. Although the cause behind this statistical association was not further explored in the present analysis, recent work by Grady (2006; 2010) and Grady et al. (2008) has demostrated that neighborhood segregation is associated with health disparities in New York and Michigan. In the present context, higher hospitalization rates for areas having a higher percentage of Black residents might point to underlying health issues that stem from neighborhood effects (Darden et al., 2010; Diez Roux et al., 2001; Oakes, 2004). Considering that metropolitan Detroit is one of the most segregated

cities in the USA (Darden et al., 2007) and a large proportion of Michigan's African American population resides in this region, our findings suggest that a more detailed analysis exploring the effects of race, segregation, and neighborhoods on hospital utilization rates in southeast Michigan is warranted.

4.6 Limitations

Our analysis did not consider alternative neighborhood structures in the EB smoothing process or the spatial regression models. Other neighborhood structures, such as those based on distance or k-nearest neighbors, require a defined threshold value for determining neighbor status. Given the large range of data configurations evaluated and their dissimilar geographic scales (for reference, see Figure 25), specifying a single distance or k threshold would not provide a consistent spatial structure throughout scales of analysis. Hence, the decision to employ a first-order neighborhood structure was considered necessary due to the multi-scalar nature of the research design. For example, if the neighborhood structure was defined using the 10 nearest neighbors, the neighborhood organization would vary considerably as the data were aggregated to more regional scales[14]. The same difficultly would manifest if a minimum distance threshold was implemented, augmented by the limitations associated with measuring distances among highly aggregated areal units (Hewko et al., 2002). For the purposes of our analysis, the first-order neighborhood structure provided a characterization of spatial structure supported by theory (Tobler, 1970) and flexible enough to accommodate the multi-scale nature of the research design.

Although the scale effect of the MAUP was explored in our analysis, the zoning effect was not explicitly examined. However, the effects of zone modification was implicitly addressed through the use of a non-agglomerative clustering methodology. Specifically, for each iteration in the clustering method, the Zip Code data were clustered, not the clusters from the previous step in the iteration. Hence, in many cases, regions were essentially "rezoned", thus

[14]Specifically, 10 neighbors may approximate first-order neighbors at low levels of aggregation, but 2nd or 3rd order neighbors at higher levels of aggregation.

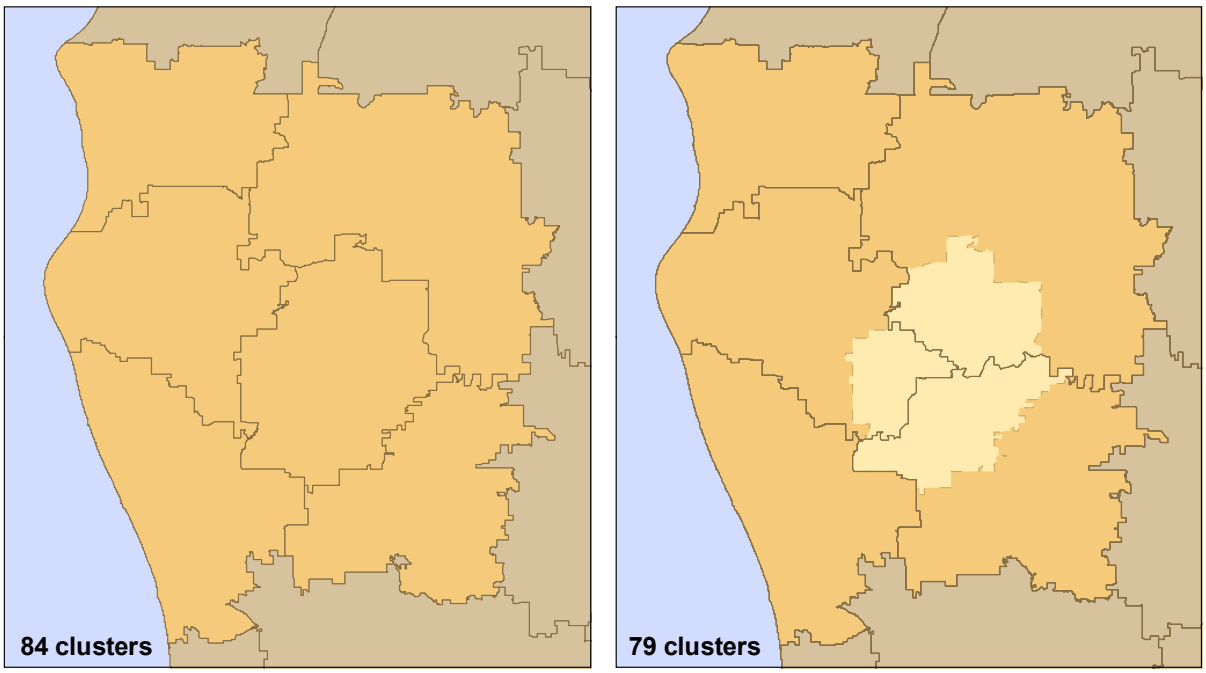

Figure 27: **Example of rezoned region.** In the 84 cluster solution (left), the region contains 6 clusters. In the 79 cluster solution (right), the same region contains 5 non-agglomerative clusters.

providing an implied examination of the zoning effect of the MAUP. To illustrate this point, Figure 27 contains an example of a small region that was rezoned rather than agglomerated as the level of aggregation changed. Given this limitation, we recommend that further consideration of the zoning effect of the MAUP to be included in future research of Roemer's Law.

4.7 Conclusions

This research found a positive, significant association between the hospital bed availability and hospital utilization rates while controlling for other determinants of hospitalization. The research design was implemented in a explicitly spatial context, incorporating the spatial and aspatial aspects of health care access and utilization along with the spatial structure of their relationship. Thus, we have provided compelling empirical evidence to support Roemer's Law.

Recent hospital construction and expansion (bypassing the CON program through legislative action) and a proposed transfer of beds into areas of the state without a demonstrated need for additional hospital beds highlight the importance of our findings in Michigan. Nationally, as health care systems and hospitals adapt to increasing health care costs, a changing economic climate, and provisions contained within the Affordable Care Act, gaining a clearer understanding of the effects of hospital bed availability on hospital utilization is paramount.

Whereas the findings of this study address the research question originally posed, they also elicit a number of new questions regarding health care policy and health services research. Perhaps, the most important question is, "what are the causal mechanisms that lead to higher hospitalization rates in areas with higher hospital bed availability?". While some have suggested that the answer lies in the clinical decision-making process of physicians (Mulley, 2009), others have suggested that it may be the hospitals themselves (Shwartz et al., 2011) and the question remains unanswered.

Acknowledgements: This research was funded by the Michigan Department of Health, Certificate of Need Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The R code used in this work can be found in Appendix E.

Conclusions

5.1 Overall contributions

This dissertation contributes significant new knowledge to the field of health services research. The specific salient outcomes include: detailing both the theoretical and applied differences between the raster and network data models for estimating travel time among locations, offering a methodology that simultaneously clusters observations based on comprehensive patterns of utilization and geographic location, and producing compelling, robust evidence that hospital availability has a positive, significant relationship with hospital utilization rates.

Study #1 explores the impacts of data model choice on travel time estimates. A hierarchical classification system is presented for assigning speed limits to roads. To explore the suitability of the assigned speed limits, travel time estimates are compared to those from Google Maps, an independent data source. Subsequently, populations and areas with limited geographic access to hospitals are identified using the raster and network data models. The regions with limited access were generally similar in shape and configuration. However, the analysis showed that the raster-based method produced larger overall regions than the network-based method, leading to a greater number of people identified as having limited areas. The major theoretical differences between the underlying data models were linked to the observed differences in an applied case study. Consequently, the network data model is suggested as preferable for estimating vehicular travel time if the topological relationships governing real-world travel are a priority in study design; these relationships are upheld within the data model itself. When estimating travel time for non-vehicular travel modes, in scenarios where travel is not restricted to roadways, and in cases where each location must be explicitly measured, the raster data model is more suitable given the unconstrained nature of movement in the data model.

Study #2 presents a clustering methodology for grouping geographically proximal hospitals with similar community utilization patterns. The methodology was specifically developed to create Hospital Groups for Michigan's CON Program. Therefore, the scientific

and political concerns encountered during the methodology's development are detailed in the study. The clustering methodology employs a K-means + Ward's clustering algorithm, simultaneously grouping hospitals based on their overall patient utilization patterns and geographic location. All possible values for the number of Hospital Groups (k) are evaluated and a hueristic is provided to select the final configuration. The methodology was designed to be repeatable, sustainable, and actionable. However, the clustering methodology can be employed to group any type of spatial observations having multiple attributes as it was built on first principles of clustering analysis. The methodology can be easily integrated for use with areal units to create small areas (or regions). In this, the clustering methodology provides a substantial theoretical advantage over the most oft-employed methods to create small areas such that it integrates *overall* patterns of health care utilization and geographic location, rather than relying on simple characterizations of utilization or relying solely on geographic location.

The final study (Study #3) in this dissertation examines the effects of Roemer's Law, a simple but powerful statement that proposes that hospital utilization will increase if the supply of hospital beds is increased. This study provides several key innovative and improved approaches to the study of health care access and utilization. The research design improves upon previous examinations of Roemer's Law by incorporating spatial factors in the analysis; the spatial nature of both utilization and access are considered, while also accounting for the spatial structure of their relationship. Secondly, the conceptual model of access is extended past the traditional *barriers only* model in which access is characterized only by the presence of factors limiting service utilization. The conceptual model is not in itself unique. CON programs are built around this theory. Yet, most applied access-related studies fail to account for this phenomenon. Third, the research design incorporates a novel, multi-scalar approach, exploring the stability of the statistical association between hospital bed availability and utilization rates. The multi-scalar approach allows for a richer understanding of the effects of Roemer's Law, while also providing a general framework for spatial regression analysis

with areal data.

The findings from Study #3 showed strong empirical evidence of the effects of Roemer's Law in Michigan. In a state-level study including the entire system of hospitals and population (including over 160 hospitals, over 1 million patient admissions, and nearly 10 million residents), the availability of hospital beds was found to have a significant, positive effect on hospital utilization rates while controlling for other determinants of variation in utilization rates. Additionally, this relationship proved to be highly stable across geographic scales. These results suggest that the effects of Roemer's Law are robust and due to health processes unrelated to the scale of analysis.

The main outcomes of the individual studies can be separated into those relating to advancement in health services research and those relating to matters of public policy. From a public policy perspective, this work offers updated methodologies to assist CON programs in their assessment and regulation of health care access. Study #1 provides a step-by-step guide for implementing the network-based method of identifying Limited Access Areas in Michigan. In addition, this study offers detailed descriptions of the theoretical differences and applied implications of the data models.

The clustering methodology detailed in Study #2 is very near final approval into Michigan's Hospital Bed Standards. The new methodology improves upon the previous methodology, which had become unusable given changes in hospital utilization patterns over time. The most important policy-related implication of the new methodology is that it provides an acceptable Hospital Group solution, thus removing the need for an expert committee to modify the automated results. Thus, from a public policy perspective, the new clustering methodology provides a greater level of objectivity.

Study #3 finds strong evidence of the effects of Roemer's Law, thus providing empirical support that areas with greater hospital bed availability have greater hospital utilization rates. Hence, this work provides empirical support for continued CON-based regulation of Michigan's hospital bed supply. Study #3 also showed that regions with a higher proportion

of Black residents had higher hospital utilization rates, even while controlling for other determinants of increased rates. This specific finding raises larger concerns about public health for disadvantaged populations in Michigan.

At a macro-level, this dissertation has provided a broadly-ranging exploration of access to hospitals and hospital utilization within a regulated health care system. The topic was approached from an explicitly spatial perspective, exposing the importance of location, geography, and distance-related factors in health services research. The research has delivered tangible research outcomes while also providing methodological advancements with the potential to improve the effectiveness of CON-based assessment and regulation of health care services. Thus, when viewed in its entirety, this dissertation provides key insights into the *relationship* between access and utilization, the *study* of access and utilization, and the *methods* used by CON programs in their mediation of health care resources.

5.2 Future research

5.2.1 Geographic accessibility

One of the most overlooked and under-reported aspects of research of geographic accessibility is the uncertainty present in population-level travel time estimates. As was discussed in Study #1, the actual travel time among locations is governed by a large number of factors including, but not limited to: individual driving characteristics, traffic volume, and the specific route chosen. Given these sources of variation, population-level models of travel time can only aim to provide generalized estimates of travel time among locations. However, the accuracy of these estimates has been largely ignored in previous health services accessibility research. The most pertinent sources of uncertainty uncovered in Study #1 are 1) the accuracy of travel speeds assigned to the roads data and 2) the completeness and/or accuracy of the roads data. The research approach in Study #1 provides an initial step toward addressing these issues. By comparing network-derived travel time estimates with estimates from Google Maps, the custom-built network dataset is evaluated against an independently

derived dataset. Given Google's lack of transparency regarding the input data and methods used for their travel time estimates, this approach cannot provide quantitative estimates of the accuracy of the network or the uncertainty present in the model.

The second source of uncertainty in population travel time estimates arises from the completeness and/or the positional accuracy of the roads data and how they affect travel time estimates. In the specific state-level case study, there were over 750,000 line segments (network edges) resulting in over 500,000 intersections (network nodes). Although these data are the most up-to-date available and are provided with metadata that include both a short description of the methods used to gather the data and the sources of the roads data, no method currently exists to evaluate the completeness or accuracy of the data themselves. This issue may be especially salient in Michigan, where a large number of private roads[15] are found. MDOT could not provide a quantitative estimate of the accuracy or completeness of the dataset when contacted directly. Given recent research illustrating the importance of roads data in health-based access studies (see Frizzelle et al., 2009), further exploration into methods that would provide quantitative estimates of the uncertainty present in large roads databases or methods to improve their accuracy is warranted.

Perhaps the most important questions raised in the geographic access study are those regarding the appropriateness and utility of Michigan's definition of limited access areas. First, the 30 minute cut-off value employed in identifying limited access areas is likely outdated. Although a number of studies have invoked 30 minutes as an appropriate travel time to discern those with geographic access from those without, like Roemer's Law, a surprisingly small amount of empirical research exists to justify this particular choice. The most cited work, that of Bosanac et al. (1976), is over 30 years old. Both the ability to travel and the *expectation* of reasonable travel to obtain health care have likely changed since since that research. As reported, for Michigan in 2010, roughly 20% of inpatient days were spent

[15]Private roads are those not maintained by local, regional, or state government agencies. Thus, they are not official roads.

in hospitals more than 30 miles from the patients' residences. Second, as was discussed in Study #3, a more comprehensive framework of spatial accessibility includes not only distance, but also supply and demand. Thus, although travel time provides a simple, easy to understand measure of geographic access, this metric alone does not incorporate the other factors known to influence spatial accessibility and is insufficient in identifying those having limited access. These findings call into question both the appropriateness of 30 minutes as a cut-off value and use of distance alone to determine access status. Importantly, I have exposed the need for future research that incorporates current patient travel patterns and expectations of health service accessibility to provide a more complete characterization of what constitutes "limited" geographic access.

5.2.2 Clustering health care observations

John Griffith, one of the architects of the previous clustering methodology employed by Michigan's CON program, delivered a positive review of the new clustering methodology presented in Study #2, stating that it was an *important new solution* and *an advance over prior work*. However, the selection of the final number of hospital groups in the clustering methodology remains essentially heuristic. Important research questions remain unanswered, specifically 1) "What is the right number of clusters?" and 2) "What is the proper balance between objective, scientific analysis and political considerations in health policy?" On the surface, the two questions may appear highly dissimilar; however, throughout the development process of the clustering methodology, they were revealed to be unequivocally linked.

Despite the large amount of literature on clustering theory, methods, and uses, a recent review from the clustering literature states (Jain, 2010, p. 654)[16]:

> *The most critical choice is K. While no perfect mathematical criterion exists, a number of heuristics are available for choosing K. Typically, K-means is run independently for different values of K and the partition that appears the most*

[16]In Jain (2010), K is the number of clusters.

meaningful to the domain expert is selected.

Therefore, given the current status of clustering methodologies, the first question above cannot be authoritatively answered. As a result, the first question becomes "What is the right number of clusters for the particular application?" which directly corresponds to the second. Professor Griffith, given his longtime research focus on policy-based issues in health care (e.g., Griffith, 1972; Griffith et al., 1981), seems to have understood this limitation in the process of identifying the number of Hospital Groups, thus viewing the heuristic employed in the clustering methodology as a step toward objectivity in a highly politicized process, not a limitation of the research. Yet, although a politically acceptable solution to this problem was delivered, further research efforts towards the statistical evaluation of cluster solutions would likely provide valuable insights toward identifying the right number of clusters in a dataset, thus marginalizing the subjectivity introduced by employing a domain expert (or set of experts) for this task.

In the case of Michigan's Hospital Groups, further evaluation of the state's proposed Hospital Group configuration may also benefit policy makers by providing a quantitative estimate of how the number of Hospital Groups affects predictions of future bed demand. As the results of Study #3 illustrated, the relationship between hospital utilization rates and hospital bed availability became undetectable at regional-level scales. Specifically, a threshold level was discovered near 90 observation units; further aggregation into fewer observation units yielded regression models with insignificant predictor variables. Because the hospitals in proposed Hospital Groups are more highly aggregated than the previous configuration (35 vs. 64 groups) and both fall under the identified threshold, the most apparent concern is that localized unmet hospital bed demand and/or excess capacity will be masked by variation within the spatially larger groups. As a result, regions with an unmet demand may go unnoticed. Furthermore, the threshold uncovered in Study #3 casts doubt upon the use of the Hospital Groups as a unit of analysis for exploring other health-related relationships in Michigan.

Study #2 also invoked limitations in the current methods used to create small areas in health services research. Small areas are created by aggregating individual areal observations into groups or regions. Studies that explore variations in health care spending, health outcomes, and health care utilization have been present and forefront in the health services literature since Wennberg and Gittlesohn's (1973) work exploring variations in utilization rates in Maine. However, despite the nearly 40 years that have passed since this seminal publication, their relatively simple method to create the small areas remains oft-used in current research. This method assigns the initial observations to groups (small areas) using only a single measure of hospital utilization, the plurality of visits (e.g., Unit A would be assigned to Hospital 1's group if more residents visited that facility than any other, regardless of the actual proportion). While this method performs well in regions where facilities are well distributed, it often requires manual adjustment of the small area membership in regions with multiple facilities. Hence, the clustering methodology developed in Study #2 provides an important advance in improving the creation of small areas. However, given that many small area studies focus on exploring differences among facilities, this methodology only delivers the first step in this process. To link the clustered regions to specific hospitals or groups of hospitals would require an additional step not explored in Study #2. One possible approach for this task is to consider attributes of *both* the facilities and population units in the cluster formation method, as suggested by Gilmour (2010).

5.2.3 Roemer's Law

The final study found evidence of a positive association between hospital bed availability and hospital utilization rates while controlling for other determinants of hospital utilization. Although this outcome does provide support for Roemer's Law, it begs the question, "What causes Roemer's Law?" Given the ecological nature of this research, an attempt to assign causation based on the findings is not justified. Hence, the logical next step to better understand the implications of Roemer's Law is to explore the causal factors that produce higher utilization rates in areas with greater hospital bed availability. Although Wennberg

(2005) points to clinical decision-making as a possible cause of Roemer's Law, the complex interaction of the actors involved in hospitalization (i.e., doctors, hospitals, and patients), in combination with the socio-demographic, environmental, and stochastic nature of ill-health events that lead to inpatient hospitalization, suggest that this explanation may be inadequate. The complexity of the process poses significant challenges to future research endeavors. Given the high costs of hospitalizations and strong evidence of Roemer's Law demonstrated by this study, efforts to understand the underlying mechanisms are clearly warranted.

The demonstrated effect of Roemer's Law in Study #3 suggests that efforts to mediate the availability of hospital beds will likely impact hospital utilization rates. These results lend support for continued CON regulation of hospital bed availability under the assumptions that 1) overutilization is present in areas with high hospital bed availability and 2) curbing overutilization of hospital services will assist in lowering overall health care costs. This dissertation provides the initial step toward evaluating the first assumption. A logical continuation of the work would be to isolate the effects of hospital bed availability on utilization rates and identify areas where high availability leads to higher than expected utilization. Additionally, exploring temporal changes in utilization rates as hospitals have opened, closed, and/or expanded their capacity would also likely provide insights towards identifying areas in which overutilization may be present, along with a more detailed understanding of the relationship between availability and utilization in these areas. Furthermore, similar inpatient hospitalization data from a state without CON regulation would provide the opportunity to build a natural experiment exploring whether the presence of a CON program had an impact on this relationship.

There has been little recent research examining the effectiveness of CON programs in controlling health care costs. The lack of research likely stems from the variation in scope and size among the state-based CON programs and the limited availability of information

regarding the specifics of CON laws and oversight (Rivers et al., 2007)[17]. This dissertation corroborates Rivers et al.'s assertion as searches for detailed information regarding nearby states' CON programs or recent research exploring specific CON programs were largely unsuccessful.

The small number of recent CON-related studies provided conflicting results regarding the programs' effectiveness. A general review determined that the programs have not reached their goal of health care cost-containment (Banks et al., 1999). Conover and Sloan (1998) found that states that had repealed their CON laws did not experience an increase in health care costs per capita. Further research found that CON regulation may actually lead to *higher* health care costs (Rivers et al., 2007) and lower levels of hospital efficiency (Granderson, 2011) by obstructing the potential for competition among hospitals. In an example of conflicting research, Ferrier et al. (2010) showed that states with CON programs have higher levels of hospital efficiency, thus improving resource allocation and lowering social costs. Additionally, recent work by Hellinger (2009) demonstrated that states with CON programs have fewer hospital beds per capita, which is associated with lower overall health care costs (however, the CON variable was not itself a significant predictor of health care costs in the model).

CON-based research papers appear to have declined in number over the last decade. Given that the general findings show that CON programs have failed to reach their aims of health care cost containment, it begs the question of why a majority of states that continue to employ them. Although efforts towards easing CON regulation or deregulation altogether have increased in recent years (see Romano, 2003; Robeznieks, 2008), CON programs persist. Perhaps this is a reflection of the duality in the overall goals of many CON programs. Outright removal of CON laws would not only remove the program tasked with regulating health services expansion, but also the program that attempts to provide equitable access

[17]Finn (2007) provides a historical overview of CON at a national level and a detailed examination of Michigan's CON program.

to health services for the population. As Rivers et al. (2007) suggest, tasking both goals to the same program is not practical. In Michigan, this sentiment was offered prior to Rivers et al.'s research by Conover and Sloan (2003) in their assessment of the state's CON program.

This dissertation found that hospital bed availability is positively related with hospital utilization rates, thus providing support for the continued regulation of hospital bed supply. However, further identifying areas with significant over- or underutilization of hospital services in the state would provide a better understanding of the effectiveness of Michigan's CON program in mediating hospital bed supply to meet population need. This work would provide the necessary foundation to a larger exploration of the effects of the state's CON program on health care costs.

5.2.4 Spatial structure

Study #3 examined the relationship between access and utilization over a broad geographic area containing large variations in demographic structure, socio-economic status, and urban/rural settings. Given the variability among areas, further explorations of specific regions of the state may offer more detailed information regarding the significance and magnitude of this relationship. The CAR and SAR regression models, although accounting for spatial structure, consider the magnitude of the observed relationships among variables to be stationary (i.e., a single coefficient describes the relationship over the entire study area). Techniques such as Geographically Weighted Regression (GWR) provide the ability to identify nonstationarity in regression-based coefficients, allowing spatial structure in the observed relationships themselves to be identified (Brunsdon et al., 1996). GWR could be implemented to identify regional characteristics of the drivers of hospital utilization, thus providing an initial step toward a more detailed examination of the determinants of hospital utilization within specific regions.

5.2.5 Health insurance

In Study #3, health insurance coverage was considered an "aspatial" measure of access to health care and only contained the percent of the population having coverage. No difference was made among public or private insurance, type of insurance (e.g., health maintenance organization (HMO), preferred provider organization(PPO)), or level of coverage. Although more detailed data regarding the nature of health insurance coverage would be useful, a more interesting avenue to explore is the potential for the "type" of insurance to impact utilization patterns. Because HMOs and PPOs offer incentives for staying within a specific network of providers, the resultant utilization patterns of the population may reflect the geographical distribution of providers within their specific network not those expected based on travel time or hospital bed availability. Additionally, differences in utilization patterns may manifest between those with public insurance (i.e., Medicare and Medicaid) and those having private insurance, as well as differences in utilization rates among public and private insurance holders. Examining the the effects of health insurance coverage on spatial patterns of utilization may provide important insights into understanding the travel behavior of patients, while also potentially offering a better understanding of how redistribution of hospital bed availability may impact hospital utilization rates.

5.2.6 Spatial accessibility

The Enhanced 2-Step Floating Catchment Area method was employed to describe the availability and accessibility of hospital beds in Study #3. The creation of improved metrics to describe spatial accessibility is an active area of health services research; FCA metrics are a relatively recent development with improvements and modifications being offered regularly. An updated method, the 3 Step Floating Catchment Area (3SFCA) was recently proposed by Wan et al. (2012), introducing potential competition among facilities into the calculation of spatial accessibility. However, preliminary experiments conducted for Study #3 (not presented in manuscript) suggest that the 3SFCA underestimates potential demand, thus

providing artificially inflated levels of spatial accessibility in areas with multiple nearby facilities. From a theoretical perspective, the inclusion of facility competition into an FCA-based metric has the potential to provide a more accurate and comprehensive characterization of availability and accessibility. Yet, the initial experiments suggest that the 3SFCA has not accomplished this goal. A better understanding of the relationship among supply, demand, and potential competition is necessary to provide an applied FCA metric that incorporates these elements simultaneously. Further attention is warranted given the widespread use of these metrics in current health services research.

5.2.7 Access and utilization in a regulated health care system

Within the last five years in Michigan, two hospitals were opened in areas without a demonstrated need for additional hospital beds. Approval for these facilities did not come from the state CON program, but through specially-drafted legislation subverting the CON process (Greene, 2012). More recently, another Michigan conglomerate hospital system filed a CON application to transfer a large number of licensed beds into an area of the state without a demonstrated need for additional hospital beds (Hopkins, 2012). If the CON application for the current request is denied, special legislative action to approve the transfer appears highly likely (Greene, 2012). Thus, control over the distribution of hospital beds in Michigan will again be removed from the state's CON program. In addition, the redistribution of the state's hospital bed supply will not follow demonstrated patterns of population need. Although these are specific examples from the particular study area explored in this research, they are a microcosm of the larger changes occurring in the US health care system. As a result of rising health care costs and a shift toward profit-maximizing behavior (Kuttner, 2008), the US health care system continues to undergo significant changes, many of which impact health care delivery and population access.

The Patient Protection Act of 2010 attempts to address "affordability" by increasing access to health insurance for the currently uninsured and underinsured populations (Schoen et al., 2011), limiting insurance companies' abilities to deny coverage, and redistributing

the burden of public health (and costs) onto all stakeholders involved (i.e., health care practitioners, health insurance companies, and health care consumers). Increases in health insurance coverage, along with a greying US population due to increased life expectancies and the aging of the baby-boom generation, have the potential to significantly raise the future demand for health care services (Hofer et al., 2011; Strunk et al., 2006). The potential burden placed on the US health care system due to increased health insurance access and population demand for services, in conjunction with the changes resulting from a shifting economic environment, highlight one piece of the uncertainty existing for the future public health of US citizens. The poor current performance of the US health care system in numerous measures of public health outcomes (Murray and Frenk, 2010) only offers further concern.

By examining the meta-relationship between health care access and utilization, this dissertation has provided important findings while also supplying a number of research pathways for future studies in health geography. This work did not consider the relationship among access, utilization, and *public health outcomes*. Thus, at a macro level, the most important question invoked by this research, but yet to be answered is, "How do access and utilization affect public health outcomes?" Given the recent changes within the US health care system and those likely forthcoming, answering this question is paramount to understanding how changes in health care access and utilization will affect population health in the US.

Appendices

Appendix A
R code to implement the Thomas Methodology

```
#############################################################################
####                                                                     ####
####  Citation information (code)                                        ####
####  Delamater PL, Shortridge AM, and JP Messina, Regional health       ####
####  care planning: a methodology to cluster facilities using           ####
####  community utilization patterns. BMC Health Services Research       ####
####                                                                     ####
####  Citation information (original methodology)                        ####
####  Thomas JW, Griffith JR, and P Durance, Defining hospital           ####
####  clusters and associated service communities in metropolitan        ####
####  areas, Socio-Economic Planning Sciences 1981, 15(2):45-51          ####
####                                                                     ####
####  Max Relevance Algorithm Clustering Algorithm                       ####
####                                                                     ####
####  Requires: Patient visits table (zip -> hospital)                   ####
####            Zip population data                                      ####
####            Hospital info table                                      ####
####            Hospital zip code table                                  ####
####                                                                     ####
####  Interpreted and converted to R code by Paul Delamater and          ####
####  Ashton Shortridge during summer, 2011 for the Michigan             ####
####  Hospital Bed Standard Advisory Committee working group.            ####
####  Funding for this research was provided by the Michigan             ####
####  Department of Community Health.                                    ####
####                                                                     ####
#############################################################################


#############################################################################
##                                                                         ##
##  Get input data                                                         ##
##                                                                         ##
#############################################################################


############################
## Read patient visits data
############################

####  Note: pv is a table with hospitals in rows and zip codes
```

```
####        in columns. Hospital identifier column should be
####        labeled "HOSP_ID". Zip code column lables should
####        be the five digit zip code (e.g., "48823"). Table
####        entries are the number of hospitalizations from
####        residents of each zip code at each hospital.

# Load data
pv <- read.csv("inputdata/hosp.zip.visits.mtx.csv")

# Ensure HOSP_ID in character format
pv$HOSP_ID <- as.character(pv$HOSP_ID)

# Remove characters from column names
# R adds an "X" to the zip code number
# Assumes all zip codes are 5 digits
names(pv)[2:ncol(pv)] <- substr(as.character(names(pv)[2:ncol(pv)]), 2, 6)

################################################################
## Convert number of visits to proportions (Relevance Index)
################################################################

# Define variable for last zip code column
n.zip <- ncol(pv)

# Sum hospital visits for each zip code
# Assumes HOSP_ID is first column
zip.visits <- colSums(pv[,2:n.zip])

# Divide each entry by summed visits to create Rij values
pv[,2:n.zip] <- pv[,2:n.zip] / rep(zip.visits, each = nrow(pv))

################################
## Read hospital attributes data
################################

####  Note: hosp.info is a table with hospitals in rows and
####        attributes in columns. In this case, the column
####        that corresponds to the patient records is "MIDB".

# Load data
hosp.info <- read.csv("inputdata/hospitals.csv")

####  Note: hosp.HAU is a table with hospitals in rows and
####        attributes in columns. In this case, the column
####        that corresponds to the patient records is "MIDB".
```

```
####        Home Areal Unit is "ZIP".

# Load home areal unit (zip code) of each hospital
hosp.HAU <- read.csv("inputdata/hospital.zipcodes.csv")

# Attach to home areal unit to patient records
pv <- merge(pv, hosp.HAU, by.x="HOSP_ID", by.y="MIDB", all.x=TRUE)

# Change column name
names(pv)[ncol(pv)] <- "HAU"

# Add column with Rij (Relevance Index) of each hospital in
# its own home areal unit
for (h in 1:nrow(pv)) pv$RiHAU[h] <- pv[h,which(names(pv)==pv$HAU[h])]

################################
## Get zip code population data
################################

####  Note: zip pop is a table with the zip code name in a
####        column, "ZIP" and the population of the zip code
####        in a column, "POP"

# Load data
zip.pop <- read.csv("inputdata/zipcode.population.csv")


#############################################################################
##                                                                         ##
## Code to implement Thomas Methodology                                    ##
##                                                                         ##
#############################################################################


#######################################
## Prepare data and create data holders
#######################################

# Add column with initial alpha values (all are set at 0.02)
# In an update of this code, initial alpha values are set at 0.05
pv$alpha <- 0.02

# Define initial values for alpha variables
alpha.1 <- 0.02
alpha.2 <- 0.125
```

```
# Add binary holder column for individuals / groups
pv$Group <- 0
# Add column to hold hospital names after clustering
pv$GrNames <- pv$HOSP_ID

#######################################################
## Calculate population weighted relevence index, Rj
#######################################################

####  Note: Pi = population of areal unit i
####        Rij = relevance index values for areal unit i
####        to hospital j

# Calculate PiRij values (Pi * Rij)
PiRij.matrix <- pv
PiRij.matrix[,2:n.zip] <- PiRij.matrix[,2:n.zip] * rep(zip.pop$POP, each
  = nrow(PiRij.matrix))

# Create holder for Rj values
Rj.all <- NULL

# Create holder for Ij zip codes
Ij.matrix <- pv
Ij.matrix[,2:n.zip] <- 0

# Calculate Rj for each hospital
for (j in 1:nrow(pv)) {

  # Get hospital j's Ri values
  hosp.j <- pv[j,2:n.zip]

  ####  Note: From Thomas et al., Ij = set of areal units
  ####        for which individual relevance values of
  ####        hospital j exceeds or equals alpha

  # Find zip codes with Rij greater than alpha
  Ij.list <- which(hosp.j >= alpha.1)+1

  # Write zip codes greater than alpha to Ij holder
  Ij.matrix[j,c(Ij.list)] <- 1

  # If no areal units in Ij, Rj value is zero
  if (length(Ij.list) == 0) {
```

```
    # Write hospital ID and 0 to Rj holder
    Rj.all <- rbind(Rj.all, cbind(as.character(pv$HOSP_ID[j]), 0))

    } else {


  # Get list of zip code names
  Ij.zips <- names(pv)[Ij.list]

  # Get numerator value for Rj
  PiRij <- sum(PiRij.matrix[j,Ij.list])

  # Get denominator value for Rj (total zip code population)
  Pi <- sum(zip.pop$POP[c(Ij.list-1)])

  ####  Note: Rj = sum(Pi(dij/Di)) / sum(Pi)
  ####        where dij/Di is Relevance Index

  # Calculate Rj (population weighted relevance index)
  Rj <- PiRij / Pi

  # Put in holder
  Rj.all <- rbind(Rj.all, cbind(as.character(pv$HOSP_ID[j]), Rj))

  }

}

# Make Rj.all into dataframe
Rj.all <- as.data.frame(Rj.all)

# Rename columns in Rj.all
names(Rj.all) <- c("HOSP_ID", "Rj")

# Convert from factor to numeric and character
Rj.all$Rj <- as.numeric(levels(Rj.all$Rj)[Rj.all$Rj])
Rj.all$HOSP_ID <- as.character(Rj.all$HOSP_ID)

#####################################################
## Remove hospitals with Rj of 0 from analysis
## These hospitals are ungroupable using the method
#####################################################

# Locate hospitals with Rj = 0
zeros <- which(Rj.all$Rj == 0)
```

```
# Get hospital ID
Ungroupable.hospitals <- Rj.all[c(zeros),1]

# Remove hospitals from matrices
pv <- pv[-c(zeros),]
Rj.all <- Rj.all[-c(zeros),]
Ij.matrix <- Ij.matrix[-c(zeros),]
PiRij.matrix <- PiRij.matrix[-c(zeros),]

# Write ungroupable hospitals info to table
Ungroupable.hospitals.info <- hosp.info[hosp.info$MIDB %in%
  Ungroupable.hospitals, ]

############################################################
##  Start iterative process part of the code and explicitly
##  state which method will be used to STOP the process
############################################################

# Create holder for grouped hospitals
Grouped.Hospitals <- NULL

# Create holder for temporary Rj.min values
Rj.temp <- NULL

####  Note: From Thomas et al., The procedure terminates
####        when one of three conditions occurs: (1) all
####        hospitals have been aggregated into a single
####        large cluster; (2) a user-specified number of
####        iterations has been completed; or (3) all
####        identified clusters are stable, i.e., no
####        cluster serves more than alpha of the patients
####        in the home areal unit of any other cluster.
####
####        These lines will make the iterative process stop
####        at a specified number of Subareas, similar to
####        option number (2) above. To choose this option
####        uncomment the following lines and comment out,
####        "run <- 1" and "while (run == 1) {"

# Select desired number of Subareas
# n.subareas <- 64
# Start grouping hospitals
# while (nrow(pv) > n.subareas) {
```

```
####        These lines will make the iterative process stop
####        when no hospital/group has greater than alpha
####        of any other hospitals/group's home area (option
####        number (3) above). These line also stops code if
####        all hospitals are aggregated into one large
####        group (option number (1) above).

# Create variable used in the iterative process for stopping
run <- 1

# Start grouping hospitals
while (run == 1) {

  ################################
  ## Find hospital with minimum Rj
  ################################

  ####  Note: Checks for hospitals in a temporary holder. This
  ####        holder is defined below. It is used in case any
  ####        hospital is the min Rj, but does not have another
  ####        hospital to group with yet

  if (length(Rj.temp) == 0) {

    # Locate hospital with minimum Rj value
    which.hosps <- which(Rj.all$Rj == min(Rj.all$Rj))

    # Get number of "minimum" Rj hosps
    n.min.hosps <- length(which.hosps)

    } else {

      # Locate hospital with minimum Rj value (minus temp)
      which.hosps <- which(Rj.all$Rj == min(Rj.all$Rj[-Rj.temp]))

      # Get number of "minimum" Rj hosps to determine if ties exist
      n.min.hosps <- length(which.hosps)

    }


  # If a tie exists, randomly select which of the hospitals is
  # is selected for aggregation.  Otherwise min.hosp is used
  if (n.min.hosps > 1) {
```

```
  # Create random variable using number of tied hospitals
  min.hosp <- round((n.min.hosps-1)*runif(1))+1

  # Select hosptial using random variable
  min.hosp <- which.hosps[min.hosp]

  } else {

    # Use the single hospital
    min.hosp <- which.hosps

  }

# Subset minimum hospital from Rj.all
Rj.min <- Rj.all[min.hosp,]

# Print to screen to display which hospital is selected
print(paste("Rj.min = ", Rj.min[2], ", HOSP_ID = ", Rj.min[1],
  sep=""))

# Get Rj.min's home areal unit (column number!)
Rj.min.Ij <- which(names(pv) == pv$HAU[min.hosp])

# Print Rj.min's home areal unit and RI
print(paste("Rj.min HAU = ", pv$HAU[min.hosp], ", RI = ",
  pv[min.hosp, Rj.min.Ij], sep=""))

#####################################################################
## Find hospital/cluster with max RI in Rj min's home areal unit
#####################################################################

####  Note: From Thomas et al., the hospital with the smallest
####        Rj is identified and grouped to form a cluster with
####        the hospital having the greatest individual
####        relevance in hospital j's home areal unit.

# Find max RI in Rj min's home areal unit
Rj.max.Rj.min <- which(pv[,Rj.min.Ij] == max(pv[,Rj.min.Ij]))

# If statement in case it selects itself
# e.g., no hospital or cluster has higher Ri in minimum's
# home area
if (Rj.max.Rj.min == min.hosp) {

  # Pick the next highest after removing min hospital
```

```
  next.Rj.max <- max(pv[-min.hosp,Rj.min.Ij])
  Rj.max.Rj.min <- which(pv[,Rj.min.Ij] == next.Rj.max)

  }

# In case of ties for Rj.max select randomly from tied hospitals
if (length(Rj.max.Rj.min) > 1) {

  # Generate random number
  rand <- round((length(Rj.max.Rj.min)-1)*runif(1))+1

  # Use random number to select
  Rj.max.Rj.min <- Rj.max.Rj.min[rand]

  }

# Get RI of Rj.max
alpha.Rj.max <- pv[Rj.max.Rj.min, Rj.min.Ij]

# Print alpha value and HOSP ID to screen
print(paste("alpha.Rj.max = ", alpha.Rj.max, ", HOSP_ID = ",
  pv$HOSP_ID[Rj.max.Rj.min], sep=""))

#####################################################################
## Big logic part of code. Determines whether to group hospitals
## or move to next minimum hospital in list
#####################################################################

####  Note: From Thomas et al., ...the hospital with the smallest
####        Rj is identified and grouped to form a cluster with
####        the hospital having the greatest individual relevance
####        in hospital j's home areal unit.
####
####        We assume that there is a 'cut-off' value in this
####        step based on the text in termination option number
####        (3), i.e., no cluster serves more than alpha of the
####        patients in the home areal unit of any other cluster.

# If the Rj value in Rj.min's home area is larger than the
# alpha cutoff of the hospital or cluster, then cluster
if (alpha.Rj.max >= pv$alpha[Rj.max.Rj.min]) {

  ########################################
  ## Update RI values to reflect clustering
  ########################################
```

```
# Sum RI values for clustered hospitals
pv[Rj.max.Rj.min,2:n.zip] <- pv[Rj.max.Rj.min,2:n.zip] +
  pv[min.hosp,2:n.zip]

# Update alpha score and group columns
pv$alpha[Rj.max.Rj.min] <- alpha.2
pv$Group[Rj.max.Rj.min] <- 1
pv$GrNames[Rj.max.Rj.min] <- paste(pv$GrNames[Rj.max.Rj.min],
  pv$GrNames[min.hosp], sep=",")

#####################################
## Update home areal unit for cluster
#####################################

####  Note: From Thomas et al., When a previously formed cluster
####        j* is identified for further clustering, its home
####       areal unit is assumed to be the home areal unit of
####        the hospital (member of j*) having the highest Rij
####        among the cluster hospitals' home areas

# If Rj min's relevance in its home area is larger than Rj max
# assign new home areal unit to newly formed cluster
if (pv$RiHAU[Rj.max.Rj.min] < pv$RiHAU[min.hosp]) {

  # Assign Rij to cluster entry
  pv$RiHAU[Rj.max.Rj.min] <- pv$RiHAU[min.hosp]

  # Assign new home areal unit to cluster entry
  pv$HAU[Rj.max.Rj.min] <- pv$HAU[min.hosp]

}

#########################################################
## Update Ij.matrix to reflect new alpha value of cluster
#########################################################

# Find zip codes above new alpha value
Ij.new <- which(pv[Rj.max.Rj.min,2:n.zip] >= alpha.2)+1

# Clear old Ij row, then write new zip codes to Ij holder
Ij.matrix[Rj.max.Rj.min, 2:n.zip] <- 0
Ij.matrix[Rj.max.Rj.min,c(Ij.new)] <- 1

######################
```

```
## Update PiRij.matrix
######################

# Sum PiRij entries
PiRij.matrix[Rj.max.Rj.min,2:n.zip] <-
  PiRij.matrix[Rj.max.Rj.min,2:n.zip] +
  PiRij.matrix[min.hosp,2:n.zip]

##############################################
## Update Rj.all with new list of Ij zip codes
##############################################

# Get numerator value for Rj
n.PiRij <- sum(PiRij.matrix[Rj.max.Rj.min,Ij.new])

# Get denominator value (total zip code population)
n.Pi <- sum(zip.pop$POP[c(Ij.new-1)])

# Calculate Rj  (population weighted relevence index)
Rj <- n.PiRij / n.Pi

# Put in holder
Rj.all$Rj[Rj.max.Rj.min] <- Rj

############################################################
## Remove Rj.min from pv, Ij.matrix, PiRij.matrix, Rj.all
## because it has now been grouped
############################################################
pv <- pv[-c(min.hosp),]
Ij.matrix <- Ij.matrix[-c(min.hosp),]
PiRij.matrix <- PiRij.matrix[-c(min.hosp),]
Rj.all <- Rj.all[-c(min.hosp),]

# Write Rj.min hosp to holder
Grouped.Hospitals <- c(Grouped.Hospitals, Rj.min$HOSP_ID)

# Print to screen which hospitals have been grouped
print(Grouped.Hospitals)

# Reset Rj.temp because current Rj.min has been grouped
Rj.temp <- NULL

} else {

  ################################################################
```

```
      ## Re-run steps with a different Rj.min because aggregation may
      ## produce clusters with home areas > alpha) in former Rj min's
      ## home area.  So we hold onto this Rj.min and re-check later
      ## List this Rj.min in holder
      ###################################################################

      Rj.temp <- c(Rj.temp, min.hosp)

    }

  # Print to screen which hospitals are in Rj.temp and
  # the length of both Rj.temp and Rj.all
  print(paste("Rj temp has: ", length(Rj.temp), " hospitals/cluster", sep=""))
  print(paste("Rj all has: ", nrow(Rj.all)-1, " hospitals/clusters remaining",
     sep = ""))

  ###################################################
  ## Determine whether to keep attempting to cluster
  ## or to terminate the iterative process
  ###################################################

  # If all the hospitals (-1) are in Rj.temp, then no
  # hospital has more than alpha of another's home area
  if (length(Rj.temp) == nrow(Rj.all)-1) {
    run <- 0
  }

  # If all hospitals are grouped Rj.all has one row
  if (nrow(Rj.all) == 1) {
    run <- 0
  }

}

###################################################
## Attach Subarea designation to hospital info file
###################################################

# Get number of Subareas
n.subareas <- dim(pv)[1]

# Make empty holder
subarea.table <- NULL

# Break apart output table from Thomas method
```

```
# and insert into holder
for (p in 1:n.subareas) {
  names <- unlist(strsplit(pv$GrNames[p], ","))
  subarea.table <- rbind(subarea.table, cbind(p, names))
}

# Rename column names
colnames(group.table) <- c("Thomas", "MIDB")

# Attach Subarea names to hospital info file
hosp.info <- merge(hosp.info, group.table, by="MIDB", all.x=TRUE)

# Name ungroupable hospitals "NG"
hosp.info$Thomas <-as.character(hosp.info$Thomas)
hosp.info$Thomas[is.na(hosp.info$Thomas)] <- "NG"
```

Appendix B
R code to implement the new clustering methodology

```
#############################################################################
####                                                                     ####
####  Citation information                                               ####
####  Delamater PL, Shortridge AM, and JP Messina, Regional health       ####
####  care planning: a methodology to cluster facilities using           ####
####  community utilization patterns. BMC Health Services Research       ####
####                                                                     ####
####  2-step K-means + Ward's Algorithm                                  ####
####                                                                     ####
####  Requires: Patient visits table (zip -> hospital)                   ####
####            Hospital travel distance table (hosp -> hosp)            ####
####            Hospital info table                                      ####
####                                                                     ####
####  Methodology developed by Paul Delamater, Ashton Shortridge,        ####
####  and Joe Messina during summer, 2011 for the Michigan Hospital      ####
####  Bed Standard Advisory Committee working group. Funding for         ####
####  this research was provided by the Michigan Department of           ####
####  Community Health.                                                  ####
####                                                                     ####
#############################################################################


#############################################################################
##                                                                         ##
##  Get input data                                                         ##
##                                                                         ##
#############################################################################


############################
## Read patient visits data
############################

#### Note:  pd.1, pd.2, pd.3 are tables with hospitals in rows
####        and zip codes in columns. Hospital identifier column
####        should be labeled "HOSP_ID". Zip code column lables
####        should be the five digit zip code (e.g., "48823").
####        Table entries are the number of patient days from
####        residents of each zip code at each hospital.
####
```

```
####        Assumes patient day matrices have similar dimensions!

# Load data
pd.1 <- read.csv("inputdata/hosp.zip.patdays.mtx.y1.csv")
pd.2 <- read.csv("inputdata/hosp.zip.patdays.mtx.y2.csv")
pd.3 <- read.csv("inputdata/hosp.zip.patdays.mtx.y3.csv")

# Create 3 year sum matrix
p.sum.3yr <- pd.1[,2:ncol(pd.1)] + pd.2[,2:ncol(pd.2)] + pd.3[,2:ncol(pd.3)]

# Re-attach hospital names column
p.sum.3yr <- cbind(pd.1[,1], p.sum.3yr)

# Rename hospital names column
names(p.sum.3yr)[1] <- "HOSP_ID"


# Ensure HOSP_ID in character format
p.sum.3yr$HOSP_ID <- as.character(p.sum.3yr$HOSP_ID)

# Remove characters from column names
# R adds an "X" to the zip code number
# Assumes all zip codes are 5 digits
names(p.sum.3yr)[2:ncol(p.sum.3yr)] <-
  substr(as.character(names(p.sum.3yr)[2:ncol(p.sum.3yr)]), 2, 6)

#############################################################
## Convert raw patient days to proportions (Commitment Index)
#############################################################

# Define variable for last zip code column
n.zip <- ncol(p.sum.3yr)

# Sum patient days for each hospital
# Assumes HOSP_ID is first column
hosp.pat <- rowSums(p.sum.3yr[,2:n.zip])

# Divide each column by total patient days
p.sum.3yr[,2:n.zip] <- p.sum.3yr[,2:n.zip] / hosp.pat

# Rename table
p.CI.3yr <- p.sum.3yr
rm(p.sum.3yr)

#########################################
```

```
## Remove hospitals with no patient visits
########################################

# Locate hospitals with zero visits
zero.pv <- which(hosp.pat == 0)

# Get names of zero hospitals (will need this later!)
zero.names <- as.character(p.CI.3yr$HOSP_ID[zero.pv])

#### Note:  p.CI.3yr is now an n x z+1 matrix of CI values.
####        The "+1" includes the identifier column (HOSP_ID).

# Remove hospitals from CI matrix
p.CI.3yr <- p.CI.3yr[-c(zero.pv),]

################################
## Read hospital attributes data
################################

#### Note:  hosp.info is a table with hospitals in rows and
####        attributes in columns. In this case, the column
####        that corresponds to the patient records is "MIDB".

# Load data
hosp.info <- read.csv("inputdata/hospitals.csv")

############################
## Read travel distance data
############################

#### Note:  od is a table with "TO", "FROM", and "DISTANCE"
####        as columns (format from ArcGIS Network Analyst).
####        This table must be re-arranged such that it is
####        an actual OD matrix (n x n dimensions). If data
####        is already arranged in an OD matrix, skip to
####        "Scale table" section.

# Load data
od <- read.csv("inputdata/travel-distance.csv")

############################
## Convert table to OD matrix
############################

# Create empty holder
```

```
dist.mat <- NULL

# Get unique FROM hospitals
f.hosp <- unique(od$FROM)

# Loop through hospitals
for (fr in 1:length(f.hosp)) {
  # Subset
  od.sm <- od[od$FROM == f.hosp[fr],]
  # Sort matrix, shouldn't be necessary... but safer
  od.sm <- od.sm[order(od.sm$TO),]
  # Append distance to holder as a ROW
  dist.mat <- rbind(dist.mat, od.sm$DISTANCE)
}

# Make into dataframe
dist.mat <- as.data.frame(dist.mat)

# Assign column names
names(dist.mat) <- f.hosp

# Assign row names
row.names(dist.mat) <- f.hosp

##########################################
## Scale table to match CI data range (0-1)
##########################################

# Get maximum distance between hospitals
max <- max(dist.mat)

# Rescale data
dist.mat <- dist.mat/max

##############################################
## Join distance data matrix to CI data matrix
##############################################

#### Note:  To create the final n x m data matrix used for
####        clustering, the n x z and n x n matrix are joined.

# Add column for table join
dist.mat$HOSP_ID <- row.names(dist.mat)

## Join tables, add distance matrix to CI data
```

```
p.CI.3yr <- merge(p.CI.3yr, dist.mat, by="HOSP_ID")




##############################################################################
##                                                                          ##
##  Custom 2-step K-means + Ward's clustering function                      ##
##                                                                          ##
##############################################################################

#### Note:  The inputs for the function are the n x m data
####        matrix (x) and the desired number of clusters (clusters).

kmeans.ward <- function(x, clusters) {

  # Create distance matrix
  d <- dist(x, "euclidean")

  # Perform Ward's clustering
  hc <- hclust(d, method="ward")

  # Get cluster members at "K" clusters
  memb <- cutree(hc, k = clusters)

  # Make empty holder for cluster center locations
  cent <- NULL

  # Get cluster centers
  for (k in 1:clusters) {
    cent <- rbind(cent, colMeans(x[memb == k,]))
  }

  # Use cluster centers from Ward's to seed K-means clustering
  k.m <- kmeans(x, cent, iter.max = 10000)

  # Return the K-means object
  return(k.m)

}


##############################################################################
##                                                                          ##
```

```
##  Create initial cluster solutions for Hospital Groups                ##
##                                                                      ##
##########################################################################

########################################
## Prepare data and create data holders
########################################

#### Note:  All possible numbers of clusters are considered
####        from 2 to n-1.

# Define the range of cluster solutions to evaluate (the set k)
cl.max <- nrow(p.CI.3yr)-1
clusters <- c(2:cl.max)

# Create an empty holder for cluster statistics
wss <- bss <- r2 <- incF <- SingHosp <- MaxSize <- rep(0, length(clusters))
k.data.pat <- cbind(clusters, wss, bss, r2, incF, SingHosp, MaxSize)

# Get number of data attributes in table (columns)
col.max <- ncol(p.CI.3yr)


#######################################################
## Conduct K-means + Ward's for all cluster solutions
#######################################################

for (K in 1:length(clusters)) {

  # Use K-means + Wards method to create clusters
  Kclust <- kmeans.ward(p.CI.3yr[,2:col.max], clusters[K])

  # Write cluster statistics to data holder
  # Within sum of squares
  k.data.pat[K,2] <- Kclust$tot.withinss

  # Between sum of squares
  k.data.pat[K,3] <- Kclust$betweenss

  # R^2
  k.data.pat[K,4] <- 1-(Kclust$tot.withinss/Kclust$totss)

  # Number of single hosp clusters
  table.c <- table(Kclust$cluster)
  k.data.pat[K,6] <- sum(table.c == 1)
```

```
  # Maximum size of any single cluster
  k.data.pat[K,7] <- max(table.c)

}

# Convert data holder to dataframe
k.data.pat <- as.data.frame(k.data.pat)

#################################
## Calculate incremental F scores
#################################

n.obs <- nrow(p.CI.3yr)

for (i in 2:length(clusters)) {

  k.data.pat$incF[i] <- ((k.data.pat$r2[i]-k.data.pat$r2[i-1])/
    (k.data.pat$clusters[i]-k.data.pat$clusters[i-1])) /
    ((1-k.data.pat$r2[i])/((n.obs)-(k.data.pat$clusters[i]-1)))

}


#############################################################################
##                                                                         ##
##  Select the number of Hospital Groups using the heuristic               ##
##                                                                         ##
#############################################################################


###########################################
## Find local maxima in incremental F scores
###########################################

# Make variable of the last candidate solution to evaluate for maxima
i <- cl.max-1



# Find the local maxima
incF.peaks <- which(k.data.pat$incF[3:i] > k.data.pat$incF[2:(i-1)] &
  k.data.pat$incF[3:i] > k.data.pat$incF[4:(i+1)])+2

# Subset initial candidate solutions
```

```
candidates <- k.data.pat[incF.peaks,]

#########################################################################
## Remove solutions wherein a single cluster has more than 20 hospitals
#########################################################################

candidates <- candidates[candidates$MaxSize < 20,]

#############################################################
## Subset solutions to those with the "minimum" number of
## single hospital Hospital Groups from remaining solutions
#############################################################

candidates <- candidates[candidates$SingHosp == min(candidates$SingHosp), ]

#####################################
## From the remaining solutions select
## the solution with most clusters, K
#####################################

solution <- candidates[candidates$clusters == max(candidates$clusters), ]

# Get number of clusters
n.clusters <- solution$clusters

#####################################################
## Use K-means + Wards method to re-create clusters
#####################################################

#### Note:  Only the cluster statistics were kept in the
####        initial clustering process. The final cluster
####        solution is recreated to extract Hospital Group
####        membership and cluster center information.
####        Because the clustering algorithm provides
####        deterministic results, this clustering
####        configuration will be identical to the one
####        formed in the intial clustering process.

HG.solution <- kmeans.ward(p.CI.3yr[,2:col.max], n.clusters)

# Attach Hospital Group number to MIDB name
HG.names <- as.data.frame(cbind(p.CI.3yr$HOSP_ID, HG.solution$cluster))
names(HG.names) <- c("MIDB", "HG")
```

```
##############################################################################
##                                                                          ##
##  Rename the Hospital Groups                                              ##
##                                                                          ##
##############################################################################


#### Note:  The K-means + Ward's names clusters using random
####        numbers. This section will re-enumerate the Hospital
####        Groups based on an existing larger regional group.
####        (HSA - Health Service Area) and the sum of the beds
####        in the Hospital Groups. This section can be omitted
####        if re-enumerating is not necessary.

####        This sections also requires that the hospital
####        information file (hosp.info) has columns named
####        "HSA" and "BEDS".


# Attach initial cluster number to hospital information table
hosp.info <- merge(hosp.info, HG.names, by="MIDB", all.x=TRUE)

# Convert cluster number column to character format
hosp.info$HG <- as.character(hosp.info$HG)

# If hospitals were removed becasue they didn not have patient
# records, assign them to "NG"
hosp.info$HG[is.na(hosp.info$HG)] <- "NG"

# For each Hospital Group, find the HSA where the max number of
# hospitals falls inside. These lines of code assumes that there is
# a column named "HSA" in the hospital information table (hosp.info)
HG.HSA <- NULL
for (hg in 1:n.clusters) {
  sub <- hosp.info$HSA[hosp.info$HG == hg]
  t.sub <- table(sub)
  HG.HSA <- c(HG.HSA, names(t.sub[t.sub == max(t.sub)]))
}

##############################################
## Rename Hospital Groups by HSA and bed count
##############################################

# Make holder
HG.NEW <- NULL
```

```
# Make counter variable. Will hold the "last" Hospital Group
# name assigned
max.hg <- 0

# Get number of "regions" (HSAs)
hsa.list <- as.numeric(sort(unique(HG.HSA)))

# Start looping through the regions
for (hsa in hsa.list) {

  # Get Hospital Groups in region
  hsa.hgs <- which(HG.HSA == hsa)

  # Subset hospital information file to only hospitals
  # in these HSAs
  sub.hosp.info <- hosp.info[hosp.info$HG %in% hsa.hgs,]

  # Aggregate the number of hospital beds in each
  # Hospital Group. Assumes there is a column in hosp.info
  # named "BEDS"
  bed.totals <- aggregate(sub.hosp.info$BEDS, by=list(HG = sub.hosp.info$HG), sum)

  # Reorder aggregated table by total number of beds
  bed.totals <- bed.totals[order(bed.totals$x, decreasing=TRUE), ]

  # Change column type to character
  bed.totals$HG <- as.character(bed.totals$HG)

  # Make numbers for first and last Hospital Group in
  # this subset. Uses counter.
  f.hg <- max.hg+1
  l.hg <- nrow(bed.totals)+max.hg

  # Make join table
  j.HG.names <- as.data.frame(cbind(bed.totals$HG, f.hg:l.hg))

  # Name columns
  names(j.HG.names) <- c("HG_O", "HG_N")

  # Append the holder table
  HG.NEW <- rbind(HG.NEW, j.HG.names)

  # Advance counter
  max.hg <- l.hg
```

```
}

#################################################################
## Attach new Hospital Group names to hospital information table
#################################################################

hosp.info <- merge(hosp.info, HG.NEW, by.x="HG", by.y="HG_O", all.x=TRUE)

# Remove old cluster numbers
hosp.info$HG <- NULL

# Rename Hospital Group column
col <- which(names(hosp.info) == "HG_N")
names(hosp.info)[col] <- "HG"


#########################################################################
##                                                                     ##
##  Assign a new hospital to existing Hospital Groups                  ##
##                                                                     ##
#########################################################################


#### Note:  This code requires a 1 x n vector of hospital distances
####        to assign a hospital to the existing Hospital Groups
####        based on location. Uses a Euclidean distance measure from
####        the new hospital to the existing cluster centers.

#####################################################
## Get original cluster centers from K-means object
#####################################################

HG.centers <- as.data.frame(HG.solution$centers)

# Subset to only "travel distance" attributes
HG.centers <- HG.centers[,n.zip:ncol(HG.centers)]

# Attach new cluster names to cluster centers
HG.centers$HG_O <- rownames(HG.centers)
HG.centers <- merge(HG.centers, HG.NEW, by="HG_O")

# Remove old names and re-sort data
HG.centers$HG_O <- NULL
rownames(HG.centers) <- HG.centers$HG_N)
```

```
HG.centers$HG_N <- NULL
HG.centers <- HG.centers[order(as.numeric(rownames(HG.centers))),]

##################################
## Get new hospital or observation
##################################

# Get travel distance for new observation
new.hosp.loc <- read.csv("inputdata/new.hospital.location.csv")

# Remove characters from column names
# R adds an "X" to the column names that
# are only numeric values.
# Assumes hospital name is 4 characters long.
fix.names <- which(nchar(names(new.hosp.loc)) > 4)
names(new.hosp.loc)[fix.names] <-
  substr(as.character(names(new.hosp.loc)[fix.names]), 2, 5)

# Test that columns match in new hospital and Hospital
# Group cluster centers
if (sum(names(new.hosp.loc) != names(HG.centers)) > 0)
  print("Columns do not match")

# Divide travel distances by the maximum travel distance
# between any hospitals in Michigan
new.hosp.loc <- new.hosp.loc / max

#####################################################
## Create function to measure Euclidean distance in
## n-dimensional space
#####################################################

euc.dist <- function(x1, x2) {
  dist <- sqrt(sum((x1-x2)^2))
  return(dist)
}

#########################################################
## Measure distance from new location to all existing
## Hospital Group centers
#########################################################

new.dists <- apply(HG.centers, 1, euc.dist, x2=new.hosp.loc)

# Get closest Hospital Group
```

```
HG.new.hosp.loc <- names(new.dists)[new.dists == min(new.dists)]

#### Note:  This is the Hospital Group that the new
####        hospital is assigned to

print(HG.new.hosp.loc)
```

Appendix C
Testimony– Blue Cross Blue Shield of Michigan/Blue Care Network

Testimony
Blue Cross Blue Shield of Michigan/Blue Care Network
CON Commission Meeting: Proposed Hospital Bed Standards
December 15, 2011

Thank you for the opportunity to provide testimony on behalf of Blue Cross Blue Shield of Michigan (BCBSM) and Blue Care Network (BCN). BCBSM/BCN supports the proposed hospital bed standards which have been submitted for Commission consideration by the Hospital Bed SAC. The proposed standards reflect months of deliberative discussions and ensure that the needs and realities of the health care marketplace in Michigan are the central tenet of the standards.

Hospital Group and Bed Need Methodolgy

The proposed methodologies developed by the workgroup and approved by the SAC were developed over a period of five months with the participation of multiple stakeholders and the assistance of the MSU Department of Geography. The workgroup focused on the goal of developing objective, replicable, and sustainable standards which could be utilized now and into the future.

The standards developed through the workgroup process accomplish these goals in the following manner:

- The proposed hospital group methodology groups hospitals based on location and utilization patterns. **This methodology will more logically group hospitals than the groupings provided by the existing methodology.**

The testimony's text formatting has been slighly modified to meet the required format for this dissertation. However, instances of boldface text are unchanged from the original document.

- The demand for bed need will be based on modeling of trends based on the previous five years of county-wide patient day data. The previous methodology relied on zip-code level data and often inaccurate population projections. **The proposed methodology will capture trends in patient day rates more effectively than the current methodology, will avoid the errors that are encountered when using small data sets, and will require the collection of dramatically less data.**

- According to MSU Geography, which has been contracted to run this data for the Department in previous years, **the methodologies "can be executed within a short time frame, using open-source code, and produces replicable results."**

When considering the tenets of cost, quality, and access, the proposed methodologies show that the current number of hospitals and hospital beds in the state are more than adequately serving the demands of Michigan's population. When run illustratively for the workgroup using 2009 MIDB data, the proposed methodologies found no areas of hospital bed need in the state and an overall excess of 6,747 hospital beds state-wide. Should patient population and utilization trends change in the future, the methodologies are equipped to reflect such changes.

Hospital Bed Reduction

BCBSM supports the proposals that emerged from the hospital bed reduction work goup as a valuable first step in addressing the excess bed capacity in Michigan's hospitals. The proposals adopted by the SAC will limit the financial incentive for hospitals to use large amounts of excess beds as a bargaining tool for their purchase. Additionally, the proposals will promote the development of capital projects that will be more reflective of a hospital's average occupancy, which could provide cost savings in the future. While BCBSM believes that the proposal is a step in the right direction, continued efforts must address excess hospital capacity on a larger scale in order to truly make a more significant impact on excess costs within the health care system.

Conclusion

BCBSM/BCN supports the Hospital Bed Standards recommended by the Hospital Bed SAC to the CON Commission. The thorough review of these standards over the past six months has resulted in significant improvements to the standards that will ensure appropriate hospital access and reflect the health care needs of the state's population for years to come.

12/15/11

Appendix D
Additional Figures and Tables

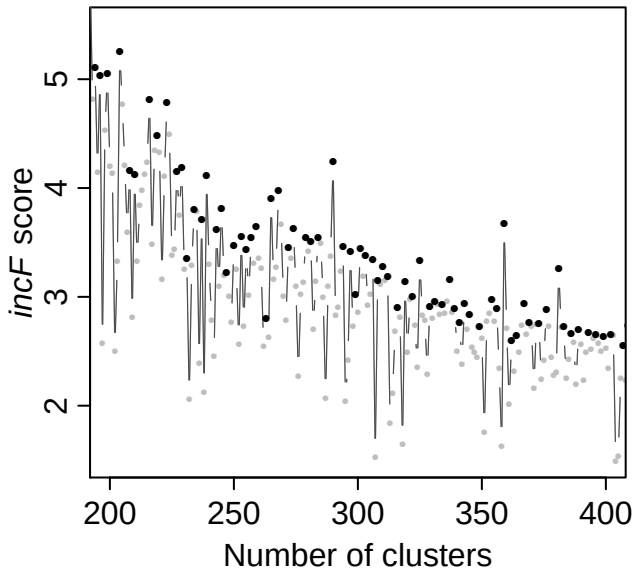

Figure D.1: ***incF* scores for cluster solutions in set S.** Black points represent peak values in *incF* scores. The data has been truncated for display purposes.

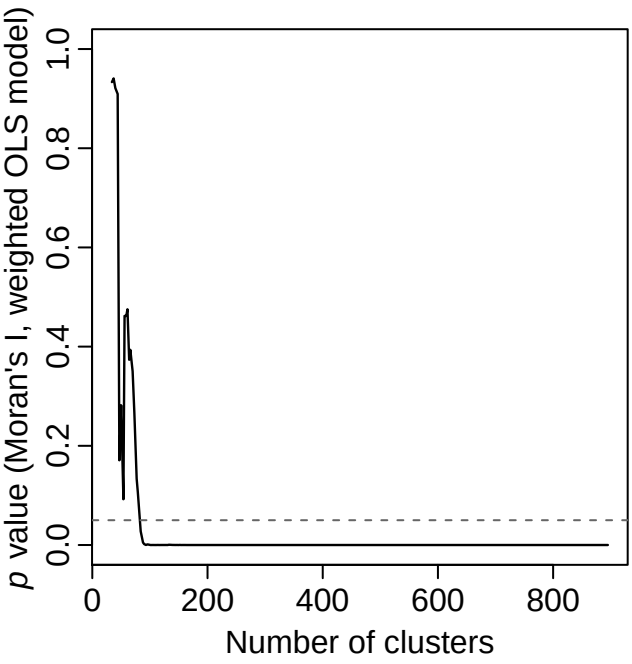

Figure D.2: **Moran's I of regression residuals for weighted OLS regression model.** All values less than 0.05 (dotted line) have significant spatial autocorrelation in the model residuals.

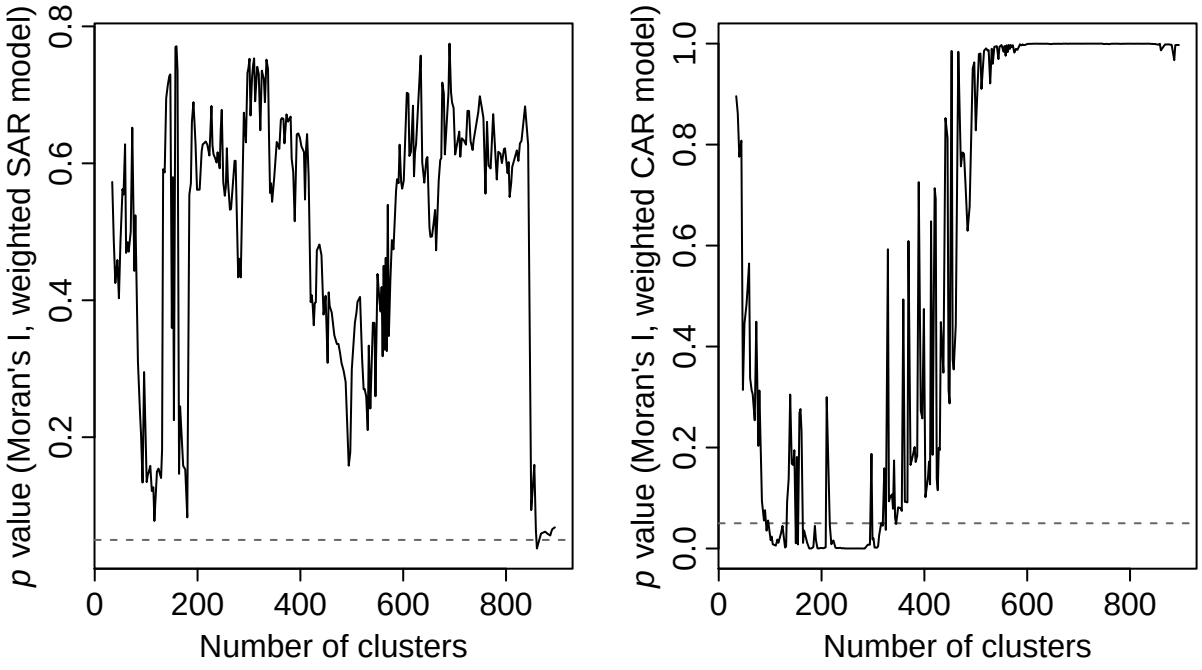

Figure D.3: **Moran's I of regression residuals for weighted SAR and CAR models.** All values less than 0.05 (dotted line) have significant spatial autocorrelation in the model residuals.

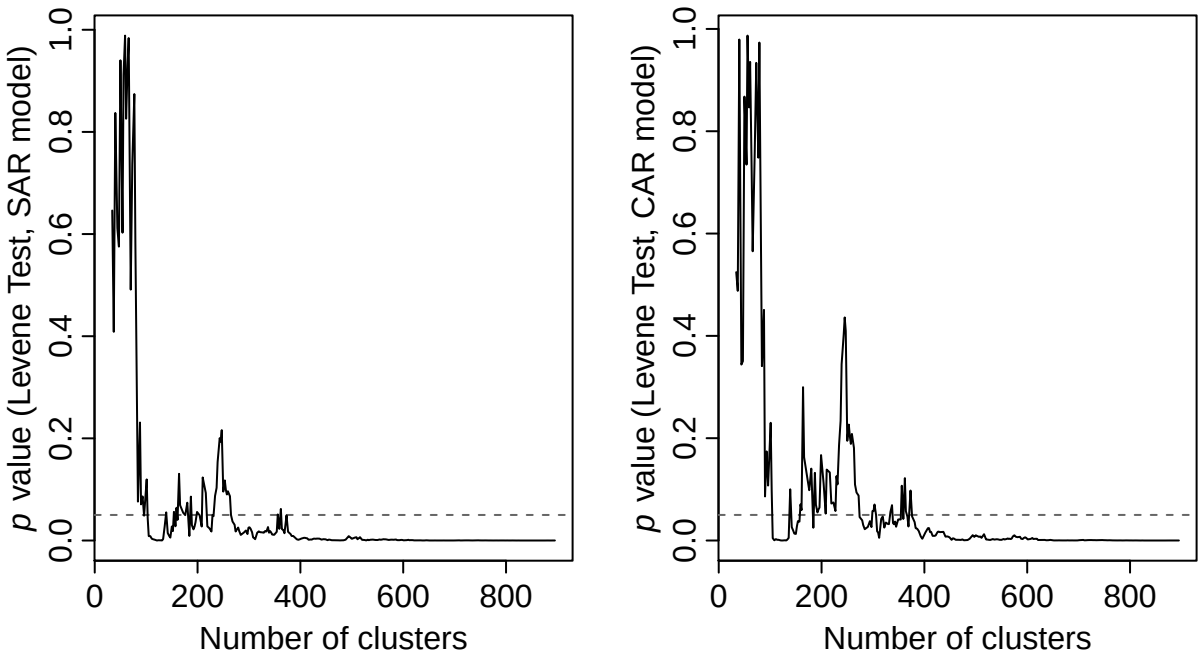

Figure D.4: **Levene Test of regression residuals for SAR and CAR models.** All values less than 0.05 (dotted line) have significant heteroscedasticity in the model residuals due to population size.

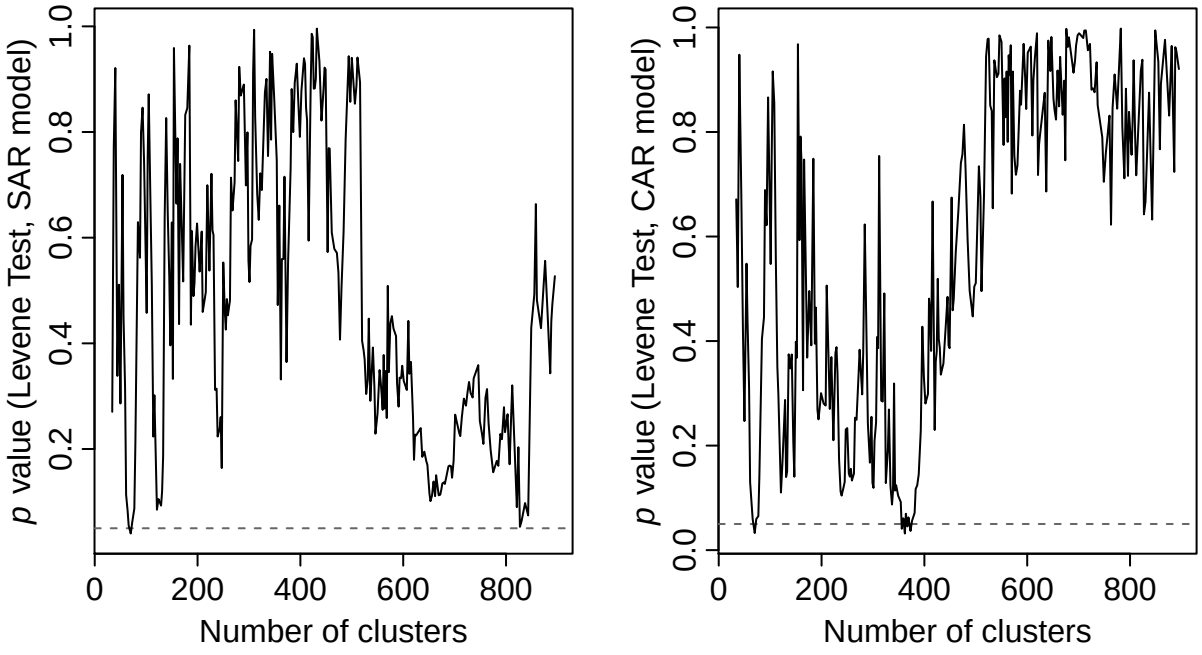

Figure D.5: **Levene Test of regression residuals for weighted SAR and CAR models.** All values less than 0.05 (dotted line) have significant heteroscedasticity in the model residuals due to population size.

Table D.1: **Cluster solutions and *incF* scores.**

CL	***incF***	**CL**	***incF***	**CL**	***incF***	**CL**	***incF***
8	173.588	219	4.478	420	2.574	634	1.950
13	117.188	223	4.785	422	2.525	637	1.936
17	57.137	227	4.155	424	2.514	641	1.968
19	46.520	229	4.191	426	2.470	645	1.957
21	41.105	231	3.352	428	2.446	647	1.928
23	35.267	234	3.802	430	2.422	651	1.915
27	28.022	237	3.709	432	2.421	653	1.896
29	27.952	239	4.119	437	2.399	656	1.890
32	21.648	243	3.621	441	2.409	659	1.876
34	32.543	245	3.812	445	2.388	662	1.869
37	22.762	247	3.220	447	2.343	664	1.901
40	24.243	250	3.472	449	2.370	669	1.887
44	24.131	253	3.558	453	2.367	672	1.882
47	22.427	255	3.436	455	2.367	674	1.865
50	17.827	257	3.541	457	2.362	676	1.838
54	17.783	259	3.644	461	2.584	679	1.839
56	23.635	263	2.805	466	2.379	681	1.850
59	15.538	265	3.905	471	2.305	688	1.787
61	17.436	268	3.972	474	2.347	690	1.787
64	10.838	272	3.450	477	2.350	693	1.779
66	18.027	274	3.630	480	2.330	695	1.771
70	11.505	279	3.544	484	2.287	698	1.769
73	20.162	281	3.508	488	2.295	701	1.773
77	13.521	284	3.550	494	2.287	709	1.799
79	14.299	290	4.243	497	2.284	711	1.784
84	8.937	294	3.461	500	2.287	714	1.789
88	9.763	297	3.419	506	2.408	718	1.806
90	12.296	299	3.022	509	2.042	722	1.803
93	8.184	301	3.446	511	2.206	725	1.765
96	8.975	303	3.384	516	2.270	728	1.770
101	10.237	306	3.340	520	2.112	731	1.789
103	8.573	308	3.149	523	2.193	735	1.734
105	8.618	310	3.277	525	2.203	737	1.699
108	9.590	312	3.186	528	2.317	746	1.752

Cont. on next page

Table D.1 – *Cont. from previous page*

CL	***incF***	**CL**	***incF***	**CL**	***incF***	**CL**	***incF***
111	6.621	316	2.903	531	2.176	749	1.683
114	10.148	319	3.137	533	2.164	753	1.681
116	9.488	322	3.003	536	2.156	756	1.697
121	7.840	325	3.333	541	2.139	760	1.679
124	6.302	329	2.908	544	2.148	763	1.680
129	10.391	331	2.958	546	2.636	766	1.682
131	7.376	334	2.929	550	2.133	770	1.680
133	8.357	337	3.159	554	2.146	775	1.670
136	7.806	339	2.895	556	2.150	782	1.672
139	6.168	341	2.760	558	2.153	785	1.735
142	5.989	343	2.942	560	1.958	789	1.658
145	6.427	345	2.834	562	2.149	792	1.654
147	6.223	349	2.725	564	2.182	796	1.597
150	5.386	354	2.978	566	2.177	798	1.593
152	6.834	356	2.891	568	2.151	803	1.562
154	6.671	359	3.673	570	2.101	805	1.564
157	6.621	362	2.597	572	2.152	807	1.533
159	5.148	364	2.648	575	2.150	810	1.528
161	6.301	367	2.936	578	2.143	812	1.589
164	6.079	369	2.762	581	2.191	821	1.544
166	5.838	373	2.758	586	2.096	824	1.521
172	7.055	376	2.884	588	2.098	827	1.532
176	6.552	381	3.257	591	2.097	830	1.480
180	6.088	383	2.725	593	2.170	837	1.515
184	5.825	386	2.659	596	2.063	843	1.484
187	4.415	389	2.697	598	2.110	849	1.463
189	4.086	393	2.673	601	2.078	855	1.415
192	6.036	396	2.651	607	2.080	858	1.361
194	5.112	399	2.638	610	2.054	860	1.373
196	5.035	402	2.658	612	1.916	868	1.418
199	5.050	407	2.551	615	2.038	876	1.546
204	5.255	409	2.736	619	2.029	881	1.600
208	4.161	411	2.895	621	2.037	886	1.751
210	4.120	413	3.031	623	2.098	888	1.764

Cont. on next page

Table D.1 – *Cont. from previous page*

CL	***incF***	**CL**	***incF***	**CL**	***incF***	**CL**	***incF***
216	4.818	416	2.443	625	2.057	890	1.665

Table D.2: **Number of components and % of the total variance explained for each functional set of variables.**

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
34	1	0.85	1	0.43	1	0.64	2	0.89	1	0.93
37	1	0.86	1	0.40	1	0.65	2	0.89	1	0.93
40	1	0.85	2	0.61	1	0.64	2	0.89	1	0.93
44	1	0.83	5	1.00	2	0.88	2	0.88	1	0.93
47	1	0.82	3	0.78	1	0.62	2	0.88	1	0.93
50	1	0.81	3	0.77	1	0.64	2	0.88	1	0.93
54	1	0.81	3	0.76	2	0.85	2	0.88	1	0.93
56	1	0.81	2	0.59	1	0.61	2	0.88	1	0.93
59	1	0.80	1	0.38	1	0.60	2	0.88	1	0.93
61	1	0.79	2	0.59	1	0.60	2	0.88	1	0.93
64	1	0.78	2	0.58	2	0.84	2	0.88	1	0.93
66	1	0.77	3	0.75	1	0.59	2	0.87	1	0.93
70	1	0.77	3	0.75	1	0.60	2	0.86	1	0.93
73	1	0.79	1	0.36	1	0.60	2	0.87	1	0.92
77	1	0.82	2	0.57	1	0.60	2	0.87	1	0.92
79	1	0.82	3	0.74	1	0.60	2	0.86	1	0.92
84	1	0.81	3	0.74	1	0.59	2	0.87	1	0.91
88	1	0.80	2	0.57	1	0.59	2	0.86	1	0.91
90	1	0.80	1	0.37	1	0.59	2	0.86	1	0.91
93	1	0.80	1	0.37	1	0.59	2	0.86	1	0.91
96	1	0.79	5	1.00	1	0.59	2	0.86	1	0.93
101	1	0.79	2	0.57	1	0.60	2	0.85	1	0.94
103	1	0.79	2	0.56	1	0.60	2	0.85	1	0.94
105	1	0.79	2	0.56	1	0.59	2	0.85	1	0.93
108	1	0.79	3	0.73	1	0.59	2	0.85	1	0.93
111	1	0.79	3	0.73	1	0.60	2	0.85	1	0.93
114	1	0.79	4	0.87	1	0.60	2	0.85	1	0.93
116	1	0.79	2	0.56	1	0.60	2	0.85	1	0.93
121	1	0.79	2	0.56	1	0.61	2	0.84	1	0.93
124	1	0.78	3	0.73	1	0.61	2	0.85	1	0.93
129	1	0.78	3	0.73	1	0.60	2	0.84	1	0.94
131	1	0.78	1	0.36	1	0.59	2	0.84	1	0.93

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
133	2	0.89	3	0.74	2	0.91	2	0.85	1	0.93
136	2	0.89	3	0.74	2	0.91	2	0.85	1	0.93
139	2	0.89	3	0.75	2	0.91	2	0.85	1	0.93
142	2	0.89	3	0.74	2	0.91	2	0.84	1	0.93
145	2	0.89	3	0.74	2	0.91	2	0.84	1	0.93
147	2	0.89	3	0.74	2	0.91	2	0.84	1	0.93
150	1	0.70	3	0.74	2	0.91	2	0.84	1	0.93
152	2	0.89	3	0.74	2	0.91	2	0.84	1	0.91
154	1	0.70	3	0.74	2	0.91	2	0.84	1	0.91
157	2	0.89	3	0.74	2	0.91	2	0.84	1	0.91
159	2	0.89	3	0.74	2	0.91	2	0.84	1	0.91
161	2	0.89	3	0.74	2	0.91	2	0.84	1	0.91
164	2	0.89	3	0.73	2	0.91	2	0.83	1	0.90
166	1	0.70	3	0.73	2	0.91	2	0.83	1	0.90
172	2	0.89	3	0.73	2	0.90	2	0.83	1	0.90
176	2	0.89	3	0.73	2	0.89	2	0.83	1	0.90
180	2	0.89	3	0.73	2	0.89	2	0.83	1	0.90
184	1	0.71	2	0.56	2	0.89	2	0.83	1	0.90
187	2	0.89	3	0.73	2	0.89	2	0.82	1	0.90
189	1	0.72	2	0.56	2	0.89	2	0.82	1	0.89
192	1	0.71	3	0.73	2	0.89	2	0.83	1	0.89
194	1	0.71	3	0.73	2	0.89	2	0.83	1	0.89
196	1	0.71	2	0.56	2	0.89	2	0.82	1	0.89
199	1	0.71	1	0.36	2	0.88	2	0.82	1	0.90
204	1	0.71	2	0.56	2	0.88	2	0.83	1	0.90
208	1	0.71	2	0.56	2	0.88	2	0.82	1	0.90
210	2	0.88	3	0.73	2	0.88	2	0.82	1	0.90
216	2	0.88	2	0.56	2	0.88	2	0.82	1	0.89
219	1	0.72	3	0.72	2	0.88	2	0.82	1	0.89
223	1	0.72	2	0.55	2	0.88	2	0.82	1	0.89
227	1	0.71	2	0.55	2	0.88	2	0.82	1	0.89
229	1	0.72	2	0.55	2	0.88	2	0.82	1	0.89
231	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
234	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88
237	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88
239	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88
243	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
245	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88
247	1	0.72	4	0.87	2	0.88	2	0.81	1	0.88
250	1	0.71	2	0.55	2	0.88	2	0.81	1	0.88
253	1	0.72	2	0.55	2	0.88	2	0.81	1	0.88
255	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
257	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
259	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
263	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
265	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
268	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
272	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
274	1	0.72	2	0.55	2	0.88	2	0.80	1	0.88
279	1	0.72	3	0.72	2	0.88	2	0.79	1	0.88
281	1	0.72	2	0.55	2	0.88	2	0.79	1	0.87
284	1	0.73	2	0.55	2	0.88	2	0.79	1	0.87
290	1	0.72	2	0.55	2	0.87	2	0.79	1	0.87
294	1	0.72	2	0.55	2	0.87	2	0.79	1	0.87
297	1	0.72	1	0.35	2	0.87	2	0.79	1	0.87
299	1	0.72	2	0.55	2	0.87	2	0.79	1	0.87
301	1	0.72	2	0.55	2	0.87	2	0.79	1	0.87
303	1	0.72	2	0.54	2	0.87	2	0.78	1	0.87
306	1	0.72	2	0.54	2	0.87	2	0.78	1	0.87
308	1	0.72	3	0.71	2	0.87	2	0.78	1	0.87
310	1	0.72	3	0.71	2	0.87	2	0.78	1	0.87
312	1	0.72	1	0.35	2	0.87	2	0.78	1	0.87
316	1	0.72	2	0.54	2	0.87	2	0.78	1	0.86
319	1	0.72	2	0.54	2	0.87	2	0.78	1	0.87
322	1	0.71	4	0.87	2	0.87	2	0.78	1	0.87
325	1	0.71	2	0.54	2	0.87	2	0.78	1	0.87

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
329	1	0.71	1	0.36	2	0.87	2	0.78	1	0.87
331	1	0.71	4	0.87	2	0.87	2	0.78	1	0.86
334	1	0.71	3	0.71	2	0.87	2	0.78	1	0.87
337	1	0.71	3	0.72	2	0.87	2	0.78	1	0.86
339	1	0.71	3	0.72	2	0.87	2	0.78	1	0.86
341	1	0.71	4	0.87	2	0.87	2	0.78	1	0.86
343	1	0.71	2	0.54	2	0.87	2	0.78	1	0.86
345	1	0.71	3	0.71	2	0.87	2	0.78	1	0.86
349	1	0.71	2	0.54	2	0.87	2	0.78	1	0.86
354	1	0.72	2	0.54	2	0.87	2	0.78	1	0.86
356	1	0.72	3	0.71	2	0.87	2	0.78	1	0.85
359	1	0.72	1	0.36	2	0.87	2	0.78	1	0.85
362	1	0.72	3	0.72	2	0.87	2	0.78	1	0.85
364	1	0.72	2	0.55	2	0.87	2	0.78	1	0.85
367	1	0.72	3	0.72	2	0.87	2	0.78	1	0.85
369	1	0.72	1	0.36	2	0.87	2	0.78	1	0.85
373	1	0.72	2	0.54	2	0.87	2	0.78	1	0.84
376	1	0.72	2	0.54	2	0.87	2	0.78	1	0.84
381	1	0.72	2	0.54	2	0.87	2	0.77	1	0.85
383	1	0.72	2	0.54	2	0.87	2	0.77	1	0.85
386	1	0.72	2	0.54	2	0.87	2	0.77	1	0.85
389	1	0.72	1	0.36	2	0.87	2	0.77	1	0.85
393	1	0.71	2	0.54	2	0.86	2	0.77	1	0.85
396	1	0.71	4	0.86	2	0.86	2	0.77	1	0.85
399	1	0.71	3	0.70	2	0.86	2	0.77	1	0.84
402	1	0.71	2	0.54	2	0.86	2	0.77	1	0.84
407	1	0.71	2	0.54	2	0.86	2	0.77	1	0.84
409	1	0.71	1	0.35	2	0.86	2	0.77	1	0.84
411	1	0.71	4	0.86	2	0.86	2	0.77	1	0.84
413	1	0.71	4	0.86	2	0.86	2	0.77	1	0.84
416	1	0.71	1	0.35	2	0.86	2	0.77	1	0.84
420	1	0.71	2	0.54	2	0.86	2	0.77	1	0.82
422	1	0.71	4	0.86	2	0.86	2	0.77	1	0.82

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
424	1	0.71	3	0.70	2	0.86	2	0.76	1	0.82
426	1	0.71	4	0.86	2	0.86	2	0.76	1	0.82
428	1	0.71	2	0.54	2	0.86	2	0.76	1	0.82
430	1	0.71	2	0.54	2	0.86	2	0.76	1	0.82
432	1	0.71	2	0.54	2	0.86	2	0.76	1	0.81
437	1	0.71	4	0.86	2	0.86	2	0.76	1	0.82
441	1	0.71	4	0.86	2	0.86	2	0.76	1	0.80
445	1	0.72	4	0.86	2	0.86	2	0.76	1	0.80
447	1	0.72	2	0.54	2	0.86	2	0.76	1	0.80
449	1	0.72	2	0.53	2	0.86	2	0.76	1	0.81
453	1	0.71	1	0.35	2	0.86	2	0.76	1	0.81
455	1	0.71	2	0.53	2	0.86	2	0.76	1	0.81
457	1	0.72	2	0.53	2	0.86	2	0.76	1	0.81
461	1	0.73	2	0.53	2	0.86	2	0.76	1	0.80
466	1	0.72	4	0.86	2	0.86	2	0.75	1	0.80
471	1	0.72	2	0.53	2	0.86	2	0.75	1	0.80
474	1	0.72	4	0.86	2	0.86	2	0.75	1	0.80
477	1	0.72	2	0.53	2	0.86	2	0.75	1	0.80
480	1	0.72	2	0.53	2	0.86	2	0.75	1	0.80
484	1	0.72	2	0.54	2	0.86	2	0.75	1	0.80
488	1	0.73	2	0.53	2	0.86	2	0.74	1	0.80
494	1	0.73	1	0.35	2	0.86	2	0.75	1	0.80
497	1	0.73	1	0.35	2	0.86	2	0.74	1	0.80
500	1	0.72	2	0.53	2	0.85	2	0.74	1	0.80
506	1	0.73	4	0.86	2	0.85	2	0.74	1	0.80
509	1	0.72	4	0.86	2	0.85	2	0.74	1	0.80
511	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
516	1	0.73	4	0.86	2	0.85	2	0.75	1	0.79
520	1	0.73	4	0.86	2	0.85	2	0.75	1	0.79
523	1	0.73	4	0.86	2	0.85	2	0.75	1	0.79
525	1	0.73	4	0.86	2	0.85	2	0.75	1	0.79
528	1	0.73	2	0.53	2	0.85	2	0.75	1	0.79
531	1	0.72	1	0.35	2	0.85	2	0.74	1	0.79

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
533	1	0.72	3	0.70	2	0.85	2	0.74	1	0.79
536	1	0.72	1	0.35	2	0.85	2	0.74	1	0.79
541	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
544	1	0.72	3	0.70	2	0.85	2	0.74	1	0.79
546	1	0.73	1	0.35	2	0.85	2	0.74	1	0.80
550	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
554	1	0.72	3	0.70	2	0.85	2	0.74	1	0.79
556	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
558	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
560	1	0.73	1	0.35	2	0.85	2	0.74	1	0.79
562	1	0.73	3	0.70	2	0.85	2	0.74	1	0.79
564	1	0.73	1	0.35	2	0.85	2	0.74	1	0.79
566	1	0.73	3	0.70	2	0.85	2	0.74	1	0.79
568	1	0.73	1	0.35	2	0.85	2	0.74	1	0.79
570	1	0.73	3	0.70	2	0.85	2	0.74	1	0.79
572	1	0.73	1	0.35	2	0.85	2	0.74	1	0.79
575	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
578	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
581	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
586	1	0.73	4	0.86	2	0.85	2	0.74	1	0.79
588	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
591	1	0.73	2	0.53	2	0.85	2	0.74	1	0.79
593	1	0.73	4	0.86	2	0.85	2	0.74	1	0.79
596	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
598	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
601	1	0.72	2	0.54	2	0.85	2	0.74	1	0.79
607	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
610	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
612	1	0.72	2	0.54	2	0.85	2	0.74	1	0.78
615	1	0.72	2	0.54	2	0.85	2	0.74	1	0.78
619	1	0.72	4	0.86	2	0.85	2	0.74	1	0.78
621	1	0.72	1	0.36	2	0.85	2	0.74	1	0.78
623	1	0.72	2	0.54	2	0.85	2	0.74	1	0.78

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
625	1	0.72	2	0.54	2	0.85	2	0.74	1	0.78
634	1	0.71	4	0.86	2	0.85	2	0.74	1	0.78
637	1	0.71	2	0.53	2	0.85	2	0.74	1	0.78
641	1	0.71	2	0.53	2	0.85	2	0.74	1	0.79
645	1	0.71	4	0.86	2	0.85	2	0.74	1	0.79
647	1	0.71	4	0.86	2	0.85	2	0.74	1	0.79
651	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
653	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
656	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
659	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
662	1	0.72	4	0.86	2	0.85	2	0.74	1	0.79
664	1	0.72	2	0.53	2	0.85	2	0.74	1	0.79
669	1	0.71	2	0.53	2	0.85	2	0.73	1	0.79
672	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
674	1	0.72	2	0.53	1	0.59	2	0.73	1	0.79
676	1	0.72	3	0.70	2	0.85	2	0.73	1	0.79
679	1	0.72	4	0.86	2	0.85	2	0.73	1	0.79
681	1	0.72	1	0.35	2	0.85	2	0.73	1	0.79
688	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
690	1	0.72	3	0.70	2	0.85	2	0.73	1	0.79
693	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
695	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
698	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
701	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
709	1	0.72	4	0.86	2	0.85	2	0.73	1	0.79
711	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
714	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
718	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
722	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
725	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
728	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
731	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
735	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		ETH		TRAN		MOB		CASE	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
737	1	0.72	2	0.53	2	0.85	2	0.73	1	0.79
746	1	0.73	2	0.53	2	0.85	2	0.73	1	0.79
749	1	0.73	4	0.86	2	0.85	2	0.73	1	0.79
753	1	0.73	4	0.86	2	0.85	2	0.73	1	0.79
756	1	0.73	4	0.86	2	0.85	2	0.73	1	0.79
760	1	0.73	4	0.86	2	0.84	2	0.73	1	0.80
763	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
766	1	0.73	1	0.35	2	0.84	2	0.73	1	0.80
770	1	0.73	1	0.35	2	0.84	2	0.73	1	0.80
775	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
782	1	0.73	4	0.86	2	0.84	2	0.73	1	0.80
785	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
789	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
792	1	0.73	4	0.86	2	0.84	2	0.73	1	0.80
796	1	0.73	3	0.70	2	0.84	2	0.73	1	0.80
798	1	0.73	3	0.70	2	0.84	2	0.73	1	0.80
803	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
805	1	0.73	2	0.53	2	0.84	2	0.73	1	0.80
807	1	0.73	4	0.86	2	0.84	2	0.73	1	0.80
810	1	0.74	3	0.70	2	0.84	2	0.73	1	0.80
812	1	0.74	3	0.70	2	0.84	2	0.73	1	0.80
821	1	0.74	3	0.70	2	0.84	2	0.73	1	0.80
824	1	0.74	4	0.86	2	0.84	2	0.73	1	0.80
827	1	0.74	4	0.86	2	0.84	2	0.73	1	0.80
830	1	0.74	4	0.86	2	0.84	2	0.73	1	0.80
837	1	0.74	2	0.53	2	0.84	2	0.73	1	0.80
843	1	0.74	4	0.86	2	0.85	2	0.73	1	0.80
849	1	0.74	2	0.53	2	0.85	2	0.73	1	0.80
855	1	0.74	4	0.86	2	0.85	2	0.73	1	0.81
858	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81
860	1	0.74	4	0.86	2	0.85	2	0.73	1	0.81
868	1	0.74	2	0.53	2	0.85	2	0.73	1	0.81
876	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81

Cont. on next page

Table D.2 – *Cont. from previous page*

	SES		**ETH**		**TRAN**		**MOB**		**CASE**	
CL	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**	**n**	**s**
881	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81
886	1	0.74	4	0.86	2	0.85	2	0.73	1	0.81
888	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81
890	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81
895	1	0.74	3	0.70	2	0.85	2	0.73	1	0.81

Appendix E
Additional R Code

```
#####################################################################
## R code to import ACS 5yr data and transfer/modify/figure out ##
## allocation issues to get data into ZIP code format           ##
#####################################################################

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)
library(spdep)

### Import Block Group data
acs.bg <- read.csv("/media/data/GISdata/acs2010/5yr/tables/
  ACS.5yr.IncEdMob.BlockGroup.csv")

### Remove leading characters from GEOID
acs.bg$GEOID <- substr(acs.bg$GEOID, 8, 19)

### First, import the Block Group shapefile and remove any BGs
### that do not have population (water!)

BG.proj.clip <- read.dbf("/media/data/GISdata/census_2010_data/
  block_groups/MI_2010_blockgroups_proj_clip.dbf")
BG.proj.clip <- BG.proj.clip$dbf

# Get unique GEOID10 for "good" Block Groups
BGs <- unique(BG.proj.clip$GEOID10)


### Subset ACS data to these Block Groups
which.bgs <- which(acs.bg$GEOID %in% BGs)
acs.bg <- acs.bg[which.bgs,]


### Read in Block Pop data
### Will use this to weight and allocate to
### block groups, tracts, and counties
bp.cent <- read.dbf("/media/data/GISdata/census_2010_data/blocks/
  MI_2010_blocks_proj_cent_pop/MI_2010_blocks_proj_cent_pop_gt0.dbf")
bp.cent <- bp.cent$db
```

```
bp.cent <- bp.cent[,c(1,17)]

bp.cent$GEOIDBG <- substr(bp.cent$GEOID, 1, 12)
bp.cent$GEOIDT <- substr(bp.cent$GEOIDBG, 1, 11)
bp.cent$GEOIDC <- substr(bp.cent$GEOIDBG, 1, 5)

### Aggregate by block group, tract, and county
BG.pop <- aggregate(bp.cent$POP100, by=list("GEOIDBG" = bp.cent$GEOIDBG),
  sum)
T.pop <- aggregate(bp.cent$POP100, by=list("GEOIDT" = bp.cent$GEOIDT),
  sum)
C.pop <- aggregate(bp.cent$POP100, by=list("GEOIDC" = bp.cent$GEOIDT),
  sum)

names(BG.pop)[2] <- names(T.pop)[2] <- names(C.pop)[2] <- "POP"

### Attach population to block groups
acs.bg <- merge(acs.bg, BG.pop, by.x="GEOID", by.y="GEOIDBG", all.x=TRUE)

sum(BG.pop$POP)
sum(acs.bg$POP, na.rm=TRUE)

### Remove any Block Groups with no pop
acs.bg <- acs.bg[!is.na(acs.bg$POP),]
sum(acs.bg$POP)


#########
### Convert counts to percentages
#########

acs.bg.pct <- acs.bg

### Education
acs.bg.pct[,11:15] <- acs.bg.pct[,11:15] / acs.bg.pct[,10]
#sum(acs.bg.pct[2,11:15])

### Mobility (Trav to work, trav time)
acs.bg.pct[,17:28] <- acs.bg.pct[,17:28] / acs.bg.pct[,16]
#sum(acs.bg.pct[1,17:21])
#sum(acs.bg.pct[1,22:28])


##########
### Import polygon file and get neighbors
```

```
##########

# Read in block group polygon file
bgs.poly <-  readOGR("/media/data/GISdata/census_2010_data/block_groups/
  MI_2010_blockgroups_proj_clip.shp", layer="MI_2010_blockgroups_proj_clip")

# Make neighbors list
## Use queen contiguity
bgs.nb <- poly2nb(bgs.poly, queen = TRUE)

na.MHI <- which(is.na(acs.bg.pct$MedHouInc))
na.ME <- which(is.na(acs.bg.pct$MedEarn16p))
na.E <- which(is.na(acs.bg.pct$EdPop25p) | acs.bg.pct$EdPop25p == 0)
na.T <- which(is.na(acs.bg.pct$TrvWorkPop16p) | acs.bg.pct$TrvWorkPop16p
  == 0)

### Get interpolation stats
#u.na.bgs <- unique(c(na.MHI, na.ME, na.E, na.T))
#length(u.na.bgs)
#nrow(acs.bg.pct)
#bg.na.pop <- sum(acs.bg.pct[u.na.bgs,ncol(acs.bg.pct)])
#bg.na.pop / sum(acs.bg$POP)
#length(u.na.bgs) / nrow(acs.bg.pct)

### Start looping through each Block Group that is missing data
### Median Household Income
for (i in 1:length(na.MHI)) {

  ## Get the Block Group ID
  na.id <- acs.bg.pct$GEOID[na.MHI[i]]

  ## Find which entry in the shapefile
  which.poly <- which(bgs.poly$GEOID10 == na.id)

  ## Get list of neighbors
  nbs <- bgs.poly$GEOID10[bgs.nb[which.poly][[1]]]

  ## Get the Median Houshold Income values of the neighbors
  ## Get the Populations of the neighbors
  nb.vals <- acs.bg.pct[acs.bg.pct$GEOID %in% nbs, c(1,8,29)]

  ## Get weighted average of neighbors
  mhi.interp <- sum(nb.vals$MedHouInc * nb.vals$POP, na.rm=TRUE) /
    sum(nb.vals$POP, na.rm=TRUE)
```

```
  ## Put interpolated value into table
  acs.bg.pct$MedHouInc[na.MHI[i]] <- mhi.interp

}

sum(is.na(acs.bg.pct$MedHouInc))


### Start looping through each Block Group that is missing data
### Median Earnings
for (i in 1:length(na.ME)) {

  ## Get the Block Group ID
  na.id <- acs.bg.pct$GEOID[na.ME[i]]

  ## Find which entry in the shapefile
  which.poly <- which(bgs.poly$GEOID10 == na.id)

  ## Get list of neighbors
  nbs <- bgs.poly$GEOID10[bgs.nb[which.poly][[1]]]

  ## Get the Median Houshold Income values of the neighbors
  ## Get the Populations of the neighbors
  nb.vals <- acs.bg.pct[acs.bg.pct$GEOID %in% nbs, c(1,9,29)]

  ## Get weighted average of neighbors
  mhi.interp <- sum(nb.vals$MedEarn16p * nb.vals$POP, na.rm=TRUE) /
    sum(nb.vals$POP, na.rm=TRUE)

  ## Put interpolated value into table
  acs.bg.pct$MedEarn16p[na.ME[i]] <- mhi.interp

}

which(is.na(acs.bg.pct$MedEarn16p))


### Start looping through each Block Group that is missing data
### Median Household Income
for (i in 1:length(na.MHI)) {

  ## Get the Block Group ID
  na.id <- acs.bg.pct$GEOID[na.MHI[i]]

  ## Find which entry in the shapefile
```

```
  which.poly <- which(bgs.poly$GEOID10 == na.id)

  ## Get list of neighbors
  nbs <- bgs.poly$GEOID10[bgs.nb[which.poly][[1]]]

  ## Get the Median Houshold Income values of the neighbors
  ## Get the Populations of the neighbors
  nb.vals <- acs.bg.pct[acs.bg.pct$GEOID %in% nbs, c(1,8,29)]

  ## Get weighted average of neighbors
  mhi.interp <- sum(nb.vals$MedHouInc * nb.vals$POP, na.rm=TRUE) /
    sum(nb.vals$POP, na.rm=TRUE)

  ## Put interpolated value into table
  acs.bg.pct$MedHouInc[na.MHI[i]] <- mhi.interp

}

sum(is.na(acs.bg.pct$MedHouInc))

### Start looping through each Block Group that is missing data
### Education
for (i in 1:length(na.E)) {

  ## Get the Block Group ID
  na.id <- acs.bg.pct$GEOID[na.E[i]]

  ## Find which entry in the shapefile
  which.poly <- which(bgs.poly$GEOID10 == na.id)

  ## Get list of neighbors
  nbs <- bgs.poly$GEOID10[bgs.nb[which.poly][[1]]]

  ## Get the Median Houshold Income values of the neighbors
  ## Get the Populations of the neighbors
  nb.vals <- acs.bg.pct[acs.bg.pct$GEOID %in% nbs, c(1,10:15)]

  ## Get weighted average of neighbors
  mhi.interp <- colSums(nb.vals[,3:7] * nb.vals$EdPop25p, na.rm=TRUE) /
    sum(nb.vals$EdPop25p, na.rm=TRUE)

  ## Put interpolated value into table
  acs.bg.pct[na.E[i],11:15] <- mhi.interp

}
```

```
which(is.na(acs.bg.pct$EdltHS))

### Start looping through each Block Group that is missing data
### Mobility
for (i in 1:length(na.T)) {

  ## Get the Block Group ID
  na.id <- acs.bg.pct$GEOID[na.T[i]]

  ## Find which entry in the shapefile
  which.poly <- which(bgs.poly$GEOID10 == na.id)

  ## Get list of neighbors
  nbs <- bgs.poly$GEOID10[bgs.nb[which.poly][[1]]]

  ## Get the Median Houshold Income values of the neighbors
  ## Get the Populations of the neighbors
  nb.vals <- acs.bg.pct[acs.bg.pct$GEOID %in% nbs, c(1,16:28)]

  ## Get weighted average of neighbors
  mhi.interp <- colSums(nb.vals[,3:14] * nb.vals$TrvWorkPop16p, na.rm=TRUE) /
    sum(nb.vals$TrvWorkPop16p, na.rm=TRUE)

  ## Put interpolated value into table
  acs.bg.pct[na.T[i],17:28] <- mhi.interp

}

which(is.na(acs.bg.pct$TrvCar))


### Write out table
write.csv(acs.bg.pct, file="/media/data/GISdata/acs2010/5yr/tables/
  ACS.5yr.IncEdMob.BlockGroup.Interpolated.csv", row.names=FALSE)

#####################################################################
#####################################################################



#############################################################
## R Code to aggregate county data from SAHIE by Zip Code ##
## Use block / zip percentages because of nesting         ##
#############################################################
```

```
options(scipen=500)


## Get joined block / zip code : age proportions table
zip.blk.age <- read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/
  tables/zipcode_blocks_age_breakdown.csv")

sum(zip.blk.age$SumAgePop)

## Make County ID column
zip.blk.age$GEOIDC <- substr(zip.blk.age$GEOID10, 1, 5)

## Aggregate to Counties
zip.c <- aggregate(zip.blk.age[,c(5:22,24)], by=list("GEOIDC" =
  zip.blk.age$GEOIDC, "ZIP" = zip.blk.age$ZIP), sum)
sum(zip.c$SumAgePop)

## Create similar age brackets
zip.c$P0_19 <- rowSums(zip.c[,3:6])
zip.c$P20_64 <- rowSums(zip.c[,7:15])
zip.c$P0_64 <- rowSums(zip.c[,3:15])
zip.c$P65p <- rowSums(zip.c[,16:20])

## Subset
zip.c <- zip.c[,c(1:2,21:25)]

sum(zip.c$SumAgePop)

##
## First, sum Zip Code 0 - 64 populations
zip.pop <- aggregate(zip.c[,6], by=list("ZIP" = zip.c$ZIP), sum)
names(zip.pop)[2] <- "POP064ZIP"


## Merge
zip.c <- merge(zip.c, zip.pop, by="ZIP", all.x=TRUE)

## Create County :: Zip percentage
zip.c$CPCT064 <- zip.c$P0_64 / zip.c$POP064ZIP


## Get SAHIE data table
sahie.c <- read.csv("/media/data/GISdata/census_SAHIE/2009/tables/
  sahie.county.health.insurance.estimates.csv")
```

```
## Remove error columns
sahie.c <- sahie.c[,c(1:4,6,8:9,11,13:14,16)]

## Fix 0-18, 18-64 :: Make into 0-18, 19-64
sahie.c$POP_19_64 <- sahie.c$POP_18_64 -
  ((sahie.c$POP_0_18 + sahie.c$POP_18_64) - sahie.c$POP)
sahie.c$INSPOP_19_64 <- sahie.c$INSPOP_18_64 -
  ((sahie.c$INSPOP_0_18 + sahie.c$INSPOP_18_64) - sahie.c$INSPOP)
sahie.c$POP_19_64<- sahie.c$POP_18_64 -
  ((sahie.c$POP_0_18 + sahie.c$POP_18_64) - sahie.c$POP)
sahie.c$UNINSPOP_19_64 <- sahie.c$POP_19_64 - sahie.c$INSPOP_19_64


## Convert SAHIEs to Percentages
sahie.c.o <- sahie.c
sahie.c[,4:5] <- sahie.c[,4:5] / sahie.c$POP
sahie.c[,7:8] <- sahie.c[,7:8] / sahie.c$POP_0_18
sahie.c[,10:11] <- sahie.c[,10:11] / sahie.c$POP_18_64
sahie.c[,13:14] <- sahie.c[,13:14] / sahie.c$POP_19_64


## Get COUNTY pop data
c.age <-  read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/
  tables/county_AGE_blockpop_adj.csv")

## Aggregate to similar age brackets
c.age$P0_19 <- rowSums(c.age[,2:5])
c.age$P20_64 <- rowSums(c.age[,6:14])
c.age$P0_64 <- rowSums(c.age[,2:14])
c.age$P65p <- rowSums(c.age[,15:19])
c.age <- c.age[,c(1,20:24)]


######################
### Use the weights to allocate COUNTY values to ZIP CODES!!
######################

zip.sahie <- NULL

u.zips <- unique(zip.c$ZIP)

for (i in 1:length(u.zips)) {

  ## Get weights
```

```
  zip.c.weights <- zip.c[zip.c$ZIP == u.zips[i], c(1,2,9)]

  ## Get SAHIE County values
  sahie.c.sub <- sahie.c[sahie.c$COUNTY %in%
    as.numeric(substr(zip.c.weights$GEOIDC, 3, 5)), c(1,2,4,7,13)]

  ## Sort each
  zip.c.weights <- zip.c.weights[order(zip.c.weights$GEOIDC),]
  sahie.c.sub <- sahie.c.sub[order(sahie.c.sub$COUNTY),]

  # if(sum(zip.bg.weights$GEOIDBG != acs.bg.sub$GEOID) > 0)
    print(paste("Something went wrong at i = ", i, sep=""))

  ## Multiply by weights and SUM
  w.sahie <- sum(sahie.c.sub$INSPOP * zip.c.weights[,3])

  ## Now, figure out overall insured rate... assuming all 65+ are insured
  zip.o <- zip.c[zip.c$ZIP == u.zips[i], c(1,2,6,7)]
  zip.64u <- sum(zip.o$P0_64 * sahie.c.sub$INSPOP)
  zip.65p <- sum(zip.o$P65p * 1)
  ov.ins.rate <- (zip.64u + zip.65p) / (sum(zip.o$P0_64) + sum(zip.o$P65p))

  ## Now, get overall insured rate... and 0-64 rate using the separate rates
  zip.o <- zip.c[zip.c$ZIP == u.zips[i], c(1,2,4,5,6,7)]
  zip.0.19 <- sum(zip.o$P0_19 * sahie.c.sub$INSPOP_0_18)
  zip.20.64 <- sum(zip.o$P20_64 * sahie.c.sub$INSPOP_19_64)
  zip.65p <- sum(zip.o$P65p * 1)
  interp.ins.rate <- (zip.0.19 + zip.20.64 + zip.65p) / (sum(zip.o$P0_19)
    + sum(zip.o$P20_64) + sum(zip.o$P65p))

  interp.ins.rate.64 <- (zip.0.19 + zip.20.64) / (sum(zip.o$P0_19)
    + sum(zip.o$P20_64))

  ## Put in holder
  zip.sahie <- rbind(zip.sahie, c(u.zips[i], w.sahie, ov.ins.rate,
    interp.ins.rate, interp.ins.rate.64))

}

class(zip.sahie)
zip.sahie <- as.data.frame(zip.sahie)
names(zip.sahie) <- c("ZIP", "HeaInsRate0_64", "HeaInsRateTot",
  "HeaInsRateIntTot", "HeaInsRateInt0_64")
```

```
### Write out file
write.csv(zip.sahie, file="/home/delamate/MDCH/data/dissertation/
  zipcodes/tables/zipcode_SAHIE_HealthInsurance.csv", row.names=FALSE)

#################################################################
#################################################################



#################################################################
## R Code to subset patient days to only those spent     ##
## in a hospital < 60 minutes travel time from residence ##
## This matches specification in E2SFCA calculation      ##
#################################################################

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)

## Get records
records <- read.csv(file="2010/2010MIDBrecords.csv")

## Get Zip / Hosp distances
## Get OD matrix for Zip Codes
## Read in origin-destination shapefile table
od <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/OD/hosps_esri2010pwcent_traveltime_od.dbf")
## Remove header info
od <- od$dbf
## Subset
od <- od[,c(7,8,5)]

# Remove weird point
# Weird point is Block centroid that didn't fall in any Zip Code!
badID <- which(od$FAC_ID %in% unique(od$FAC_ID)[170:338])
od <- od[-c(badID),]

## Get hospital info file and attach FAC_ID
hosp.info <- read.csv("/media/data/GISdata/hospitals/csv/
  2011-hosps-beds.csv")
od <- merge(od, hosp.info[,c(1,4)], by="FAC_ID", all.x=TRUE)

## Aggregate distances by ZIPCODE / MIDB
od <- aggregate(od$Total_MinT, by=list("ZIPCODE" = od$ZIPCODE,
```

```
  "MIDB" = od$MIDB), mean)
names(od)[3] <- "MIN"

length(unique(od$ZIPCODE))
length(unique(od$MIDB))


### Attach travel time to patient records
names(records)[c(2,4)] <- c("MIDB", "ZIPCODE")
records <- merge(records, od, by=c("ZIPCODE", "MIDB"), all.x=TRUE)


### Subset
records.60 <- records[records$MIN <= 60, ]
no.dist <- which(is.na(records.60$LOS))
records.60 <- records.60[-no.dist,]
records <- records.60

#write.csv(records, file="2010/2010MIDBrecords.60min.csv",
  row.names=FALSE)

### Aggregate by age!
### Bin into AGE categories
records0004 <- records[as.numeric(as.character(records$AGE)) < 5, ]
records0509 <- records[as.numeric(as.character(records$AGE)) >= 5 &
  as.numeric(as.character(records$AGE)) < 10, ]
records1014 <- records[as.numeric(as.character(records$AGE)) >= 10 &
  as.numeric(as.character(records$AGE)) < 15, ]
records1519 <- records[as.numeric(as.character(records$AGE)) >= 15 &
  as.numeric(as.character(records$AGE)) < 20, ]
records2024 <- records[as.numeric(as.character(records$AGE)) >= 20 &
  as.numeric(as.character(records$AGE)) < 25, ]
records2529 <- records[as.numeric(as.character(records$AGE)) >= 25 &
  as.numeric(as.character(records$AGE)) < 30, ]
records3034 <- records[as.numeric(as.character(records$AGE)) >= 30 &
  as.numeric(as.character(records$AGE)) < 35, ]
records3539 <- records[as.numeric(as.character(records$AGE)) >= 35 &
  as.numeric(as.character(records$AGE)) < 40, ]
records4044 <- records[as.numeric(as.character(records$AGE)) >= 40 &
  as.numeric(as.character(records$AGE)) < 45, ]
records4549 <- records[as.numeric(as.character(records$AGE)) >= 45 &
  as.numeric(as.character(records$AGE)) < 50, ]
records5054 <- records[as.numeric(as.character(records$AGE)) >= 50 &
  as.numeric(as.character(records$AGE)) < 55, ]
records5559 <- records[as.numeric(as.character(records$AGE)) >= 55 &
```

```
  as.numeric(as.character(records$AGE)) < 60, ]
records6064 <- records[as.numeric(as.character(records$AGE)) >= 60 &
  as.numeric(as.character(records$AGE)) < 65, ]
records6569 <- records[as.numeric(as.character(records$AGE)) >= 65 &
  as.numeric(as.character(records$AGE)) < 70, ]
records7074 <- records[as.numeric(as.character(records$AGE)) >= 70 &
  as.numeric(as.character(records$AGE)) < 75, ]
records7579 <- records[as.numeric(as.character(records$AGE)) >= 75 &
  as.numeric(as.character(records$AGE)) < 80, ]
records8084 <- records[as.numeric(as.character(records$AGE)) >= 80 &
  as.numeric(as.character(records$AGE)) < 85, ]
records85p <- records[as.numeric(as.character(records$AGE)) >= 85, ]

records0004 <- aggregate(records0004[,4], by=list(ZIP =
  records0004$ZIPCODE), FUN=sum)
records0509 <- aggregate(records0509[,4], by=list(ZIP =
  records0509$ZIPCODE), FUN=sum)
records1014 <- aggregate(records1014[,4], by=list(ZIP =
  records1014$ZIPCODE), FUN=sum)
records1519 <- aggregate(records1519[,4], by=list(ZIP =
  records1519$ZIPCODE), FUN=sum)
records2024 <- aggregate(records2024[,4], by=list(ZIP =
  records2024$ZIPCODE), FUN=sum)
records2529 <- aggregate(records2529[,4], by=list(ZIP =
  records2529$ZIPCODE), FUN=sum)
records3034 <- aggregate(records3034[,4], by=list(ZIP =
  records3034$ZIPCODE), FUN=sum)
records3539 <- aggregate(records3539[,4], by=list(ZIP =
  records3539$ZIPCODE), FUN=sum)
records4044 <- aggregate(records4044[,4], by=list(ZIP =
  records4044$ZIPCODE), FUN=sum)
records4549 <- aggregate(records4549[,4], by=list(ZIP =
  records4549$ZIPCODE), FUN=sum)
records5054 <- aggregate(records5054[,4], by=list(ZIP =
  records5054$ZIPCODE), FUN=sum)
records5559 <- aggregate(records5559[,4], by=list(ZIP =
  records5559$ZIPCODE), FUN=sum)
records6064 <- aggregate(records6064[,4], by=list(ZIP =
  records6064$ZIPCODE), FUN=sum)
records6569 <- aggregate(records6569[,4], by=list(ZIP =
  records6569$ZIPCODE), FUN=sum)
records7074 <- aggregate(records7074[,4], by=list(ZIP =
  records7074$ZIPCODE), FUN=sum)
records7579 <- aggregate(records7579[,4], by=list(ZIP =
  records7579$ZIPCODE), FUN=sum)
```

```
records8084 <- aggregate(records8084[,4], by=list(ZIP =
  records8084$ZIPCODE), FUN=sum)
records85p <- aggregate(records85p[,4], by=list(ZIP =
  records85p$ZIPCODE), FUN=sum)

## Get Zip Code age breakdown
zip.age <- read.csv(file="/home/delamate/MDCH/data/dissertation/
  zipcodes/tables/zipcode_AGE_blockpop_adj.csv")

zip.days <- merge(records0004, records0509, by="ZIP", all.x=TRUE,
  all.y=TRUE)
zip.days <- merge(zip.days, records1014, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]
zip.days <- merge(zip.days, records1519, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records2024, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records2529, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]
zip.days <- merge(zip.days, records3034, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records3539, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records4044, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]
zip.days <- merge(zip.days, records4549, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records5054, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records5559, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]
zip.days <- merge(zip.days, records6064, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records6569, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records7074, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]
zip.days <- merge(zip.days, records7579, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records8084, by="ZIP", all.x=TRUE, all.y=TRUE)
zip.days <- merge(zip.days, records85p, by="ZIP", all.x=TRUE, all.y=TRUE)
names(zip.days)[2:ncol(zip.days)] <- names(zip.age)[2:ncol(zip.days)]

zip.days[is.na(zip.days)] <- 0

names(zip.days)

### Clean up file to account for funky Zip Codes ##################
### See dissertation data notes ###################################

### Remove islands
#islands <- which(zip.days$ZIP %in% c(48028, 49726, 49757, 49775, 49782))
#airport <- which(zip.days$ZIP == 48242)
#zip.days <- zip.days[-c(islands, airport), ]
```

```
### Reassign zips
zip.days$ZIP[zip.days$ZIP == 48710] <- 48706
zip.days$ZIP[zip.days$ZIP == 48743] <- 48739
zip.days$ZIP[zip.days$ZIP == 48824 | zip.days$ZIP == 48825] <- 48823
zip.days$ZIP[zip.days$ZIP == 49104] <- 49103
zip.days$ZIP[zip.days$ZIP == 49792] <- 49765
zip.days$ZIP[zip.days$ZIP == 49873] <- 49807

zip.days <- aggregate(zip.days[,2:ncol(zip.days)], by=list("ZIP" =
  zip.days$ZIP), sum)
dim(zip.days)

### 3 Zip Codes have NO patient days within 60 minutes!!
### 892 records... add zero entries for these  49725, 49858, 49893
zip.days <- rbind(zip.days, c(49725, rep(0,18)))
zip.days <- rbind(zip.days, c(49858, rep(0,18)))
zip.days <- rbind(zip.days, c(49893, rep(0,18)))
zip.days <- zip.days[order(zip.days$ZIP), ]
dim(zip.days)

zip.days$TotDays <- rowSums(zip.days[2:19])

## Write records to file
write.csv(zip.days, file="/home/delamate/MDCH/data/dissertation/
  zipcodes/tables/zipcode_2010_UTILIZATION_age_breakdown.60min.csv",
  row.names=FALSE)

#############################################################
#############################################################



#############################################################
##  R Code to estimate distance decay in a gravity model ##
##  Uses MIDB patient days and network GIS travel data   ##
#############################################################

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)
library(classInt)

## Get table of patient days
```

```
pd <- read.csv("/home/delamate/MDCH/data/dissertation/utilization/2010/
  2010MIDBrecords.csv")

## Subset to needed data
pd <- pd[,c(2,4,3)]

## Get travel distance data
## Get OD matrix for Zip Codes
## Read in origin-destination shapefile table
od <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/OD/hosps_esri2010pwcent_traveltime_od.dbf")
## Remove header info
od <- od$dbf
## Subset
od <- od[,c(7,8,5)]

# Get hosp info and attach MIDB number
hosp.info <- read.csv("/media/data/GISdata/hospitals/csv/
  2011-hosps-beds.csv")

# Attach to OD matrix
od <- merge(od, hosp.info[,c(1,4)], by="FAC_ID", all.x=TRUE)

## Get mean of distances to grouped hospitals
od <- aggregate(od$Total_MinT, by=list("ZIPCODE" = od$ZIPCODE, "MIDB" =
  od$MIDB), mean)
names(od)[c(1,3)] <- c("ZIP", "Min")


## Get total number of beds per MIDB hospital (SUPPLY)
n.beds <- aggregate(hosp.info$BEDS2010, by=list(MIDB = hosp.info$MIDB),
  sum)
names(n.beds)[2] <- "BEDS"

## Get total number of patient days per MIDB hospital (DEMAND)
n.patient.days <- aggregate(pd$LOS, by=list(MIDB = pd$HOSP_ID), sum)
names(n.patient.days)[2] <- "PDj"

## Get total number of patient days per ZIP
n.pat.days.zip <- aggregate(pd$LOS, by=list(ZIP = pd$MIDB_ZIP), sum)
names(n.pat.days.zip)[2] <- "PDi"

## Get patient flows from each Zip to each hospital
zip.pd <- aggregate(pd$LOS, by=list(ZIP = pd$MIDB_ZIP, MIDB = pd$HOSP_ID),
  sum)
```

```
names(zip.pd)[3] <- "PDij"


## Aggregate data into single table
data <- zip.pd
data <- merge(data, n.beds, by="MIDB", all.x=TRUE)
data <- merge(data, n.patient.days, by="MIDB", all.x=TRUE)
data <- merge(data, n.pat.days.zip, by="ZIP", all.x=TRUE)
data <- merge(data, od, by=c("ZIP", "MIDB"), all.x=TRUE)

####
#### Distance decay function is cumulative by distance traveled!!
#### It's a probability that people travel X dist or less
####

#### Get total days
## Remove "NAs" in distance
nas <- which(is.na(data$Min))
data <- data[-nas,]
t.pat.days <- sum(data$PDij)

#### Bin probability by Distance
max.d <- max(data$Min)
dist.bin <- 0:floor(max.d)

#### Make holder
Prob <- as.data.frame(cbind(dist.bin, rep(0, length(dist.bin))))
names(Prob) <- c("Min", "Prob")

#### Loop through and get probabilities
for (d in dist.bin) {

  ## Get records greater than or equal to distance bin
  sub.data <- data[data$Min >= d,]

  ## Get patient days
  d.pat.days <- sum(sub.data$PDij)

  ## Get probability
  Prob$Prob[d+1] <- d.pat.days / t.pat.days

}


nlc <- nls.control(maxiter = 1000000)
```

```
nlsfit.d.log.logistic <- nls(Prob ~ 1 / (1 + (Min/a)^b), data=Prob,
  start=list(a=1, b=1), control=nlc, trace=TRUE)
summary(nlsfit.d.log.logistic)
a <- summary(nlsfit.d.log.logistic)$coefficients[1,1]
b <- summary(nlsfit.d.log.logistic)$coefficients[2,1]


## Write out data for making figures
fig.dat <- as.data.frame(cbind(Prob, fitted(nlsfit.d.log.logistic)))
names(fig.dat) <- c("Min", "Util", "dll")
write.csv(fig.dat, file="distance.decay.utilization.csv", row.names=FALSE)

## Write out weights table
weight.table <- as.data.frame(cbind(0:700, 1 / (1 + ((0:700)/a)^b)))
names(weight.table) <- c("Min", "dllWgt")
write.csv(weight.table, file="/home/delamate/MDCH/data/dissertation/
  Distance.decay/decreasing.log.likelihood.weights.empirical.2010.csv",
  row.names=FALSE)

#############################################################
#############################################################



#####################################################
## Code to calculate demand for E2SFCA           ##
## Get poplulation and allocate based on weights ##
#####################################################

options(scipen=999)

## Get block / ring data
blcks <- read.csv("/home/delamate/MDCH/data/dissertation/E2SFCA/tables/
  block.centroid.rings.csv")

## Get 2010 weights
weights <- read.csv("/home/delamate/MDCH/data/dissertation/
  Distance.decay/decreasing.log.likelihood.weights.empirical.2010.csv")

## Convert aggregate weights to "ring" structure

w.5 <- mean(weights$dllWgt[1:6])
w.10 <- mean(weights$dllWgt[6:11])
w.15 <- mean(weights$dllWgt[11:16])
```

```
w.20 <- mean(weights$dllWgt[16:21])
w.25 <- mean(weights$dllWgt[21:26])
w.30 <- mean(weights$dllWgt[26:31])
w.35 <- mean(weights$dllWgt[31:36])
w.40 <- mean(weights$dllWgt[36:41])
w.45 <- mean(weights$dllWgt[41:46])
w.60 <- mean(weights$dllWgt[46:51])

weights.mean <- c(w.5, w.10, w.15, w.20, w.25, w.30, w.35, w.40, w.45,
  w.60)
nms <- seq(5,60,5)
nms <- nms[c(1:9,12)]
names(weights.mean) <- c(paste("W", nms, sep=""))


weights.mat <- as.data.frame(matrix(rep(weights.mean, nrow(blcks)),
  nrow=nrow(blcks), ncol=length(weights.mean), byrow=TRUE))
names(weights.mat) <- names(weights.mean)

## Make presence / absence table
blcks.p <- blcks
blcks.p[,3:12] <- as.numeric(blcks.p[,3:12] > 0)


## Multiply p/a by weights
wght.blck <- blcks.p
wght.blck[,3:12] <- wght.blck[,3:12]*weights.mat


## Multiply pop by weights
pop.wght <- wght.blck
wght.blck$Min5 <- wght.blck$Min5 * wght.blck$POP100
wght.blck$Min10 <- wght.blck$Min10 * wght.blck$POP100
wght.blck$Min15 <- wght.blck$Min15 * wght.blck$POP100
wght.blck$Min20 <- wght.blck$Min20 * wght.blck$POP100
wght.blck$Min25 <- wght.blck$Min25 * wght.blck$POP100
wght.blck$Min30 <- wght.blck$Min30 * wght.blck$POP100
wght.blck$Min35 <- wght.blck$Min35 * wght.blck$POP100
wght.blck$Min40 <- wght.blck$Min40 * wght.blck$POP100
wght.blck$Min45 <- wght.blck$Min45 * wght.blck$POP100
wght.blck$Min60 <- wght.blck$Min60 * wght.blck$POP100


library(rgdal)
library(sp)
```

```
library(maptools)
library(shapefiles)

## Read in 5 minute file

dbf5 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_5.dbf")

dbf5 <- dbf5$dbf

## Subset to needed columns
dbf5 <- dbf5[,c(names(dbf5) == "GEOID10" | names(dbf5) == "FAC_ID")]

## Attach demand per hospital!
dbf5 <- merge(dbf5, wght.blck[,c(1,3)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand <- aggregate(dbf5$Min5, by=list(FAC_ID = dbf5$FAC_ID), sum)
names(hosp.demand)[2] <- "Min5"

rm(dbf5)

## Read in 10 minute file

dbf10 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_10.dbf")

dbf10 <- dbf10$dbf

## Subset to needed columns
dbf10 <- dbf10[,c(names(dbf10) == "GEOID10" | names(dbf10) == "FAC_ID")]

## Attach demand per hospital!
dbf10 <- merge(dbf10, wght.blck[,c(1,4)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf10$Min10, by=list(FAC_ID = dbf10$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min10"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf10)
```

```
## Read in 15 minute file

dbf15 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_15.dbf")

dbf15 <- dbf15$dbf

## Subset to needed columns
dbf15 <- dbf15[,c(names(dbf15) == "GEOID10" | names(dbf15) == "FAC_ID")]

## Attach demand per hospital!
dbf15 <- merge(dbf15, wght.blck[,c(1,5)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf15$Min15, by=list(FAC_ID = dbf15$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min15"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf15)


## Read in 20 minute file

dbf20 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_20.dbf")

dbf20 <- dbf20$dbf

## Subset to needed columns
dbf20 <- dbf20[,c(names(dbf20) == "GEOID10" | names(dbf20) == "FAC_ID")]

## Attach demand per hospital!
dbf20 <- merge(dbf20, wght.blck[,c(1,6)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf20$Min20, by=list(FAC_ID = dbf20$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min20"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf20)
```

```
## Read in 25 minute file

dbf25 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_25.dbf")

dbf25 <- dbf25$dbf

## Subset to needed columns
dbf25 <- dbf25[,c(names(dbf25) == "GEOID10" | names(dbf25) == "FAC_ID")]

## Attach demand per hospital!
dbf25 <- merge(dbf25, wght.blck[,c(1,7)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf25$Min25, by=list(FAC_ID = dbf25$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min25"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf25)


## Read in 30 minute file

dbf30 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_30.dbf")

dbf30 <- dbf30$dbf

## Subset to needed columns
dbf30 <- dbf30[,c(names(dbf30) == "GEOID10" | names(dbf30) == "FAC_ID")]

## Attach demand per hospital!
dbf30 <- merge(dbf30, wght.blck[,c(1,8)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf30$Min30, by=list(FAC_ID = dbf30$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min30"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf30)
```

```
## Read in 35 minute file

dbf35 <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_35.dbf")

dbf35 <- dbf35$dbf

## Subset to needed columns
dbf35 <- dbf35[,c(names(dbf35) == "GEOID10" | names(dbf35) == "FAC_ID")]

## Attach demand per hospital!
dbf35 <- merge(dbf35, wght.blck[,c(1,9)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf35$Min35, by=list(FAC_ID = dbf35$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min35"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf35)


### Read in 40 minute files

dbf40a <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_40a.dbf")
dbf40b <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_40b.dbf")

dbf40a <- dbf40a$dbf
dbf40b <- dbf40b$dbf

## Subset to needed columns
dbf40a <- dbf40a[,c(names(dbf40a) == "GEOID10" | names(dbf40a) == "FAC_ID")]
dbf40b <- dbf40b[,c(names(dbf40b) == "GEOID10" | names(dbf40b) == "FAC_ID")]

## Combine tables
dbf40 <- rbind(dbf40a, dbf40b)

## Attach demand per hospital!
dbf40 <- merge(dbf40, wght.blck[,c(1,10)], by="GEOID10", all.x=TRUE)
```

```
## Aggregate by hospital
hosp.demand.t <- aggregate(dbf40$Min40, by=list(FAC_ID = dbf40$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min40"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf40a)
rm(dbf40b)
rm(dbf40)


### Read in 45 minute files

dbf45a <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_45a.dbf")
dbf45b <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_45b.dbf")

dbf45a <- dbf45a$dbf
dbf45b <- dbf45b$dbf

## Subset to needed columns
dbf45a <- dbf45a[,c(names(dbf45a) == "GEOID10" | names(dbf45a) == "FAC_ID")]
dbf45b <- dbf45b[,c(names(dbf45b) == "GEOID10" | names(dbf45b) == "FAC_ID")]

## Combine tables
dbf45 <- rbind(dbf45a, dbf45b)

## Attach demand per hospital!
dbf45 <- merge(dbf45, wght.blck[,c(1,11)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf45$Min45, by=list(FAC_ID = dbf45$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min45"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf45a)
rm(dbf45b)
rm(dbf45)


### Read in 60 minute files
```

```
dbf60a <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_60a.dbf")
dbf60b <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/Service-areas/join/MI_2010_blocks_proj_cent_pop_gt0_join_60b.dbf")

dbf60a <- dbf60a$dbf
dbf60b <- dbf60b$dbf

## Subset to needed columns
dbf60a <- dbf60a[,c(names(dbf60a) == "GEOID10" | names(dbf60a) == "FAC_ID")]
dbf60b <- dbf60b[,c(names(dbf60b) == "GEOID10" | names(dbf60b) == "FAC_ID")]

## Combine tables
dbf60 <- rbind(dbf60a, dbf60b)

## Attach demand per hospital!
dbf60 <- merge(dbf60, wght.blck[,c(1,12)], by="GEOID10", all.x=TRUE)

## Aggregate by hospital
hosp.demand.t <- aggregate(dbf60$Min60, by=list(FAC_ID = dbf60$FAC_ID), sum)
names(hosp.demand.t)[2] <- "Min60"

## Merge
hosp.demand <- merge(hosp.demand, hosp.demand.t, by="FAC_ID", all.x=TRUE)

rm(dbf60a)
rm(dbf60b)
rm(dbf60)



### Sum
hosp.demand$DemandSum <- rowSums(hosp.demand[,2:11])


### Get bed info

## Import hospital bed numbers
hosp.info <- read.csv("/media/data/GISdata/hospitals/csv/
  2011-hosps-beds.csv")

## Attach beds to demand table
hosp.demand <- merge(hosp.demand, hosp.info[,c(4,8)], by="FAC_ID",
  all.x=TRUE)
```

```
## Reorder
hosp.demand <- hosp.demand[,c(1,13,2:12)]


####
#### Calculate Hospital Supply!
####

hosp.demand$HospSupply <- hosp.demand$BEDS2010 / hosp.demand$DemandSum
# hosp.demand[,c(1,2,14)]



####
#### Calculate E2SFCA
####

## Get OD matrix for Zip Codes
## Read in origin-destination shapefile table
od <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/OD/hosps_esri2010pwcent_traveltime_od.dbf")
## Remove header info
od <- od$dbf
## Subset
od <- od[,c(7,8,5)]

# Remove weird point
# Weird point is Block centroid that didn't fall in any Zip Code!
badID <- which(od$FAC_ID %in% unique(od$FAC_ID)[170:338])
od <- od[-c(badID),]

# Calculate weights
### Because we have the actual distance, we can calculate this using
### the formula from estimate.decay.parameter.R
#2009 data# weight = (1 / (1 + (Min/13.77)^1.83127))

od$dllWgt <- (1 / (1 + (od$Total_MinT/13.885928)^1.817622))

#########################
############
############ Important!
############
############ Because we only considered demand out to 60 minutes,
############ change any weight above 60 minutes to 0 (zero)
```

```
############
########################

od$dllWgt[od$Total_MinT > 60] <- 0


## Attach supply at each hospital
od <- merge(od, hosp.demand[,c(1,14)], by="FAC_ID", all.x=TRUE)

## Multiply supply by weight
od$Supply <- od$dllWgt*od$HospSupply


## Aggregate for each ZIP CODE
E2SFCA <- aggregate(od$Supply, by=list(ZIP = od$ZIPCODE), sum)
names(E2SFCA)[2] <- "E2SFCA"


### Clean up file to account for funky Zip Codes ##################
### See dissertation data notes ###################################

### Get zip pop data :: WE NEED ORIGINAL DATA HERE! ::
### Not "adjusted for bad zips" data!
zip.pop <- read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/
  tables/zipcode_blockpop_adj_ALLZIPS.csv")

### Attach to 3SFCA
E2SFCA <- merge(E2SFCA, zip.pop, by="ZIP", all.x=TRUE)

### Remove islands
#islands <- which(od$ZIPCODE %in% c(48028, 49726, 49757, 49775, 49782))
#od <- od[-islands, ]

### Reassign zips
E2SFCA$ZIP[E2SFCA$ZIP == 48710] <- 48706
E2SFCA$ZIP[E2SFCA$ZIP == 48743] <- 48739
E2SFCA$ZIP[E2SFCA$ZIP == 48824 | E2SFCA$ZIP == 48825] <- 48823
E2SFCA$ZIP[E2SFCA$ZIP == 49104] <- 49103
E2SFCA$ZIP[E2SFCA$ZIP == 49792] <- 49765
E2SFCA$ZIP[E2SFCA$ZIP == 49873] <- 49807

###############
### Do a weighted mean by zip population
###############
```

```
E2SFCA$WgtE2SFCA <- E2SFCA$E2SFCA*E2SFCA$BkPopAdj2010
E2SFCA <- aggregate(E2SFCA[,3:4], by=list("ZIP" = E2SFCA$ZIP), sum)
E2SFCA$E2SFCA <- E2SFCA$WgtE2SFCA / E2SFCA$BkPopAdj2010

### Clean up file to account for funkty Zip Codes ##################
### See dissertation data notes ####################################


E2SFCA <- E2SFCA[,-3]


### Test to see if math worked
sum(hosp.info$BEDS2010)
sum(E2SFCA$E2SFCA * E2SFCA$BkPopAdj2010)



## Write out to table
write.csv(E2SFCA, file="/home/delamate/MDCH/data/dissertation/zipcodes/
  tables/zipcode_E2SFCA_2010_new.csv", row.names=FALSE)



## Attach to Zip Polys and map (updated clusters!)
## Import Zip Code file and attach
 zip.poly <- readOGR("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d/895.shp", layer="895")

## Join
zip.poly@data <- cbind(zip.poly@data, E2SFCA)
zip.poly$E31000 <- zip.poly$E2SFCA * 1000


 writeOGR(zip.poly, dsn="/media/data/Project Files/Delamater/Dissertation/
  Utilization/E2SFCA", layer="MI_CL895_E2SFCA_new", driver="ESRI Shapefile")

####################################################
####################################################



####################################################
## R Code to cluster Zip Codes into regions      ##
## Built upon code from Michigan Hospital Groups ##
####################################################
```

```
### Read in 2010 patient records
pd <- read.csv("/home/delamate/MDCH/data/dissertation/utilization/2010/
  2010MIDBrecords.csv")

### Drop age column
pd <- pd[,-1]

### Clean up file to account for funky Zip Codes ###################
### See dissertation data notes ####################################

### Remove islands
islands <- which(pd$MIDB_ZIP %in% c(48028, 49726, 49757, 49775, 49782))
pd <- pd[-islands, ]

### Reassign zips
pd$MIDB_ZIP[pd$MIDB_ZIP == 48710] <- 48706
pd$MIDB_ZIP[pd$MIDB_ZIP == 48743] <- 48739
pd$MIDB_ZIP[pd$MIDB_ZIP == 48824 | pd$MIDB_ZIP == 48825] <- 48823
pd$MIDB_ZIP[pd$MIDB_ZIP == 49104] <- 49103
pd$MIDB_ZIP[pd$MIDB_ZIP == 49792] <- 49765
pd$MIDB_ZIP[pd$MIDB_ZIP == 49873] <- 49807

### Clean up file to account for funkty Zip Codes ##################
### See dissertation data notes ####################################

### Convert to OD matrix
out <- aggregate(pd$LOS, by=list("ZIP" = pd$MIDB_ZIP, "MIDB" =
  pd$HOSP_ID), sum)
names(out)[3] <- "PD"

pd <- reshape(out, direction='wide',idvar='ZIP', timevar='MIDB')
rm(out)
pd[is.na(pd)] <- 0

### Sort... rename columns
pd <- pd[order(pd$ZIP), ]
names(pd)[2:ncol(pd)] <- paste("PD", substr(names(pd)[2:ncol(pd)], 4, 7),
  sep="")

### Convert patient days to Commitment Index values
z.days <- rowSums(pd[,2:ncol(pd)])
ci.pd <- pd[,2:ncol(pd)] / z.days
ci.pd$ZIP <- pd$ZIP
ci.pd <- ci.pd[,c(159,1:158)]
```

```
# write.csv(pd, file="datatables/2010.patient.days.csv",
  row.names=FALSE)
# write.csv(ci.pd, file="datatables/2010.patient.days.CI.csv",
  row.names=FALSE)

# pd <- read.csv(file="datatables/2010.patient.days.csv")
# ci.pd <- read.csv(file="datatables/2010.patient.days.CI.csv")
#sum(ci.pd[3,2:ncol(ci.pd)])


### Read in Distance data from Zip Codes to hospitals
library(rgdal)
library(sp)
library(maptools)
library(shapefiles)

## Get OD matrix for Zip Codes
## Read in origin-destination shapefile table
od <- read.dbf("/media/data/Project Files/Delamater/Dissertation/
  Utilization/OD/hosps_esri2010pwcent_traveltime_od.dbf")
## Remove header info
od <- od$dbf
## Subset
od <- od[,c(7,8,5)]

### Have to attach MIDB number to OD matrix
## Import hospital info
hosp.info <- read.csv("/media/data/GISdata/hospitals/csv/
  2011-hosps-beds.csv")

## Attach MIDB number to OD matrix
od <- merge(od, hosp.info[,c(4,1)], by="FAC_ID", all.x=TRUE)
sum(is.na(od$MIDB))

### Clean up file to account for funky Zip Codes ##################
### See dissertation data notes ###################################

### Remove islands
#islands <- which(od$ZIPCODE %in% c(48028, 49726, 49757, 49775, 49782))
#od <- od[-islands, ]

### Reassign zips
od$ZIPCODE[od$ZIPCODE == 48710] <- 48706
od$ZIPCODE[od$ZIPCODE == 48743] <- 48739
```

```
od$ZIPCODE[od$ZIPCODE == 48824 | od$ZIPCODE == 48825] <- 48823
od$ZIPCODE[od$ZIPCODE == 49104] <- 49103
od$ZIPCODE[od$ZIPCODE == 49792] <- 49765
od$ZIPCODE[od$ZIPCODE == 49873] <- 49807

### Clean up file to account for funkty Zip Codes ##################
### See dissertation data notes ####################################

## Because some hospitals have the same MIDB, but are in different
## locations, aggregate records (mean) by BOTH : FROM and TO
## this gives the "mean" distance for these locations
od <- aggregate(od[,3], by=list("ZIP" = od$ZIPCODE, "MIDB" = od$MIDB),
  mean)
names(od)[3] <- "Min"

### Convert to OD matrix
od <- reshape(od, direction='wide',idvar='ZIP', timevar='MIDB')
od[is.na(od)] <- 0
names(od)[2:ncol(od)] <- paste("D", substr(names(od)[2:ncol(od)], 5, 8),
  sep="")

# write.csv(od, file="datatables/traveltime.zipcodes.hospitals.csv",
  row.names=FALSE)

### Scale distance matrix from 0-1
max.dist <- max(od[,2:ncol(od)])

# Divide by maximum travel time btwn any two hospitals
od[,2:ncol(od)] <- od[,2:ncol(od)] / max.dist

# write.csv(od, file="datatables/traveltime.zipcodes.hospitals.scaled.csv",
  row.names=FALSE)

######################
## Attach utilization matrix to distance matrix
######################

data <- merge(ci.pd, od, by="ZIP")
dim(data)
sum(is.na(data))

# write.csv(data, file="datatables/cluster.data.2010.csv", row.names=FALSE)
# data <- read.csv("datatables/cluster.data.2010.csv")
```

```
#################################################################
## Function to seed Kmeans cluster algorithm with centers     ##
## provided by a Ward's cluster output.  Stabilizes results   ##
## and produces better results                                ##
#################################################################

kmeans.ward <- function(x, clusters) {
  d <- dist(x, "euclidean") # create distance matrix
  hc <- hclust(d, method="ward") # initial clusters
  memb <- cutree(hc, k = clusters) # get 'n' clusters
  cent <- NULL  # make holder
  for (k in 1:clusters) {   # get cluster centers
    cent <- rbind(cent, colMeans(x[memb == k,]))
  }
  k.m <- kmeans(x, cent, iter.max = 10000) # seed kmeans with ward's
  return(k.m)
}


## Define the range of solutions to evaluate
cl.max <- nrow(data)-1
clusters <- c(2:cl.max)

## Create a holder for cluster statistics
wss <- bss <- r2 <- incF <- rep(0, length(clusters))
k.data.pat <- cbind(clusters, wss, bss, r2, incF)

## Create a holder for cluster membership
membership <- data.frame(data$ZIP)
names(membership) <- "ZIP"

# Get number of columns in data
col.max <- ncol(data)

start <- Sys.time()
count <- seq(0, cl.max, 25)

# Loop through clusters
for (z in 1:length(clusters)) {

  ## Use K-means + Wards method to create clusters
  kmeans <- kmeans.ward(data[,2:col.max], clusters[z])

  ## Write cluster stats to data holder
  k.data.pat[z,2] <- kmeans$tot.withinss
```

```
  k.data.pat[z,3] <- kmeans$betweenss
  k.data.pat[z,4] <- 1-(kmeans$tot.withinss/kmeans$totss)

  ## Write cluster membership to data holder
  membership$clusters <- kmeans$cluster
  names(membership)[z+1] <- paste("CL", z+1, sep="")

  if (z %in% count) print(paste("Cluster: ", z, " at ",
    Sys.time()-start, sep=""))

}

print(Sys.time() - start)

## Convert data holder to data frame
k.data.pat <- as.data.frame(k.data.pat)

## Calculate incremental F score
for (i in 2:length(clusters)) {
  k.data.pat$incF[i] <- ((k.data.pat$r2[i]-k.data.pat$r2[i-1])/
    (k.data.pat$clusters[i]-k.data.pat$clusters[i-1])) /
    ((1-k.data.pat$r2[i])/((nrow(data))-(k.data.pat$clusters[i]-1)))
    }

## Write data to file
write.csv(k.data.pat, file="datatables/cluster.stats.csv", row.names=FALSE)


## Find peaks in incremental F score
incF.peaks <- which(k.data.pat$incF[3:(cl.max-1)] >
  k.data.pat$incF[2:(cl.max-2)] & k.data.pat$incF[3:(cl.max-1)] >
  k.data.pat$incF[4:cl.max])+2

## Subset results
cluster.groups <- k.data.pat[incF.peaks,]

membership <- membership[,c(1,cluster.groups$clusters)]

## Write out cluster membership to file
write.csv(membership, file="datatables/cluster.membership.incF.peaks.csv",
  row.names=FALSE)

#####################################################
#####################################################
```

```
##########################################
## R Code to create data for regressions ##
##########################################

start <- Sys.time()

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)
library(spdep)
library(gpclib)
library(plotrix)
# This line allows maptools to use gpc lib
gpclibPermit()
library(car)
library(MASS)
library(psych)

### Read in cluster information
cl <- read.csv("/home/delamate/MDCH/data/dissertation/clustering/
     datatables/cluster.membership.incF.peaks.csv")
   ### Add "non-clustered" column
   cl[,ncol(cl)+1] <- seq(1:895)
   names(cl)[ncol(cl)] <- "CL895"

### Get total number of clusterings to be evaluated
e <- ncol(cl)-1

# Get list of dissolved shapefiles
cs <- list.files("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d", pattern='.shp')
cs <- cs[seq(1,length(cs),2)]

# Get hospitalization information
pd <- read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/tables/
  zipcode_2010_UTILIZATION_age_breakdown.csv")

# Get zip age breakdown
age <- read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/tables/
  zipcode_AGE_blockpop_adj.csv")

# Get large data table
```

```
data <- read.csv("/home/delamate/MDCH/data/dissertation/zipcodes/tables/
  zipcode_all_variables.csv")

   ## Add age categories for income, insurance variables
   data$Pop0_64 <- rowSums(age[,2:14])
   data$Pop16p <- rowSums(age[,5:19])
   data$Pop25p <- rowSums(age[,7:19])

## Make STANDARD population (state totals)
std.pop <- colSums(age[,-1])


## For PCA random variable
set.seed(1)

tolerance <- function (x) {
  1/vif(x)
}


####################
### START ITERATION
####################

for (i in e:1) {

  ######################
  ### Attach cluster membership to files
  ######################
  pd.i <- merge(pd, cl[,c(1,i+1)], by="ZIP", all.x=TRUE)
  age.i <- merge(age, cl[,c(1,i+1)], by="ZIP", all.x=TRUE)
  data.i <- merge(data, cl[,c(1,i+1)], by="ZIP", all.x=TRUE)

  ######################
  ### Get dissolved shapefile, make neighbors
  ######################
  shp <- readOGR(paste("/media/data/Project Files/Delamater/Dissertation/
    Clustering/Cluster_shapefiles/d/", cs[i], sep=""), layer=substr(cs[i],
    1, nchar(cs[1])-4), verbose=FALSE)

  ### Get neighbors
  nb <- poly2nb(shp)

  ### One island poly doesn't have neighbors in the original file...
  ### It is connected by a bridge to a single zip code
```

```
### Assign neighbors to it manually
nb.mat <- nb2mat(nb,style="B",zero.policy=TRUE)
### Find regions with zero neighbors
w.zero <- as.numeric(which(rowSums(nb.mat) == 0))
### If island poly is not grouped, assign neighbor as 48193
if (length(w.zero) > 0) {
  near.clust <- data.i[data.i$ZIP == 48193, 42]
  nb[[w.zero]] <- near.clust
  nb[[near.clust]] <- as.integer(c(w.zero, nb[[near.clust]]))
}


######################
### Aggregate attributes
######################

### First, those columns that use FULL population
data.i[,c(6,8:9,11,30:38)] <- data.i[,c(6,8:9,11,30:38)] *
  data.i$BkPopAdj2010
### Next, those columns that use 0-64
data.i[,c(7,10)] <- data.i[,c(7,10)] * data.i$Pop0_64
### Next, those columns that use 16+
data.i[,c(12,18:29)] <- data.i[,c(12,18:29)] * data.i$Pop16p
### Finally, those columns that use 25+
data.i[,13:17] <- data.i[,13:17] * data.i$Pop25p

### Now, sum by CLUSTER
data.cl <- aggregate(data.i[,2:41], by=list("CL" = data.i[,42]), sum)

### Now, divide by appropriate "summed" population
### First, those columns that use FULL population
data.cl[,c(6,8:9,11,30:38)] <- data.cl[,c(6,8:9,11,30:38)] /
  data.cl$BkPopAdj2010
### Next, those columns that use 0-64
data.cl[,c(7,10)] <- data.cl[,c(7,10)] / data.cl$Pop0_64
### Next, those columns that use 16+
data.cl[,c(12,18:29)] <- data.cl[,c(12,18:29)] / data.cl$Pop16p
### Finally, those columns that use 25+
data.cl[,13:17] <- data.cl[,13:17] / data.cl$Pop25p

names(data.cl)[1] <- names(cl)[i+1]

######################
### Aggregate hospitalization rates
### Calculate Empirical Bayes estimates of patient day rates
```

```
######################

## First, aggregate patient days and age population by CLUSTER
pd.i <- aggregate(pd.i[,2:19], by=list("CL" = pd.i[,21]), sum)
age.i <- aggregate(age.i[,2:19], by=list("CL" = age.i[,21]), sum)

# Make holders
EB <- data.frame(CL = pd.i$CL)
CR <- data.frame(CL = pd.i$CL)
EB.phi <- data.frame(CL = pd.i$CL)
EB.gamma <- data.frame(CL = pd.i$CL)

## Loop through each age group
for (z in 2:19) {
if (sum(age[,z] == 0) > 0) {
  age.zero <- age.i[,z]
  age.zero[which(age.zero == 0)] <- 1
  eb <- EBlocal(pd.i[,z], age.zero, nb, zero.policy=TRUE)
} else {
  eb <- EBlocal(pd.i[,z], age.i[,z], nb, zero.policy=TRUE)
}
eb[is.na(eb)] <- 0
EB[,z] <- eb$est
CR[,z] <- eb$raw
EB.phi[,z] <- attributes(eb)$parameters$a
EB.gamma[,z] <- attributes(eb)$parameters$m
names(EB)[z] <- names(CR)[z] <- names(EB.phi)[z] <- names(EB.gamma)[z]
  <- names(pd.i)[z]
}

### Write out age-specific rates
write.csv(CR, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates/Crude/", names(cl)[i+1],
  ".csv", sep=""), row.names=FALSE)
write.csv(EB, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates/EBsmooth/", names(cl)[i+1],
  ".csv", sep=""), row.names=FALSE)
write.csv(EB.phi, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates/EBphi/", names(cl)[i+1],
  ".csv", sep=""), row.names=FALSE)
write.csv(EB.gamma, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates/EBgamma/", names(cl)[i+1],
  ".csv", sep=""), row.names=FALSE)

### Get crude population rate
```

```
pop.days <- sum(pd.i[,2:19])
pop.c.r <- pop.days / std.pop[19]

### Calculate overall EB adj hospitalization rates
eb.adj <- NULL

for (z in 1:nrow(pd.i)) {

  ## Multiply zip-specific rates by std pop
  z.r.eb.adj <- EB[z,2:19] * std.pop[1:18]

  ## Sum and divide by total population
  z.r.eb.adj <- sum(z.r.eb.adj) / std.pop[19]

  ## Attach to holder
  eb.adj <- rbind(eb.adj, c(as.numeric(pd.i$CL[z]),
    as.numeric(z.r.eb.adj)))

}

### Insert into data table
data.cl$AgeAdjPatDayRateEBadj <- eb.adj[,2]

### Calculate Standardized Rate Ratio and Standardized Rate Difference
data.cl$StRateRatio <- data.cl$AgeAdjPatDayRateEBadj / pop.c.r
data.cl$StRateDif <- data.cl$AgeAdjPatDayRateEBadj - pop.c.r

### Write out the Aggregated data table
write.csv(data.cl, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/all_data/", names(cl)[i+1], ".csv", sep=""),
  row.names=FALSE)


######################
### Principal Components
######################

data.cl$RANDOM <- runif(nrow(data.cl))

### SES: Education, Income
cor.ses <- cor(data.cl[,c(11:14,16:17,42)])
pca.ses <- principal(cor.ses, nfactors=7, rotate="none", scores=FALSE)
pca.ses.loadings <- unclass(pca.ses$loadings)
m <- apply(abs(pca.ses.loadings), 2, FUN=max)
n.ses.pc <- as.numeric(which(abs(pca.ses.loadings[7,]) == m))
```

```
  if (length(n.ses.pc) == 0) n.ses.pc <- which(abs(pca.ses.loadings[7,])
    == max(abs(pca.ses.loadings[7,])))
  # Reconduct PCA without random variable to get "pct of variance explained"
  cor.ses <-  cor(data.cl[,c(11:14,16:17)])
  pca.ses <- principal(cor.ses, nfactors=6, rotate="none", scores=FALSE)
  pca.ses.loadings <- unclass(pca.ses$loadings)
  pca.var <- apply(pca.ses.loadings^2, 2, sum) / 6
  write.csv(pca.var[1:(n.ses.pc-1)], file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_variance/SES/",names(cl)[i+1],
    ".csv", sep=""))
  # Do actual PCA with rotation and # of components
  pca.ses <- principal(data.cl[,c(11:14,16:17)], nfactors=n.ses.pc[1]-1,
    rotate="varimax", scores=TRUE)
  pca.ses.loadings <- unclass(pca.ses$loadings)
#  round(pca.ses.loadings, 2)
  ### Write loadings to file
  write.csv(pca.ses.loadings, file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_loadings/SES/",names(cl)[i+1],
    ".csv", sep=""))

  ### ETH: Ethnicity
  pca.eth <- principal(data.cl[,c(32:36,38,42)], nfactors=7, rotate="none",
    scores=FALSE)
  pca.eth.loadings <- unclass(pca.eth$loadings)
  m <- apply(abs(pca.eth.loadings), 2, FUN=max)
  n.eth.pc <- as.numeric(which(abs(pca.eth.loadings[7,]) == m))
  if (length(n.eth.pc) == 0) n.eth.pc <- which(abs(pca.eth.loadings[7,]) ==
    max(abs(pca.eth.loadings[7,])))
  # Reconduct PCA without random variable to get "pct of variance explained"
  pca.eth <- principal(data.cl[,c(32:36,38)], nfactors=6, rotate="none",
    scores=FALSE)
  pca.eth.loadings <- unclass(pca.eth$loadings)
  pca.var <- apply(pca.eth.loadings^2, 2, sum) / 6
  write.csv(pca.var[1:(n.eth.pc-1)], file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_variance/ETH/",names(cl)[i+1],
    ".csv", sep=""))
  pca.eth <- principal(data.cl[,c(32:36,38)], nfactors=n.eth.pc[1]-1,
    rotate="varimax", scores=TRUE)
  pca.eth.loadings <- unclass(pca.eth$loadings)
#  round(pca.eth.loadings, 2)
  ### Write loadings to file
  write.csv(pca.eth.loadings, file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_loadings/ETH/",names(cl)[i+1],
    ".csv", sep=""))
```

```
  ### MOBILITY 1
  pca.mob <- principal(data.cl[,c(18:20,22,42)], nfactors=5, rotate="none",
    scores=FALSE)
  pca.mob.loadings <- unclass(pca.mob$loadings)
  m <- apply(abs(pca.mob.loadings), 2, FUN=max)
  n.mob.pc <- as.numeric(which(abs(pca.mob.loadings[5,]) == m))
  if (length(n.mob.pc) == 0) n.mob.pc <- which(pca.mob.loadings[5,] ==
    max(pca.mob.loadings[5,]))
  # Reconduct PCA without random variable to get "pct of variance explained"
  pca.mob <- principal(data.cl[,c(18:20,22)], nfactors=4, rotate="none",
    scores=FALSE)
  pca.mob.loadings <- unclass(pca.mob$loadings)
  pca.var <- apply(pca.mob.loadings^2, 2, sum) / 4
  write.csv(pca.var[1:(n.mob.pc-1)], file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_variance/MOB/",names(cl)[i+1],
    ".csv", sep=""))
  pca.mob <- principal(data.cl[,c(18:20,22)], nfactors=n.mob.pc[1]-1,
    rotate="varimax", scores=TRUE)
  pca.mob.loadings <- unclass(pca.mob$loadings)
  #round(pca.mob.loadings, 2)
  ### Write loadings to file
  write.csv(pca.mob.loadings, file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_loadings/MOB/",names(cl)[i+1],
    ".csv", sep=""))

  ### MOBILITY 2
  pca.mob2 <- principal(data.cl[,c(23:28,42)], nfactors=7, rotate="none",
    scores=FALSE)
  pca.mob2.loadings <- unclass(pca.mob2$loadings)
  m <- apply(abs(pca.mob2.loadings), 2, FUN=max)
  n.mob2.pc <- as.numeric(which(abs(pca.mob2.loadings[7,]) == m))
  if (length(n.mob2.pc) == 0) n.mob2.pc <- which(pca.mob2.loadings[7,] ==
    max(pca.mob2.loadings[7,]))
  # Reconduct PCA without random variable to get "pct of variance explained"
  pca.mob2 <- principal(data.cl[,c(23:28)], nfactors=6, rotate="none",
    scores=FALSE)
  pca.mob2.loadings <- unclass(pca.mob2$loadings)
  pca.var <- apply(pca.mob2.loadings^2, 2, sum) / 6
  write.csv(pca.var[1:(n.mob2.pc-1)], file=paste("/home/delamate/MDCH/data/
    dissertation/regressions/output_tables/pca_variance/MOB2/",names(cl)[i+1],
    ".csv", sep=""))
  pca.mob2 <- principal(data.cl[,c(23:28)], nfactors=n.mob2.pc[1]-1,
    rotate="varimax", scores=TRUE)
  pca.mob2.loadings <- unclass(pca.mob2$loadings)
#  round(pca.mob2.loadings, 2)
```

```
### Write loadings to file
write.csv(pca.mob2.loadings, file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/pca_loadings/MOB2/",names(cl)[i+1],
  ".csv", sep=""))

### Aggregate PCA scores into table
pca.scores <- as.data.frame(cbind(pca.ses$scores, pca.eth$scores,
  pca.mob$scores, pca.mob2$scores))

### Rename columns
names(pca.scores) <- c(paste("SESPC", 1:(n.ses.pc[1]-1), sep=""),
  paste("ETHPC", 1:(n.eth.pc[1]-1), sep=""), paste("MOBPC", 1:(n.mob.pc[1]-1),
  sep=""), paste("MOB2PC", 1:(n.mob2.pc[1]-1), sep=""))

write.csv(pca.scores, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/pca_data/",names(cl)[i+1],".csv", sep=""))

######
### Remove MOBPC2 if it is in data (no theory!)
######
if ("MOBPC2" %in% names(pca.scores)) pca.scores <-
  pca.scores[,-(which(names(pca.scores) == "MOBPC2"))]

### Make new data
data.pc <- as.data.frame(cbind(data.cl[,c(1:4,6,10,31)], pca.scores))

listw <- nb2listw(nb, style = "B", zero.policy=TRUE)

### Regress E2SFCA and Ethnicity here
if (cor(data.pc$E2SFCA, data.pc$ETHPC1) >= 0.4) {
  lm.E2SFCA <- lm(data.pc$E2SFCA ~ data.pc$ETHPC1)
  resid.E2SFCA <- residuals(lm.E2SFCA)
  data.pc[,5] <- resid.E2SFCA
  names(data.pc)[5] <- "E2SFCAresid"
  # write out regression statistics
  write.csv(summary(lm.E2SFCA)$coefficients, file=paste
    ("/home/delamate/MDCH/data/dissertation/regressions/output_tables/
    pre.regression.stats/E2SFCA/", names(cl)[i+1],".csv", sep=""))

}

### Regress Health Insurance and SES here
if (cor(data.pc$HeaInsRateInt0_64, data.pc$SESPC1) >= 0.4) {
  lm.hi <- lm(data.pc$HeaInsRateInt0_64 ~ data.pc$SESPC1)
  resid.HI <- residuals(lm.hi)
```

```
  data.pc[,6] <- resid.HI
  names(data.pc)[6] <- "HeaInsRateInt0_64resid"
  write.csv(summary(lm.hi)$coefficients, file=paste("/home/delamate/
    MDCH/data/dissertation/regressions/output_tables/pre.regression.stats/
    HealIns/", names(cl)[i+1],".csv", sep=""))
}

cor.table <- abs(cor(data.pc[,5:ncol(data.pc)]))
write.csv(round(cor.table, 6),  file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/correlation_all/",names(cl)[i+1],
  ".csv", sep=""))

names.lm <- paste(names(data.pc)[5:ncol(data.pc)], collapse=" + ")
vif.lm <- vif(lm(formula(paste("StRateDif ~ ", names.lm), sep=""),
  data=data.pc))
write.csv(vif.lm, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/vif_all/",names(cl)[i+1],".csv", sep=""))

### Remove correlated variables here
### If there's too much correlation... don't proceed!!
while (max(vif.lm) > 2) {

  ### Find max
  which.vif <- names(vif.lm)[(max(which(vif.lm > 2)))]
  which.max <- which(names(data.pc) == which.vif)

  ### Remove correlated variable
  data.pc <- data.pc[,-which.max]

  names.lm <- paste(names(data.pc)[5:ncol(data.pc)], collapse=" + ")
  vif.lm <- vif(lm(formula(paste("StRateDif ~ ", names.lm), sep=""), data=data.pc))

}

  write.csv(vif.lm, file=paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/vif/",names(cl)[i+1],".csv", sep=""))

g.vars <- which(colnames(cor.table) %in% names(data.pc))

write.csv(round(cor.table[g.vars,g.vars], 6), file=paste("/home/delamate/
  MDCH/data/dissertation/regressions/output_tables/correlation/",
  names(cl)[i+1],".csv", sep=""))

lm.dat <- as.matrix(data.pc[,5:ncol(data.pc)])
```

```
  write.csv(lm.dat, file=paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/lm_data/", names(cl)[i+1],".csv", sep=""),
    row.names=FALSE)

}

print(Sys.time() - start)

##########################################
##########################################



#########################################################################
## R Code to conduct state-level simple linear regressions         ##
## at many scales of analysis to show AUTOCORRELATION of residuals ##
#########################################################################

start <- Sys.time()

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)
library(spdep)
library(gpclib)
library(plotrix)
# This line allows maptools to use gpc lib
gpclibPermit()
library(car)
library(MASS)
library(psych)


### Read in cluster information
cl <- read.csv("/home/delamate/MDCH/data/dissertation/clustering/
  datatables/cluster.membership.incF.peaks.csv")
   ### Add "non-clustered" column
   cl[,ncol(cl)+1] <- seq(1:895)
   names(cl)[ncol(cl)] <- "CL895"

# Get list of population files
pop.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/all_data")
order <- unlist(strsplit(pop.files, ".csv"))
```

```
order <- order(as.numeric(substr(order, 3, 5)))
pop.files <- pop.files[order]

# Get list of ACS files
acs.files <-  list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.ACS/data")
order <- unlist(strsplit(acs.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
acs.files <- acs.files[order]

# Get list of Low Variation files
lv.files <-  list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.LowV/data")
order <- unlist(strsplit(lv.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
lv.files <- lv.files[order]

# Get list of dissolved shapefiles
cs <- list.files("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d", pattern='.shp')
cs <- cs[seq(1,length(cs),2)]

# Get list of utilization rates and variables
u.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.60min/data")
order <- unlist(strsplit(u.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
u.files <- u.files[order]
i.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/lm_data")
order <- unlist(strsplit(i.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
i.files <- i.files[order]

matches <- which(u.files %in% i.files)
u.files <- u.files[matches]

pop.files <- pop.files[matches]
acs.files <- acs.files[matches]
lv.files <- lv.files[matches]

shp.match <- which(as.numeric(unlist(strsplit(cs, ".shp"))) %in%
  as.numeric(substr(unlist(strsplit(i.files, ".csv")), 3, 5)))
cs <- cs[shp.match]
```

```
cl.match <- which(as.numeric(substr(names(cl)[2:ncol(cl)], 3, 5))
  %in% as.numeric(substr(unlist(strsplit(i.files, ".csv")), 3, 5)))
cl <- cl[,c(1,cl.match+1)]

## Choose significance for regression models
sig <- 0.05
# sig <- 0.1

### Get total number of clusterings to be evaluated
e <- length(u.files)

####################
### START ITERATION
####################

for (i in e:1) {

  ######################
  ### Get data
  ######################
  ut <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/patdayrates.60min/data/", u.files[i],
    sep=""))
  lm <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/lm_data/", i.files[i], sep=""))
  acs <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/patdayrates.ACS/data/", i.files[i],
    sep=""))
  lv <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/patdayrates.LowV/data/", i.files[i],
    sep=""))
  pop <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/all_data/", pop.files[i], sep=""))
  pop <- pop$BkPopAdj2010
  i.pop <- 1/sqrt(pop)

  pop.med <- median(pop)
  fits.groups <- pop <= pop.med
  qx <- quantile(pop, probs=seq(0,1,0.2))
  q.fits.groups <- cut(pop, qx, include.lowest = TRUE)

  ######
  ### Remove MOBPC2 if it is in data (no theory!)
  ######
  if ("MOBPC2" %in% names(lm)) lm <- lm[,-(which(names(lm) == "MOBPC2"))]
```

```
######################
### Get dissolved shapefile, make neighbors
######################
shp <- readOGR(paste("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d/", cs[i], sep=""), layer=substr(cs[i],
  1, nchar(cs[1])-4), verbose=FALSE)

### Get neighbors
nb <- poly2nb(shp)

### One island poly doesn't have neighbors in the original file...
### It is connected by a bridge to a single zip code
### Assign neighbors to it manually
nb.mat <- nb2mat(nb,style="B",zero.policy=TRUE)
### Find regions with zero neighbors
w.zero <- as.numeric(which(rowSums(nb.mat) == 0))
### If island poly is not grouped, assign neighbors as 48193 (only for one)
if (length(w.zero) > 0) {
  near.clust <- cl[cl$ZIP == 48193, i+1]
  nb[[w.zero]] <- near.clust
  nb[[near.clust]] <- as.integer(c(w.zero, nb[[near.clust]]))
}

listw <- nb2listw(nb, style = "B", zero.policy=TRUE)

### Do PCA on ACS and LV
pca.h <- principal(cbind(acs$AgeAdjPatDayRateEBadj,
  lv$AgeAdjPatDayRateEBadj), nfactors=1, rotate="varimax", scores=TRUE)
acs.lv.scores <- pca.h$scores

pca.h2 <- principal(cbind(acs$AgeAdjPatDayRateEBadj,
  lv$AgeAdjPatDayRateEBadj), nfactors=2, rotate="none", scores=FALSE)
pca.h2.loadings <- unclass(pca.h2$loadings)
pca.var <- apply(pca.h2.loadings^2, 2, sum) / 2

# Regress on ETHPC1
acs.lv.lm <- lm(acs.lv.scores ~ lm$ETHPC1)


######
###### LM weighted, normal
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
```

```
colnames(lm.dat)[1] <- "ACSLowVPCresid"

## Regression
milm <- lm(ut$StRateDif ~ lm.dat, weights=i.pop)

## If there are non-significant terms in the model...
while (sum(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(milm)$coefficients[2:(ncol(lm.dat)+1),
    4] == max(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  milm <- lm(ut$StRateDif ~ lm.dat, weights=i.pop)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(milm)$coefficients
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/weighted/
  n.scale/betas/",names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(milm), listw, randomisation=FALSE)

resid.med <- median(residuals(milm))
r.fits.groups <- residuals(milm) <= resid.med

stats <- c(summary(milm)$adj.r.squared, summary(milm)$fstatistic[1],
  pf(summary(milm)$fstatistic[1], summary(milm)$fstatistic[2],
  summary(milm)$fstatistic[3], lower.tail = FALSE),
  leveneTest(residuals(milm),factor(fits.groups))[1,3],
  leveneTest(residuals(milm),factor(r.fits.groups))[1,3], mt$statistic,
  mt$p.value)
names(stats) <- c("adjR2", "F", "F.p", "LevenePOP2", "LeveneRESID2",
  "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
```

```
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/weighted/
  n.scale/sig/",names(cl)[i+1],".csv", sep=""))


######
###### LM weighted, scaled
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)
ut <- as.data.frame(scale(ut))

## Regression
milm <- lm(ut$StRateDif ~ lm.dat, weights=i.pop)

## If there are non-significant terms in the model...
while (sum(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(milm)$coefficients[2:(ncol(lm.dat)+1),
    4] == max(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  milm <- lm(ut$StRateDif ~ lm.dat, weights=i.pop)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(milm)$coefficients
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/weighted/
  scale/betas/",names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(milm), listw, randomisation=FALSE)

resid.med <- median(residuals(milm))
r.fits.groups <- residuals(milm) <= resid.med
```

```
stats <- c(summary(milm)$adj.r.squared, summary(milm)$fstatistic[1],
  pf(summary(milm)$fstatistic[1], summary(milm)$fstatistic[2],
  summary(milm)$fstatistic[3], lower.tail = FALSE),
  leveneTest(residuals(milm),factor(fits.groups))[1,3],
  leveneTest(residuals(milm),factor(r.fits.groups))[1,3], mt$statistic,
  mt$p.value)
names(stats) <- c("adjR2", "F", "F.p", "LevenePOP2", "LeveneRESID2",
  "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/weighted/
  scale/sig/",names(cl)[i+1],".csv", sep=""))

######
###### LM non-weighted, normal
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"

## Regression
milm <- lm(ut$StRateDif ~ lm.dat)

## If there are non-significant terms in the model...
while (sum(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(milm)$coefficients[2:(ncol(lm.dat)+1),
    4] == max(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  milm <- lm(ut$StRateDif ~ lm.dat)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(milm)$coefficients
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name
```

```
write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/n.weighted/
  n.scale/betas/",names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(milm), listw, randomisation=FALSE)

resid.med <- median(residuals(milm))
r.fits.groups <- residuals(milm) <= resid.med

stats <- c(summary(milm)$adj.r.squared, summary(milm)$fstatistic[1],
  pf(summary(milm)$fstatistic[1], summary(milm)$fstatistic[2],
  summary(milm)$fstatistic[3], lower.tail = FALSE),
  leveneTest(residuals(milm),factor(fits.groups))[1,3],
  leveneTest(residuals(milm),factor(r.fits.groups))[1,3], mt$statistic,
  mt$p.value)
names(stats) <- c("adjR2", "F", "F.p", "LevenePOP2", "LeveneRESID2",
  "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/regression.stats.ACS.LowV.dep.lm/n.weighted/
  n.scale/sig/",names(cl)[i+1],".csv", sep=""))


######
###### LM non-weighted, scaled
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)
ut <- as.data.frame(scale(ut))

## Regression
milm <- lm(ut$StRateDif ~ lm.dat)

## If there are non-significant terms in the model...
while (sum(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(milm)$coefficients[2:(ncol(lm.dat)+1),
    4] == max(summary(milm)$coefficients[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")
```

```
    lm.dat <- lm.dat[,-bad.t]

    milm <- lm(ut$StRateDif ~ lm.dat)

    if (is.vector(lm.dat) == TRUE) break

  }

  lm.sum <- summary(milm)$coefficients
  if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

  write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/regression.stats.ACS.LowV.dep.lm/n.weighted/
    scale/betas/",names(cl)[i+1],".csv", sep=""))

  mt <- moran.test(residuals(milm), listw, randomisation=FALSE)

  resid.med <- median(residuals(milm))
  r.fits.groups <- residuals(milm) <= resid.med

  stats <- c(summary(milm)$adj.r.squared, summary(milm)$fstatistic[1],
    pf(summary(milm)$fstatistic[1], summary(milm)$fstatistic[2],
    summary(milm)$fstatistic[3], lower.tail = FALSE),
    leveneTest(residuals(milm),factor(fits.groups))[1,3],
    leveneTest(residuals(milm),factor(r.fits.groups))[1,3], mt$statistic,
    mt$p.value)
  names(stats) <- c("adjR2", "F", "F.p", "LevenePOP2", "LeveneRESID2",
    "Moran", "Moran.p")

  write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
    regressions/output_tables/regression.stats.ACS.LowV.dep.lm/n.weighted/
    scale/sig/",names(cl)[i+1],".csv", sep=""))

}

print(Sys.time() - start)

#########################################################################
#########################################################################



##############################################
## R Code to conduct state-level regressions ##
## at many scales of analysis                ##
```

```
################################################

start <- Sys.time()

library(rgdal)
library(sp)
library(maptools)
library(shapefiles)
library(spdep)
library(gpclib)
library(plotrix)
# This line allows maptools to use gpc lib
gpclibPermit()
library(car)
library(MASS)
library(psych)

### Read in cluster information
cl <- read.csv("/home/delamate/MDCH/data/dissertation/clustering/
  datatables/cluster.membership.incF.peaks.csv")
   ### Add "non-clustered" column
   cl[,ncol(cl)+1] <- seq(1:895)
   names(cl)[ncol(cl)] <- "CL895"

# Get list of population files
pop.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/all_data")
order <- unlist(strsplit(pop.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
pop.files <- pop.files[order]

# Get list of ACS files
acs.files <-  list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.ACS/data")
order <- unlist(strsplit(acs.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
acs.files <- acs.files[order]

# Get list of Low Variation files
lv.files <-  list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.LowV/data")
order <- unlist(strsplit(lv.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
lv.files <- lv.files[order]
```

```
# Get list of dissolved shapefiles
cs <- list.files("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d", pattern='.shp')
cs <- cs[seq(1,length(cs),2)]

# Get list of utilization rates and variables
u.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.60min/data")
order <- unlist(strsplit(u.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
u.files <- u.files[order]
i.files <- list.files("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/lm_data")
order <- unlist(strsplit(i.files, ".csv"))
order <- order(as.numeric(substr(order, 3, 5)))
i.files <- i.files[order]

matches <- which(u.files %in% i.files)
u.files <- u.files[matches]

pop.files <- pop.files[matches]
acs.files <- acs.files[matches]
lv.files <- lv.files[matches]

shp.match <- which(as.numeric(unlist(strsplit(cs, ".shp"))) %in%
  as.numeric(substr(unlist(strsplit(i.files, ".csv")), 3, 5)))
cs <- cs[shp.match]

cl.match <- which(as.numeric(substr(names(cl)[2:ncol(cl)], 3, 5))
  %in% as.numeric(substr(unlist(strsplit(i.files, ".csv")), 3, 5)))
cl <- cl[,c(1,cl.match+1)]

## Choose significance for regression models
sig <- 0.05
# sig <- 0.1

### Get total number of clusterings to be evaluated
e <- length(u.files)

####################
### START ITERATION
####################

for (i in e:1) {
```

```
######################
### Get data
######################
ut <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.60min/data/", u.files[i],
  sep=""))
lm <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/lm_data/", i.files[i], sep=""))
acs <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.ACS/data/", i.files[i],
  sep=""))
lv <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/patdayrates.LowV/data/", i.files[i],
  sep=""))
pop <- read.csv(paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/all_data/", pop.files[i], sep=""))
pop <- pop$BkPopAdj2010
i.pop <- 1/sqrt(pop)

pop.med <- median(pop)
fits.groups <- pop <= pop.med
qx <- quantile(pop, probs=seq(0,1,0.2))
q.fits.groups <- cut(pop, qx, include.lowest = TRUE)

######
### Remove MOBPC2 if it is in data (no theory!)
######
if ("MOBPC2" %in% names(lm)) lm <- lm[,-(which(names(lm) == "MOBPC2"))]

######################
### Get dissolved shapefile, make neighbors
######################
shp <- readOGR(paste("/media/data/Project Files/Delamater/Dissertation/
  Clustering/Cluster_shapefiles/d/", cs[i], sep=""), layer=substr(cs[i],
  1, nchar(cs[1])-4), verbose=FALSE)

### Get neighbors
nb <- poly2nb(shp)

### One island poly doesn't have neighbors in the original file...
### It is connected by a bridge to a single zip code
### Assign neighbors to it manually
nb.mat <- nb2mat(nb,style="B",zero.policy=TRUE)
### Find regions with zero neighbors
w.zero <- as.numeric(which(rowSums(nb.mat) == 0))
```

```
### If island poly is not grouped, assign neighbors as 48193 (only for one)
if (length(w.zero) > 0) {
  near.clust <- cl[cl$ZIP == 48193, i+1]
  nb[[w.zero]] <- near.clust
  nb[[near.clust]] <- as.integer(c(w.zero, nb[[near.clust]]))
}

listw <- nb2listw(nb, style = "B", zero.policy=TRUE)

### Do PCA on ACS and LV
pca.h <- principal(cbind(acs$AgeAdjPatDayRateEBadj,
  lv$AgeAdjPatDayRateEBadj), nfactors=1, rotate="varimax",
  scores=TRUE)
acs.lv.scores <- pca.h$scores

# Regress on ETHPC1
acs.lv.lm <- lm(acs.lv.scores ~ lm$ETHPC1)
write.csv(summary(acs.lv.lm)$coefficients, file=
  paste("/home/delamate/MDCH/data/dissertation/regressions/output_tables/
  pre.regression.stats/ACS.LowV/", names(cl)[i+1],".csv", sep=""))


######
###### SAR
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)
ut <- as.data.frame(scale(ut))

write.csv(cor(lm.dat), file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/
  correlation/",names(cl)[i+1],".csv", sep=""))
if (sum(abs(cor(lm.dat)) > 0.5) > ncol(lm.dat)) write.table
  (as.character(names(cl)[i+1]), file="/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/bad.correlation.regressions.txt",
  col.names = FALSE, row.names=FALSE, append=TRUE)

## Spatial regression
misar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw)

## If there are non-significant terms in the model...
while (sum(summary(misar)$Coef[2:(ncol(lm.dat)+1),4] > sig) > 0) {
```

```
  ## Remove and remodel
  bad.t <- as.numeric(which(summary(misar)$Coef[2:(ncol(lm.dat)+1),
    4] == max(summary(misar)$Coef[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  misar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(misar)$Coef
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/SAR/
  betas/",names(cl)[i+1],".csv", sep=""))

write.csv(residuals(misar), file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/SAR/residuals/",names(cl)[i+1],
  ".csv", sep=""))

write.csv(misar$fit$signal_trend, file=paste("/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/SAR/fitted/effect/",
  names(cl)[i+1],".csv", sep=""))

write.csv(misar$fit$signal_stochastic, file=paste("/home/delamate/
  MDCH/data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/SAR/fitted/spatial/",
  names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(misar), listw, randomisation=FALSE)

stats <- c(misar$lambda, as.numeric(summary(misar)$LR1$p.value),
  misar$LL, misar$LL0, misar$fit$s2, AIC(misar),
  as.numeric(summary(misar, Nagel=TRUE)$NK), mt$statistic, mt$p.value)
names(stats) <- c("lambda", "lambda.p", "LL", "LL0", "s2", "AIC",
  "NagelR2", "Moran", "Moran.p")
```

```
write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/SAR/
  sig/",names(cl)[i+1],".csv", sep=""))

## Test SAR for heteroskedasticity
write.csv(leveneTest(residuals(misar),factor(fits.groups))[1,],
   file=paste("/home/delamate/MDCH/data/dissertation/regressions/
   output_tables/scale.regression.stats.ACS.LowV.dep/levene/SAR/2/",
   names(cl)[i+1],".csv", sep=""))
write.csv(leveneTest(residuals(misar),factor(q.fits.groups))[1,],
   file=paste("/home/delamate/MDCH/data/dissertation/regressions/
   output_tables/scale.regression.stats.ACS.LowV.dep/levene/SAR/5/",
   names(cl)[i+1],".csv", sep=""))
resid.med <- median(residuals(misar))
r.fits.groups <- residuals(misar) <= resid.med
write.csv(leveneTest(residuals(misar),factor(r.fits.groups))[1,],
   file=paste("/home/delamate/MDCH/data/dissertation/regressions/
   output_tables/scale.regression.stats.ACS.LowV.dep/levene/SAR/
   resid2/",names(cl)[i+1],".csv", sep=""))
r.qx <- quantile(residuals(misar), probs=seq(0,1,0.2))
rq.fits.groups <- cut(residuals(misar), r.qx, include.lowest = TRUE)
write.csv(leveneTest(residuals(misar),factor(rq.fits.groups))[1,],
   file=paste("/home/delamate/MDCH/data/dissertation/regressions/
   output_tables/scale.regression.stats.ACS.LowV.dep/levene/SAR/
   resid5/",names(cl)[i+1],".csv", sep=""))

######
###### weighted SAR
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)

## Spatial regression
w.misar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, weights=i.pop)

## If there are non-significant terms in the model...
while (sum(summary(w.misar)$Coef[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(w.misar)$Coef[2:(ncol(lm.dat)+1),
    4] == max(summary(w.misar)$Coef[2:(ncol(lm.dat)+1),4])))
```

```
  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  w.misar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, weights=i.pop)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(w.misar)$Coef
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/wSAR/
  betas/",names(cl)[i+1],".csv", sep=""))

write.csv(residuals(w.misar), file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wSAR/residuals/",names(cl)[i+1],
  ".csv", sep=""))

write.csv(w.misar$fit$signal_trend, file=paste("/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wSAR/fitted/effect/",names(cl)[i+1],
  ".csv", sep=""))

write.csv(w.misar$fit$signal_stochastic, file=paste("/home/delamate/
  MDCH/data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wSAR/fitted/spatial/",names(cl)[i+1],
  ".csv", sep=""))

mt <- moran.test(residuals(w.misar), listw, randomisation=FALSE)

stats <- c(w.misar$lambda, as.numeric(summary(w.misar)$LR1$p.value),
  w.misar$LL, w.misar$LL0, w.misar$fit$s2, AIC(w.misar),
  as.numeric(summary(w.misar, Nagel=TRUE)$NK), mt$statistic,
  mt$p.value)
names(stats) <- c("lambda", "lambda.p", "LL", "LL0", "s2", "AIC",
  "NagelR2", "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/wSAR/
  sig/",names(cl)[i+1],".csv", sep=""))
```

```
## Test wSAR for heteroskedasticity
write.csv(leveneTest(residuals(w.misar),factor(fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/wSAR/2/",
  names(cl)[i+1],".csv", sep=""))
write.csv(leveneTest(residuals(w.misar),factor(q.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/wSAR/5/",
  names(cl)[i+1],".csv", sep=""))
resid.med <- median(residuals(w.misar))
r.fits.groups <- residuals(w.misar) <= resid.med
write.csv(leveneTest(residuals(w.misar),factor(r.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/wSAR/
  resid2/",names(cl)[i+1],".csv", sep=""))
r.qx <- quantile(residuals(w.misar), probs=seq(0,1,0.2))
rq.fits.groups <- cut(residuals(w.misar), r.qx, include.lowest = TRUE)
write.csv(leveneTest(residuals(w.misar),factor(rq.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/wSAR/resid5/",
  names(cl)[i+1],".csv", sep=""))


######
###### CAR
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)

## Spatial regression
micar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, family="CAR")

## If there are non-significant terms in the model...
while (sum(summary(micar)$Coef[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(micar)$Coef[2:(ncol(lm.dat)+1),
    4] == max(summary(micar)$Coef[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")
```

```
  lm.dat <- lm.dat[,-bad.t]

  micar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, family="CAR")

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(micar)$Coef
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/CAR/
  betas/",names(cl)[i+1],".csv", sep=""))

write.csv(residuals(micar), file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/CAR/residuals/",
  names(cl)[i+1],".csv", sep=""))

write.csv(micar$fit$signal_trend, file=paste("/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/CAR/fitted/effect/",names(cl)[i+1],
  ".csv", sep=""))

write.csv(micar$fit$signal_stochastic, file=paste("/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/CAR/fitted/spatial/",
  names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(micar), listw, randomisation=FALSE)

stats <- c(micar$lambda, as.numeric(summary(micar)$LR1$p.value),
  micar$LL, micar$LL0, micar$fit$s2, AIC(micar),
  as.numeric(summary(micar, Nagel=TRUE)$NK), mt$statistic, mt$p.value)
names(stats) <- c("lambda", "lambda.p", "LL", "LL0", "s2", "AIC",
  "NagelR2", "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/CAR/
  sig/",names(cl)[i+1],".csv", sep=""))

## Test CAR for heteroskedasticity
write.csv(leveneTest(residuals(micar),factor(fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
```

```
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/CAR/2/",
  names(cl)[i+1],".csv", sep=""))
write.csv(leveneTest(residuals(micar),factor(q.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/CAR/5/",
  names(cl)[i+1],".csv", sep=""))
resid.med <- median(residuals(micar))
r.fits.groups <- residuals(micar) <= resid.med
write.csv(leveneTest(residuals(micar),factor(r.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/CAR/resid2/",
  names(cl)[i+1],".csv", sep=""))
r.qx <- quantile(residuals(micar), probs=seq(0,1,0.2))
rq.fits.groups <- cut(residuals(micar), r.qx, include.lowest = TRUE)
write.csv(leveneTest(residuals(micar),factor(rq.fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/CAR/resid5/",
  names(cl)[i+1],".csv", sep=""))


######
###### weighted CAR
######

lm.dat <- as.matrix(cbind(residuals(acs.lv.lm), lm))
colnames(lm.dat)[1] <- "ACSLowVPCresid"
lm.dat <- scale(lm.dat)

## Spatial regression
w.micar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, family="CAR",
  weights=i.pop)

## If there are non-significant terms in the model...
while (sum(summary(w.micar)$Coef[2:(ncol(lm.dat)+1),4] > sig) > 0) {

  ## Remove and remodel
  bad.t <- as.numeric(which(summary(w.micar)$Coef[2:(ncol(lm.dat)+1),
    4] == max(summary(w.micar)$Coef[2:(ncol(lm.dat)+1),4])))

  if (ncol(lm.dat) == 2) name <- paste("lm.dat", colnames(lm.dat)[-bad.t],
    sep="")

  lm.dat <- lm.dat[,-bad.t]

  w.micar <- spautolm(ut$StRateDif ~ lm.dat, listw = listw, family="CAR",
```

```
     weights=i.pop)

  if (is.vector(lm.dat) == TRUE) break

}

lm.sum <- summary(w.micar)$Coef
if (is.vector(lm.dat) == TRUE) rownames(lm.sum)[2] <- name

write.csv(lm.sum, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/wCAR/
  betas/",names(cl)[i+1],".csv", sep=""))

write.csv(residuals(w.micar), file=paste("/home/delamate/MDCH/data/
  dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wCAR/residuals/",names(cl)[i+1],
  ".csv", sep=""))

write.csv(w.micar$fit$signal_trend, file=paste("/home/delamate/MDCH/
  data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wCAR/fitted/effect/",
  names(cl)[i+1],".csv", sep=""))

write.csv(w.micar$fit$signal_stochastic, file=paste("/home/delamate/
  MDCH/data/dissertation/regressions/output_tables/
  scale.regression.stats.ACS.LowV.dep/wCAR/fitted/spatial/",
  names(cl)[i+1],".csv", sep=""))

mt <- moran.test(residuals(w.micar), listw, randomisation=FALSE)

stats <- c(w.micar$lambda, as.numeric(summary(w.micar)$LR1$p.value),
  w.micar$LL, w.micar$LL0, w.micar$fit$s2, AIC(w.micar),
  as.numeric(summary(w.micar, Nagel=TRUE)$NK), mt$statistic, mt$p.value)
names(stats) <- c("lambda", "lambda.p", "LL", "LL0", "s2", "AIC",
  "NagelR2", "Moran", "Moran.p")

write.csv(stats, file=paste("/home/delamate/MDCH/data/dissertation/
  regressions/output_tables/scale.regression.stats.ACS.LowV.dep/wCAR/sig/",
  names(cl)[i+1],".csv", sep=""))

## Test wCAR for heteroskedasticity
write.csv(leveneTest(residuals(w.micar),factor(fits.groups))[1,],
  file=paste("/home/delamate/MDCH/data/dissertation/regressions/
  output_tables/scale.regression.stats.ACS.LowV.dep/levene/wCAR/2/",
  names(cl)[i+1],".csv", sep=""))
```

```
  write.csv(leveneTest(residuals(w.micar),factor(q.fits.groups))[1,],
    file=paste("/home/delamate/MDCH/data/dissertation/regressions/
    output_tables/scale.regression.stats.ACS.LowV.dep/levene/wCAR/5/",
    names(cl)[i+1],".csv", sep=""))
  resid.med <- median(residuals(w.micar))
  r.fits.groups <- residuals(w.micar) <= resid.med
  write.csv(leveneTest(residuals(w.micar),factor(r.fits.groups))[1,],
    file=paste("/home/delamate/MDCH/data/dissertation/regressions/
    output_tables/scale.regression.stats.ACS.LowV.dep/levene/wCAR/resid2/",
    names(cl)[i+1],".csv", sep=""))
  r.qx <- quantile(residuals(w.micar), probs=seq(0,1,0.2))
  rq.fits.groups <- cut(residuals(w.micar), r.qx, include.lowest = TRUE)
  write.csv(leveneTest(residuals(w.micar),factor(rq.fits.groups))[1,],
    file=paste("/home/delamate/MDCH/data/dissertation/regressions/
    output_tables/scale.regression.stats.ACS.LowV.dep/levene/wCAR/resid5/",
    names(cl)[i+1],".csv", sep=""))


}

print(Sys.time() - start)
```

REFERENCES

REFERENCES

Aday, L.A., Andersen, R., 1974. A framework tor the study of access to medical care. Health Services Research 9, 208–220.

Alexander, J.A., Lee, S.D., Griffith, J.R., Mick, S.S., Lin, X., Banaszak-Holl, J., 1999. Do Market-Level hospital and physician resources affect small area variation in hospital use? Medical Care Research and Review 56, 94–117.

Andersen, R., Newman, J.F., 1973. Societal and individual determinants of medical care utilization in the united states. The Milbank Memorial Fund Quarterly: Health and Society 51, 95–124.

Angell, M., 2008. Privatizing health care is not the answer: lessons from the united states. Canadian Medical Association Journal 179, 916–919.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.

Anselin, L., 2003. Spatial externalities, spatial multipliers, and spatial econometrics. International Regional Science Review 26, 153–166.

Apparicio, P., Abdelmajid, M., Riva, M., Shearmur, R., 2008. Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. International Journal of Health Geographics 7, 1–14.

Arcury, T.A., Gesler, W.M., Preisser, J.S., Sherman, J., Spencer, J., Perin, J., 2005. The effects of geography and spatial behavior on health care utilization among the residents of a rural region. Health Services Research 40, 135–156.

Banks, D.A., Foreman, S.E., Keeler, T.E., 1999. Cross-subsidization in hospital care: some lessons from the law and economics of regulation. Health Matrix 9, 1–35.

Bay, K.S., Nestman, L.J., 1984. The use of bed distribution and service population indexes for hospital bed allocation. Health Services Research 19, 141–160.

van Bemmelen, J., Quak, W., van Hekken, M., van Oosterom, P., 1993. Vector vs. raster-based algorithms for cross country movement planning, in: Proceedings Auto-Carto, pp. 304–317.

Berke, E., Shi, X., 2009. Computing travel time when the exact address is unknown: a comparison of point and polygon ZIP code approximation methods. International Journal of Health Geographics 8, 1–9.

Berry Jr., R.E., 1973. On grouping hospitals for economic analysis. Inquiry 10, 5–12.

Bindman, A.B., Grumbach, K., Osmond, D., Komaromy, M., Vranizan, K., Lurie, N., Billings, J., Stewart, A., 1995. Preventable hospitalizations and access to health care. JAMA: The Journal of the American Medical Association 274, 305–311.

Birkmeyer, J.D., Siewers, A.E., Marth, N.J., Goodman, D.C., 2003. Regionalization of high-risk surgery and implications for patient travel times. The Journal of the American Medical Association 290, 2703–2708.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. Use R!, Springer, New York, NY.

Bosanac, E.M., Parkinson, R.C., Hall, D.S., 1976. Geographic access to hospital care: A 30-Minute travel time standard. Medical Care 14, 616–624.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis 28, 281–298.

BuiltWith Trends, 2012. Top in mapping. http://trends.builtwith.com/mapping/top, last accessed: 2012.

Carr, B.G., Branas, C.C., Metlay, J.P., Sullivan, A.F., Camargo Jr., C.A., 2009. Access to emergency care in the united states. Annals of Emergency Medicine 54, 261–269.

Chi, G., Zhu, J., 2008. Spatial regression models for demographic analysis. Population Research and Policy Review 27, 17–42.

Clark, J.D., 1990. Variation in michigan hospital use rates: Do physician and hospital characteristics provide the explanation? Social Science & Medicine 30, 67–82.

Conover, C.J., Sloan, F.A., 1998. Does removing Certificate-of-Need regulations lead to a surge in health care spending? Journal of Health Politics, Policy and Law 23, 455–481.

Conover, C.J., Sloan, F.A., 2003. Evaluation of Certificate of Need in Michigan. Technical Report. Center for Health Policy, Law and Management, Duke University.

Couclelis, H., 1992. People manipulate objects (but cultivate fields): Beyond the raster-vector debate in gis, in: Frank, A.V., Campari, I., Formentini, U. (Eds.), Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Springer Berlin / Heidelberg. volume 639 of *Lecture Notes in Computer Science*, pp. 65–77.

Cromley, E.K., McLafferty, S., 2002. GIS and Public Health. Guilford Press, New York.

Cunningham, P.J., 2010. The growing financial burden of health care: National and state trends, 2001–2006. Health Affairs 29, 1037–1044.

Current, J.R., Schilling, D.A., 1990. Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems. Geographical Analysis 22, 116–126.

Dai, D., 2010. Black residential segregation, disparities in spatial access to health care facilities, and late-stage breast cancer diagnosis in metropolitan detroit. Health & Place 16, 1038–1052.

Darden, J., Rahbar, M., Jezierski, L., Li, M., Velie, E., 2010. The measurement of neighborhood socioeconomic characteristics and black and white residential segregation in metropolitan detroit: Implications for the study of social disparities in health. Annals of the Association of American Geographers 100, 137–158.

Darden, J.T., Stokes, C., Thomas, R.W., 2007. The state of Black Michigan, 1967-2007. Michigan State University Press, East Lansing.

Delamater, P.L., Messina, J.P., Shortridge, A.M., Grady, S.C., 2012. Measuring geographic access to health care: raster and network-based methods. International Journal of Health Geographics (In press).

Delamater, P.L., Shortridge, A.M., Messina, J.P., Under review. Regional health care planning: a methodology to cluster facilities using community utilization patterns. BMC Health Services Research .

Diez Roux, A.V., Merkin, S.S., Arnett, D., Chambless, L., Massing, M., Nieto, F.J., Sorlie, P., Szklo, M., Tyroler, H.A., Watson, R.L., 2001. Neighborhood of residence and incidence of coronary heart disease. New England Journal of Medicine 345, 99–106.

Donabedian, A., 1972. Models for organizing the delivery of personal health services and criteria for evaluating them. The Milbank Memorial Fund Quarterly 50, 103–154.

van Doorslaer, E., van Vliet, R., 1989. "A built bed is a filled bed?" an empirical reexamination. Social Science & Medicine 28, 155–164.

Drineas, P., Frieze, A., Kannan, R., Vempala, S., Vinay, V., 2004. Clustering large graphs via the singular value decomposition. Machine Learning 56, 9–33.

Dubbs, N.L., Bazzoli, G.J., Shortell, S.M., Kralovec, P.D., 2004. Reexamining organizational configurations: An update, validation, and expansion of the taxonomy of health networks and systems. Health Services Research 39, 207–220.

Dubes, R., Jain, A.K., 1976. Clustering techniques: The user's dilemma. Pattern Recognition 8, 247–260.

Elayat, H., Murphy, B., Prabhakar, N., 1978. Entropy in the hierarchical cluster analysis of hospitals. Health Services Research 13, 395–403.

ESRI, 2010. Algorithms used by network analyst. http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Algorithms_used_by_Network_Analy

Feldstein, P.J., 1966. Research on the demand for health services. The Milbank Memorial Fund Quarterly 44, 128–165.

Ferrier, G., Leleu, H., Valdmanis, V., 2010. The impact of CON regulation on hospital efficiency. Health Care Management Science 13, 84–100.

Finn, M., 2007. Health care demand in Michigan: An examination of the Michigan Certificate of Need acute care bed need methodology. Ph.D. thesis. Michigan State University.

Fisher, E.S., Wennberg, D.E., Stukel, T.A., Gottlieb, D.J., Lucas, F.L., Pinder, E.L., 2003. The implications of regional variations in medicare spending. part 1: The content, quality, and accessibility of care. Annals of Internal Medicine 138, 273–287.

Fisher, E.S., Wennberg, J.E., Stukel, T.A., Sharp, S.M., 1994. Hospital readmission rates for cohorts of medicare beneficiaries in boston and new haven. New England Journal of Medicine 331, 989–995.

Folland, S., Stano, M., 1990. Small area variations: A critical review of propositions, methods, and evidence. Medical Care Research and Review 47, 419–465.

Fotheringham, A.S., Wong, D.W.S., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environment and Planning A 23, 1025–1044.

Freeman, J.D., Kadiyala, S., Bell, J.F., Martin, D.P., 2008. The causal effect of health insurance on utilization and outcomes in adults: A systematic review of US studies. Medical Care 46, 1023–1032.

Frizzelle, B., Evenson, K., Rodriguez, D., Laraia, B., 2009. The importance of accurate road data for spatial applications in public health: customizing a road network. International Journal of Health Geographics 8, 1–24.

Garnick, D., Luft, H., Robinson, J., Tetreault, J., 1987. Appropriate measures of hospital market areas. Health Services Research 22, 69–89.

Gilmour, S.J., 2010. Identification of hospital catchment areas using clustering: an example from the NHS. Health Services Research 45, 497–513.

Ginsburg, P.B., Koretz, D.M., 1983. Bed availability and hospital utilization: estimates of the "Roemer effect". Health Care Finance Review 5, 87–92.

Goodchild, M.F., Yuan, M., Cova, T.J., 2007. Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science 21, 239–260.

Goodman, D.C., Fisher, E., Stukel, T.A., Chang, C., 1997. The distance to community medical care and the likelihood of hospitalization: is closer always better? American Journal of Public Health 87, 1144–1150.

Goodman, D.C., Mick, S.S., Bott, D., Stukel, T., Chang, C.h., Marth, N., Poage, J., Carretta, H.J., 2003. Primary care service areas: A new tool for the evaluation of primary care services. Health Services Research 38, 287–309.

Grady, S.C., 2006. Racial disparities in low birthweight and the contribution of residential segregation: A multilevel analysis. Social Science & Medicine 63, 3013–3029.

Grady, S.C., 2010. Racial residential segregation impacts on low birth weight using improved neighborhood boundary definitions. Spatial and Spatio-temporal Epidemiology 1, 239–249.

Grady, S.C., Ramírez, I.J., 2008. Mediating medical risk factors in the residential segregation and low birthweight relationship by race in new york city. Health & Place 14, 661–677.

Graham, M.H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84, 2809–2815.

Granderson, G., 2011. The impacts of hospital alliance membership, alliance size, and repealing certificate of need regulation, on the cost efficiency of non-profit hospitals. Managerial and Decision Economics 32, 159–173.

Green, L.A., Fryer, G.E., Yawn, B.P., Lanier, D., Dovey, S.M., 2001. The ecology of medical care revisited. New England Journal of Medicine 344, 2021–2025.

Greene, J., 2012. McLaren's plan b if CON is rejected: Legislators. http://www.crainsdetroit.com/article/20120212/HEALTH/302129950/mclaren-s-plan-b-if-con-is-rejected-legislators.

Griffith, J.R., 1972. Quantitative Techniques for Hospital Planning and Control. Lexington Books, Lexington, MA.

Griffith, J.R., Restuccia, J.D., Tedeschi, P.J., Wilson, P.A., Zuckerman, H.S., 1981. Measuring community hospital service in michigan. Health Services Research 16, 135–160.

Guagliardo, M., 2004. Spatial accessibility of primary care: concepts, methods and challenges. International Journal of Health Geographics 3, 1–13.

Gujarati, D.N., 1988. Basic Econometrics. McGraw-Hill.

Harris, D.M., 1975. An elaboration of the relationship between general hospital bed supply and general hospital utilization. Journal of Health and Social Behavior 16, 163–172.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108.

Haynes, R., Jones, A., Sauerzapf, V., Zhao, H., 2006. Validation of travel times to hospital estimated by gis. International Journal of Health Geographics 5, 1–8.

Hellinger, F., 2009. The effect of certificate-of-need laws on hospital beds and healthcare expenditures: an empirical analysis. American Journal of Managed Care 15, 737–744.

Hewko, J., Smoyer-Tomic, K.E., Hodgson, M.J., 2002. Measuring neighbourhood spatial accessibility to urban amenities: does aggregation error matter? Environment and Planning A 34, 1185–1206.

Higgs, G., 2004. A literature review of the use of gis-based measures of access to health care services. Health Services and Outcomes Research Methodology 5, 119–139.

Higgs, G., 2009. The role of GIS for health utilization studies: literature review. Health Services and Outcomes Research Methodology 9, 84–99.

Hofer, A.N., Abraham, J.M., Moscovice, I., 2011. Expansion of coverage under the patient protection and affordable care act and primary care utilization. Milbank Quarterly 89, 69–89.

Hopkins, C., 2012. McLaren's request to transfer hospital beds from pontiac causes divide. The Oakland Press .

Hunter, J.M., Shannon, G.W., Sambrook, S.L., 1986. Rings of madness: Service areas of 19th century asylums in north america. Social Science & Medicine 23, 1033–1050.

Illinois General Assembly, 2012. Illinois health care facilities plan: Narrative and planning policies: Introduction formula components, planning area development policies, and normal travel time determinations. http://www.ilga.gov/commission/jcar/admincode/077/077011000D05100R.html, last accessed: 2012.

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Series in Statistics, Springer. 2nd edition.

Jones, S.G., Ashby, A.J., Momin, S.R., Naidoo, A., 2010. Spatial implications associated with using euclidean distance measurements and geographic centroid imputation in health care research. Health Services Research 45, 316–327.

Joseph, A.E., Phillips, D.R., 1984. Accessibility and utilization: geographical perspectives on health care delivery. Harper & Row Ltd, London.

Kaiser Family Foundation, 2009. Health Care Costs: A Primer. Technical Report 7670-02.

Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience.

Khan, A.A., 1992. An integrated approach to measuring potential spatial access to health care services. Socio-Economic Planning Sciences 26, 275–287.

Klastorin, T., Watts, C., 1981. The determination of alternative hospital classifications. Health Services Research 16, 205–220.

Klastorin, T.D., Watts, C.A., 1982. A current reappraisal of berry's hospital typology. Medical Care 20, 441–449.

Klauss, G., Staub, L., Widmer, M., Busato, A., 2005. Hospital service areas - a new tool for health care planning in switzerland. BMC Health Services Research 5, 1–15.

Klein, R.J., Proctor, S.E., Boudreault, M.A., Turczyn, K.M., 2002. Healthy People 2010 Criteria for Data Suppression. Technical Report 24. National Center for Health Statistics. Hyattsville, Maryland.

Kravet, S.J., Shore, A.D., Miller, R., Green, G.B., Kolodner, K., Wright, S.M., 2008. Health care utilization and the proportion of primary care physicians. The American Journal of Medicine 121, 142–148.

Kroneman, M., Siegers, J.J., 2004. The effect of hospital bed reduction on the use of beds: A comparative study of 10 european countries. Social Science & Medicine 59, 1731–1740.

Kuttner, R., 2008. Market-Based failure — a second opinion on U.S. health care costs. New England Journal of Medicine 358, 549–551.

Kwan, M.P., Hong, X.D., 1998. Network-based constraints-oriented choice set formation using gis. Geographical Systems , 139–162.

Langley, S., Fuller, S., Messina, J., Shortridge, A., Grady, S., 2010. A methodology for projecting hospital bed need: a michigan case study. Source Code for Biology and Medicine 5, 1–10.

Longley, P.A., Goodchild, M., Maguire, D.J., Rhind, D.W., 2010. Geographic Information Systems and Science. Wiley. 3rd edition.

Lopez-Quilez, A., Munoz, F., 2009. Geostatistical computing of acoustic maps in the presence of barriers. Mathematical and Computer Modelling 50, 929–938.

Luginaah, I., Jerrett, M., Elliott, S., Eyles, J., Parizeau, K., Birch, S., Abernathy, T., Veenstra, G., Hutchinson, B., Giovis, C., 2001. Health profiles of hamilton: Spatial characterisation of neighbourhoods for health investigations. GeoJournal 53, 135–147.

Luke, R.D., 2006. Taxonomy of health networks and systems: A reassessment. Health Services Research 41, 618–628.

Luo, W., Qi, Y., 2009. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. Health & Place 15, 1100–1107.

Luo, W., Wang, F., 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, 865–884.

Marshall, R.J., 1991. Mapping disease and mortality rates using empirical bayes estimators. Journal of the Royal Statistical Society. Series C (Applied Statistics) 40, 283–294.

Martin, D., Wrigley, H., Barnett, S., Roderick, P., 2002. Increasing the sophistication of access measurement in a rural healthcare study. Health & Place 8, 3–13.

McGinley, P.J., 1995. Beyond health care reform: Reconsidering certificate of need laws in a "Managed competition" system. Florida State University Law Review 23, 141–188.

McGrail, M.R., Humphreys, J.S., 2009. Measuring spatial accessibility to primary care in rural areas: Improving the effectiveness of the two-step floating catchment area method. Applied Geography 29, 533–541.

McGuirk, M.A., Porell, F.W., 1984. Spatial patterns of hospital utilization: the impact of distance and time. Inquiry 21, 84–95.

McLafferty, S.L., 2003. Gis and health care. Annual Reviews in Public Health 24, 25–42.

Meade, M., Emch, M., 2010. Medical Geography. Guilford Publications, New York. third edition.

Messina, J.P., Shortridge, A.M., Groop, R.E., Varnakovida, P., Finn, M.J., 2006. Evaluating michigan's community hospital access: spatial methods for decision support. International Journal of Health Geographics 5, 1–18.

Michigan Department of Community Health, 2009. Certificate of need review standards for hospital beds.

Michigan Office of Highway Safety Planning, . Establishing Realistic Speed Limits. Technical Report OHSP 894. Lansing, MI.

Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.

Milligan, G.W., Cooper, M.C., 1987. Methodology review: Clustering methods. Applied Psychological Measurement 11, 329–354.

Milligan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. Journal of Classification 5, 181–204.

Mobley, L.R., Kuo, T.M., Andrews, L., 2008. How sensitive are multilevel regression findings to defined area of context? Medical Care Research and Review 65, 315–337.

Mulley, A.G., 2009. Inconvenient truths about supplier induced demand and unwarranted variation in medical practice. BMJ 339, 1007–1009.

Murray, C.J.L., Frenk, J., 2010. Ranking 37th — measuring the performance of the U.S. health care system. New England Journal of Medicine 362, 98–99.

Nallamothu, B.K., Bates, E.R., Wang, Y., Bradley, E.H., Krumholz, H.M., 2006. Driving times and distances to hospitals with percutaneous coronary intervention in the united states: Implications for prehospital triage of patients with st-elevation myocardial infarction. Circulation 113, 1189–1195.

National Conference of State Legislatures, 2011. CON-Certificate of need state laws. http://www.ncsl.org/issues-research/health/con-certificate-of-need-state-laws.aspx.

New York State Department of Health, 2012. Title 10 NYCRR: part 709 - determination of public need for medical facility construction. http://www.health.state.ny.us/nysdoh/rules/709.htm, last accessed: 2012.

Ngui, A., Apparicio, P., 2011. Optimizing the two-step floating catchment area method for measuring spatial accessibility to medical clinics in montreal. BMC Health Services Research 11, 1–12.

Norris, J.C., Van der laan, M.J., Lane, S., Anderson, J.N., Block, G., 2003. Nonlinearity in demographic and behavioral determinants of morbidity. Health Services Research 38, 1791–1818.

North Carolina Department of Health and Human Services, 2012. North carolina 2012 state medical facilities plan. http://www.ncdhhs.gov/dhsr/ncsmfp/2012/2012smfp.pdf, last accessed: 2012.

Oakes, J.M., 2004. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social Science & Medicine 58, 1929–1952.

Odoi, A., Martin, S.W., Michel, P., Holt, J., Middleton, D., Wilson, J., 2003. Geographical and temporal distribution of human giardiasis in ontario, canada. International Journal of Health Geographics 2, 1–13.

Oleske, D.M., 2009. An epidemiologic framework for the delivery of health care services, in: Epidemiology and the Delivery of Health Care Services. Springer US, pp. 3–30.

Onega, T., Duell, E.J., Shi, X., Wang, D., Demidenko, E., Goodman, D., 2008. Geographic access to cancer care in the U.S. Cancer 112, 909–918.

Openshaw, S., 1984. The modifiable areal unit problem. GeoBooks, Norwich, UK.

Paez, A., Mercado, R., Farber, S., Morency, C., Roorda, M., 2010. Accessibility to health care facilities in montreal island: an application of relative accessibility indicators from the perspective of senior and non-senior residents. International Journal of Health Geographics 9, 1–15.

Pasley, B.H., Lagoe, R.J., Marshall, N.O., 1995. Excess acute care bed capacity and its causes: the experience of new york state. Health Services Research 30, 115–131.

Paul-Shaheen, P., Carpenter, E.S., 1982. Legislating hospital bed reduction: The michigan experience. Journal of Health Politics, Policy and Law 6, 653–675.

Pedigo, A.S., Odoi, A., 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, 924–930.

Penchansky, R., Thomas, J.W., 1981. The concept of access: Definition and relationship to consumer satisfaction. Medical Care 19, 127–140.

Phibbs, C.S., Luft, H.S., 1995. Correlation of travel time on roads versus straight line distance. Medical Care Research and Review 52, 532–542.

Price, M., 2008. Slopes, sharp turns, and speed. ArcUser , 50–55.

Price, M., 2009. Convincing the chief. ArcUser , 50–54.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. Technical Report. Vienna, Austria.

Radke, J., Mu, L., 2000. Spatial decompositions, modeling and mapping service regions to predict access to social programs. Annals of GIS 6, 105–112.

Ray, N., Ebener, S., 2008. Accessmod 3.0: computing geographic coverage and accessibility to health care services using anisotropic movement of patients. International Journal of Health Geographics 7, 1–17.

Rey, S.J., Anselin, L., Folch, D.C., Arribas-Bel, D., Sastré Gutiérrez, M.L., Interlante, L., 2011. Measuring spatial dynamics in metropolitan areas. Economic Development Quarterly 25, 54–64.

Ricketts, T.C., Randolph, R., Howard, H.A., Pathman, D., Carey, T., 2001. Hospitalization rates as indicators of access to primary care. Health & Place 7, 27–38.

Rivers, P.A., Fottler, M.D., Younis, M.Z., 2007. Does certificate of need really contain hospital costs in the united states? Health Education Journal 66, 229–244.

Robeznieks, A., 2008. Site under construction. Modern Healthcare 38, 6–7, 16.

Roemer, M.I., 1961. Bed supply and hospital utilization: a natural experiment. Hospitals 35, 36–42.

Rogerson, P.A., 2006. Statistical Methods for Geography: A Student's Guide. SAGE Publications. 2nd edition.

Rohrer, J., 1990. Supply-induced demand for hospital care. Health Services Management Research 3, 41–48.

Romano, M., 2003. Pros and cons of certificates. Modern Healthcare 33, 4.

Rosenthal, M.B., Zaslavsky, A., Newhouse, J.P., 2005. The geographic distribution of physicians revisited. Health Services Research 40, 1931–1952.

Sander, H.A., Ghosh, D., van Riper, D., Manson, S.M., 2010. How do you measure distance in spatial models? an example using open-space valuation. Environment and Planning B: Planning and Design 37, 874–894.

Schoen, C., Doty, M.M., Robertson, R.H., Collins, S.R., 2011. Affordable care act reforms could reduce the number of underinsured US adults by 70 percent. Health Affairs 30, 1762 –1771.

Schuurman, N., Berube, M., Crooks, V.A., 2010. Measuring potential spatial access to primary health care physicians using a modified gravity model. Canadian Geographer 54, 29–45.

Schuurman, N., Fiedler, R., Grzybowski, S., Grund, D., 2006. Defining rational hospital catchments for non-urban areas based on travel-time. International Journal of Health Geographics 5, 1–8.

Shahid, R., Bertazzon, S., Knudtson, M., Ghali, W., 2009. Comparison of distance measures in spatial analytical modeling for health service planning. BMC Health Services Research 9, 1–14.

Shain, M., Roemer, M.I., 1959. Hospital costs relate to the supply of beds. Modern Hospital 92, 71–73.

Shannon, G.W., Skinner, J.L., Bashshur, R.L., 1973. Time and distance: the journey for medical care. International Journal of Health Services 3, 237–244.

Shortt, N.K., Moore, A., Coombes, M., Wymer, C., 2005. Defining regions for locality health care planning: a multidimensional approach. Social Science & Medicine 60, 2715–2727.

Shwartz, M., Payne, S.M., Restuccia, J.D., Ash, A.S., 2001. Does it matter how small geographic areas are constructed? ward's algorithm versus the plurality rule. Health Services and Outcomes Research Methodology 2, 5–18.

Shwartz, M., Peköz, E.A., Labonte, A., Heineke, J., Restuccia, J.D., 2011. Bringing responsibility for small area variations in hospitalization rates back to the hospital: The propensity to hospitalize index and a test of the roemer's law. Medical Care 49, 1062–1067.

Sparks, P.J., Sparks, C.S., 2010. An application of spatially autoregressive models to the study of US county mortality rates. Population, Space and Place 16, 465–481.

Steinley, D., 2003. Local optima in K-Means clustering: What you don't know may hurt you. Psychological Methods 8, 294–304.

Steinley, D., 2006. K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology 59, 1–34.

Steinley, D., Banks, D., House, L., McMorris, F.R., Arabie, P., Gaul, W., 2004. Standardizing variables in k-means clustering, in: Classification, Clustering, and Data Mining Applications. Springer Berlin Heidelberg. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 53–60.

Strunk, B.C., Ginsburg, P.B., Banker, M.I., 2006. The effect of population aging on future hospital demand. Health Affairs 25, w141–w149.

Swift, A., Liu, L., Uber, J., 2008. Reducing MAUP bias of correlation statistics between water quality and GI illness. Computers, Environment and Urban Systems 32, 134–148.

Tanser, F., Gijsbertsen, B., Herbst, K., 2006. Modelling and understanding primary health care accessibility and utilization in rural south africa: An exploration using a geographical information system. Social Science & Medicine 63, 691–705.

Thomas, J.W., 1979. Techniques for defining geographic boundaries for health regions. Socio-Economic Planning Sciences 13, 321–326.

Thomas, J.W., Berki, S.E., Wyszewianski, L., Ashcraft, M.L.E., 1983. Classification of hospitals based on measured output: The VA system. Medical Care 21, 715–733.

Thomas, J.W., Griffith, J.R., Durance, P., 1981. Defining hospital clusters and associated service communities in metropolitan areas. Socio-Economic Planning Sciences 15, 45–51.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240.

Upchurch, C., Kuby, M., Zoldak, M., Barranda, A., 2004. Using GIS to generate mutually exclusive service areas linking travel on and off a network. Journal of Transport Geography 12, 23–33.

US Federal Trade Commission, 2004. Improving Health Care: A Dose of Competition. Technical Report. US Government Printing Office. Washington, D.C.

Vertrees, J.C., Manton, K.G., 1986. A multivariate approach for classifying hospitals and computing blended payment rates. Medical Care 24, 283–300.

de Vries, J.J., Nijkamp, P., Rietveld, P., 2009. Exponential or power distance-decay for commuting? an alternative specification. Environment and Planning A 41, 461–480.

Vyas, S., Kumaranayake, L., 2006. Constructing socio-economic status indices: how to use principal components analysis. Health Policy and Planning 21, 459–468.

Wan, N., Zhan, F.B., Zou, B., Chow, E., 2011. A relative spatial access assessment approach for analyzing potential spatial access to colorectal cancer services in texas. Applied Geography 32, 291–299.

Wan, N., Zou, B., Sternberg, T., 2012. A three-step floating catchment area method for analyzing spatial access to health services. International Journal of Geographical Information Science .

Wang, F., Xu, Y., 2011. Estimating O–D travel time matrix by google maps API: implementation, advantages, and implications. Annals of GIS 17, 199–209.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

Welch, H.G., Sharp, S.M., Gottlieb, D.J., Skinner, J.S., Wennberg, J.E., 2011. Geographic variation in diagnosis frequency and risk of death among medicare beneficiaries. JAMA: The Journal of the American Medical Association 305, 1113–1118.

Wennberg, J., 2005. Variation in Use of Medicare Services Among Regions and Selected Academic Medical Centers: Is More Better? Technical Report 874. The Commonwealth Fund.

Wennberg, J., Cooper, M., Dartmouth Atlas of Health Care Working Group, 1999. The Quality of Medical Care in the United States: A Report on the Medicare Program. Technical Report. American Hospital Association. Chicago, IL.

Wennberg, J., Gittelsohn, A., 1973. Small area variations in health care delivery. Science 182, 1102–1108.

White, K.L., Williams, T.F., Greenberg, B.G., 1961. The ecology of medical care. New England Journal of Medicine 265, 885–892.

Witlox, F., 2007. Evaluating the reliability of reported distance data in urban travel behaviour analysis. Journal of Transport Geography 15, 172–183.

Wright, D.B., Ricketts III, T.C., 2010. The road to efficiency? re-examining the impact of the primary care physician workforce on health care utilization rates. Social Science & Medicine 70, 2006–2010.

Young, T.K., 2005. Population Health: Concepts and Methods. Oxford University Press, New York.

Zwanziger, J., Khan, N., 2008. Safety-Net hospitals. Medical Care Research and Review 65, 478–495.