

THE INFLUENCE OF DISEASE MAPPING METHODS ON SPATIAL PATTERNS AND
NEIGHBORHOOD CHARACTERISTICS FOR HEALTH RISK

Warangkana Ruckthongsook

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2017

APPROVED:

Joseph R. Oppong, Major Professor
Chetan Tiwari, Committee Member
Prathiba Natesan, Committee Member
Sam Atkinson, Committee Member
Armin R. Mikler, Committee Member
Art Goven, Chair of the Department of
Biological Sciences

Su Gao, Dean of the College of Science
Victor Prybutok, Dean of the Toulouse
Graduate School

Ruckthongsook, Warangkana. *The Influence of Disease Mapping Methods on Spatial Patterns and Neighborhood Characteristics for Health Risk*. Doctor of Philosophy

(Environmental Science), December 2017, 103 pp., 10 tables, 13 figures, chapter references.

This thesis addresses three interrelated challenges of disease mapping and contributes a new approach for improving visualization of disease burdens to enhance disease surveillance systems. First, it determines an appropriate threshold choice (smoothing parameter) for the adaptive kernel density estimation (KDE) in disease mapping. The results show that the appropriate threshold value depends on the characteristics of data, and bandwidth selector algorithms can be used to guide such decisions about mapping parameters. Similar approaches are recommended for map-makers who are faced with decisions about choosing threshold values for their own data. This can facilitate threshold selection. Second, the study evaluates the relative performance of the adaptive KDE and spatial empirical Bayes for disease mapping. The results reveal that while the estimated rates at the state level computed from both methods are identical, those at the zip code level are slightly different. These findings indicate that using either the adaptive KDE or spatial empirical Bayes method to map disease in urban areas may provide identical rate estimates, but caution is necessary when mapping diseases in non-urban (sparsely populated) areas. This study contributes insights on the relative performance in terms of accuracy of visual representation and associated limitations. Lastly, the study contributes a new approach for delimiting spatial units of disease risk using straightforward statistical and spatial methods and social determinants of health. The results show that the neighborhood risk map not only helps in geographically targeting where but also in tailoring interventions in those areas to those high risk populations. Moreover, when health data is limited, the neighborhood risk map alone is adequate for identifying where and which populations are at risk. These findings will benefit public health tasks of planning and targeting appropriate intervention even in areas with limited

and poor-quality health data. This study not only fills the identified gaps of knowledge in disease mapping but also has a wide range of broader impacts. The findings of this study improve and enhance the use of the adaptive KDE method in health research, provide better awareness and understanding of disease mapping methods, and offer an alternative method to identify populations at risk in areas with limited health data. Overall, these findings will benefit public health practitioners and health researchers as well as enhance disease surveillance systems.

Copyright 2017

By

Warangkana Ruckthongsook

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from my friends, and support from my family. First, I would like to express my deep gratitude to my major professor, Dr. Joseph Oppong, for his excellent guidance, patience, caring, enthusiastic encouragement, and providing me with an excellent atmosphere for doing research. Thanks for being a great role model as a researcher and an educator. I would also like to thank Dr. Chetan Tiwari for his valuable and constructive suggestions during the development of this research work. His willingness to give his time so generously has been very much appreciated. My grateful thanks are also extended to Dr. Prathiba Natesan for her advice and furthering my thinking about research methodologies and to Drs. Armin Mikler and Sam Atkinson for their useful critiques of this research work. I would also like to express my gratitude to every professor and staff member who contributed to this research and providing me with an excellent education during my study at UNT.

I would like to also thank to my friends and colleagues in the College of Science, the College of Liberal Arts & Social Sciences, the College of Education, and the Health and Medical Geography Research Group at UNT, whom I have interacted with during my study. We were not only able to support each other by deliberating over our problems and findings, but also happily by talking about things other than just our papers. My special thanks to my friends, Phra Viroj Opasso, Supasinee Vongkulbhisal, and Chuleeporn Broussard for providing the inspiration and pushing me to complete this work and standing by my side through all the good and bad times. Also, I would like to thank all of those who helped me directly or indirectly to complete this work.

Lastly, but most importantly, I wish to thank my parents, sister, and grandparents for their unending love, support and encouragement throughout my study. Thanks for your faith in me and for allowing me to be as ambitious as I wanted.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1. INTRODUCTION	1
1.1 Background	1
1.2 Research Gaps	4
1.3 Goals and Objectives	5
1.3.1 Objective 1: Improve the KDE Method in Disease Mapping Context	5
1.3.2 Objective 2: Evaluate the Relative Performance of the KDE and Spatial Empirical Bayes Methods	6
1.3.3 Objective 3: Provide an Alternative Approach for Defining Neighborhood of Risk and Identifying At Risk Populations in those Spatial Units	6
1.4 Study Design	7
1.5 Organization of the Thesis	8
1.6 References	9
CHAPTER 2. EVALUATION OF THRESHOLD SELECTION METHODS FOR ADAPTIVE KERNEL DENSITY ESTIMATION IN DISEASE MAPPING.....	12
2.1 Abstract	12
2.2 Introduction	13
2.3 Data and Methods	17
2.3.1 Methods for Objective 1	19
2.3.2 Methods for Objective 2	20
2.4 Results and Discussion	24
2.4.1 The Impact of Threshold Choice on Population Density Estimates	24
2.4.2 Impact of Threshold Choice on the Distribution of Rate Estimates	29
2.4.3 Desirable Threshold Choices	32
2.5 Conclusion	34
2.6 References	34

CHAPTER 3. RELATIVE EFFICACY OF ADAPTIVE KERNEL DENSITY ESTIMATION AND SPATIAL EMPIRICAL BAYES METHODS FOR DISEASE MAPPING	37
3.1 Abstract	37
3.2 Introduction.....	38
3.3 Data	40
3.4 Methods.....	42
3.4.1 Methods to Generate Simulated Data	42
3.4.2 Methods to Compute Estimated Rates	42
3.4.3 Evaluation	44
3.5 Results and Discussion	46
3.5.1 Global Difference.....	46
3.5.2 Local Difference	48
3.6 Conclusion	52
3.7 References.....	53
CHAPTER 4. DEFINING SPATIAL UNITS FOR MAPPING HIV AND TARGETING INTERVENTION IN TEXAS	56
4.1 Abstract	56
4.2 Introduction.....	56
4.3 Study Areas	60
4.4 Data and Methods	62
4.5 Results.....	65
4.5.1 Classifying Characteristics.....	65
4.5.2 Defining Spatial Units.....	69
4.5.3 The Use of HIV and Disease Specific Neighborhood Maps	73
4.6 Discussion	74
4.6.1 Poverty and Black Populations	74
4.6.2 Low Education and Hispanic Population.....	77
4.6.3 The Use of Health and Disease Specific Neighborhood Maps.....	77
4.7 Limitations	78
4.8 Conclusion	79
4.9 References.....	79
CHAPTER 5. SUMMARY AND CONCLUSIONS	84

5.1	Introduction.....	84
5.2	Summary of Findings.....	86
5.2.1	Objective 1: To Improve Kernel Density Estimation Method in Disease Mapping Context	86
5.2.2	Objective 2: To Evaluate the Relative Efficacy of the Adaptive Kernel Density Estimation and Spatial Empirical Bayes Approaches in Disease Mapping	87
5.2.3	Objective 3: To Provide an Alternative Approach for Defining Neighborhood Of Risk and Identifying At Risk Populations in those Spatial Units.....	88
5.3	Broader Impacts	90
5.4	References.....	91
APPENDIX A. SUPPLEMENTARY DATA.....		94
APPENDIX B. SUPPLEMENTARY RESULTS.....		98

LIST OF TABLES

	Page
Table 2.1. Age-adjusted and age-specific heart disease death rates for males in Texas by age group, 2009-2013 (CDC NCHS, 2015), and population distribution from the 2010 U.S. Census Bureau (U.S. Census Bureau, 2010).	19
Table 2.2. Summaries of characteristics of simulated baseline rate distribution.....	23
Table 2.3. Descriptive results and calculated threshold stratified by age groups	25
Table 2.4. Characteristics of the population density curve estimates from various thresholds stratified by age groups	29
Table 3.1. Population distribution (U.S. Census Bureau, 2010) and age-specific heart disease death rate among male in Texas, 2009-2013 (CDC NCHS, 2015).....	42
Table 4.1. A summary of study area characteristics	62
Table 4.2. Social determinants of health variables used in the study obtained from 2013 ACS 5-year estimates at the block group level.	64
Table 4.3. Factor loadings of selected variables in Public Health Region 3	67
Table 4.4. Factor loadings of selected variables in Public Health Region 4	68
Table 4.5. The cluster means for each of the significant factors of Public Health Regions 3 and 4	70

LIST OF FIGURES

	Page
Figure 2.1. Geographic distribution of age-specific heart disease mortality rates for males aged 65 years and older between 2009 and 2013	15
Figure 2.2. Flow chart of methodology showing the steps using in this study.....	18
Figure 2.3. The running RMSE between the simulated baseline rates and the true value as a function of the number of replicates (L) for all age groups.	23
Figure 2.4. Density curves overlaid on population distribution (age > 35; ZCTA level).....	27
Figure 2.5. The distribution of estimated rates of each bandwidth from 100 repetitions: (A) Aged 35 to 44 years; (B) Aged 45 to 54 years; (C) Aged 55 to 64 years; (D) Aged 65 years and older; (E) Aged 35 years and older.	31
Figure 2.6. Geographic distribution of age-specific heart disease mortality rates for males aged 35-44, 45-54, 55-64, 65 years and older, and 35 years and older	33
Figure 3.1. Spatial distribution of heart disease in Texas, aged 35 years and over, 2010-2012: (A) crude death rate; (B) spatially smoothed death rate computed using a local empirical Bayes algorithm proposed by Marshall (1991); (C) spatially smoothed death rate computed using KDE method.....	41
Figure 3.2. The distribution of estimated state rates computed from the adaptive KDE and spatial empirical Bayes methods ($l = 100$ simulations): (A) Age 35 to 44 years; (B) Age 45 to 54 years; (C) Age 55 to 64 years; (D) Age 65 years and older; (E) Age 35 years and older (overall).	47
Figure 3.3. The distribution of RMSE (the difference of rate estimates between the adaptive KDE and spatial empirical Bayes methods) ($L = 100$ simulations): (A) Age 35 to 44 years; (B) Age 45 to 54 years; (C) Age 55 to 64 years; (D) Age 65 years and older; (E) Age 35 years and older (overall).....	49
Figure 4.1. Study area	61
Figure 4.2. Disease specific neighborhood maps of: (A) Public Health Region 3; (B) Public Health Region 4.	72
Figure 4.3. The use of HIV and neighborhood characteristics maps of Public Health Region 3: (A) The spatial distribution of HIV incidence between 2009 and 2011; (B) A neighborhood map with a transparent overlay of high HIV incidence rates.	75
Figure 4.4. The use of HIV and neighborhood characteristics maps of Public Health Region 4: (A) The spatial distribution of HIV incidence between 2009 and 2011; (B) A neighborhood map with a transparent overlay of high HIV incidence rates.	76

CHAPTER 1

INTRODUCTION

This dissertation examines three interrelated aspects of the complex questions of disease mapping methods and definition of neighborhood units for health research. This chapter provides a brief overview of the thesis and its organization as well as the research gaps that this study seeks to fill and study's objectives.

1.1 Background

Disease mapping has long been a part of public health and epidemiology since John Snow's map of cholera deaths in the 1850s. It is considered to be an exploratory analysis to illustrate the geographic distribution of disease and offers an alternative approach for public health to address and specify the hypotheses of disease etiology and causation since it is easier, quicker, and less expensive to conduct than case-control and cohort studies. Using GIS and spatial analysis methods, disease maps are produced not only to visualize the spatial distribution of disease burden but also to emphasize geographically-defined clusters of disease i.e. hotspots (Cromley & McLafferty, 2012; McLafferty, 2015) which helps public health practitioners to monitor and prevent disease outbreaks as well as target vulnerable populations for intervention.

Geographic health data come in two forms – point and areal data (Bithell, 2000; Cromley & McLafferty, 2012), and different data types may use different disease mapping methods. Due to privacy and confidentiality concerns, however, most health data are commonly released as areal data. This study, therefore, focuses only on methods for areal data. When mapping diseases, rates are usually used since they facilitate comparison across different population sizes, unlike the number of cases. Because they are easy to compute, crude rates – the ratio of number

of cases in a spatial unit divided by its total population – are commonly used to map disease, and the resulting spatial distribution is presented as a choropleth map – showing abrupt changes at each boundary. In reality, however, the spatial distribution of disease changes gradually and is not confined to artificial boundaries such as zip codes, counties or states. To address this limitation of choropleth maps, continuous surface or isopleth maps are recommended for displaying the spatial distribution of disease (Cromley & McLafferty 2012; Beyer, Tiwari, & Rushton, 2012) since these kinds of maps illustrate gradual changes which reflect realistic disease distribution. Moreover, using crude rates to map disease requires considerable care because of the small numbers problem – a well-known issue in disease mapping. The small numbers problem can impact the effectiveness and accuracy of disease maps for disease presentation. Any slight change in the number of cases generates wide variations in rates, making the resulting rates unstable. Generally, it occurs in areas with small populations. To solve this problem, smoothed rate methods have been implemented in disease mapping including kernel density estimation and spatial empirical Bayes methods.

The kernel density estimation (KDE) method, a nonparametric disease mapping method, was introduced in the late 1980s (Silverman, 1986; Kelsall & Diggle, 1995). It uses a surface model and extrapolates point data over an area of interest without relying on fixed boundaries such as political boundaries. To compute disease rates, a spatial window or kernel (a circle) is moved across the study area, and the ratio between density of events (i.e., cases) and the density of the background (i.e., population) is calculated within this kernel. The size of the kernel, called bandwidth or threshold, is crucial since it determines the smoothness of results (Silverman, 1986; Kelsall & Diggle, 1995; Waller & Gotway, 2004). Different bandwidths produce different maps, and an inappropriate value can lead to misinterpretation. Current approaches to defining an

appropriate bandwidth rely on somewhat arbitrary choices made by experts who understand the disease being mapped or by the map-makers themselves. The resulting pattern of disease maps depends on the size (radius) of the kernel. Currently, no clear guidance exists for how this may or should be done, which is a research gap. In fact, methods for automatically selecting bandwidth have been proposed to help users find a bandwidth value that is reasonable for a wide range of data distributions, but without any mathematical guarantees of being close to the optimal bandwidth (Wand & Jones, 1995). While these methods have been proposed and used with non-spatial data, their relative utility in disease mapping remains unknown. To fill this research gap and improve the KDE method, this study assesses the relative performance of these methods in terms of resolution and reliability for disease mapping.

Spatial empirical Bayes method, a model-based approach, is one popular disease mapping method. The Bayesian approach is attractive to statisticians and geostatisticians for two main reasons. First, the model itself is flexible enough to include related parameters and account for all inferences that may occur from those parameters. Second, the Bayes procedure offers “borrowing of strength” (Best, Richardson, & Thomson, 2005) across space (in spatial analysis) either using global or local means. The latter reason allows researchers to use Bayes in small areas with limited samples. The primary objective of both KDE and Bayesian approaches is to explore the spatial distribution of disease. Both are widely used since they address the issue of the small numbers problem. Since they have different theoretical approaches, several questions arise: (1) Does producing disease maps using the same dataset but different methods (KDE or Bayesian approach) yield different results? (2) If the two methods provide different results, which method should be used and in what context? These questions are important and need to be answered because using an inappropriate method may lead to failures in interventions. To the

best of my knowledge, these approaches (Bayesian and KDE methods) have never been compared, side-by-side, and there is a need to better understand their merits in disease mapping and clarify the methods with regard to their use in public health and decision-making.

The final motivation of this thesis stems from a limitation of disease maps and restrictions imposed by the availability of health data. Though disease maps help us better understand the distribution of disease across space and pinpoint *where* intervention is most needed, they alone cannot identify *who* or *which populations* are most at risk and need the interventions the most. Moreover, in areas with limited and poor-quality health data, such as many developing countries, producing reliable representations of disease burdens is problematic. Thus, public health tasks of planning and targeting appropriate intervention (McLafferty 2015; Cromley & McLafferty, 2012) become extremely difficult. Currently, studies of neighborhood boundaries in health research are soaring, and some cross-sectional studies have examined the impact of neighborhood characteristics on health and indicated that neighborhood affluence has a positive relationship with health (Chaix, Merlo, Evans, Leal, & Havard, 2009; Cockings & Martin, 2005; Lebel, Pampalon, & Villeneuve 2007; Coulton, Korbin, Chan, & Su, 2001; Browning & Cagney, 2003). Can we identify sub areas, neighborhoods of risk, based on internal consistency and external difference with regards to known risk factors? Such maps would be invaluable for disease burden assessment and targeting interventions. This is the goal of this motivation. This study seeks to define and identify high-risk spatial units for health research. The approach is applied to the study of HIV/AIDS in Texas.

1.2 Research Gaps

The key gaps in research that this study seeks to fill are as follows:

- For the KDE method, what is an appropriate bandwidth (or threshold) value for constructing disease maps? How do you determine it?
- How do different disease mapping methods such as KDE and spatial empirical Bayes influence the perception of disease burdens? How effective are these methods for representing disease distribution?
- When should the KDE or Bayesian method be used to produce disease maps and in what situation? What are the criteria to select a disease mapping method?
- When health data is sparse and has poor quality, how do you produce reliable representations of disease burdens? Can the spatial description of place effects help to identify where a population at risk resides?
- What is an appropriate method to define spatial units for health research? What variables (e.g., income and education) should be considered and selected? What criteria should be used to select these variables?

1.3 Goals and Objectives

The overall goal of the thesis is to provide a better understanding of the importance of disease mapping methods in visualization of disease burdens and enhance disease surveillance systems.

Within this larger goal, the specific objectives of the study are to:

1. Improve the KDE method by illustrating an approach to determine the desirable threshold choice.
2. Evaluate the relative performance of the kernel density estimation and spatial empirical Bayes in disease mapping.
3. Provide an alternative approach for defining neighborhood of risk and identifying at risk populations in those spatial units.

1.3.1 Objective 1: Improve the KDE Method in Disease Mapping Context

Most researchers agree that the selection of bandwidth (or threshold) is crucial, as it affects the degree of smoothing that occurs on the map (Cromley & McLafferty, 2012; Bithell, 2000; Carlos, Shi, Sargent, Tanski, & Berke, 2010; Beyer et al., 2012; Chi, Wang, Li, Zheng, &

Liao, 2007; Shi, 2010; Talbot, Kulldorff, Forand, & Haley, 2000; Cai, 2007; Rushton & Lolonis, 1996; Tiwari & Rushton, 2005; Silverman, 1986; Wand & Jones, 1995). Recent studies in disease mapping used a knowledge-based judgement to arbitrarily selected threshold values. In fact, methods for automatically selecting bandwidths have been proposed and used produce distribution of non-spatial data. However, their application for threshold selection in the disease mapping context has not been evaluate. Relevant research questions are:

- How can we apply automatically selecting bandwidth methods in the disease mapping context?
- Do threshold values from a knowledge-based judgement and bandwidth selectors significantly represent different spatial distributions of disease?
- How do different bandwidth values impact the perception of disease burdens?
- What is a desirable threshold choice to construct disease maps?

1.3.2 Objective 2: Evaluate the Relative Performance of the KDE and Spatial Empirical Bayes Methods

In recent years, while many new disease mapping methods have been developed, the comparison between disease mapping methods is limited, and no systematic evaluation of the mapping methods exists. Relevant research questions are:

- How do these different mapping methods influence the outcome of maps?
- What appropriate disease mapping method should be used in what settings (rural/urban areas)?
- What are the criteria and guidelines for selecting an appropriate method?

1.3.3 Objective 3: Provide an Alternative Approach for Defining Neighborhood of Risk and Identifying At Risk Populations in those Spatial Units

In areas with limited and poor-quality health data, such as many developing countries, producing

reliable representations of disease burdens is problematic. Defining spatial units for health research can help public health practitioners reframe approaches for spatial targeting of intervention.

- How do you identify possible locations and configurations of population at risk?
- What is an appropriate method to define spatial units for health research?
- What variables should be considered and included in the model?
- How effective is the proposed method for identifying the possible locations of population at risk?

1.4 Study Design

To examine the applicability of automatically selecting bandwidths methods and a desirable threshold value for the KDE method (objective 1) and to assess the relative performance of disease mapping methods (objective 2), this study uses a synthetic dataset generated under specific scenarios for known underlying risks and population sizes. A simulated dataset, l ($l = 1, 2, 3... 100$ replications) is generated based on population data from the 2010 U.S. Census Bureau (U.S. Census Bureau, 2010) and heart disease mortality data between 2009 and 2013 obtained from the Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2015).

To determine a desirable threshold value, this study includes the total of ten thresholds – six arbitrary and four calculated thresholds. Six arbitrary thresholds are mostly used in previous studies (Rushton & Lolonis 1996; Oppong, Tiwari, Ruckthongsook, Huddleston & Arbona, 2012). Four calculated thresholds are derived from the population data. Using a simulated dataset, each threshold value is used to compute disease rate resulting in 100 replications per threshold. Then, the root-mean-square-error (RMSE) is employed to measure the accuracy

between estimated state rates computed from thresholds using the KDE method and state rates computed directly from the simulated dataset by dividing the total number of simulated cases by the total population. The same set of simulated datasets is also used to assess the performance of disease mapping methods while computing estimated disease rates using spatial empirical Bayes method. Then, the root-mean-square-error (RMSE) is employed to measure the difference between rate estimates from the KDE and spatial empirical Bayes methods.

To identify high-risk spatial units for HIV infection (objective 3), a list of HIV-related variables is selected based on the Centers for Disease Control and Prevention's (CDC) National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) guidelines (2010). These variables are obtained from the 2013 ACS 5-year estimates dataset at the block group level. Factor analysis is employed to reduce the large numbers of HIV-related variables into numbers of identifiable dimension. Then, block groups are grouped and assigned into k partitions using k -mean clustering method. The resulting k types of clusters are imported to ArcGIS 10.2 and mapped out to illustrate the delineation of homogeneous zones.

1.5 Organization of the Thesis

This thesis consists of five chapters including this introductory chapter. The first manuscript (Chapter 2) focuses on the improvement of the KDE method for disease mapping. Using ten thresholds (six arbitrary and four calculated thresholds), the study seeks to find a desirable threshold choice and evaluate the relative performance of these thresholds. The second manuscript (Chapter 3) examines and evaluates the relative efficacy of the KDE and spatial empirical Bayes approaches in disease mapping. This study seeks to answer the differences and similarities of rate estimates computed from these two methods within the same conditions –

using the same dataset and consistent variables. The third manuscript (Chapter 4) illustrates how to classify and construct the health-risk map using straightforward statistical and spatial methods. This manuscript attempts to portray the usefulness of health-risk maps when health data is sparse and/or has poor quality; it also illustrates the powerful effectiveness of the health-risk map when combined with health data. The final chapter (Chapter 5) summarizes all the findings, contributions, and broader impacts made by this study as well as directions for future research.

1.6 References

- Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research, 14*, 35-59.
doi:10.1191/0962280205sm388oa
- Beyer, K. M., Tiwari, C., & Rushton, G. (2012). Five essential properties of disease maps. *Annals of the Association of American Geographers, 102*(5), 1067-1075.
doi:10.1080/00045608.2012.659940
- Bithell, J. F. (2000). A classification of disease mapping methods. *Statist. Med., 19*, 2203-2215.
- Browning, C. R., & Cagney, K. A. (2003). Moving beyond poverty: neighborhood structure, social processes, and health. *Journal of Health and Social Behavior, 44*(4), 552-571.
- Cai, Q. (2007). *Mapping disease risk using spatial filtering methods* (Unpublished doctoral dissertation). The University of Iowa, Iowa City, Iowa.
- Carlos, H. A., Shi, X., Sargent, J., Tanski, S., & Berke, E. M. (2010). Density estimation and adaptive bandwidths: a primer for public health practitioners. *International Journal of Health Geographics, 9*(39). doi:10.1186/1476-072X-9-39
- Centers for Disease Control and Prevention, National Center for Health Statistics. (2015). *Underlying cause of death - heart disease mortality data set, 2009 to 2013* [Data set]. Retrieved from CDC WONDER Online Database: <https://wonder.cdc.gov/ucd-icd10.html>
- Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2010). *Establishing a holistic framework to reduce inequities in HIV, viral hepatitis, STDs, and tuberculosis in the United States*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved from <https://www.cdc.gov/socialdeterminants/docs/SDH-White-Paper-2010.pdf>

- Chaix, B., Merlo, J., Evans, D., Leal, C., & Havard, S. (2009). Neighbourhoods in eco-epidemiologic research: Delimiting personal exposure areas. A response to Riva, Gauvin, Apparicio and Brodeur. *Social Science & Medicine*, *69*(9), 1306-1310. doi:10.1016/j.socscimed.2009.07.018
- Chi, W., Wang, J., Li, X., Zheng, X., & Liao, Y. (2007). Application of GIS-based spatial filtering method for neural tube defects disease mapping. *Wuhan University Journal of Natural Sciences*, *12*(6), 1125-1130. doi:10.1007/s11859-007-0097-6
- Cockings, S., & Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine*, *60*(12), 2729-2742.
- Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping residents' perceptions of neighborhood boundaries: a methodological note. *American Journal of Community Psychology*, *29*(2), 371-383.
- Cromley, E. K., & McLafferty, S. L. (2012). *GIS and public health* (2nd ed.). New York, NY: The Guilford Press.
- Kelsall, J. E., & Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, *1*(1), 3-16.
- Lebel, A., Pampalon, R., & Villeneuve, P. Y. (2007). A multi-perspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region. *International Journal of Health Geographics*, *6*, 27. doi:10.1186/1476-072X-6-27
- McLafferty, S. (2015). Disease cluster detection methods: recent developments and public health implications. *Annals of GIS*, *21*(2), 127-133. doi:10.1080/19475683.2015.1008572
- Oppong, J. R., Tiwari, C., Ruckthongsook, W., Huddleston, J., & Arbona, S. (2012). Mapping late testers for HIV in Texas. *Health & Place*, *18*, 568-575. doi:10.1016/j.healthplace.2012.01.008
- Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, *15*, 717-726.
- Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over in homogeneous backgrounds. *International Journal of Geographical Information Science*, *24*(5), 643-660. doi:10.1080/13658810902950625
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York, NY: Chapman and Hall.
- Talbot, T. O., Kulldorff, M., Forand, S. P., & Haley, V. B. (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine*, *19*, 2399-2408.
- Tiwari, C., & Rushton, G. (2005). Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In P. Fisher, *Developments in spatial data handling* (pp. 665-676). Berlin: Springer-Verlag.

U.S. Census Bureau. (2010). *Profile of general population and housing characteristics: 2010 - 2010 Census Summary File 1* [Data set]. Retrieved from American FactFinder:
<https://factfinder.census.gov/>

Waller, L., & Gotway, C. (2004). *Applied spatial statistics for public health data*. Hoboken, NJ: John Wiley & Sons, Inc.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Boca Raton, FL: Chapman & Hall/CRC.

CHAPTER 2

EVALUATION OF THRESHOLD SELECTION METHODS FOR ADAPTIVE KERNEL DENSITY ESTIMATION IN DISEASE MAPPING

2.1 Abstract

Maps of disease rates produced without careful consideration of the underlying population distribution may be unreliable due to the well-known small numbers problem. Smoothing methods such as Kernel Density Estimation (KDE) are employed to control the population basis of spatial support used to calculate each disease rate. The degree of smoothing is controlled by a user-defined parameter (bandwidth or threshold) which influences the resolution of the disease map and the reliability of the computed rates. Methods for automatically selecting a smoothing parameter such as normal scale, plug-in, and smoothed cross validation bandwidth selectors have been proposed for use with non-spatial data, but their relative utility remains unknown. This study assesses the relative performance of these methods in terms of resolution and reliability for disease mapping. Using a simulated dataset of heart disease mortality among males aged 35 years and older in Texas, we assess methods for automatically selecting a smoothing parameter. The results show that while all parameter choices accurately estimate the overall state rates, they vary in terms of the degree of spatial resolution. Further, parameter choices resulting in desirable characteristics for one sub group of the population (e.g., a specific age-group) may not necessarily be appropriate for other groups. In this research, the findings shows that the appropriate threshold value depends on the characteristics of the data, and that bandwidth selector algorithms can be used to guide such decisions about mapping parameters. An unguided choice may produce maps that distort the balance of resolution and statistical reliability.

Keywords: kernel density estimation, bandwidth selection, threshold, Bayesian, disease mapping

2.2 Introduction

Disease maps, an essential component of epidemiological surveillance, are used to illustrate the geographic distributions of diseases. Disease outcomes are typically represented as rates which are computed by dividing the number of disease cases by the population contained within some geographic region such as zip code or county. Rates that are computed without careful consideration of the underlying population distribution may be unreliable due to the well-known small numbers problem (Cromley & McLafferty, 2012). For example, areas with small populations are more likely to produce unstable rate estimates compared to areas with larger population sizes. Smoothing methods including kernel density estimation are commonly used to address the problem of unstable rates (Cromley & McLafferty, 2012; Clayton & Kaldor, 1987; Marshall, 1991; Rushton & Lolonis, 1996; Bithell, 2000; Talbot, Kulldorff, Forand, & Haley, 2000; MacNab, Farrell, Gustafson, & Wen, 2004; Best, Richardson, & Thomson, 2005; Tiwari & Rushton, 2005; Waller & Carlin, 2010; Beyer, Tiwari, & Rushton, 2012).

Kernel density estimation (KDE) is a non-parametric method that can be used to explore the spatial density of point data (Cromley & McLafferty, 2012). In the context of disease mapping, KDE methods operate by computing rates within a moving spatial window or kernel (typically a circle) placed across the entire study area. A ratio of the density of events (i.e., cases) and the density of the background (i.e., the population) is calculated within each kernel (Shi, 2010). Another KDE method computes the rate by dividing the number of cases that fall inside a kernel by the population that is contained within the same kernel (Rushton & Lolonis, 1996; Tiwari & Rushton, 2005).

The size of the kernel, bandwidth, is a crucial parameter that influences the degree of smoothing on the map in KDE (Silverman, 1986; Kelsall & Diggle, 1995; Waller & Gotway, 2004). The bandwidth can be either fixed or variable (adaptive). For the fixed bandwidth approach, the kernel has a fixed-size radius, and all kernels (circles) have the same radii. In health studies, the fixed bandwidth approach may not be suitable since populations are not evenly distributed across space. Moreover, unstable rates may result if the circle falls in low population-density areas. Similarly, in the adaptive bandwidth approach, the kernel radius grows or shrinks to accommodate varying population size. The minimum population size that is used to define the kernel bandwidth, and consequently the degree of smoothing on a map, is a user-defined parameter. In this study, it is referred as the threshold value (h).

Figure 2.1 illustrates the spatial distribution of heart disease mortality rates for males aged 65 years and older using data obtained from the Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2015). These maps were produced using the adaptive bandwidth kernel density estimation method with different threshold values. As shown in Figure 2.1A, when using the smallest threshold value ($h = 50$), the resulting map portrays high levels of geographic detail in the estimated rate. However, as the thresholds increase, the resulting maps show lower levels of geographic detail (Figure 2.1B-D). Further, maps produced using small threshold values tend to display greater fluctuations in rate estimates ($\mu = 1,330$ per 100,000 population, $\sigma = 639.9$ at $h = 50$). In contrast, maps produced using larger threshold values tend to show lower levels of fluctuation ($\mu = 1,209.5$ per 100,000 population, $\sigma = 268.4$ at $h = 1000$). The trade-off between geographic detail and reliability depends on the choice of the threshold value. A value that is too small may result in under-smoothing, i.e., high levels of geographic detail but greater fluctuation in rate estimates (Figure 2.1A). Conversely, a

value that is too large will result in over-smoothing, i.e., low levels of geographic detail but less fluctuation in rate estimates (Figure 2.1D).

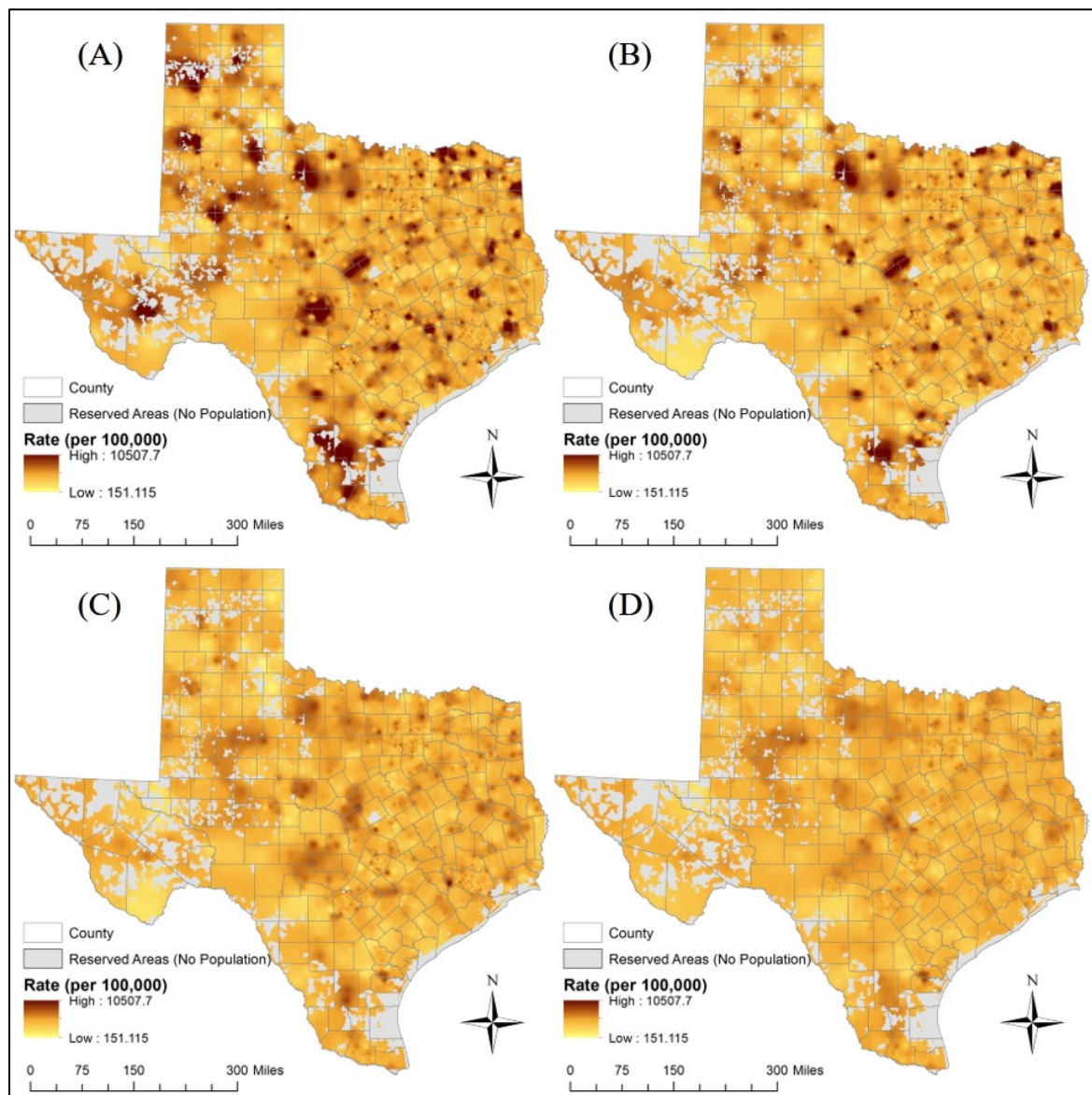


Figure 2.1 Geographic distribution of age-specific heart disease mortality rates for males aged 65 years and older between 2009 and 2013. Maps were created using the adaptive bandwidth kernel density estimation method with various bandwidths (h): (A) $h = 50$; (B) $h = 100$; (C) $h = 500$; (D) $h = 1000$. (Note: the data were obtained from CDC NCHS (2015))

The problem of choosing an appropriate smoothing parameter – bandwidth or threshold – has been discussed in previous studies (Cromley & McLafferty, 2012; Bithell, 2000; Carlos, Shi, Sargent, Tanski, & Berke, 2010; Beyer et al., 2012; Chi, Wang, Li, Zheng, & Liao, 2007; Shi, 2010; Talbot et al., 2000; Cai, 2007; Rushton & Lolonis, 1996; Tiwari & Rushton, 2005; Silverman, 1986; Wand & Jones, 1995). Silverman (1986) and Wand and Jones (1995) recommend subjective selection of the bandwidth parameter based on visual inspection. The process of visual evaluation of the bandwidth parameter begins with examining several plots of the data and selecting the density that is the “most pleasing” in some sense (Wand & Jones, 1995, p. 58). Although this approach has been used by others (Shi, 2010, p. 654), the process can be time-consuming if many density estimates are required. In other cases, map-makers may not utilize information about the structure of the data to inform choice of threshold value.

Many bandwidth selectors available for use with non-spatial data could potentially be adapted for spatial data. However, their applicability for this purpose has not been evaluated. Non-spatial bandwidth selectors may be grouped into two classes – (a) quick and simple, and (b) hi-tech bandwidth selectors (Wand & Jones, 1995). Quick and simple bandwidth selectors aim to find a threshold value that is reasonable for a wide range of data distributions. One such method is the normal scale bandwidth selector (Wand & Jones, 1995). This method recommends a bandwidth value which can be used as a starting point or a “first guess” (Wand & Jones, 1995). The bandwidth is calculated by referencing a standard distribution that is derived from the data (see Silverman, 1986 and Wand & Jones, 1995 for details). In contrast, hi-tech bandwidth selectors, which are data-driven, seek to find an optimal bandwidth by minimizing the mean integrated square error (MISE) of the kernel density estimator (Chiu, 1992; Wand & Jones, 1995). For example, plug-in (Wand & Jones, 1994) and smoothed cross-validation (Hall &

Marron, 1991) bandwidth selectors estimate a pre-smoothing parameter based on the pairwise differences of the observations obtained using the pilot bandwidth value. The pre-smoothing parameter is then used to find the optimal bandwidth value (Wand & Jones, 1995; Hall, Sheather, Jones, & Marron, 1991). Additional information on the theoretical basis of these methods are found in Silverman (1986), Wand and Jones (1994; 1995), Chiu (1992), Hall and Marron (1991), Hall et al. (1991).

In summary, while these methods have been used to produce distributions of non-spatial data, their suitability for threshold selection in disease mapping remains unknown. Using a simulated dataset, this study illustrates and examines the applicability of such methods for disease mapping.

2.3 Data and Methods

The methods used in this study were presented in two parts (Figure 2.2). First, this study examined the applicability of the visual and subjective methods for choosing a threshold value (Objective 1). Initially, threshold values ranging from 50 to 10,000 were used under the assumption that map-makers will select threshold values based on arbitrary choices or some prior knowledge of the disease. Subsequently, bandwidth selection methods – normal scale (h_{ns}), plug-in (h_{pi}), smoothed cross-validation (h_{scv}), and *median* – were used for comparison. This study used the 10-year age-stratified population data for males in Texas obtained from the 2010 U.S. Census Bureau at the zip code level (Table 2.1) (U.S. Census Bureau, 2010).

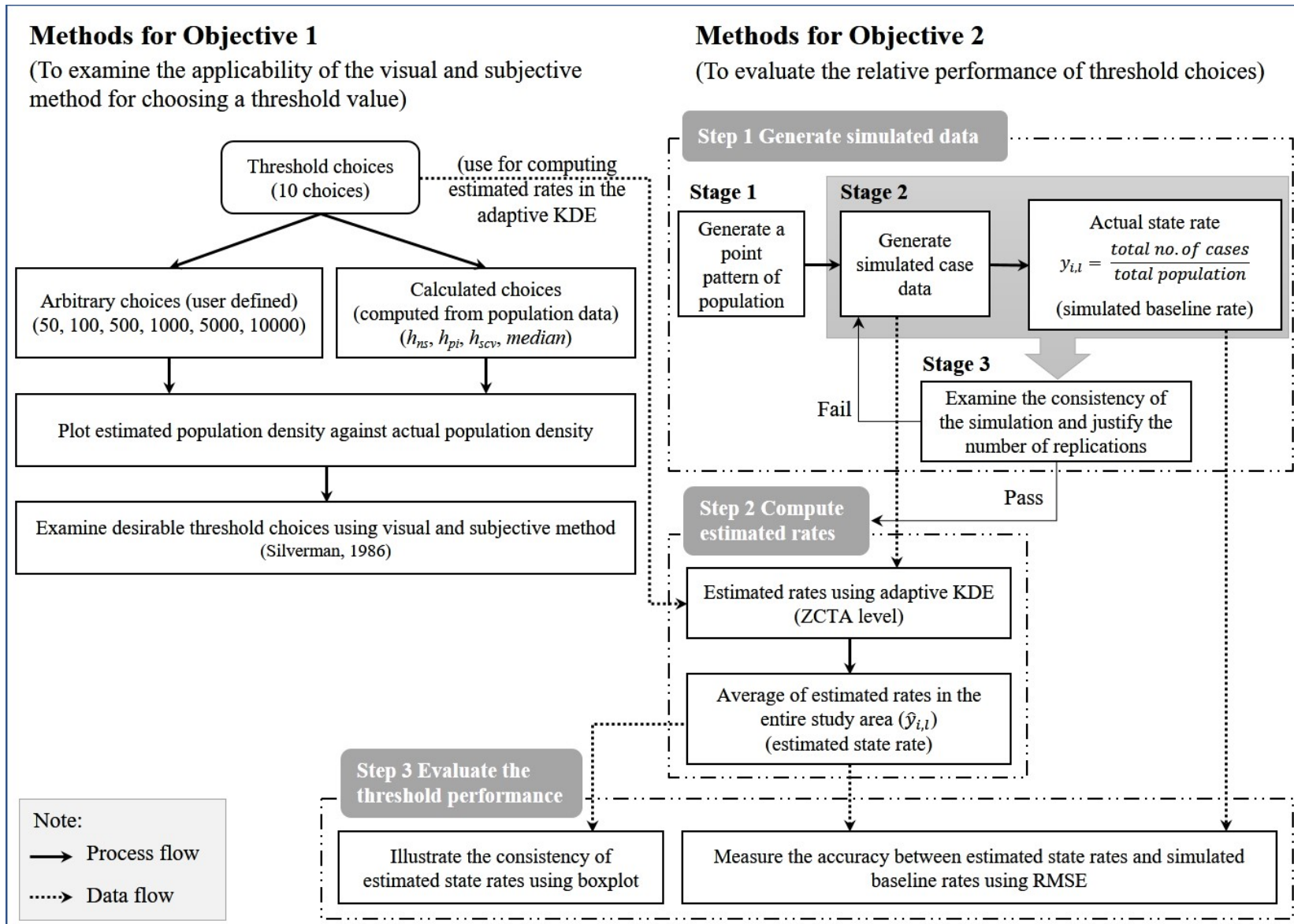


Figure 2.2 Flow chart of methodology showing the steps using in this study.

Second, the relative performance of different threshold values for disease maps was evaluated using the same dataset (Objective 2). We first generated a simulated dataset of age-specific heart disease mortality rates among males aged 35 years and older in Texas. Statewide rates for generating the simulated case counts in each age-group were obtained from the CDC NCHS (2015) (Table 2.1).

Table 2.1 Age-adjusted and age-specific heart disease death rates for males in Texas by age group, 2009-2013 (CDC NCHS, 2015), and population distribution from the 2010 U.S. Census Bureau (U.S. Census Bureau, 2010).

Age	Population	Range of aggregated population at the ZCTA level	Rate (per 100,000)
35-44	1,722,904	[1, 7925]	33.87
45-54	1,702,639	[1, 7407]	115.15
55-64	1,256,976	[1, 4948]	297.36
65+	1,135,517	[1, 4792]	1,245.93
Total (35+)	5,818,036	[1, 22555]	351.15

2.3.1 Methods for Objective 1

A total of ten thresholds were used in this study. Six thresholds were a series of arbitrary choices – 50, 100, 500, 1000, 5000, and 10000. These six thresholds remained constant for all age groups. The remaining four were calculated based on population data aggregated at the zip code level using *median* and three bandwidth selectors – the normal scale (h_{ns}), the plug-in (h_{pi}), and the smoothed cross-validation (h_{scv}). The *median* threshold was determined by computing the median population value across all zip codes. Threshold values from the three bandwidth selectors – h_{ns} , h_{pi} , and h_{scv} – were computed using the *ks*-package in *R* (Duong, 2017). Since these four thresholds were calculated based on data, their values varied among the age groups. Desirable threshold options were selected using visual and subjective examination of the data as suggested by Silverman (1986). This involved generating plots of estimated population density

against the actual population density. For each age group, estimated population densities were computed using a kernel function with each of the ten thresholds. The actual population density was generated from the population data using a gamma distribution. The gamma distribution was used to ensure that the values are not negative. It consists of two positive parameters – shape (α) and scale (β) parameters. These two parameters were calculated using mean and standard deviation of the population data (equations 2.1 and 2.2). This process was also performed in R using probability density function (equation 2.3).

$$\alpha = \left(\frac{\mu}{\sigma}\right)^2 \quad (2.1)$$

$$\beta = \frac{\sigma^2}{\mu} \quad (2.2)$$

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad (2.3)$$

where α and β are shape and scale parameters respectively, μ and σ are respectively mean and standard deviation of the population, $\Gamma(\alpha)$ is the gamma function.

2.3.2 Methods for Objective 2

2.3.2.1 Step 1 Methods to Generate Simulated Data

The aim of this step was to generate simulated case data. This step comprised of three stages:

- Stage 1: Generate a Point Pattern of Male Populations by Age at the ZCTA Level

The data used in this process are (1) male population data stratified by age as shown in Table 1, and (2) a ZIP Code Tabulation Areas (ZCTAs) cartographic boundary file obtained from Topologically Integrated Geographic Encoding and Referencing (TIGER) (U.S. Census Bureau, 2016). Note that the ZCTAs are created by the U.S. Census Bureau and approximate the

spatial boundaries of postal zip code service areas (U.S. Census Bureau, 2015). The age-stratified population data were joined to the ZCTA cartographic boundary file in ArcGIS 10.2. Then, the random point generation tool in ArcGIS 10.2 was used to create a point distribution, where each point $(x_{i,s})$ represents a simulated individual in age group i residing in ZCTA s . The age group $i \in I$ where $I = \{35-44, 45-54, 55-64, 65+\}$.

- Stage 2: Generate Simulated Cases from a Point Pattern of Male Populations Obtained from Stage 1

To classify a simulated individual as a case, a random number was assigned to each point in the random point pattern generated from stage 1. The random number was generated from a uniform distribution on the interval $(0, 1)$ under the assumption that each person has an equal probability of being designated as a case. The probability that a simulated point, $x_{i,s}$, would be classified as a case $(c_{i,s})$ was determined using observed age-specific heart disease mortality rates (Table 2.1). For example, the observed age-specific heart disease death rate for males aged 35 to 44 years old in Texas was 33.87 per 100,000 (0.0003387) (Table 2.1). If a random number generated was in the range 0.0000001 to 0.0003387, it was classified as a simulated case. This process was replicated 100 times to produce 100 different instances of the case distribution – i.e., a 100 simulated maps of heart disease mortality could be produced from this data. For each simulated dataset (l , where $l = 1, 2, \dots, 100$), state rates, called simulated baseline rates, were computed for each age group as well as for all-groups combined. The rate $(y_{i,l})$ was computed using:

$$y_{i,l} = \frac{C_{i,l}}{P_i} \quad (2.4)$$

where $C_{i,l}$ was the total number of simulated cases for age group i at the l th simulation, and P_i was the total population for age group i ,

- Stage 3: Examine the Consistency of the Simulation and Justify the Number of Replications

For each age group, the consistency of the simulated baseline rates ($y_{i,l}$) and the justification of the number of replications were examined using a scatter plot of the running root-mean-square-error ($RMSE^M$) against the total number of replications (Koehler, Brown, & Haneuse, 2009). The $RMSE^M$ measured the differences between the simulated baseline rates and the true value, i.e., CDC's heart disease mortality rate (Table 2.1) using the following formula:

$$RMSE_{i,L}^M = \sqrt{\frac{1}{L} \sum_{l=1}^L (y_{i,l} - Y_i)^2} \quad (2.5)$$

where L was the total number of replications, $y_{i,l}$ was the simulated baseline rate of age group i at l th simulation, and Y_i was the true rate of age group i . Figure 2.3 illustrates that the magnitude of the difference between the simulated baseline rate and the true value ($RMSE$) stabilizes as L increases. In this study, when $L > 50$, the stable state was achieved for all age groups. In this study, we used 100 replications.

Based on recommendations by Natesan (2015), the coverage rate and bias of interval estimates were also examined (Table 2.2). Coverage rate is defined as the percentage of statistical estimate intervals that contain the true values. Bias of interval estimates is computed as the percentage of the statistical estimate intervals that overestimate and underestimate the true value (Natesan, 2015). While the coverage rates for the 95% interval estimates are typically expected to be around 95%, our results show that they were extremely low – less than 20% for all age groups (Table 2.2). This indicates the extreme conservative estimates of uncertainty (large standard deviations) which may result from the uniform distribution that we used to generate random numbers in Stage 2. Although the coverage rate was considerably low, the

percentages of over- and under-estimates were likely to be equal which indicated that the simulation was unbiased.

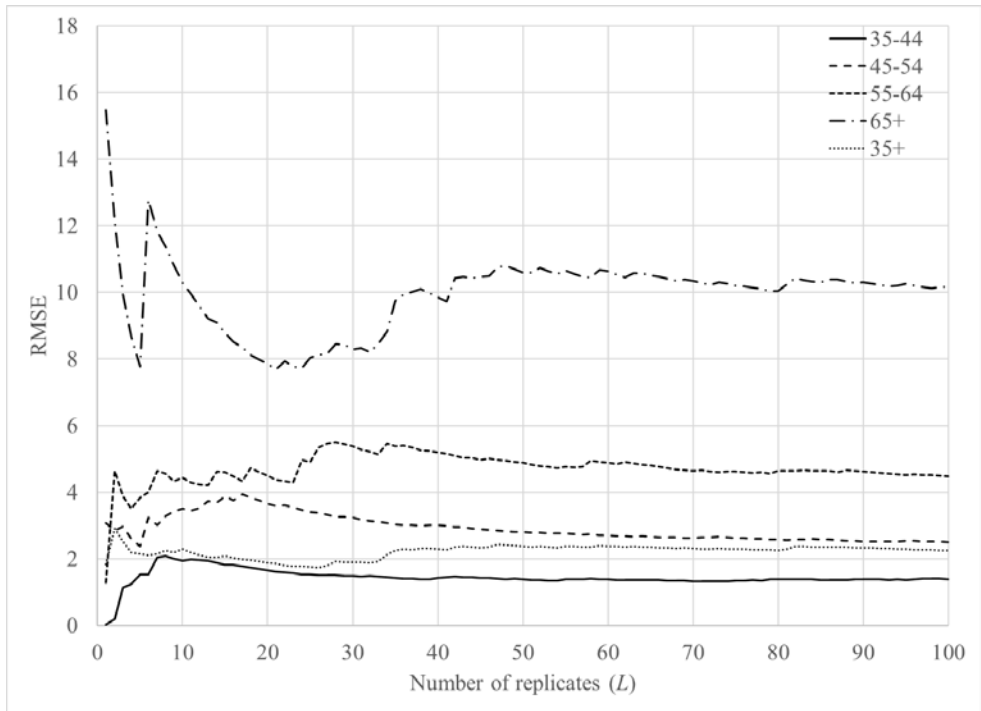


Figure 2.3 The running RMSE between the simulated baseline rates and the true value as a function of the number of replicates (L) for all age groups.

Table 2.2 Summaries of characteristics of simulated baseline rate distribution

Age group	Mean	SD	Coverage rate (%)	Over-estimated (%)	Under-estimated (%)
35-44	33.92	1.40	17	50.6	49.4
45-54	115.17	2.52	11	49.4	50.6
55-64	297.60	4.49	20	56.2	43.8
65+	1245.93	10.21	16	47.6	52.4
35+	351.12	2.27	14	52.3	47.7

2.3.2.2 Step 2 Methods to Compute Estimated Rates

For each age group i at the l th simulation, estimated rates were computed using the KDE method with aggregated simulated cases as the numerators, and the population data as the denominators. The KDE method was applied to all ten threshold values. This process was

performed using the Web-based Disease Mapping Analysis Program (WebDMAP) (Tiwari C. , 2008) and custom code written in Python. As a result, 100 estimated rates were produced for each threshold and for each age group i . These rates, which were obtained at the ZCTA level, were aggregated to the state level (called estimated state rate, $\hat{y}_{i,l}$).

2.3.2.3 Step 3 Methods to Evaluate Threshold Performance

To evaluate the relative performance of each threshold choice, the estimated state rates ($\hat{y}_{i,l}$) resulting from different thresholds were compared to the simulated baseline rates in each age group i ($y_{i,l}$ from equation 2.4). The root-mean-square-error (RMSE) was employed to measure the accuracy between estimated state rates and the simulated baseline rates using the following formula:

$$RMSE_{i,t} = \sqrt{\frac{1}{100} \sum_{l=1}^{100} (\hat{y}_{i,t,l} - y_{i,l})^2} \quad (2.6)$$

where $RMSE_{i,t}$ was the RMSE of age group i and threshold t , $\hat{y}_{i,t,l}$ was the estimated state rate of age group i and threshold t at the l th simulation, and $y_{i,l}$ was the simulated baseline state rate of age group i at the l th simulation. Further, to illustrate the consistency of the rates computed from each threshold ($\hat{y}_{i,t,l}$), a box-plot was generated to display the variation of 100 estimated state rates for each age group.

2.4 Results and Discussion

2.4.1 The Impact of Threshold Choice on Population Density Estimates

The calculated thresholds for the three selectors (plug-in (h_{pi}), smoothed cross-validation (h_{scv}), normal scale (h_{ns})) and *median* are shown in Table 2.3. The h_{pi} and h_{scv} selectors result in the smallest threshold values. In contrast, the h_{ns} and *median* selectors are approximately 4 and 8

times larger for the age groups 55 to 64, 65 years and older, and the overall population (aged 35 years and older). Further, the h_{ns} and *median* selectors are also approximately 5 and 7 times larger for the two youngest groups – 35 to 44 and 45 to 54. These results indicate that for the same data, different bandwidth selectors provide different threshold values. For this data, the h_{pi} and h_{scv} recommendations produce maps that provide greater geographic detail (lower levels of smoothing), but also larger fluctuations in estimated rates. Conversely, the other two bandwidth selectors produce greater levels of smoothing, but fewer fluctuations in rates.

Table 2.3 Descriptive results and calculated threshold stratified by age groups

Age groups	Total Population	Range	No. of ZCTAs	Calculated thresholds				% ZCTAs with specified minimum population	
				h_{pi}	h_{scv}	h_{ns}	<i>median</i>	≤ 100	≤ 300
35-44	1,722,904	[1, 7925]	1911	53	56	280	327	32%	48%
45-54	1,702,639	[1, 7407]	1910	57	55	255	399	28%	45%
55-64	1,256,976	[1, 4948]	1906	44	41	177	342	30%	48%
65 and older	1,135,517	[1, 4792]	1902	41	40	156	330	28%	48%
Total (35+)	5,818,036	[1, 22555]	1920	200	189	837	1411	14%	25%

In Figure 2.4, the density curves for populations obtained after applying each threshold (h_{pi} , h_{scv} , h_{ns} , *median*, and six arbitrary choices – 50, 100, 500, 1000, 5000, 10000) are compared to the actual population distribution (see 2.2.1 Methods for Objective 1). For each chart, the X-axis represents population with a bin size of 200 and the Y-axis is the density of ZCTAs.

The actual population density (Figure 2.4 column A) tends to follow a gamma distribution (the black line) for all age groups, which indicates that the population is not evenly distributed. Thus, many ZCTAs have low populations, and the number of ZCTAs with large populations is small. This is indicated by a long tail to the right of the distribution. Figure 2.4 column B illustrates the population density estimates computed from all ten thresholds. For all age groups, the population density estimates computed from thresholds, $h = 50, 100, h_{pi}$, and h_{scv} , provide similar density curve characteristics. The density estimates have a sharp peak and closely

match the actual gamma distribution. The resulting density curves from these four thresholds contain fluctuations at the tail end of the distribution. This suggests that these four thresholds may be too small for all age groups. For maps produced using these threshold values, the Washington State Department of Health guidelines (2012) suggest extreme caution with interpretation since the population (denominator) values are less than 100. In fact, the guidelines recommend interpretation with caution for maps produced using populations less than 300. Thus, thresholds, $h \leq 100$ may not be an appropriate choice to use. This is also true of h_{pi} , and h_{scv} for age specific groups in this study. In contrast to these small thresholds, larger thresholds provide more smoothed estimates and will not capture adequate geographic detail on a map. For example, $h = 5000$, and 10000 , may be too large for all age groups since the density curve estimates are almost flat (Figure 2.4 column B).

While six thresholds result in similar density curve characteristics for all age groups and may be considered too small ($h = 50, 100, h_{pi}$, and h_{scv}) or too large ($h = 5000$ and 10000), the remaining thresholds – h_{ns} , *median*, 500 , and 1000 – provide slightly different density curve characteristics between age 35 years and older and other age groups. For the age groups 35 to 44, 45 to 54, 55 to 64, and 65 years and older, the population density estimates computed from $h = h_{ns}$, *median*, and 500 (arbitrary choice) provide similar density curve characteristics. Thus, when $h = h_{ns}$, the density estimates are smoother and the fluctuations in the tails cease to exist. When threshold values increase ($h = \textit{median}$ and 500), the density estimates retain the modal structure of $h = h_{ns}$ but are more smoothed. This density curve characteristic is the most desirable compared to the others as it offers a reasonable compromise between smoothing of the mode and tail.

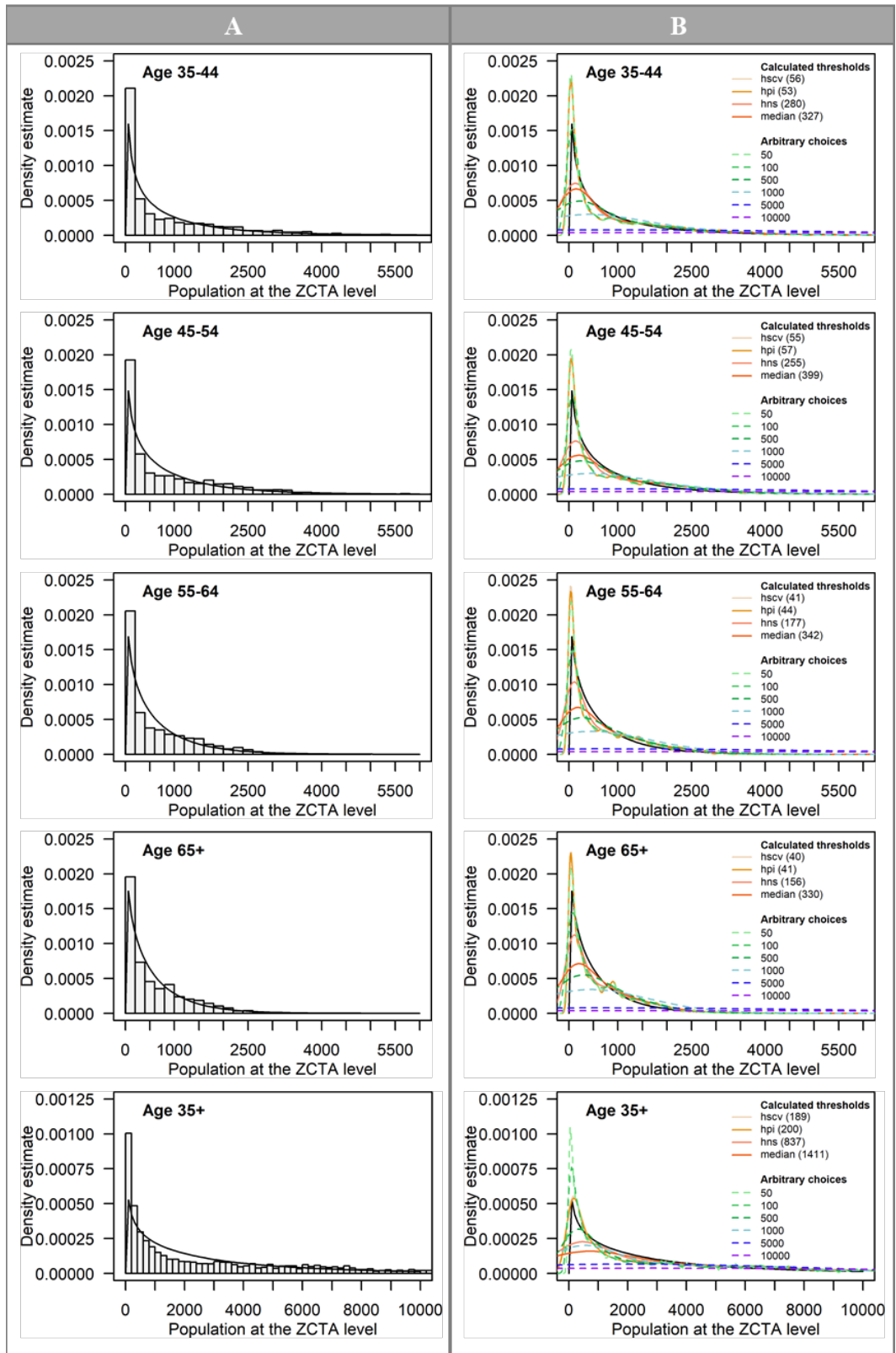


Figure 2.4 Density curves overlaid on population distribution (age > 35; ZCTA level). Column A describes the gamma distribution. Column B describes threshold choices.


The thresholds for age 35 years and older that fit this characteristic are $h = h_{ns}$, 500, and 1000. Although thresholds $h = 500$ and h_{ns} fall in the most desirable characteristic for all age groups, their density curve characteristics between age 35 years and older and other age groups are slightly different. When $h = 500$, the density curve is smoother than $h = h_{ns}$ for the age groups 35 to 44, 45 to 54, 55 to 64, and 65 years and older. In contrast, $h = h_{ns}$ offers a smoother density curve than $h = 500$ for age 35 years and older. Moreover, when $h = 1000$, the resulting density curve retains the same modal structure as $h = h_{ns}$ and 500 (which may be considered as a desirable choice for age 35 years and older), but may be too large for other age groups since the density curves are even more smoother and almost flat. Differences in population size and distribution between age 35 years and older and other age groups are probably the reason. This explanation also applies to $h = median$, which may be considered as one of desirable choices for age specific groups but may be too large threshold for ages 35 years and older.

These findings suggest that thresholds that produce desirable characteristics for one group may not necessarily work for other groups possibly due to differences in population size and distribution. For producing disease maps that incorporate the population age structure, e.g., directly age-adjusted maps, the map-makers must be careful not to choose different threshold values for each age strata as this could lead to the use of inconsistent spatial supports. Generally, spatial supports must be consistently applied across the entire map (Haining, 2003; Gotway & Young, 2002; Beyer & Rushton, 2009). In such circumstances, the map-makers may choose a threshold value that best fits a majority of the age groups.

Table 2.4 summarizes the characteristics of density curve estimates from various thresholds by age groups. The thresholds that provide the most desirable density curve characteristics are $h = h_{ns}$, *median* and 500 for age groups 35 to 44, 45 to 54, 55 to 64, and 65

years and older and $h = h_{ns}, 500,$ and 1000 for age 35 years and older. This is consistent with the recommendation of Silverman (1986), to use values that best replicate the population distribution. Additional considerations may include a comparison of the estimated rates, obtained from various threshold estimators, to the actual state rates

Table 2.4 Characteristics of the population density curve estimates from various thresholds stratified by age groups

Desirable characteristics	Density curve characteristics	Age groups				
		35-44	45-54	55-64	65+	35+
Most  Least	Density curve is smoother, and fluctuations in the tail ceases to exist.	h_{ns} <i>median</i> 500	h_{ns} <i>median</i> 500	h_{ns} <i>median</i> 500	h_{ns} <i>median</i> 500	500 h_{ns} 1000
	Density curve closely matches to the actual gamma distribution and contains fluctuations at the tail.	100	100	100	100	h_{pi} h_{scv}
	The highest density estimates of density curve is greater than that of the actual gamma distribution, and the density curve contains high fluctuations at the tail.	50 h_{pi} h_{scv}	50 h_{pi} h_{scv}	50 h_{pi} h_{scv}	50 h_{pi} h_{scv}	50 100
	Density curve is smoother and difficult to distinguish between the mode and tail.	1000	1000	1000	1000	<i>median</i>
	Density curve is flat and cannot distinguish between the mode and tail.	5000 10000	5000 10000	5000 10000	5000 10000	5000 10000

2.4.2 Impact of Threshold Choice on the Distribution of Rate Estimates

Figure 2.5 illustrates the distribution of the estimated state rates ($\hat{y}_{i,t}$) of each threshold from 100 repetitions. Since h_{pi} and h_{scv} provided almost identical values for all age groups, only h_{pi} was used in this study. For each chart, the X-axis represents the thresholds that were used to compute the estimated rates ordered from the smallest to the largest, and the number in the bracket is the RMSE of each threshold ($RMSE_{i,t}$ from equation 2.6). The Y-axis shows heart disease mortality rates (per 100,000 population) obtained from the simulated dataset, and each dot represents the estimated state rate for each simulation ($\hat{y}_{i,t}$). The simulated baseline rate (y_i)

and the crude rate are also included in each chart for reference. A crude rate was computed as the average of the ratio of simulated cases to population for each individual ZCTA. Note that the scale of the Y-axis is different for each chart – this was done to account for the large differences in heart disease risk between age groups (e.g., the average heart disease death rates for age groups 35 to 44 and 65 years and older are 33.87 and 1,245.93 per 100,000 population, respectively). Also, the crude rate (second boxplot in each panel in Figure 2.5) shows greater variation in estimated rates compared to all other boxplots. Moreover, the results show that the variation in rates decreases as threshold increase. The smaller box plots indicate that the estimated state rates for each map resulting from each simulation tends to be more consistent, and vice versa. The details of average, standard deviation, and RMSE of estimated rates are shown in Table B.1 (see Appendix B).

For the age group from 35 to 44 (Figure 2.5A), the median rate (the middle line in the boxplot) obtained for each threshold is similar. However, the width of the boxes shrinks towards the center in both the upper and lower quartiles when thresholds increase. This indicates that the estimated state rates are more consistent. Further, the boxplots tend to be similar in structure for thresholds greater than 300 ($h_{ns} \geq h \geq 1000$), in which desirable thresholds are included. The patterns of boxplots in Figure 2.5B (45-54 age group), 2.5C (55-64 age group), and 2.5D (65 years and older) follow similar trends while the boxplots for age group 35 years and older follow slightly different trends (Figure 2.5E). Thus, although the overall width of each boxplot decreases with increasing threshold, the median values also decline as thresholds increase.

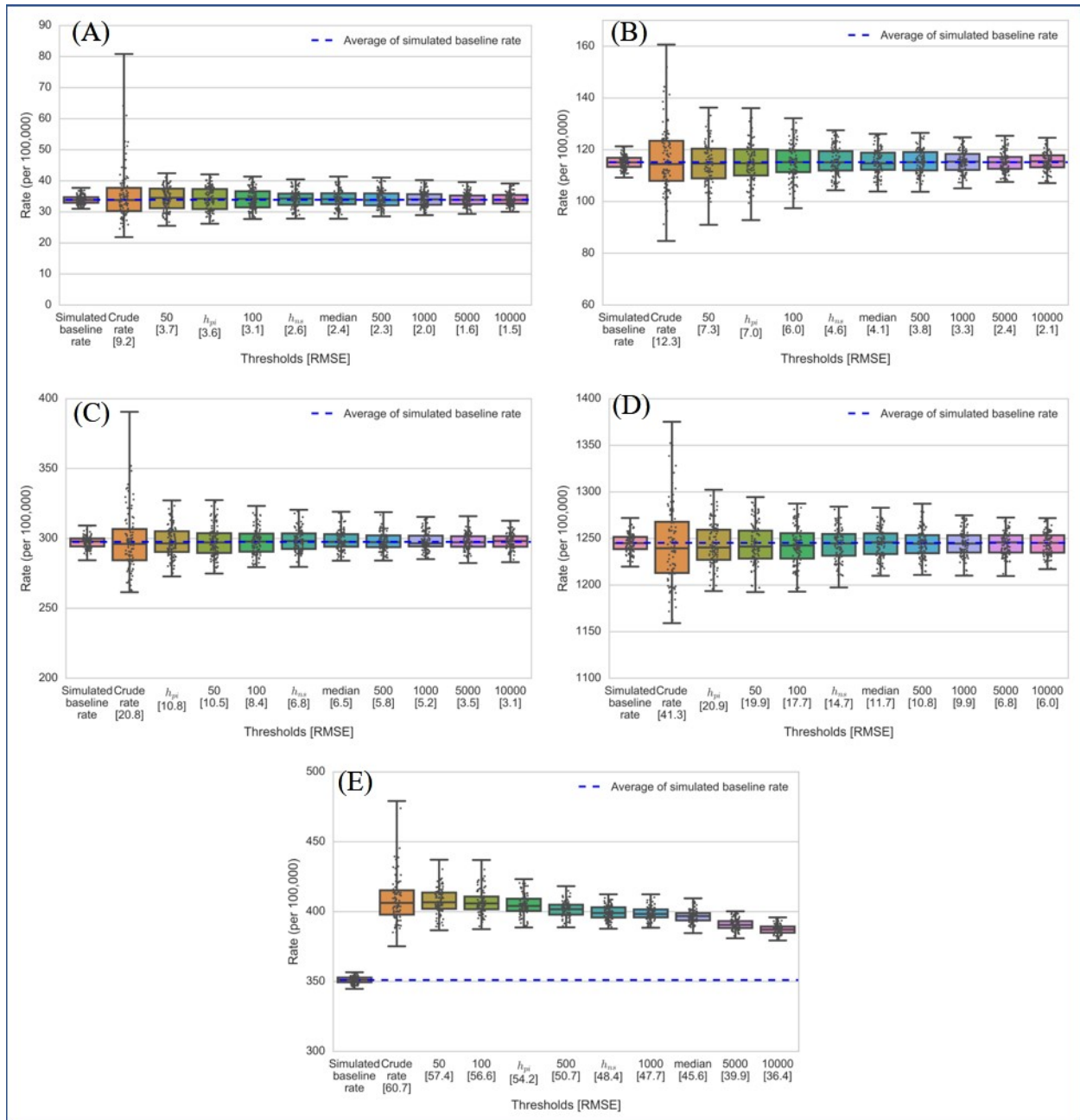


Figure 2.5 The distribution of estimated rates of each bandwidth from 100 repetitions: (A) Aged 35 to 44 years; (B) Aged 45 to 54 years; (C) Aged 55 to 64 years; (D) Aged 65 years and older; (E) Aged 35 years and older.

The inconsistency of estimated state rates for small threshold values is probably due to the small numbers problem, specifically when $h \leq 100$. This is to be expected since the threshold values (h) used to compute the estimated rates in this study are the minimum population size (denominator). Using small threshold values can result in unstable and unreliable rate estimates in spatial units with small population sizes. These unstable rates can affect the estimated state rates since they are aggregated from the smaller spatial units – ZCTA in this study. These results also suggest that $h \leq 100$ may not be an appropriate choice to use.

2.4.3 Desirable Threshold Choices

The three desirable thresholds (Table 2.4) – h_{ns} , *median*, and 500 for the age stratifications and 500, h_{ns} , and 1000 for ages 35 years and older – provide RMSEs that are not noticeably different. However, the boxplot widths are different suggesting different levels of consistency in average rate estimates in maps (Figure 2.5). For producing disease maps, there is a need to balance the amount of geographic detail portrayed on the map and accuracy of estimated rates. While the RMSE suggests similar degrees of accuracy between the maps produced using the three desirable thresholds, the remaining key factor to consider in selecting an appropriate threshold is geographic variation. When the geographic variation is the highest priority, h_{ns} may be the most desirable threshold choice for all age groups since it provides the greatest variation (more geographic detail) among the candidate thresholds, but still produces accurate rates (Figure 2.6). Moreover, compared to arbitrary choices, the h_{ns} provides a consistent way to estimate the appropriate threshold value.

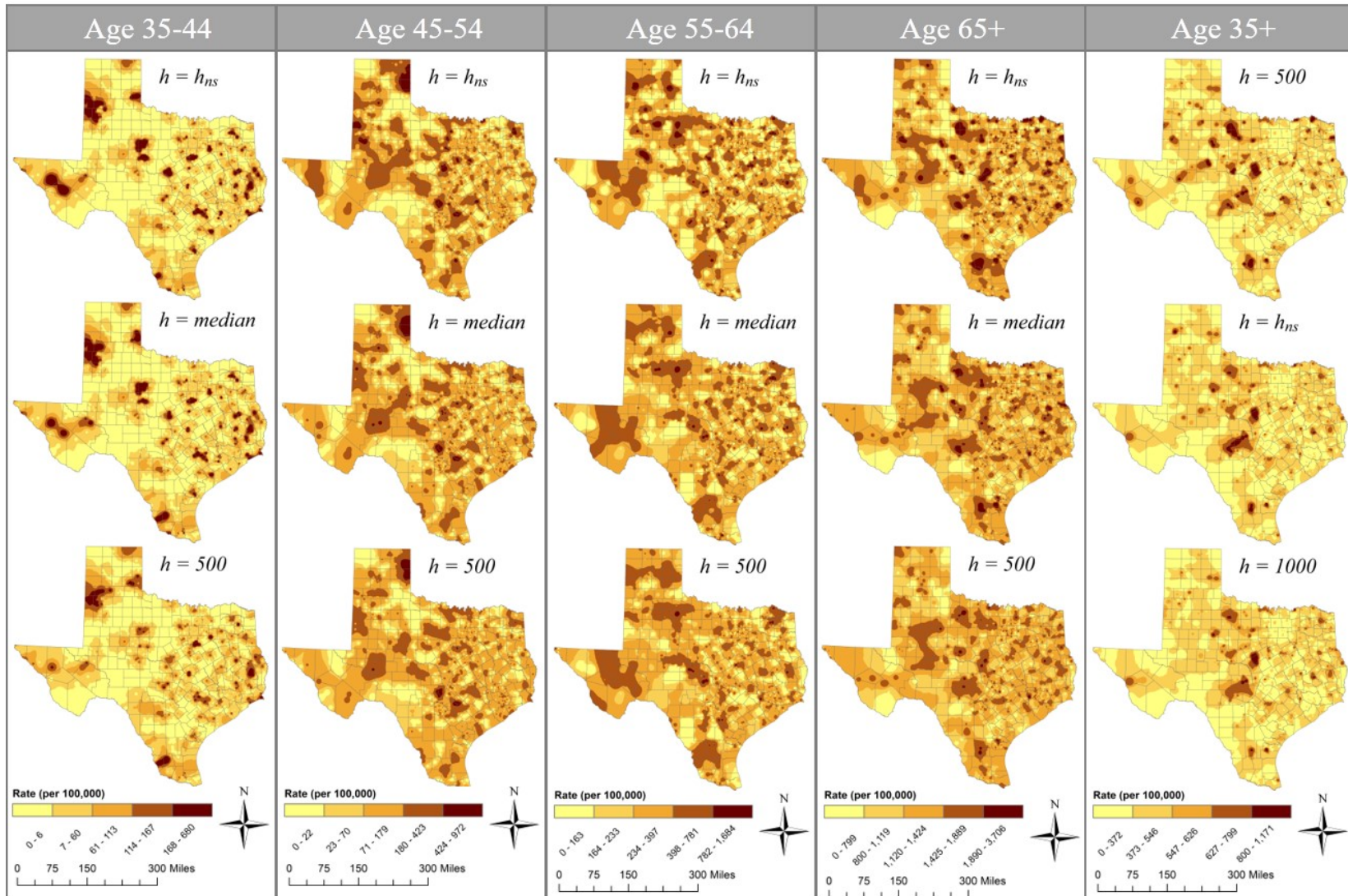


Figure 2.6 Geographic distribution of age-specific heart disease mortality rates for males aged 35-44, 45-54, 55-64, 65 years and older, and 35 years and older. Maps were created using the adaptive KDE method with simulated cases as numerators, population data as denominator, and threshold choices (h) derived from the bandwidth selector methods and arbitrary choices.

2.5 Conclusion

Determining the appropriate threshold value is essential for disease mapping because it affects the degree of smoothing that occurs on the map. In this research, methods for selecting threshold values for a synthetic dataset on heart disease mortality among males aged 35 years and older in Texas are compared using existing bandwidth selectors. The results suggest that h_{ms} is the most desirable threshold for all age-specific groups and the overall population because it provides greater spatial variation while maintaining accuracy in estimated rates. While this is true only for the case data used in this study, our findings underscore the importance of carefully choosing the threshold values to use in disease mapping. Inappropriate thresholds can produce misleading conclusions.

2.6 References

- Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research, 14*, 35-59.
doi:10.1191/0962280205sm388oa
- Beyer, K. M., & Rushton, G. (2009). Mapping cancer for community engagement. *Prev Chronic Dis, 6*(1).
- Beyer, K. M., Tiwari, C., & Rushton, G. (2012). Five essential properties of disease maps. *Annals of the Association of American Geographers, 102*(5), 1067-1075.
doi:10.1080/00045608.2012.659940
- Bithell, J. F. (2000). A classification of disease mapping methods. *Statist. Med., 19*, 2203-2215.
- Cai, Q. (2007). Mapping disease risk using spatial filtering methods (Unpublished doctoral dissertation). The University of Iowa, Iowa City, Iowa.
- Carlos, H. A., Shi, X., Sargent, J., Tanski, S., & Berke, E. M. (2010). Density estimation and adaptive bandwidths: a primer for public health practitioners. *International Journal of Health Geographics, 9*(39). doi:10.1186/1476-072X-9-39
- Centers for Disease Control and Prevention, National Center for Health Statistics. (2015). *Underlying cause of death - heart disease mortality data set, 2009 to 2013* [Data set]. Retrieved from CDC WONDER Online Database: <https://wonder.cdc.gov/ucd-icd10.html>

- Chi, W., Wang, J., Li, X., Zheng, X., & Liao, Y. (2007). Application of GIS-based spatial filtering method for neural tube defects disease mapping. *Wuhan University Journal of Natural Sciences*, 12(6), 1125-1130. doi:10.1007/s11859-007-0097-6
- Chiu, S. T. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika*, 79, 771-782.
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3), 671-681.
- Cromley, E. K., & McLafferty, S. L. (2012). *GIS and public health* (2nd ed.). New York, NY: The Guilford Press.
- Duong, T. (2017). *Kernel smoothing*. R package version 1.10.7.
- Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *J Am Stat Assoc*, 97(458), 632-648.
- Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge, UK: Cambridge University Press.
- Hall, P., & Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, 90, 149-173.
- Hall, P., Marron, J. S., & Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, 92, 1-20.
- Hall, P., Sheather, S. J., Jones, M. C., & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78, 263-269.
- Kelsall, J. E., & Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, 1(1), 3-16.
- Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009, May 1). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat.*, 63(2), 155-162. doi:10.1198/tast.2009.0030
- MacNab, Y. C., Farrell, P. J., Gustafson, P., & Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics*, 60, 865-873.
- 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.
- Natesan, P. (2015, October). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology*, 6(1599). doi:10.3389/fpsyg.2015.01599
- Park, B. U., & Turlach, B. A. (1992). Practical performance of several data-driven bandwidth selectors. *Computational statistics*, 7, 251-270.

- Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, 15, 717-726.
- Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over in homogeneous backgrounds. *International Journal of Geographical Information Science*, 24(5), 643-660. doi:10.1080/13658810902950625
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York, NY: Chapman and Hall.
- Talbot, T. O., Kulldorff, M., Forand, S. P., & Haley, V. B. (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine*, 19, 2399-2408.
- Tiwari, C. (2008). A spatial analysis system for environmental health surveillance [Unpublished doctoral dissertation]. Iowa City, Iowa: The University of Iowa.
- Tiwari, C., & Rushton, G. (2005). Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In P. Fisher, *Developments in spatial data handling* (pp. 665-676). Berlin: Springer-Verlag.
- U.S. Census Bureau. (2016). *2010 ZIP Code Tabulation Areas [Data set]*. Retrieved from TIGER/Line Shapefiles: <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>
- U.S. Census Bureau. (2010). *Profile of general population and housing characteristics: 2010 - 2010 Census Summary File 1 [Data set]*. Retrieved from American FactFinder: <https://factfinder.census.gov/>
- U.S. Census Bureau. (2015). *ZIP Code Tabulation Areas (ZCTAs)*. Retrieved from Geography: <https://www.census.gov/geo/reference/zctas.html>
- Waller, L. A., & Carlin, B. (2010). Disease mapping. In A. E. Gelfand, P. J. Diggle, M. Fuentes, & P. Guttorp, *Handbook of spatial statistics* (pp. 217-243). Boca Raton, FL: CRC Press.
- Waller, L., & Gotway, C. (2004). *Applied spatial statistics for public health data*. Hoboken, NJ: John Wiley & Sons, Inc.
- Wand, M. P., & Jones, M. C. (1994). Multivariate plugin bandwidth selection. *Computational Statistics*, 9, 97-116.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Boca Raton, FL: Chapman & Hall/CRC.
- Washington State Department of Health. (2012, October). *Guidelines for working with small numbers*. Retrieved from <http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf>

CHAPTER 3

RELATIVE EFFICACY OF ADAPTIVE KERNEL DENSITY ESTIMATION AND SPATIAL EMPIRICAL BAYES METHODS FOR DISEASE MAPPING

3.1 Abstract

Disease maps are invaluable for monitoring spatial patterns of disease and targeting interventions. While many disease mapping methods are available, their relative efficacy and limitations remain unclear. Using simulated data generated under specific scenarios for known underlying risk and population sizes, this study evaluates the two prominent smoothing methods in disease mapping – the adaptive kernel density estimation (KDE) and spatial empirical Bayes. The relative performance of these two methods are determined using two measures: (1) global measure of difference – comparing the average rate estimates in the entire study area between two methods; (2) local measure of difference – comparing rate estimates between two methods at a local scale, i.e., the zip code level. The findings illuminate the similarities and differences between these two methods and provide hitherto unknown insights into the two approaches. The most important finding of the study is that while both methods provide identical rate estimates at the global scale, differences in rate estimates occur at the local scale. Significant differences occur in sparsely populated, non-urban zip codes. Thus, using either the adaptive KDE or spatial empirical Bayes method to map diseases in densely populated (urban) areas may provide identical rate estimates, but caution is necessary when mapping disease in sparsely populated areas. Further research is required to determine the optimal disease mapping method for low population density areas.

Keywords: disease mapping, kernel density estimation, Bayesian, smoothing methods, health GIS,

3.2 Introduction

Disease maps have become a powerful tool in epidemiological surveillance. They not only illustrate the geographic distribution of diseases but also offer an alternative approach for public health to address and specify the hypotheses of disease etiology and causation (Cromley & McLafferty, 2012). To produce maps, disease rates, commonly computed by dividing the number of disease cases by the total population in the same geographic area, are recommended.

However, computing rates requires considerable care since they may lead to the small numbers problem – a well-known issue in disease mapping (Cromley & McLafferty, 2012). Basically, in areas with small populations, any change in number of cases can generate wide variations in the estimated rates. As a result, these maps may contain a combination of rates that are generated using stable and unstable population sizes. To address this issue, smoothing methods such as kernel density estimation (KDE) and spatial empirical Bayes methods are often used.

Kernel density estimation (KDE), a non-parametric method, is a generic approach that can be used to calculate disease rates and illustrate them as a continuous surface (Cromley & McLafferty, 2012). In the disease mapping context, kernel density estimation methods operate by computing rates within a spatial kernel (a circle) that is moved across the study area. Using this approach, there are two common ways to apply the KDE method in disease mapping: (1) a ratio of the density of events (i.e., cases) within each kernel to the density of the background (i.e., population) within the same-sized kernel (Shi, 2010); (2) a ratio of the number of cases to the population size within a kernel (Rushton & Lolonis, 1996). Kernel density estimation can employ either a fixed or adaptive approach. For a fixed approach, the kernel has a fixed radius for all circles. In the adaptive approach, however, the kernel has a variable radius that grows or shrinks to accommodate a defined population size. This is more suitable for health studies than

the fixed approach since it reflects differences in underlying risk population (Shi, 2010; Carlos, Shi, Sargent, Tanski, & Berke, 2010; Tiwari, 2008). In the last decades, KDE methods have been used in many studies to explore the geographic distributions of health outcomes (Rushton & Lolonis, 1996; Shi, 2010; Carlos et al., 2010; Oppong, Tiwari, Ruckthongsook, Huddleston, & Arbona, 2012; Oppong, Kutch, Tiwari, & Arborna, 2014).

Spatial empirical Bayes method is a model-based approach that accounts for the small numbers problem by aggregating data using some locally-defined neighborhood function (Clayton & Kaldor, 1987). In this approach, rates in the areas with small sample sizes are smoothed towards some local mean generated by the neighbor (Clayton & Kaldor, 1987; Waller & Carlin, 2010). Other approaches include global methods where rates in the areas with small populations are equally smoothed toward a global mean (average rate of the entire study area) without regard to their relative spatial location (Waller & Carlin, 2010). The rates computed using these methods are typically illustrated as a choropleth map – showing abrupt rate changes at the boundary. Spatial empirical Bayes applications in disease mapping include Ghosh, Natarajan, Waller, and Kim (1999), Morris, Whittaker, and Balding (2000), Meza (2003), MacNab, Farrell, Gustafson, and Wen (2004), and Schur et al. (2013).

Although producing disease maps using the same dataset but different smoothing methods may result in different spatial patterns of disease, no systematic evaluation of the efficacy of these approaches currently exists. For example, Figure 3.1A-C, illustrates the spatial patterns of heart disease mortality rates in Texas. These maps were created using the same dataset obtained from the Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2015) but different methods. Figures 3.1A were produced using crude death rates as a statistical measure. Figure 3.1B and 3.1C were respectively created using

empirical Bayes and the adaptive KDE methods. Figure 3.1A, an unsmoothed choropleth map, indicates that high disease rates seem to cluster only in East Texas; Figure 3.1B, a smoothed choropleth map, shows that high disease death rates are clustered in the Texas Panhandle, East Texas, and the west of Dallas-Fort Worth Metroplex areas, and; Figure 3.1C, a smoothed continuous surface map, suggests that the higher disease rates mostly cluster in the west of Dallas Fort-Worth Metroplex areas. The conflicting results from these three maps clearly demonstrate that the chosen disease mapping method influences the perceived spatial distribution patterns of disease rates. This study adopts the simulation-based approach from Goovaerts and Gebreab (2008) to evaluate the relative performance of the adaptive KDE and spatial empirical Bayes methods.

3.3 Data

The evaluation of the relative performance of disease mapping methods between adaptive KDE and spatial empirical Bayes in this study was illustrated using age-specific mortality rates for heart disease in Texas (see Table A.1 in Appendix A for a list of the International Classification of Diseases (ICD) for heart diseases used in this study). The data was obtained from the CDC NCHS (2015). This study focused on male aged 35-44, 45-54, 55-64, and 65 years and older recorded between 2009 and 2013. The age-specific rates were computed using the 2010 population, obtained from the 2010 U.S. Census Bureau (U.S. Census Bureau, 2010) (Table 3.1). The ZIP Code Tabulation Areas (ZCTAs) cartographic boundary files were obtained from Topologically Integrated Geographic Encoding and Referencing (TIGER) (U.S. Census Bureau, 2016). Note that ZCTAs were created by the U.S. Census Bureau and generalized areal representation of USPS ZIP Code service areas (U.S Census Bureau, 2015).

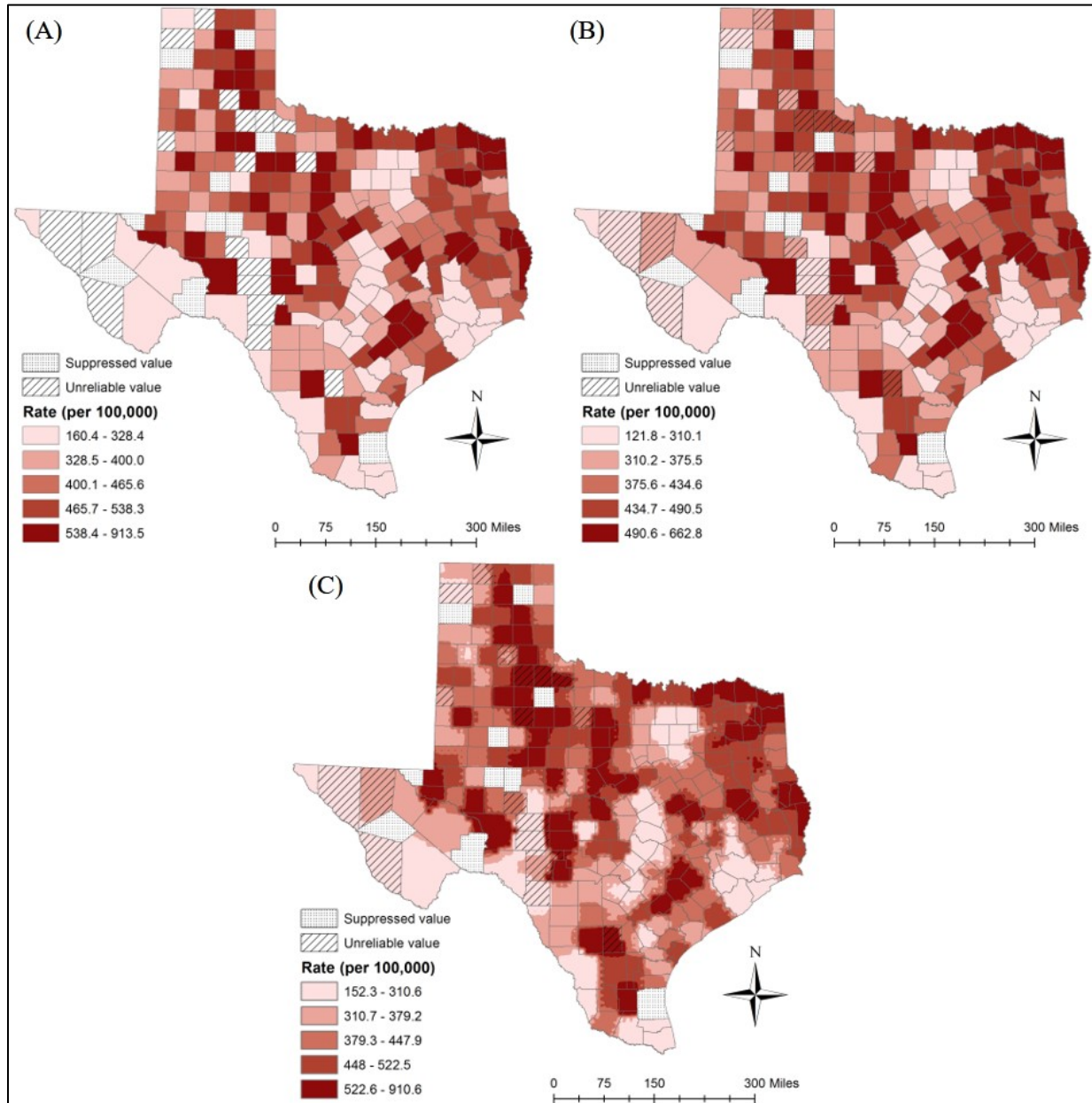


Figure 3.1 Spatial distribution of heart disease in Texas, aged 35 years and over, 2010-2012: (A) crude death rate; (B) spatially smoothed death rate computed using a local empirical Bayes algorithm proposed by Marshall (1991); (C) spatially smoothed death rate computed using KDE method (Note: all maps were created using: (1) a quantile classification, and; (2) the same dataset from CDC NCHS (2015))

Table 3.1 Population distribution (U.S. Census Bureau, 2010) and age-specific heart disease death rate among male in Texas, 2009-2013 (CDC NCHS, 2015)

Age	Male		
	2010 Population	Case (5 yrs)	Rate
35-44	1,722,904	2,918	33.87
45-54	1,702,639	9,803	115.15
55-64	1,256,976	18,689	297.36
65+	1,135,517	70,739	1,245.93
Total (35+)	5,818,036	102,149	351.15

3.4 Methods

3.4.1 Methods to Generate Simulated Data

Generating the simulated data comprised of two steps. First, population distribution for males was generated by the age groups specified in Table 3.1, and distributed using random point distribution at the ZCTA level. Next, a simulated case dataset was generated. For each age group, all points were assigned a random number generated from the uniform distribution in the interval (0,1). This assumes that each person has an equal probability of being infected. The probability that a simulated point would be classified as a case was determined using observed age-specific mortality rates (Table 3.1). For example, the observed age-specific heart disease death rate for males aged 35 to 44 years old in Texas was 33.87 per 100,000 (0.0003387) (Table 3.1). If a random number generated was in the range 0.0000001 to 0.0003387, it was classified as a simulated case. This step was processed 100 replications (*I*) for each age group.

3.4.2 Methods to Compute Estimated Rates

3.4.2.1 Adaptive Kernel Density Estimation

Web-based Disease Mapping Analysis Program (WebDMAP), an open source and web-based GIS software (Tiwari, 2008), was used to compute estimated heart disease mortality rates

using the adaptive KDE method. WebDMAP requires three separate files – the grid, control, and case files. The grid file contained geographic coordinates, which was ZCTA centroids generated in ArcGIS 10.2; the control file contained the population in each spatial unit (i.e., ZCTA); and the case file contained the number of cases of the disease in each spatial unit (i.e., simulated cases in this study). This process was repeated for 100 replications of the case data. Note that the grid and control files remained the same for all simulations. The simulations were programmed in Python and the analysis was conducted in R. The data for all simulations were stored in a PostgreSQL database server.

3.4.2.2 Spatial Empirical Bayes

Spatial empirical Bayes method was implemented using Python Spatial Analysis Library (PySAL), an open source library of computational tools for spatial analysis developed by Rey and Anselin (2010). In addition to case and population data, the spatial empirical Bayes method requires a spatial weights file. This file defines the local neighborhood used in the calculation of the smoothed rates. All spatial weight matrices were constructed to reflect the same spatial basis of support (i.e., same size of spatial unit and population in each spatial unit) used by the adaptive KDE method. This was done to ensure consistency in the spatial supports used in the calculation of rates between the adaptive KDE and the spatial empirical Bayes methods. The spatial weights file was generated using PostgreSQL and R and were converted to the GWT structure for use in PySAL. Note that GWT is a weight matrix structure that PySAL allows users to manipulate, modify and create their own spatial weight matrix (Anselin, 2003). It comprises of three columns – observation (origin), observation’s neighbor(s) and distance.

3.4.2.3 Smoothing Parameters

To compute smoothed rates, both the adaptive KDE and spatial empirical Bayes require smoothing parameters – bandwidth (threshold) and prior distribution (defined by spatial weights matrix) respectively. These are crucial parameters that influence the degree of smoothing that occurs on the map (Silverman, 1986; Kelsall & Diggle, 1995; Waller and Gotway, 2004). To account for different sized spatial supports, this study uses a variety of different threshold sizes and corresponding spatial weights files. In all, nine thresholds were selected. Six of them were arbitrary thresholds - 50, 100, 500, 1000, 5000 and 10000, and the remaining were variable thresholds computed based on population data – *median* (median population), normal scale selector (h_{ns}) and plug-in selector (h_{pi}). The latter were computed using ks-package in R. A normal scale selector is a quick and simple selector that aims to find a threshold value that is reasonable for a wide range of data distributions (see Wand & Jones, 1995 for details). A plug-in selector is a hi-tech selector that seeks to provide an answer for very general classes of underlying functions which involves more mathematical arguments and requires more computational effort (see Wand & Jones 1994 for details).

3.4.3 Evaluation

The relative performance of these two different disease mapping methods was determined using two measures: a global and local measure of difference. The global measure of difference is estimated by comparing the estimated state rate obtained from the KDE method to that obtained from the spatial empirical Bayes method. This measure was computed for each of the 100 replications. Boxplots were used to illustrate the variation in estimated state rates

between these two methods. The root-mean-square-error (RMSE) was used to measure similarity in estimated state rates obtained using these methods (equation 3.1).

$$RMSE_{i,t}^{global} = \sqrt{\frac{1}{100} \sum_{l=1}^{100} (Y_{i,t,l}^K - Y_{i,t,l}^B)^2} \quad (3.1)$$

where $RMSE_{i,t}^{global}$ was RMSE at the global difference of age group i at threshold t when i and t represented age groups and thresholds used in this study, and $Y_{i,t,l}^K$ and $Y_{i,t,l}^B$ were respectively the estimated state rate computed from the adaptive KDE and spatial empirical Bayes methods for age group i and threshold t at the l th simulation.

The local test checks for consistency in rate estimates between the two methods at a local scale, i.e., at the ZCTA level. The reasoning behind this test is based on the assumption that similar spatial supports should result in similar rate estimates. For each rate estimate in each of the 100 replications, the spatial basis of support for the KDE method was identical to the spatial basis of support for the spatial empirical Bayes method. Therefore, the case and population counts used in the calculation of each rate estimate was consistent between the two methods. The only difference in rate estimates is due to differences in how these two methods smooth the data. The RMSE value was used to measure this similarity or dissimilarity between the two methods (equation 3.2). Further, boxplots were used to illustrate the variation of RMSEs.

$$RMSE_{i,t,l}^{local} = \sqrt{\frac{1}{N} \sum_{s=1}^N (y_{i,t,l,s}^K - y_{i,t,l,s}^B)^2} \quad (3.2)$$

where $RMSE_{i,t,l}^{local}$ was RMSE at the l th simulation of age group i and threshold t when i and t represented age groups and thresholds used in this study, $y_{i,t,l,s}^K$ and $y_{i,t,l,s}^B$ were respectively the estimated rates from the adaptive KDE and spatial empirical Bayes methods of age group i and threshold t at the spatial unit s in the l th simulation, and N was a total number of ZCTAs at age group i and threshold t .

3.5 Results and Discussion

The state rate computed from the simulation data, called simulated baseline rate in this study, is mostly identical to the CDC state rate and has very low RMSE for all age groups (Table B.2 in Appendix B). Note that the simulated baseline rate was calculated by dividing the total number of simulated cases by total population while the estimated state rate refers to the average of rate estimates computed from the adaptive KDE and spatial empirical Bayes methods.

3.5.1 Global Difference

Figure 3.2 illustrates the distribution of state rate estimates from the adaptive KDE and spatial empirical Bayes methods for each threshold from 100 replications. For each chart, X-axis represents threshold values that were used to compute the rate estimates ordered from the smallest to largest thresholds, Y-axis is heart disease death rates (per 100,000 population), and each dot represents the rate estimates for each replication. Because of the vast difference of heart disease risk between age groups, the scale of Y-axis in each chart was adjusted in regard to its rates. Overall, the estimated state rates from both methods are identical to each other in all age groups and thresholds (Figure 3.2). Thus, we can conclude that the estimated state rate is consistent between the two methods.

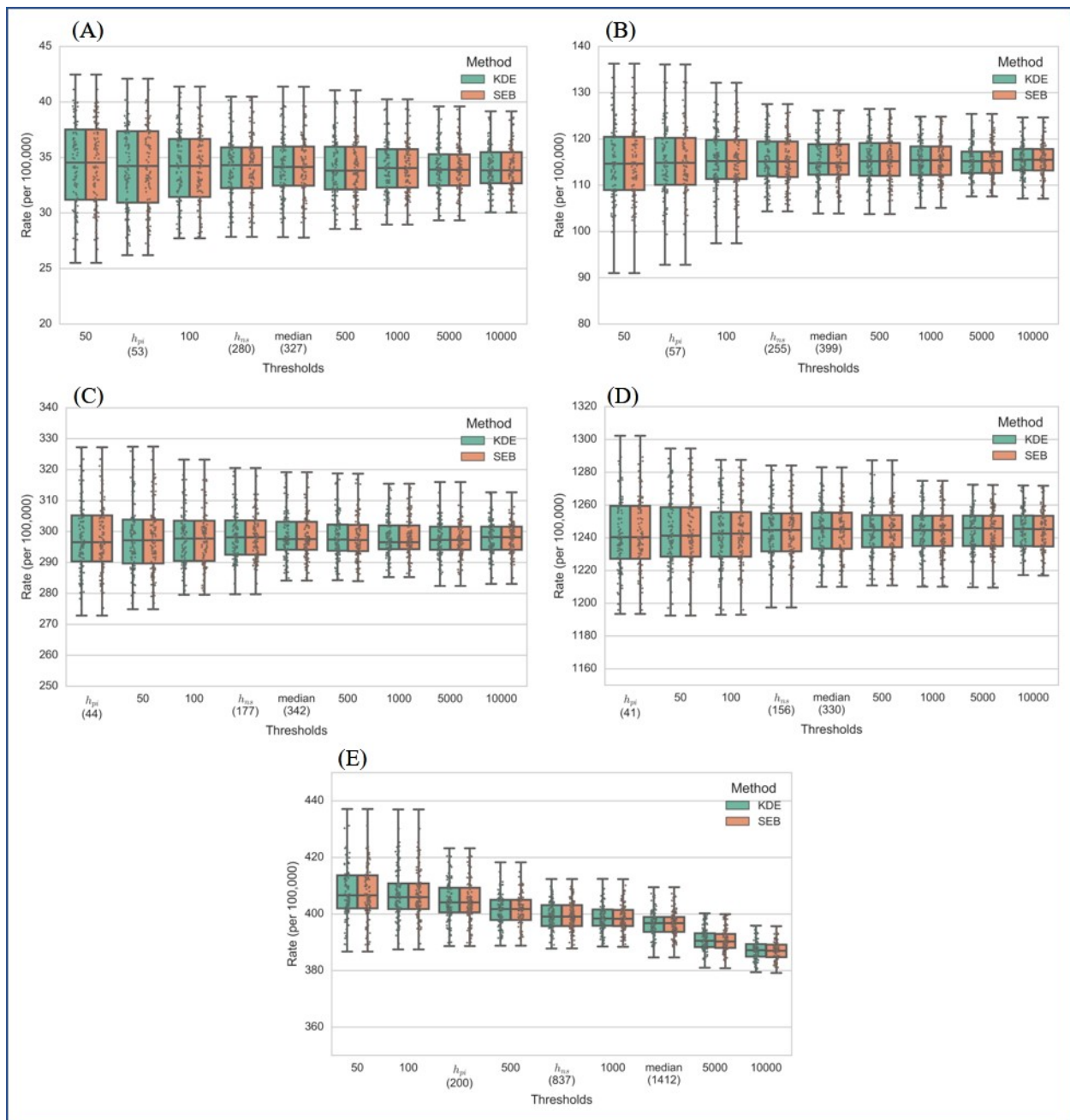


Figure 3.2 The distribution of estimated state rates computed from the adaptive KDE and spatial empirical Bayes methods ($l = 100$ simulations): (A) Age 35 to 44 years; (B) Age 45 to 54 years; (C) Age 55 to 64 years; (D) Age 65 years and older; (E) Age 35 years and older (overall).

3.5.2 Local Difference

Figure 3.3 illustrates the variation of 100 RMSEs in each threshold. The lower RMSEs indicate that the estimated rates computed from the adaptive KDE and spatial empirical Bayes at the ZCTA levels have small differences in rate estimates, and vice versa. Although we have a few extremely high RMSEs, each panel in Figure 3.3 provides a similar pattern when $h \leq 100$. Most RMSEs are less than 0.01 (Table B.3 in Appendix B). This indicates that the estimated rates computed from the adaptive KDE and spatial empirical Bayes approaches are mostly identical.

When $h > 100$, the results show greater differences of rate estimates between the two methods in all age-specific groups, but the degree of differences varies. The greater differences in rate estimates when thresholds increase is probably due to the population distribution and different smoothing approaches. Note that the threshold value is a minimum value used to compute the estimated rate. For example, when $h = 100$, the ZCTAs that have population equal or greater than 100 have their own rates: the population size in these ZCTAs reach the minimum, so the estimated rate – the ratio of a number of cases to population in those ZCTAs – can be directly computed. According to the 2010 U.S. Census Bureau (2010), about 70% of the total ZCTAs contains population equal or greater than 100, and approximately 50% of the total ZCTA has greater than 300 population when stratified by age (Table 2.3 in section 2.3.1). That is, when $h = 100$ and 300 , 70% and 50% of the total ZCTAs have their own rates (non-smoothed ZCTAs) while about 30% and 50% of total ZCTAs are smoothed, respectively. The percentages of smoothed ZCTAs increase as the threshold increases.

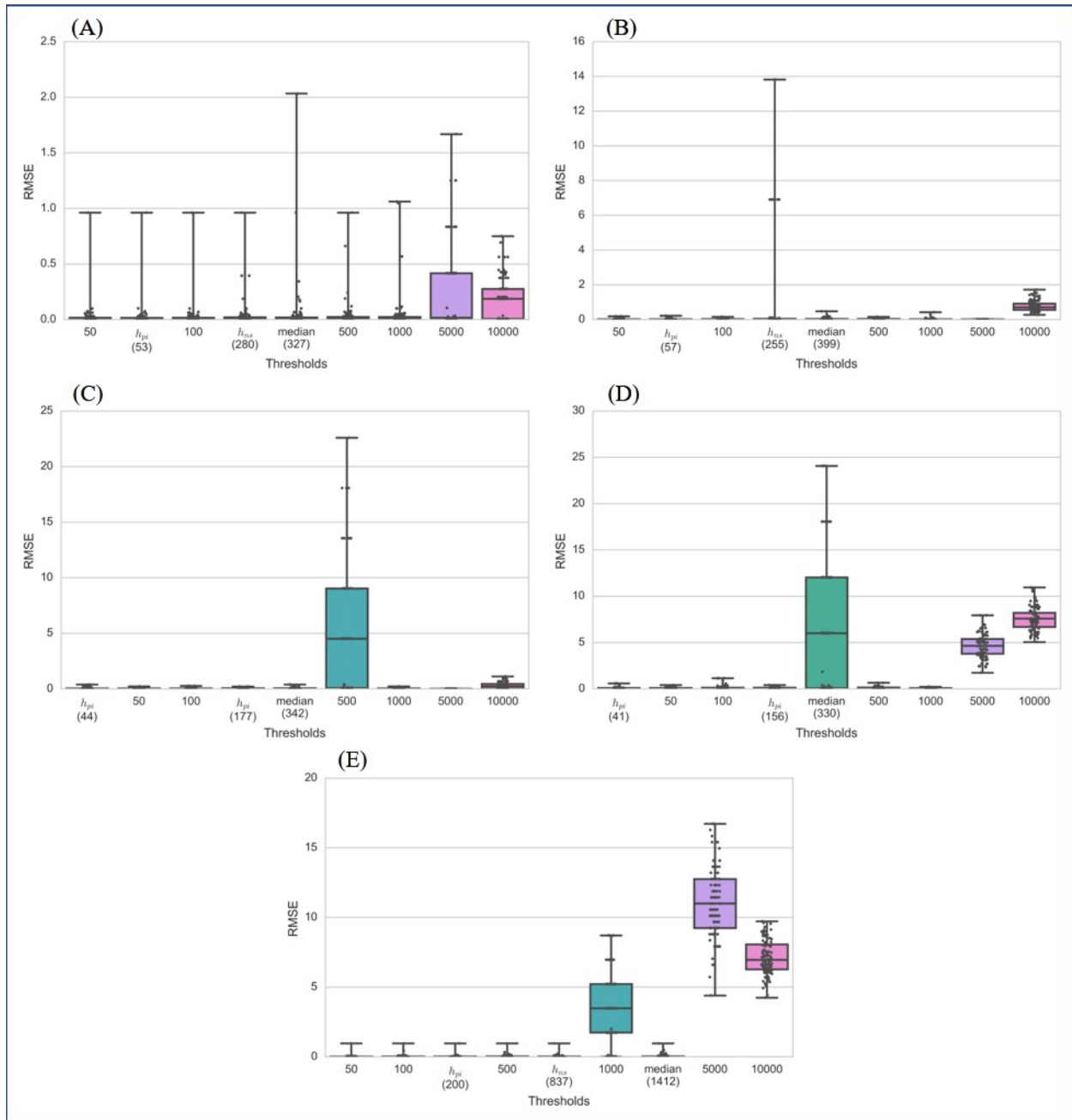


Figure 3.3 The distribution of RMSE (the difference of rate estimates between the adaptive KDE and spatial empirical Bayes methods) ($L = 100$ simulations): (A) Age 35 to 44 years; (B) Age 45 to 54 years; (C) Age 55 to 64 years; (D) Age 65 years and older; (E) Age 35 years and older (overall)

When $h = 10000$ – the largest threshold in this study, the difference in rate estimates between the adaptive KDE and spatial empirical Bayes methods has the greatest dissimilarity for all age-specific groups. The results found that there is no ZCTA that contains population greater than 10,000 in this study. As a result, to compute rate estimates when $h = 10000$, all ZCTAs are smoothed to reach the minimum requirement (threshold). This may indicate that the more ZCTAs are smoothed, the more variations and differences may occur.

Considering the unusual high RMSEs (Figure 3.3), there are some similarities between the two younger groups and between the older groups. For the younger groups (age 35 to 44 and 45 to 54) (Figure 3.3A and 3.3B), when $h < 10000$, all median RMSEs are less than 0.05 while they are greater than 0.1 at the largest threshold in both groups. The unusually high RMSEs when $h < 5000$ for both age groups are less than 10% of the total simulations. In other words, the rate estimates from the adaptive KDE and spatial empirical Bayes are identical and accounted for 90% of the total simulations for both age groups. When $h < 5000$, the results show that there are a few simulations that provide greater differences in rate estimates between the adaptive KDE and spatial empirical Bayes methods. For each threshold in both age groups, these unusually high RMSEs are caused by one ZCTA. Note that the causing ZCTA is not necessarily the same ZCTA between thresholds. The similarities of the unusually high RMSEs of these two age groups are: (1) when $h < 1000$, the ZCTAs that have the greatest difference in estimated rates are in non-urban areas, and; (2) when $h = 10000$, the ZCTA that provides the greatest difference in estimated rates in both age groups is the same ZCTA and located in urban areas. In addition to the population distribution discussed above, one possible reason for the greatest difference in rate estimates in urban areas may be due to the low number of cases in the ZCTA and its neighbors. According to the Washington State Department of Health (2012), disease rates are unreliable

when each spatial unit has less than 20 cases. The ratio of heart disease mortality of these two age groups are about 1:2,950 and 1:870 respectively (CDC NCHS, 2015). This indicates that although there is an adequate population size for computing disease rates, unstable rates may occur if the spatial unit has a low number of cases. However, this scenario only occurs when $h > 1000$ in this study.

Unlike the younger age groups, the older age groups – 55 to 64 and 65 years and older – have a great number of high differences in rate estimates at the mid-size threshold – $h = 500$ for age 55-64 and $h = median$ for age 65 years and older (Figure 3.3C and 3.3D). For both age groups, the number of these high differences are greater than 50% of the total simulations. For each age group, the results found that these high differences are caused by only one ZCTA, and all simulations that have the greatest difference in rate estimates are from the same ZCTA. The ZCTAs for both age groups that causes this circumstance are located in non-urban areas and contain population less than 15. Similar to age-specific groups, the unusual differences in rate estimates for age 35 years and older when $h = 1,000$ and $5,000$ (Figure 3.3E) are also from the ZCTAs that contain a small population size in non-urban areas.

In summary, when using small thresholds ($h \leq 100$ in this study), the estimated rates computed from the adaptive KDE and spatial empirical Bayes are most identical in greater than 90% of the total simulations. When thresholds increase, the degree of differences in rate estimates between the two methods increase in all age groups. For all thresholds, the cause of the greatest difference in rate estimates is from only one ZCTA, and this ZCTA is mostly located in non-urban areas and has a small population size. This also applies to a large number of unusual differences in rate estimates when $h = 500$ and *median* in age groups 55 to 64 and 65 years and older respectively. Since the spatial basis of support for both methods are identical in this study,

the cause of greater difference in rate estimates, in addition to population distribution, is probably due to the difference of smoothing mechanism behind the adaptive KDE and spatial empirical Bayes methods. Moreover, the number of cases may be another key factor that causes this circumstance.

3.6 Conclusion

Disease maps play a significant role in disease surveillance systems to help public health practitioners to monitor and prevent disease outbreaks as well as target vulnerable people for intervention. Even though various smoothing methods have been developed and used, no systematic evaluation of such approaches currently exists. Using a simulated disease dataset, this study evaluated the relative performance of the adaptive KDE and spatial empirical Bayes methods. For all age groups, the results found that both methods provide identical estimated state rates in all thresholds. This is one of important findings in this study.

The estimated rates at the ZCTA level computed from the adaptive KDE and spatial empirical Bayes are identical when $h \leq 100$, and the difference in rate estimates increases when thresholds increase for all age groups. For the older age groups – 55 to 64 and 65 years and older – a large number of unusual high differences in rate estimates are found in the mid-size threshold values. In this study, the cause of these high RMSEs is mostly due to a ZCTA with small population size located in non-urban areas. This also applies to some unusually high differences in rate estimates for all age groups. Since the cause of extremely high differences in rate estimates for all age groups is from the small population ZCTA in non-urban areas, this implies that using either the adaptive KDE and spatial empirical Bayes approach to map disease in urban areas may provide identical rate estimates, and caution is needed when mapping disease in non-

urban areas. Note that the scenarios found in this study may be suitable only for degenerative diseases such as heart disease in Texas. Other diseases and areas may provide different scenarios due to the difference of population distribution and disease characteristics.

3.7 References

- Anselin, L. (2003). *GeoDa 0.9 user's guide*. Retrieved August 8, 2016, from http://la1.rcc.uchicago.edu/media/geoda_files/docs/geoda093.pdf
- Carlos, H. A., Shi, X., Sargent, J., Tanski, S., & Berke, E. M. (2010). Density estimation and adaptive bandwidths: a primer for public health practitioners. *International Journal of Health Geographics*, 9(39). doi:10.1186/1476-072X-9-39
- Centers for Disease Control and Prevention, National Center for Health Statistics. (2015). *Underlying cause of death - heart disease mortality data set, 2009 to 2013 [Data set]*. Retrieved from CDC WONDER Online Database: <https://wonder.cdc.gov/ucd-icd10.html>
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3), 671-681.
- Cromley, E. K., & McLafferty, S. L. (2012). *GIS and public health* (2nd ed.). New York, NY: The Guilford Press.
- Ghosh, M., Natarajan, K., Waller, L. A., & Kim, D. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.
- Goovaerts, P., & Gebreab, S. (2008). How does Poisson kriging compare to the popular BYM model for mapping disease risks? *International Journal of Health Geographics*, 7(6).
- Kelsall, J. E., & Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, 1(1), 3-16.
- MacNab, Y. C., Farrell, P. J., Gustafson, P., & Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics*, 60, 865-873.
- 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.
- Meza, J. L. (2003). Empirical Bayes estimation smoothing of relative risks in disease mapping. *Journal of Statistical Planning and Inference*, 112, 43-62.
- Morris, A., Whittaker, J., & Balding, D. (2000). Bayesian fine-scale mapping of disease loci, by hidden Markov models. *The American Journal of Human Genetics*, 67, 155-169.

- Oppong, J. R., Kutch, L., Tiwari, C., & Arbona, S. (2014). Vulnerable places: prison locations, socioeconomic status, and HIV infection in Texas. *The Professional Geographer*, 66(4), 653-663. doi:10.1080/00330124.2013.852040
- Oppong, J. R., Tiwari, C., Ruckthongsook, W., Huddleston, J., & Arbona, S. (2012). Mapping late testers for HIV in Texas. *Health & Place*, 18, 568-575. doi:10.1016/j.healthplace.2012.01.008
- Rey, S., & Anselin, L. (2010). PySAL: a python library of spatial analytical methods. In M. Fischer, & A. Getis, *Handbook of applied spatial analysis: software tools, methods, and applications*. Berlin, Germany: Springer.
- Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, 15, 717-726.
- Schur, N., Hurlimann, E., Stensgaard, A., Chimfwembe, K., Mushinge, G., Simoonga, C., . . . Vounatsou, P. (2013). Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Tropica*, 128, 365-377.
- Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over in homogeneous backgrounds. *International Journal of Geographical Information Science*, 24(5), 643-660. doi:10.1080/13658810902950625
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York, NY: Chapman and Hall.
- Tiwari, C. (2008). *A spatial analysis system for environmental health surveillance* [Unpublished doctoral dissertation]. The University of Iowa, Iowa City, Iowa.
- U.S. Census Bureau. (2016). *2010 ZIP Code Tabulation Areas* [Data set]. Retrieved from TIGER/Line Shapefiles: <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>
- U.S. Census Bureau. (2010). *Profile of general population and housing characteristics: 2010 - 2010 Census Summary File 1* [Data set]. Retrieved from American FactFinder: <https://factfinder.census.gov/>
- U.S. Census Bureau. (2015). *ZIP Code Tabulation Areas (ZCTAs)*. Retrieved from Geography: <https://www.census.gov/geo/reference/zctas.html>
- Waller, L. A., & Carlin, B. (2010). Disease mapping. In A. E. Gelfand, P. J. Diggle, M. Fuentes, & P. Guttorp, *Handbook of spatial statistics* (pp. 217-243). Boca Raton, FL: CRC Press.
- Waller, L., & Gotway, C. (2004). *Applied spatial statistics for public health data*. Hoboken, NJ: John Wiley & Sons, Inc.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Boca Raton, FL: Chapman & Hall/CRC.

Wand, M. P., & Jones, M. C. (1994). Multivariate plugin bandwidth selection. *Computational Statistics*, 9, 97-116.

Washington State Department of Health. (2012, October). *Guidelines for working with small numbers*. Retrieved from <http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf>

CHAPTER 4
DEFINING SPATIAL UNITS FOR MAPPING HIV AND
TARGETING INTERVENTION IN TEXAS

4.1 Abstract

Disease mapping has become a powerful tool for understanding the spatial patterns of outcomes and pinpointing *where* intervention is most needed. Yet, disease maps alone cannot identify *who* or *which populations* are most at risk, and may need interventions the most. This limitation could hamper disease prevention and control efforts. Informed by the social determinants of health (SDH) framework, this study examines the spatial units of HIV risk in Texas. By combining SDH with these maps, we are able to not only geographically target *where* but also tailor interventions in those areas to *those high-risk populations*. Moreover, even when health data is limited, the health risk map alone is adequate for identifying *where* and *which populations are at risk* since it is constructed from publicly accessible disease-related data. The findings have implications for public health policy targeted at HIV risk communication and management.

Keywords: defining spatial units, mapping health risk, social determinants of health, neighborhood characteristics, HIV/AIDS, Texas

4.2 Introduction

Disease mapping is routinely used to visualize not only the spatial patterns of disease burdens but also to target and evaluate interventions. However, because diseases do not follow neat standard boundaries such as counties or even zip codes – the typical spatial unit used for reporting disease data – the usefulness of such maps depends critically on the spatial unit

employed during mapping. Kernel density estimation and kriging are frequently used to address the most common problems, but they fail to address the fundamental question of *who* within that micro spatial unit has the highest disease burden and should be prioritized for intervention. Hence, the question is whether we can identify sub areas, neighborhoods of risk, based on internal consistency and external difference with regards to known risk factors. Such maps would be invaluable for disease burden assessment and targeting interventions but are rarely developed in this study context. Consequently, this paper identifies neighborhoods of HIV/AIDS risk in Texas and the subgroups that may be prioritized for intervention.

A neighborhood describes an area or spatial unit where people interact with each other. People tend to have a good perception of what they view as *their* neighborhood, but its boundaries may be different from other people in their neighborhood. Because the perceived boundaries are not universally accepted, delimiting neighborhood boundaries is difficult. In health research, the term *neighborhood* has been used loosely to refer to “a person’s immediate residential environment, which is hypothesized to have both material and social characteristics potentially related to health” (Diez Roux, 2001). In other words, neighborhood environments and/or characteristics can influence individual health (Emch, Root, & Carrel, 2017). For example, individuals living in affluent places tend to have better health than those living in impoverished areas (Diez Roux, Mujahid, Hirsch, Moore, & Moore, 2016; Chaix, Merlo, Evans, Leal, & Havard, 2009; Cockings & Martin, 2005; Lebel, Pampalon, & Villeneuve, 2007; Coulton, Korbin, Chan, & Su, 2001; Browning & Cagney, 2003). This is known as social gradient in health (Emch et al., 2017) and is also relevant to the concept of place vulnerability – “where people live matters in their vulnerability to disease” (Oppong & Harold, 2009). Regardless of whether people are healthy or unhealthy, rich or poor, those who live in vulnerable

neighborhoods (e.g., areas with poor sanitation, food desert environments, with high poverty and high crime rates) tend to have poorer. In contrast, vulnerable people who live in healthier places (less vulnerable e.g., with access to recreation and healthy food) generally have better health. This implies that where people live affects not only their habits and behaviors, but also shapes their choices and exposures. Furthermore, vulnerable people are attracted to, and tend to live in, vulnerable neighborhoods because that is the only place they can afford. Finally, a high concentration of vulnerable people in any neighborhood can make those neighborhoods more vulnerable and less able to resist disease, and increases the vulnerability of all who live in that location.

In neighborhood health studies, administrative units have often been used as rough proxies for defining neighborhoods (Emch et al., 2017; Diez Roux, 2001; Rauh, Andrews, & Garfinkel, 2001; Pearl, Braveman, & Abrams, 2001). Emch et al. (2017) suggested using local-level areal patterns such as census blocks, block groups or tracts would be more appropriate than using coarser scales such as the health service region (HSR) or county-level. Moreover, using the smallest available geographic unit for defining neighborhoods enhances the identification of local assets and gaps (CDC, 2015).

The characteristics to be included in defining a neighborhood in health studies varies with the research focus and the preferences of the researchers. Lebel et al. (2007) suggested that both geographical scale (as suggested by Emch et al., 2017) and inner characteristics, which are defined as an important element that is able to characterize a neighborhood such as structural, infrastructural, demographic, proximity, political, public services, environmental, social-interactive, class status, and sentimental characteristics, should be considered when defining neighborhood characteristics (Lebel, et al., 2007; Galster, 2001). In practice, choice of inner

characteristics, however, is based on the purpose and rationale for describing the neighborhood. In this study, social determinants of health (SDH) for HIV/AIDS are used to define neighborhood characteristics. The definition of social determinants of health is “the circumstances in which people are born, grow up, live, work and age, and the systems put in place to deal with illness” (WHO, 2016a). Five determinants of population health are generally recognized in the scientific literature – biology and genetics, individual behavior, social environment, physical environment, and health services (Tarlov, 1999). They consist of a wider set of forces such as economics, social policies, and politics and have been acknowledged as a critical component of the post-2015 sustainable development global agenda and of the push towards progressive achievement of universal health coverage (UHC) (WHO, 2016a). The World Health Organization (WHO) encourages members to address SDH in order to improve health outcomes and reduce health inequity as well as help to identify entry points for intervention (WHO, 2016a; CDC NCHHSTP, 2010). In the U.S., the Centers for Disease Control and Prevention’s (CDC) National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) has adopted WHO’s social determinants of health conceptual framework to better analyze and understand the drivers of health and health inequities of persons infected with HIV, viral hepatitis, sexually transmitted diseases (STDs), and tuberculosis (TB) as well as determine priorities and target to refocus intervention efforts (CDC NCHHSTP, 2010). While many public health efforts historically focused on individual behaviors, NCHHSTP points out that SDH typically refers to three categories including social environment (e.g., incomes, education), physical environment (e.g., place of residence, built environment), and health services (e.g., access to care, insurance status) (CDC NCHHSTP, 2010). Many studies have revealed that social determinants of health such as poor access to care, low education, and

economic status have a great influence on health disparities in HIV/AIDS (CDC, 2016; Goswami et al., 2016; Denning & DiNenno, 2015; Vaughan, Rosenberg, Shouse, & Sullivan, 2014; Reif et al., 2013). Yet, methods to define neighborhoods of risk for specific diseases such as HIV/AIDS are limited. To fill this research gap, this study develops and tests a methodology for delimiting areas with high risk of HIV/AIDS using SDH.

4.3 Study Areas

Texas, the second most populous and second largest state in area of the U.S., is located in the south-central part of the country and comprises of 254 counties with fairly similar size and shape. It ranked third among 50 states in the number of HIV diagnoses in 2013 (CDC NCHHSTP, 2015), and the lifetime risk of HIV diagnosis was 1 in 81, which was higher than the national risk (1 in 99) (CDC NCHHSTP, 2016).

Texas Department of State Health Services (DSHS) has divided the state into 11 public health regions for health services programming and administration including preventive, protective, regulatory, and preparedness functions, and all regions are held together by the DSHS Division for Regional and Local Health Services (RLHS) (DSSH, 2016a). Each county is assigned to 1 of the 11 public health regions to provide resources for health promotion and disease prevention and control (DSHS, 2016b). This study focuses on two public health regions including Public Health Regions (PHR) 3 and 4 (Figure 4.1 and Table 4.1).

PHR 3 comprises of 19 counties with a total of nearly 7 million population centered on the Dallas-Fort Worth Metroplex, the fourth largest metropolitan area in the U.S. after Chicago, Los Angeles, and New York City (U.S. Census Bureau, 2016); According to the 2013 National Center for Health Statistics (NCHS) Urban-Rural Classification Scheme for Counties (Ingram &

Franco, 2014), 13 out of the 19 counties in PHR 3 are considered as large metropolitan counties and part of 1 county is a small micro area; 5 of them are characterized as nonmetropolitan (Ingram & Franco, 2014).

In contrast, PHR 4 contains 23 mostly rural counties with approximately 1 million population in total. Only 5 of 23 counties are small metros; the rest are defined as nonmetropolitan (Ingram & Franco, 2014). Table 4.1 shows a summary of the characteristics of both regions.

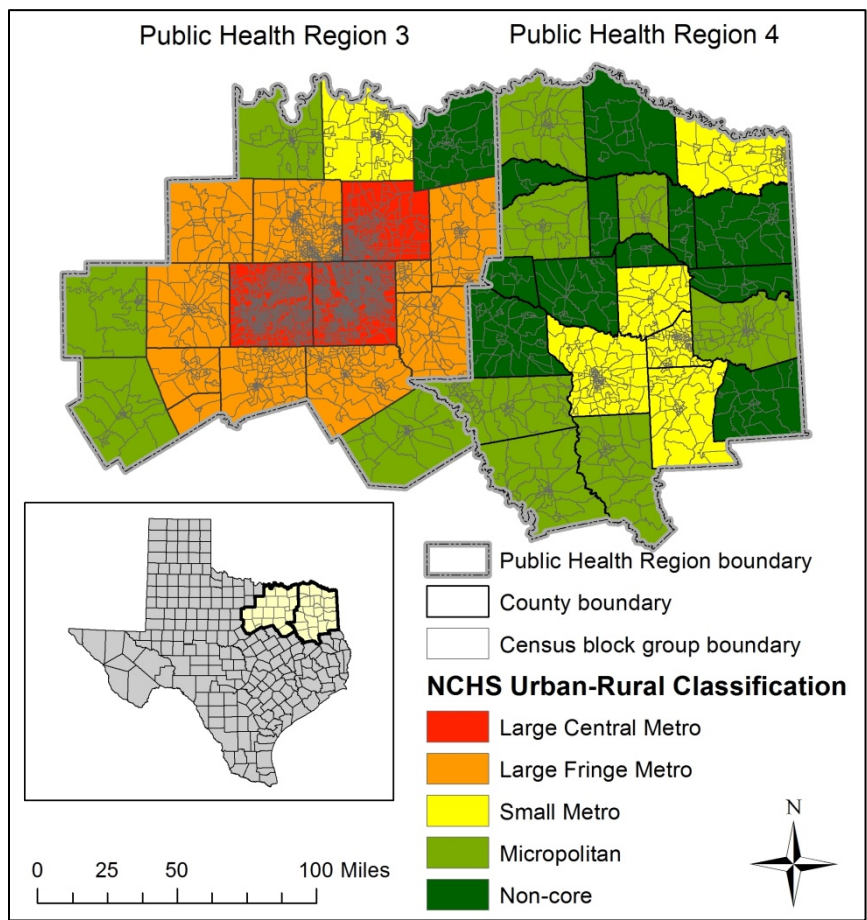


Figure 4.1 Study area

Table 4.1 A summary of study area characteristics

Characteristics	Public Health Region 3	Public Health Region 4
Area (square miles)	15,522.0	15,887.8
Total population	6,884,072	1,116,214
Total number of counties	19	23
- Large central or fringe metros	13	
- Micropolitan	4	7
- Small metros	1	5
- Non-metros	1	11
Total number of block groups (BGs)	4,403	812
- BG with no population	5	2
- Range of population	[34, 9615]*	[225, 6337]

* Note: There is only 1 block group that contains less than 100 population.

4.4 Data and Methods

This study used census-defined block groups as the basic unit for analysis. A census block group is the smallest geographical unit that the U.S. Census Bureau publishes sample data on demographic and socioeconomic indicators. Block Groups generally contain between 600 and 3,000 people and never cross state, county, or census tract boundaries (U.S. Census Bureau, 2012). The block group cartographic boundary files were obtained from 2013 Topologically Integrated Geographic Encoding and Referencing (TIGER/Line) files (U.S. Census Bureau, 2013).

Social determinants of health variables, selected based on CDC's NCHHSTP guidelines (2010), were obtained from 2013 ACS 5-year estimates dataset at the block group level. These guidelines provide three broad categories of the social determinants of health – social environment, physical environment, and health services (CDC NCHHSTP, 2010). For a complete list of SDH variables and their classes, see Table A.2 in Appendix A. To reduce the

large number of variables into a smaller number of identifiable dimensions, factor analysis, an explanatory analysis to maximize the homogeneity of variables (Vyas & Kumaranayake, 2006), was performed in SPSS. Theoretically, highly correlated variables, whether positive or negative, are basically influenced by the same factors and are combined into an identifiable dimension known as a factor. To consider whether factor analysis is an acceptable process, Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity were used to examine factorability of a set of variables, in which KMO should be greater than |0.45| and Bartlett's test should be significant ($p < 0.05$) (Kaiser & Rice, 1974). However, prior to performing factor analysis, all 66 variables were tested to determine if they met the criterion for the factor analysis (Yong & Pearce, 2013). Following standard practices, the variables that contained missing values greater than 40%, low variance, and/or extremely high correlation ($> |0.900|$) were dropped. For extremely highly correlated variables, one was included in the analysis and another was dropped since the extremely high correlation indicates both explain the same thing. As a result, a total of 26 variables were selected and entered into a Principal Component Factor Analysis to create composite variables describing demographic and socioeconomic characteristics (Table 4.2). Then, factor scores were computed for each block group using a least squares regression approach (DiStefano, Zhu, & Mindrila, 2009) and used in subsequent analysis – the cluster analysis.

Using the k -means clustering method, block groups were partitioned into k different clusters based on their socio-demographic characteristics. K -means, a distance-based method for partitional clustering, uses an allocation/re-allocation algorithm to optimally reassign objects to the nearest cluster centroid and is largely employed to classify areas (Riva, Apparicio, Gauvin, & Brodeur, 2008). The distances within cluster (intra-cluster distances) are minimized while the

distances between clusters (inter-cluster distances) are maximized. In other words, the block groups with similar characteristics are grouped together while those that are dissimilar are put in other groups.

For cluster analysis, input variables have to be somewhat different and need to be standardized (Milligan & Cooper, 1988). In this study, the factor scores obtained from the previous step were used as input variables. This ensures that there is no multicollinearity between input variables since homogeneous SDH variables were grouped within the same dimension factor.

Table 4.2 Social determinants of health variables used in the study obtained from 2013 ACS 5-year estimates at the block group level.

SDH variables	Census Table ID	Table Title
<i>Social environment</i>		
Race/ethnicity	B03002	Hispanic or Latino origin by race
Education attainment	B15003	Education attainment for the population 25years and older
Income/poverty	C17002	Ratio of income to poverty level in the past 12 months
	B19001	Household income in the past 12 months (in 2013 inflation-adjusted dollars)
	B25044	Tenure by vehicles available
Employment status	B23025	Employment status for the population 16 years and older
Occupation	C24010	Occupation for the civilian employed population 16 years and older
<i>Physical environment</i>		
Housing structure	B25024	Units in structure
Values of house	B25075	Value
Housing spaces	B25017	Rooms
<i>Health services</i>		
Language barrier (may indicate access to care)	B16002	Household language by household limited English speaking status
Health insurance status	B27010	Type of health insurance coverage by age

Prior to k -means analysis, the number of clusters, k , must be determined. The sum of squared errors (SSE), a common measure, was used to determine an appropriate k (equation 4.1) (Everitt et al., 2011). The *error* is the distance between each object (i.e., block group in the

study) and its nearest cluster. SSE is the summation of these squared errors. A small SSE indicates that the objects are close to the cluster centroid – thus, less variation within the cluster, indicating homogeneity.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} dist^2(m_i, x) \quad (4.1)$$

where k is a number of clusters; C_i is a cluster i when $i = 1, 2, \dots, k$; $x \in C_i$ is a data point (i.e., block group) in cluster C_i ; m_i is the centroid of cluster C_i .

SSE tends to decrease when k increases and approaches 0 when k is equal to the total number of objects in the dataset. To determine k , the dataset was first run through k -means clustering for a range of k values from 1 to 20 using factor scores as input variables in R. The sum of squared errors (SSE) for each k was then calculated and were plotted against the number of clusters, k . Using the elbow method, the optimal k was selected at the ‘elbow’ position in the graph, where the SSE decreases abruptly.

Subsequently, the value of k was used to perform k -mean clustering analysis in R. The k types of clusters resulting from the analysis were imported to ArcGIS 10.2 and mapped out to illustrate the delineation of homogeneous zones.

4.5 Results

4.5.1 Classifying Characteristics

Five composite factors were obtained from the factor analysis for both Public Health Regions (PHRs) 3 and 4, in which the total percentage of explained variance in each region is 71.5 and 61.8 respectively. Though their composite factors are slightly different between two regions, two of five composite factors in both regions are strongly influenced by race/ethnicity, specifically Black and Hispanic, as well as low socioeconomic variables such as low education

and income. However, the three remaining composite factors contributed by other socioeconomic variables do not provide a readily meaningful explanation (see Table B.4 in Appendix B). Moreover, the Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity test could not be computed due to non-positive definite (NPD) matrix, and the determinant of the correlation matrix is zero. The cause of zero determinant and NPD may be due to linear dependencies among the variables (Wothke, 1993; Joreskog & Yang, 1996), which could lead to unreliable results. One solution to this problem is to change the set of variables (Wothke, 1993). Therefore, instead of including all 26 variables at the same time, these variables were divided into two groups: (1) racial/ethnic groups and low socioeconomic variables, and; (2) socioeconomic variables. Each group, then, was used to perform the factor analysis. Note that the variables in each group were selected based on the initial results as discussed above.

The results show that all factor analysis performed in both PHRs 3 and 4 have KMOs greater than 0.7 and significant Bartlett's test ($p < 0.001$). This indicates that the variables included in each group – Group 1 racial/ethnic groups included and Group 2 racial/ethnic groups excluded – are correlated highly enough to provide a reasonable basis and are well predicted by each factor (Kaiser & Rice, 1974). In this study, the variables with loading at least $|0.45|$ (fair) are retained in the factor and indicate the factor(s) they belong to (Comrey & Lee, 1992; Tabachnick & Fidell, 2007). Then, each composite factor is named corresponding to those high positive loadings. In PHR 3, a total of six composite factors were extracted from both groups – three factors each (Table 4.3) while a total of five composite factors were extracted in PHR 4 – three factors from Group 1 and two factors from Group 2 (Table 4.4).

Table 4.3 Factor loadings of selected variables in Public Health Region 3

Variables	Race/ethnicity included				Race/ethnicity excluded			
	Communities	Low Education –Hispanic	Low SES	Extreme Poverty–Black	Communities	High SES	Upper Middle SES	Lower Middle SES
<i>Percent of variance</i>	-	30.5	24.0	17.1	-	33.5	22.8	16.7
<i>Eigenvalues</i>	-	8.367	1.718	1.379	-	4.945	2.056	1.039
<i>Race/ethnicity</i>								
White population	0.736	-0.608		-0.571				
Black population	0.712			0.809				
Hispanic population	0.856	0.888						
<i>Education attainment</i>								
Less than high school	0.868	0.822	0.377					
High school degrees	0.810		0.889					
Some college and associate degrees					0.520			0.663
Bachelor degrees and higher	0.887	-0.396	-0.829		0.835	0.792	0.453	
<i>Income to poverty ratio</i>								
< 0.50	0.493			0.647				
Between 0.50 and 0.99	0.508	0.497		0.417				
≥ 2.00	0.846	-0.591	-0.467	-0.528				
<i>Household incomes</i>								
\$50,000–99,999					0.569			0.743
\$100,000-149,999					0.655		0.706	
\$150,000-199,999					0.619	0.527	0.570	
≥ \$200,000					0.862	0.891		
<i>Value of house</i>								
< \$100,000	0.705	0.401	0.674					
\$100,000-149,999					0.709			0.788
\$150,000-299,999					0.855		0.924	
≥ \$300,000					0.853	0.874		
Housing ≥ 8 rooms					0.743	0.723	0.458	
<i>Occupations</i>								
Construction	0.588	0.621	0.441					
Management	0.858	-0.519	-0.732		0.818	0.732	0.520	
Production	0.516	0.317	0.626					
No vehicle own	0.571			0.730				
No health insurance	0.750	0.720	0.423					
Limited speaking English	0.759	0.862						

Note: All variables are in the percentage unit; **Boldfaced** values indicated variables that contribute in the composite factor (factor loadings ≥ |0.45|); Factor loadings < |0.3| are suppressed.

Table 4.4 Factor loadings of selected variables in Public Health Region 4

Variables	Race/ethnicity included				Race/ethnicity excluded		
	Communities	Poverty-Black	Low Education-Hispanic	Low SES	Communities	High SES	Middle SES
<i>Percent of variance</i>		28.3	22.7	16.6		35.7	23.4
<i>Eigenvalues</i>		5.484	2.133	1.846		3.829	1.494
<i>Race/ethnicity</i>							
White population	0.815	-0.733	-0.498				
Black population	0.733	0.838					
Hispanic population	0.832		0.900				
<i>Education attainment</i>							
Less than high school	0.713	0.350	0.730				
High school degrees	0.684	0.321		0.710			
Some college and associate degrees					0.450		0.670
Bachelor degrees and higher	0.789	-0.467		-0.707	0.722	0.765	0.370
<i>Income to poverty ratio</i>							
Less than 1.00	0.613	0.714					
≥ 2.00	0.771	-0.764	-0.382				
<i>Household incomes</i>							
\$75,000-149,999					0.554	0.355	0.654
≥ \$200,000					0.675	0.818	
<i>Value of house</i>							
< \$100,000	0.669	0.637		0.449			
\$100,000-149,999					0.510		0.714
\$150,000-299,999					0.529	0.536	0.492
≥ \$300,000					0.645	0.781	
Housing ≥ 8 rooms					0.599	0.733	
Living in mobile home	0.568			0.692			
<i>Occupations</i>							
Construction	0.507			0.628			
Management					0.640	0.632	0.491
No vehicle own	0.535	0.724					
No health insurance	0.525	0.318	0.602				
Limited speaking English	0.708		0.839				

Note: All variables are in the percentage unit; **Boldfaced** values indicated variables that contribute in the composite factor (factor loadings ≥ |0.45|); Factor loadings < |0.3| are suppressed.

The results from factor analysis (Table 4.3 and 4.4) show that there is a strong association between Black populations and poverty in both regions – ‘Extreme Poverty-Black’ and ‘Poverty-Black’ for PHRs 3 and 4 respectively. According to variable loadings in these factors, it indicates that Blacks living in urbanized areas (PHR 3) tend to have a greater depth of poverty than those

living in non-urbanized areas (PHR 4). ‘Low Education-Hispanic’ is characterized in both regions and indicates that Hispanic populations tend to have education attainment less than high school and no health insurance. ‘Low SES’ (SES stands for socioeconomic status) is also categorized in both regions. The variables loading in this factor indicates that having low education (high school degrees) influences their jobs and living standard regardless of living in urbanized or non-urbanized areas. ‘Middle SES’ is a socio-economic characteristic identified in both regions, but the construct of this characteristic is quite complex in PHR 3. While PHR 4 can be simply classified as ‘Middle SES’, Public Health Region 3 is sub-classified as ‘Lower Middle SES’ and ‘Upper Middle SES’. While other socioeconomic characteristics are associated with different variables, the composite variables of ‘High SES’ are the same in both regions.

4.5.2 Defining Spatial Units

Consequently, *k*-mean clustering analysis was performed for each region to determine homogeneous zones using factor scores of those characteristics. In this study, the result suggests that the optimal number of clusters, *k*, is 5 for both PHRs 3 and 4 which provide 60 and 61 percent of total variance explained respectively. The results from ANOVA confirm that all six characteristics in PHR 3 and five characteristics in PHR 4 play significant roles to perform clusters ($p < 0.01$, $d.f. = 4, 4144$ for PHR 3 and $p < 0.01$, $d.f. = 4, 800$ for PHR 4). For PHR 3, Cluster 1, 2, 4, and 5 are clearly dominated by ‘Upper Middle SES’, ‘Extreme Poverty–Black’, ‘High SES’, and ‘Low Education–Hispanic’ respectively (indicated by the highest cluster means in Table 4.5). Therefore, these clusters are used to name the respective characteristics. Unlike other clusters, Cluster 3 is likely to be dominated by two characteristics – ‘Lower Middle SES’ and ‘Low SES’ since the cluster means of these two components in this cluster are close to each

other (Table 4.5). Therefore, Cluster 3 is named ‘Low to Lower Middle SES’. For PHR 4, Clusters 1 to 5 are dominated by ‘Low SES’, ‘Poverty-Black’, ‘High SES’, ‘Middle SES’, and ‘Low Education-Hispanic’ respectively (Table 4.5).

Table 4.5 The cluster means for each of the significant factors of Public Health Regions 3 and 4

	Cluster				
	1	2	3	4	5
<i>Public Health Region 3</i>					
High SES	.09968	-.67021	-.15113	2.43336	-.73405
Upper middle SES	1.44908	-.57530	-.26326	-.42219	-.71676
Lower middle SES	-.10361	-.64226	.79461	-.76131	-.74833
Low SES	-.78614	.47174	.55392	-1.46901	.35375
Low education- Hispanic	-.42244	-.29237	-.37631	-.41011	1.69820
Extreme poverty - Black	-.26689	2.21242	-.35275	-.30626	-.02628
<i>Public Health Region 4</i>					
High SES	-0.09878	-0.59530	3.08343	0.05508	-0.54469
Middle SES	-0.15383	-0.86838	-0.24834	1.25751	-0.86858
Low SES	0.71673	-0.26389	-1.82365	-0.81876	-0.01914
Low education-Hispanic	-0.23543	-0.33056	-0.37247	-0.28968	2.25152
Poverty-Black	-0.27499	1.93282	-1.08914	-0.45018	0.30707

Figure 4.2A illustrates cluster zones in PHR 3. As can be seen, there are a lot of zone variations around the Dallas-Fort Worth metroplex. In the metroplex, both ‘Extreme Poverty–Black’ and ‘Low Education–Hispanic’ characteristics mostly cluster and occur in the vicinity of Dallas and Fort Worth areas. These two characteristics seem to be adjacent to each other and account for 41.6 and 24.4 percent of the total population in Dallas and Tarrant Counties respectively. The results also show that the next adjacent zones to these poverty and low education areas mostly are ‘Low to Lower Middle SES’, ‘Upper Middle SES’, and ‘High SES’ zones respectively. In contrast, Collin County (northeast of Dallas County) and Denton County (northwest of Dallas County) are dominated by ‘Upper Middle SES’ and ‘High SES’ while other counties are dominated by ‘Low to Lower Middle SES’ characteristic – accounted for 37.4% of the total population.

In PHR 4, zone variations mostly occur in small-metro counties such as Smith and Gregg Counties (Figure 4.2B). In these counties, ‘Poverty-Black’ and ‘Low Education-Hispanic’ characteristics are concentrated in the vicinity of cities while ‘Middle SES’ and ‘High SES’ seem to occur in the suburbs. In contrast, the geographic patterns of neighborhood characteristics in non-metropolitan counties vary. In fact, ‘Low SES’, the dominant characteristics in these counties, accounted for 47.0% of the total population.

In summary, the disease specific neighborhood maps show that ‘Poverty-Black’ tends to cluster in the vicinity of cities regardless of urbanized or non-urbanized areas. Yet, there are some spatial differences between urban and non-urban areas. In large metropolitan areas such as Dallas and Fort Worth, the extent of ‘Poverty-Black’ and ‘Low Education-Hispanic’ neighborhoods are larger than in smaller metropolitan areas, and they are adjacent to each other resulting in large areas of social disadvantage. In contrast, ‘Low Education-Hispanic’ neighborhoods in non-urbanized areas seem to be scattered and are not necessary in the vicinity of cities.

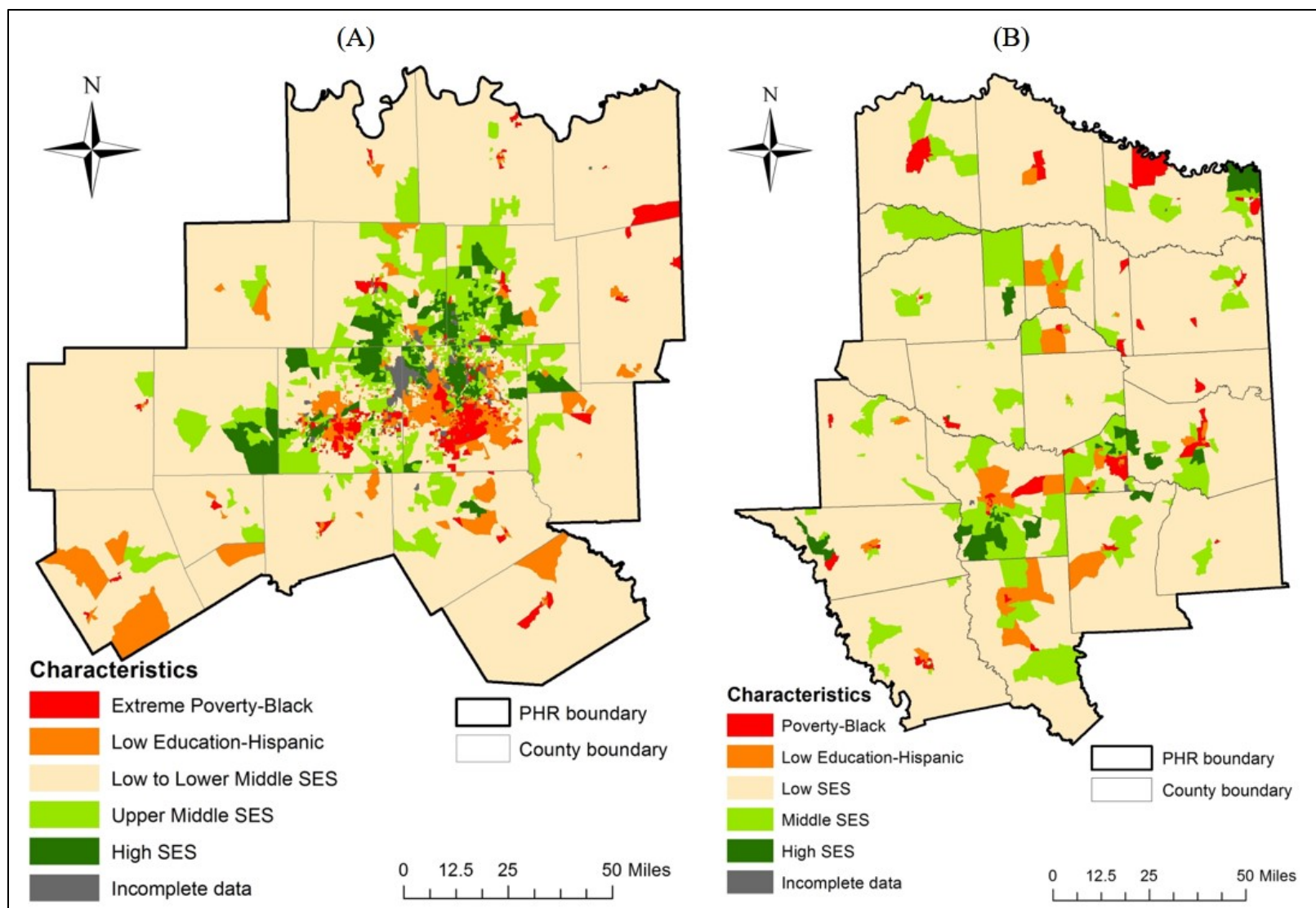


Figure 4.2 Disease specific neighborhood maps of: (A) Public Health Region 3; (B) Public Health Region 4.

4.5.3 The Use of HIV and Disease Specific Neighborhood Maps

The HIV/AIDS data used in this analysis came from a de-identified database obtained from Texas Department of State Health Services that includes all persons newly diagnosed with HIV/AIDS in these two regions between 1 January 2009 and 31 December 2013. Cases were aggregated at the zip code level. Then, to ensure spatially continuous representations of HIV/AIDS in PHRs 3 and 4, HIV/AIDS map was produced using the spatially adaptive filters method (Tiwari & Rushton, 2005) in Web-based Disease Mapping and Analysis Program (WebDMAP) and ArcGIS 10.2. To facilitate visualization, the HIV/AIDS map was overlaid on the disease specific neighborhood maps.

In PHR 3, the areas with higher HIV/AIDS rates (greater than state rate) mostly occur in the central region specifically in Dallas-Fort Worth areas (Figure 4.3A). Considering the moderate high and greater HIV/AIDS rates (> 59.73 per 100,000 population) (Figure 4.3B – striped line), ‘Extreme Poverty-Black’ neighborhoods tend to have a greater risk than other neighborhoods in the center of Dallas while ‘High SES’ seems to have a greater risk in the north of Dallas. In those highest incidence areas, poverty and Black population in the heart of Dallas and high SES in north Dallas should be prioritized for targeting intervention. While the higher HIV/AIDS rates in PHR 3 are mostly concentrated in the Dallas-Fort Worth metroplex – the most urbanized areas, those in PHR 4 are clustered in the cities and have higher concentration specifically in the south and east of the region (Figure 4.4A), where those counties are characterized as non-urban areas (Ingram & Franco, 2014). The areas with high incidence of HIV/AIDS in PHR 4 mostly occur in ‘Poverty-Black’ and ‘Low Education-Hispanic’ neighborhoods which is similar to those in PHR 3. Figure 4.4B (solid line) indicates that ‘Low Education-Hispanic’ in the south and ‘Poverty-Black’ in the north and east of the region should

be prioritized for targeting interventions. These examples illustrate that combining HIV/AIDS and neighborhood characteristics maps can identify not only *where* the higher rates occur but also *who* or *which populations* have a greater risk and should be prioritized for control and treatment.

4.6 Discussion

4.6.1 Poverty and Black Populations

The focus of this paper was to identify high risk areas for HIV/AIDS. The results found a strong association between Black populations and poverty in urban areas of both public health regions. It is well-known that poverty is an important factor that increases the chance of poor health (Conway, 2016; WHO, 2016b) and the spread of diseases such as HIV/AIDS (Denning & DiNenno, 2015; Parkhurst, 2010; International Labor Office, 2005). This is not confined to only poor people. People living in poor areas tend to have poorer health and are at greater risk than those living in better places (Emch et al., 2017; Oppong & Harold, 2009; Chaix, et al., 2009; Cockings & Martin, 2005; Cawthorne, 2010). Moreover, in the U.S., the Black population has the highest rate of HIV/AIDS among all racial/ethnic groups (CDC, 2016). The combination of these two risk markers compounds the problem. Thus, spatial units that are dominated by this characteristic within the cities of Dallas, Fort Worth, and Tyler should be targeted for intervention.

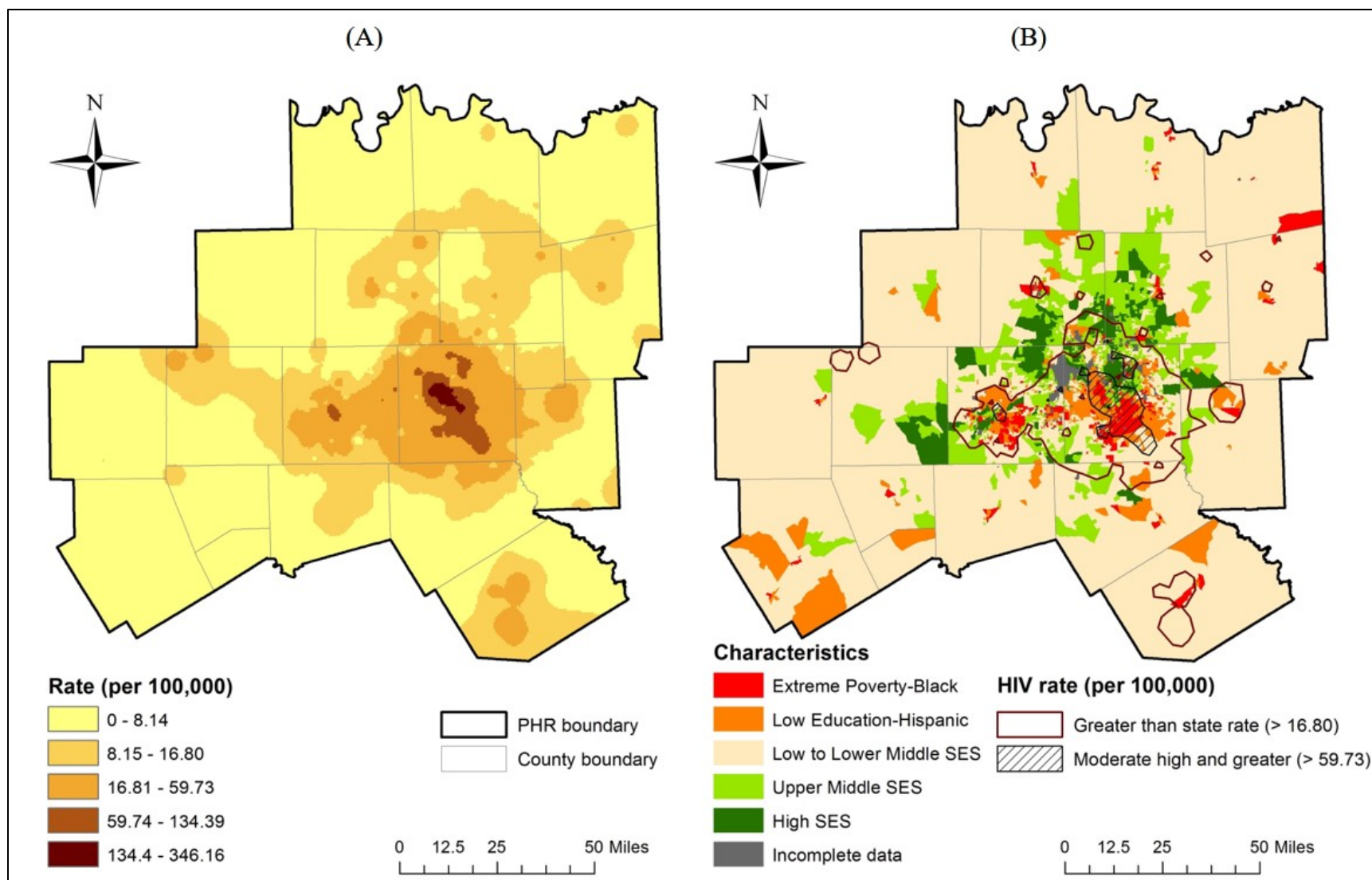


Figure 4.3 The use of HIV and neighborhood characteristics maps of Public Health Region 3: (A) The spatial distribution of HIV incidence between 2009 and 2011; (B) A neighborhood map with a transparent overlay of high HIV incidence rates.

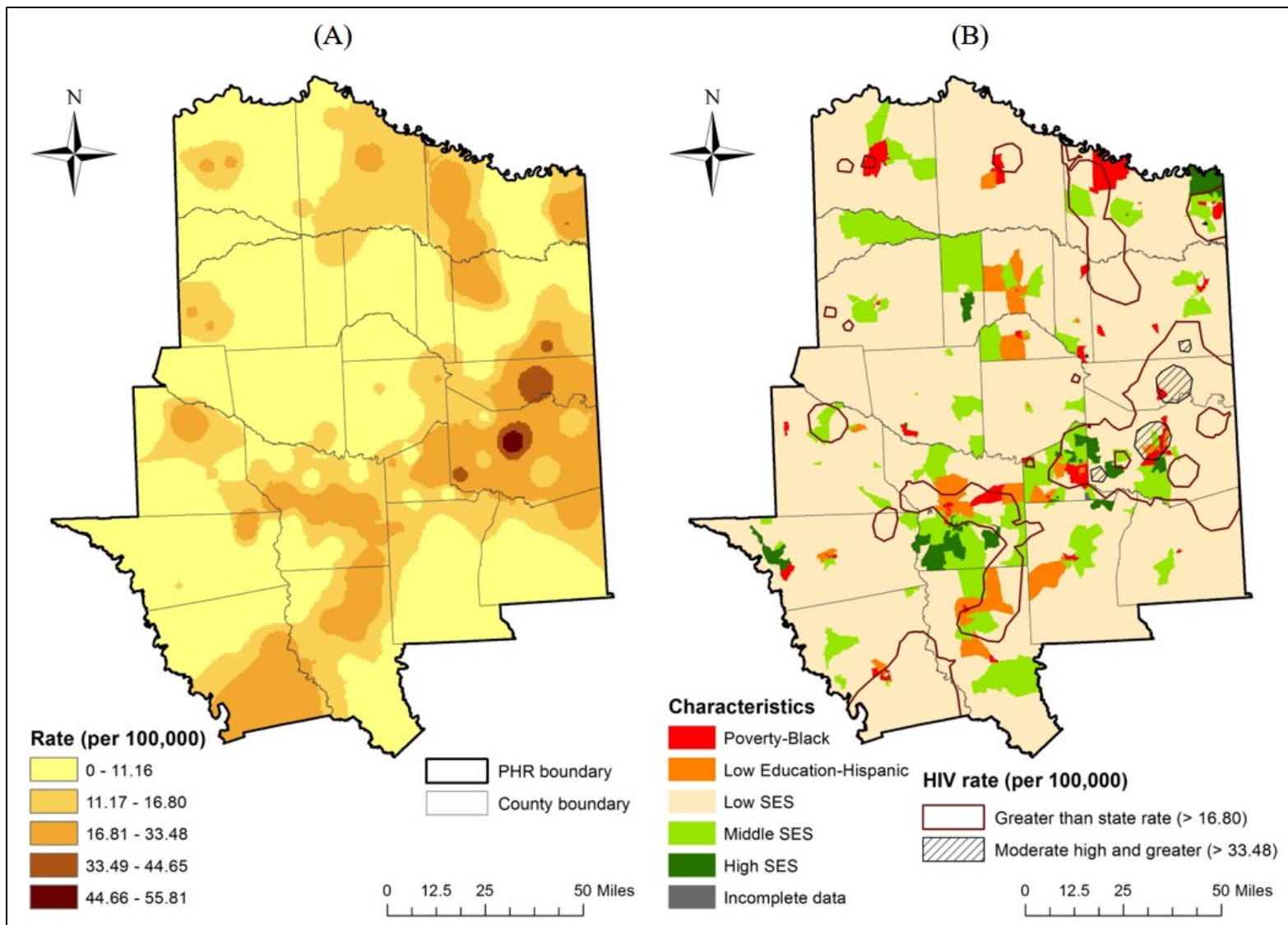


Figure 4.4 The use of HIV and neighborhood characteristics maps of Public Health Region 4: (A) The spatial distribution of HIV incidence between 2009 and 2011; (B) A neighborhood map with a transparent overlay of high HIV incidence rates.

4.6.2 Low Education and Hispanic Population

In both regions, the results indicate that Hispanic population has a positive association with social disadvantage variables including low education attainment and high rate of uninsured – one of the six key health and health care indicators (Kaiser Family Foundation, 2009; U.S. Census Bureau, 2015). The results imply that in addition to potentially being uninsured, Hispanics appear to have SES that limits their access to health care. Even though they may not be poor, lack of health insurance directly relates to limited access to care resulting in late disease diagnoses as well as delaying appropriate care. This is important. Specifically, for HIV infection, late diagnosis not only affects the effectiveness of treatment but also is an important factor in HIV spread. Approximately 70,000 people living with HIV were uninsured in 2009 (CDC NCHHSTP, 2016; Kate, et al., 2014), thus limiting their access to HIV therapy and resulting in overall poorer health outcomes. In fact, HIV symptoms in the early stages are flu-like symptoms and may take several years to present in some individuals. Lacking routine care, the uninsured are more likely to be diagnosed with HIV during a health emergency, and thus may miss the benefits of early treatment. In addition to being uninsured, Hispanic is also a risk marker for late HIV diagnosis. In Texas, approximately two in five Hispanics with HIV are diagnosed late (CDC NCHHSTP, 2015; DSHS, 2013). Consequently, spatial units dominated by this characteristic - ‘Low Education-Hispanic’ should be prioritized for testing and other appropriate interventions.

4.6.3 The Use of Health and Disease Specific Neighborhood Maps

Though disease maps can help in identifying areas with the highest disease burdens and pinpoint locations that should be prioritized for targeted intervention, they do not provide the information on *who* or *which populations* have the greatest risk. Without that information,

generating effective control plans becomes extremely difficult. Disease specific neighborhood maps help in identifying and delineating the high-risk neighborhoods that possibly have high rates of disease (HIV in this case) – spatial units dominated by ‘Poverty-Black’ and ‘Low Education Hispanic’. Thus, these spatial units potentially have higher risk than others, although they may or may not have the highest rates. In the situation that health data are not available, disease specific neighborhood maps can be used to identify potential areas and populations at risk to prioritize for prevention and control. If health data is available, such disease specific neighborhood maps can become even more powerful and meaningful. In this study, although the spatial mismatch precludes a statistical comparison of HIV disease burdens and the disease specific neighborhood maps, the simple comparison (overlay) between HIV/AIDS and disease specific neighborhood maps shows strong evidence of the utility of these maps. Combining these two maps not only helps in determining *where* but also *who* or *which populations* should be prioritized for intervention and control.

4.7 Limitations

Socio-demographic characteristics in this study are constructed from selected social determinants of health for HIV/AIDS. Other diseases could possibly provide a different set of characteristics. Due to the spatial mismatch between SDH and HIV/AIDS data, statistical cross-validation between these two different data is not feasible. In fact, socio-demographic characteristics and neighborhood zones are constructed using the data at the census block group level, and HIV data used in this study are at the zip code level.

4.8 Conclusion

Disease maps are a powerful tool for disease surveillance systems and very useful for monitoring and preventing disease outbreaks. While they can be used to pinpoint the areas that have high disease burdens and show where the population at risk is, they cannot identify *who* or *which populations* in those areas is most at risk. Using SDH variables for HIV/AIDS, this study has defined neighborhood units and identified who in those areas has the most need for targeted intervention and where they live.

Disease specific neighborhood maps not only illustrate the spatial distribution of socio-demographic characteristics but also identify the characteristics of the population at risk. Moreover, combining and overlaying these disease specific neighborhood maps with disease maps can help in geographically targeting and tailoring policy to both places and populations most in need. Since this study focuses on HIV surveillance, these disease specific neighborhood maps may not be appropriate for other diseases that may have different social determinants of health. Nevertheless, the straightforward statistical and spatial methods illustrated in this study can be simply applied to construct other specific disease neighborhood maps.

4.9 References

- Browning, C. R., & Cagney, K. A. (2003). Moving beyond poverty: neighborhood structure, social processes, and health. *Journal of Health and Social Behavior*, 44(4), 552-571.
- Cawthorne, A. (2010, July 21). *Poverty is driving an HIV epidemic*. Retrieved from <https://www.americanprogress.org/issues/poverty/news/2010/07/21/8101/poverty-is-driving-an-hiv-epidemic/>
- Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2010). *Establishing a holistic framework to reduce inequities in HIV, viral hepatitis, STDs, and tuberculosis in the United States*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved from <https://www.cdc.gov/socialdeterminants/docs/SDH-White-Paper-2010.pdf>

- Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2015, December 22). *Texas - 2015 State Health Profile*. Retrieved December 30, 2015, from https://www.cdc.gov/nchhstp/stateprofiles/pdf/texas_profile.pdf
- Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2016). *NCHHSTP: annual report 2015*. Retrieved April 10, 2017, from <https://www.cdc.gov/nchhstp/publications/docs/nchhstp-annual-report-2015.pdf>
- Centers for Disease Control and Prevention. (2015, November 9). *State, Tribal, Local & Territorial Public Health Professionals Gateway: Data & benchmarks*. Retrieved December 30, 2016, from <https://www.cdc.gov/stltpublichealth/cha/data.html>
- Centers for Disease Control and Prevention. (2016, August). *CDC fact sheet: Today's HIV/AIDS epidemic*. Retrieved December 30, 2016, from <https://www.cdc.gov/nchhstp/newsroom/docs/factsheets/todaysepidemic-508.pdf>
- Chaix, B., Merlo, J., Evans, D., Leal, C., & Havard, S. (2009). Neighbourhoods in eco-epidemiologic research: Delimiting personal exposure areas. A response to Riva, Gauvin, Apparicio and Brodeur. *Social Science & Medicine*, 69(9), 1306-1310. doi:10.1016/j.socscimed.2009.07.018
- Cockings, S., & Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine*, 60(12), 2729-2742.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Conway, C. (2016, January 6). *Poor health: When poverty become disease*. Retrieved September 29, 2016, from <https://www.ucsf.edu/news/2016/01/401251/poor-health>
- Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping residents' perceptions of neighborhood boundaries: a methodological note. *American Journal of Community Psychology*, 29(2), 371-383.
- Denning, P., & DiNenno, E. (2015, June 23). *Communities in crisis: Is there a generalized HIV epidemic in impoverished urban areas of the United States?* Retrieved September 29, 2016, from <https://www.cdc.gov/hiv/group/poverty.html>
- Diez Roux, A. V. (2001). Investigating neighborhood and area effects on health. *American Journal of Public Health*, 91(11), 1783-1789.
- Diez Roux, A. V., Mujahid, M. S., Hirsch, J. A., Moore, K., & Moore, L. V. (2016). The impact of neighborhoods on CV risk. *Global Heart*, 11(3), 353-363. doi:10.1016/j.gheart.2016.08.002

- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20). Retrieved from <http://pareonline.net/getvn.asp?v=14&n=20>
- Emch, M., Root, E. D., & Carrel, M. (2017). *Health and medical geography* (4th ed.). New York, NY: The Guilford Press.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). London, UK: John Wiley & Sons, Ltd.
- Galster, G. (2001). On the nature of neighbourhood. *Urban Studies*, 38(12), 2111-2124.
- Goswami, N. D., Schmitz, M. M., Sanchez, T., Dasgupta, S., Sullivan, P. D., Cooper, H., . . . Waller, L. (2016). Understanding local spatial variation along the care continuum: the potential impact of transportation vulnerability on HIV linkage to care and viral suppression in high-poverty areas, Atlanta, Georgia. *Journal of Acquired Immune Deficiency Syndromes*, 72(1), 65-72.
- Ingram, D. D., & Franco, S. J. (2014). 2013 NCHS urban-rural classification scheme for counties. *National Center for Health Statistics. Vital Health Stat*, 2(166). Retrieved September 29, 2016, from https://www.cdc.gov/nchs/data/series/sr_02/sr02_166.pdf
- International Labour Office. (2005). *HIV/AIDS and poverty: The critical connection*. Retrieved from http://www.ilo.org/wcmsp5/groups/public/@ed_protect/@protrav/@ilo_aids/documents/publication/wcms_120468.pdf
- Joreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 57-88). Mahwah, NJ: Lawrence Erlbaum.
- Kaiser Family Foundation. (2009, April 1). *Key of health and health care indicators*. Retrieved from <http://kff.org/disparities-policy/fact-sheet/key-health-and-health-care-indicators-by/>
- Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, 34(1), 111-117. doi:10.1177/001316447403400115
- Kates, J., Garfield, R., Young, K., Quinn, K., Frazier, E., & Skarbinski, J. (2014, January 7). *Assessing the impact of the Affordable Care Act on health insurance coverage of people with HIV*. Retrieved from <http://kff.org/report-section/assessing-the-impact-of-the-affordable-care-act-on-health-insurance-coverage-of-people-with-hiv-issue-brief/>
- Lebel, A., Pampalon, R., & Villeneuve, P. Y. (2007). A multi-perspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region. *International Journal of Health Geographics*, 6, 27. doi:10.1186/1476-072X-6-27

- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5(2), 181-204. doi:10.1007/BF01897163
- Oppong, J. R., & Harold, A. (2009). Disease, Ecology, and Environment. In T. Brown, S. McLafferty, G. Moon, T. Brown, S. McLafferty, & G. Moon (Eds.), *A companion to health and medical geography* (pp. 81-95). Oxford, UK: Wiley-Blackwell.
- Parkhurst, J. O. (2010). Understanding the correlations between wealth, poverty and human immunodeficiency virus infection in African countries. *Bulletin of the World Health Organization*, 88, 519-526. doi:10.2471/BLT.09.070185
- Pearl, M., Braveman, P., & Abrams, B. (2001). The relationship of neighborhood socioeconomic characteristics to birthweight among five ethnic groups in California. *American Journal of Public Health*, 91, 1808-1824.
- Rauh, V. A., Andrews, H. F., & Garfinkel, R. (2001). The contribution of maternal age to racial disparities in birthweight: a multi-level perspective. *American Journal of Public Health*, 91, 1815-1824.
- Reif, S. S., Whetten, K., Wilson, E. R., McAllaster, C., Pence, B. W., Legrand, S., & Gong, W. (2013). HIV/AIDS in the Southern USA: a disproportionate epidemic. *AIDS Care*, 26(3), 351-359. doi:10.1080/09540121.2013.824535
- Riva, M., Apparicio, P., Gauvin, L., & Brodeur, J.-M. (2008). Establishing the soundness of administrative spatial units for operationalising the active living potential of residential environments: An exemplar for designing optimal zones. *International Journal of Health Geographics*, 7(43). doi:10.1186/1476-072X-7-43
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.
- Tarlov, A. R. (1999). Public policy frameworks for improving population health. *Annals of the New York Academy of Sciences*, 896(1), 281-293.
- Texas Department of State Health Services. (2013, October). *Hispanics in Texas: Late HIV diagnosis and out of care*. Retrieved from <https://www.dshs.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85899418>
10
- Texas Department of State Health Services. (2016a). *Center for Health Statistics Texas County Numbers and Public Health Regions*. Retrieved September 29, 2016, from https://www.dshs.texas.gov/chs/info/info_txco.shtm
- Texas Department of State Health Services. (2016b). *Division for Regional and Local Health Services*. Retrieved September 29, 2016, from <http://www.dshs.texas.gov/rls/default.shtm>

- Tiwari, C., & Rushton, G. (2005). Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In P. Fisher, *Developments in spatial data handling* (pp. 665-676). Berlin: Springer-Verlag.
- U.S. Census Bureau. (2012). *Geographic terms and concepts - Block Groups*. Retrieved September 29, 2016, from Geography: https://www.census.gov/geo/reference/gtc/gtc_bg.html
- U.S. Census Bureau. (2013). *2013 TIGER/Line shapefiles* [Dataset]. Retrieved from <https://www.census.gov/geo/maps-data/data/tiger-line.html>
- U.S. Census Bureau. (2015, September 16). *Income, poverty and health insurance coverage in the United State: 2014*. Retrieved from <http://www.census.gov/newsroom/press-releases/2015/cb15-157.html>
- U.S. Census Bureau. (2016). *2016 Population estimates - annual estimates of the resident population: April 1, 2010 to July 1, 2016 - United States -- Metropolitan Statistical Area; and for Puerto Rico*. Retrieved from American FactFinder: <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>
- Vaughan, A. S., Rosenberg, E., Shouse, R. L., & Sullivan, P. S. (2014). Connecting race and place: a county-level analysis of White, Black, and Hispanic HIV prevalence, poverty, and level of urbanization. *American Journal of Public Health, 104*(7), e77-e84. doi:10.2105/AJPH.2014.301997
- Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning, 21*(6), 459-468.
- World Health Organization. (2016a). *Social determinants of health*. Retrieved December 30, 2016, from http://www.who.int/social_determinants/en/
- World Health Organization. (2016b). *Health and development: Poverty and health*. Retrieved September 29, 2016, from <http://www.who.int/hdp/poverty/en/>
- Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Newbury Park, NJ: Sage Publications.
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94. Retrieved from <http://tqmp.org/Content/vol09-2/p079/p079.pdf>

CHAPTER 5

SUMMARY AND CONCLUSIONS

This chapter summarizes the major findings of this thesis and contributions with regard to kernel density estimation method in disease mapping and defining spatial units for health research. It also provides a discussion of the implications of the study for practice.

5.1 Introduction

Disease mapping has become a powerful tool for understanding the distribution of disease across space. It offers an alternative approach for public health to specify and address hypotheses of disease etiology and causation as well as plan and target appropriate interventions. Yet, some limitations remain to be addressed in order to improve their effectiveness for disease mapping as well as control efforts.

To map disease, smoothed disease rate methods are recommended since they permit comparison across different population sizes, unlike the number of cases, and also address the small numbers problem – a well-known issue in disease mapping (Cromley & McLafferty, 2012). Though many smoothing methods are available, two of them – kernel density estimation (KDE) and spatial empirical Bayes – are most commonly used (Rushton & Lolonis, 1996; Clayton & Kaldor, 1987; Besag, York, & Mollie, 1991; Berke, 2004; Best, Richardson, & Thomson, 2005; Tiwari & Rushton, 2005; Oppong, Tiwari, Ruckthongsook, Huddleston, & Arbona, 2012; Moraga & Lawson, 2012). For smoothed disease rates, these two methods require a smoothing parameter called bandwidth/threshold and spatial weight in KDE and spatial empirical Bayes methods respectively. The selection of smoothing parameters is crucial because using different smoothing parameters may affect the degree of smoothing that occurs on the map.

While the selection of smoothing parameters in spatial empirical Bayes is overlooked, that of KDE has been identified repeatedly in many studies (Cromley & McLafferty, 2012; Bithell, 2000; Carlos, Shi, Sargent, Tanski, & Berke, 2010; Beyer, Tiwari, & Rushton, 2012; Chi, Wang, Li, Zheng, & Liao, 2007; Shi, 2010; Talbot, Kulldorff, Forand, & Haley, 2000; Cai, 2007; Rushton & Lolonis, 1996; Tiwari & Rushton, 2005; Silverman, 1986; Wand & Jones, 1995). In fact, several approaches to select smoothing parameters for KDE have been proposed (Silverman, 1986; Wand & Jones, 1995), but the application of those methods in disease mapping is limited. Thus, for both methods, current approaches use knowledge-based judgement made by experts who understand the disease being mapped or by the map-makers themselves (Beyer et al., 2012; Chi et al., 2007; Shi, 2010; Talbot et al., 2000; Cai, 2007; Rushton & Lolonis, 1996; Tiwari & Rushton, 2005). The questions of what is an appropriate smoothing parameter and how to select one remain unanswered. Moreover, since both KDE and spatial empirical Bayes are commonly used to map smoothed rates, the following questions have been raised: “Do different methods provide different results?” and “Which method should be used to map disease?” These questions are important, and there is a need to better understand their merits in disease mapping and clarify the methods with regard to their use in public health and decision-making.

In addition to improving disease mapping methods, this study also explores an alternative approach to enhance the effectiveness of disease prevention and control efforts as well as improve approaches for spatial targeting of intervention. Though disease maps allow public health practitioners to pinpoint *where* intervention is most needed by illustrating the geographical pattern of disease, they alone cannot identify *who* or *which populations* are most at risk and need the interventions the most. Moreover, in areas with limited and poor-quality health data, such as

many developing countries, producing reliable representations of disease burdens is problematic. Thus, public health tasks of planning and targeting appropriate intervention (McLafferty, 2015; Cromley & McLafferty, 2012) become extremely difficult. To fill this research gap, this study develops and tests a methodology for defining spatial units with high risk for HIV/AIDS in Texas and identifying the characteristics of the people groups at highest risk within those areas.

The overall goals of this study are to provide better understanding of the importance of disease mapping methods and enhance disease surveillance systems. The primary objectives of the research are as follows:

1. To improve the KDE method in the disease mapping context by illustrating how to determine a desirable smoothing parameter.
2. To examine and evaluate the relative efficacy of KDE and Bayesian approaches in disease mapping and determine whether they provide different results.
3. To provide an alternative approach for defining neighborhood of risk and identifying at risk populations in those spatial units.

5.2 Summary of Findings

5.2.1 Objective 1: To Improve Kernel Density Estimation Method in Disease Mapping Context

Using a simulated dataset generated based on age-specific heart disease death rates among males aged 35 years and older, the study examines the applicability of bandwidth selection methods which calculate a threshold value based on the population data. For each age group, this study provides comparisons between maps produced using these methods and using arbitrary threshold choices and assesses the relative performance of these methods in terms of resolution and reliability (Chapter 2).

Our findings show that each bandwidth selector provides a different value. Threshold values calculated from plug-in and smoothed cross validation selectors, which are less than 100

in all age groups, may not be an appropriate choice to use in a large study area, e.g., the State of Texas (Washington State Department of Health, 2012). To consider which one is a desirable choice, the estimated state rates computed from other thresholds (including arbitrary choices) are observed for relative performance by comparing them to the actual state rates (i.e., simulated baseline rate in this study). In all age-specific groups, all thresholds accurately give the estimated state rates but provide different geographic detail. As threshold values increase, the variation in rates between geographic units decreases, i.e., less geographic detail. That is there is a trade-off between statistical stability of estimates and geographic precision. In disease mapping, we prefer the maps that provide more geographic detail but also need to compromise between precision and spatial variation. Since their precisions are similar, the remaining key factor for selecting a desirable threshold is geographic variation. When geographic variation is the highest priority, the results suggest that recommendations made by the normal scale selector, an automated bandwidth selector algorithm, generally provide significantly better performance when compared to other choices because it provides a consistent way to estimate the appropriate threshold value. In effect, the appropriate threshold value depends on the data distribution. Unguided choice of thresholds could produce misleading conclusions.

5.2.2 Objective 2: To Evaluate the Relative Efficacy of the Adaptive Kernel Density Estimation and Spatial Empirical Bayes Approaches in Disease Mapping

Using the same parameters and simulated dataset created from Chapter 2, this study examined and compared the rate estimates computed from two different disease mapping methods – adaptive kernel density estimation (KDE) and spatial empirical Bayes (Chapter 3).

Considering global rate (estimated state rates), the results show that both the adaptive KDE and spatial empirical Bayes provide identical rate estimates at all thresholds and all age

groups. While there is no difference of the global rate estimates between the adaptive KDE and spatial empirical Bayes methods, the local rate estimates show some differences and vary among thresholds and age groups. The results indicate that when the given threshold is less than 100, both methods mostly provide identical rate estimates in all focus groups, and over 90% of total simulations in each threshold have RMSEs less than 0.5. When thresholds are greater than 100, the degree of difference varies among the focus groups, in which the largest threshold illustrates the highest difference. However, some extremely high differences and variations occur at least one mid-size threshold in all focus groups. The results found that the cause of high RMSEs is mostly influenced by lower population ZCTA in non-urban areas.

The most important finding of the study is the rate estimates from both methods are mostly identical, especially disregarding the extreme differences. Since the cause of extreme differences in all focus groups is from the ZCTA in non-urban areas, the results suggest that using either method to map diseases in densely populated (urban) areas may provide identical rate estimates, and caution is necessary when mapping disease in sparsely populated areas. Further research is required to determine the optimal method for low population density areas.

5.2.3 Objective 3: To Provide an Alternative Approach for Defining Neighborhood Of Risk and Identifying At Risk Populations in those Spatial Units

Using social determinants of health (SDH), this study sought to identify the predictive characteristics of high-risk neighborhoods for HIV infection. It illustrates how to classify and construct the health risk map using straightforward statistical and spatial methods (Chapter 4). Instead of the entire state of Texas, this study focused on Public Health Regions (PHRs) 3 and 4. These regions are designated by Texas Department of State Health Services (DSHS). PHRs 3 and 4 are respectively located in North Central Texas and North-East Texas and considered as

urban and non-urban areas according to the 2013 NCHS Urban-Rural Classification Scheme for Counties (Ingram & Franco, 2014). The social determinants of health variables were selected based on CDC NCHHSTP guidelines (2010) and obtained from the 2013 ACS 5-year estimates dataset at the block group level.

Due to a large number of SDH variables, factor analysis was employed, and the results indicated that socio-demographic characteristics are slightly different between these two regions. PHR 3 has more complexity of characteristics (six groups of socio-demographic characteristics) and higher depth of poverty than PHR 4 (five groups of socio-demographic characteristics). When constructing homogeneous zones using these characteristics, each characteristic in PHR 4 represents its own cluster while one cluster zone in PHR 3 – ‘Low to Lower-middle SES’ - is the combination of two characteristics.

When mapping these clusters, the results illustrate that the ‘Poverty-Black’ characteristic, considered as a higher risk of HIV infection, mostly clusters and concentrates in the vicinity of cities in both regions regardless of small or large cities and urban or non-urban areas. Moreover, the findings highlight that ‘Low Education-Hispanic’ characteristic, a risk marker of HIV infection, mostly occurs adjacent to spatial units dominated by ‘Poverty-Black’ in the urbanized cities in both regions resulting in large areas of social disadvantages. In contrast to these urbanized cities, spatial units dominated by ‘Low Education-Hispanic’ in non-urbanized areas seem to be scattered and are not necessarily in the vicinity of cities.

These health risk maps help in identifying and delineating the high-risk spatial units that possibly have high rates of disease (HIV/AIDS in this case) – spatial units dominated by ‘Poverty-Black’ and ‘Low Education-Hispanic’. These spatial units potentially have higher risks than the others, although they may or may not have the highest rates. In the situation that health

data are not available, these health risk maps can be used to identify potential areas and populations at risk to prioritize for prevention and control. If health data is available, such these maps can become even more powerful and meaningful. In this study, although the spatial mismatch precludes a statistical comparison of HIV disease burdens and the health risk maps, the simple comparison (overlay) between HIV/AIDS and health risk maps shows strong evidence of the utility of these maps. Combining these two maps not only helps in determining *where* but also *who* or *which populations* should be prioritized for intervention and control.

5.3 Broader Impacts

The study not only fills the identified gaps of knowledge in disease mapping but also has a wide range of broader impacts. Moreover, it impacts health disparities research and benefits public health practitioners and related health organizations in many ways. First, since the choice of threshold affects the degree of smoothing that occurs on the map, the approach to threshold selection for the KDE to determine a desirable threshold with statistical supports can help in reducing bias and constraint on the issue of threshold selection. Moreover, our findings underscore the importance of carefully choosing the threshold values to use in disease mapping. Inappropriate thresholds can produce misleading conclusions.

Second, the findings on the relative efficacy of the KDE and spatial empirical Bayes approaches will help public health practitioners and related health organizations to better understand the merits of these disease mapping methods in terms of accuracy of visual representation while recognizing the limitations and importance of selected disease mapping methods.

Lastly, the contribution of neighborhood health and place vulnerability concepts to define neighborhood of risk and identify at risk populations in those spatial units will benefit public health tasks of planning and targeting appropriate intervention even in areas with limited and poor-quality health data.

Dissemination of results is also extremely important to the success of this study. The findings of each manuscript (Chapter 2-4) have been presented at the national meetings of the American Association of Geographers and the International Medical Geography Symposium to map-makers, GIS specialists, public health workers and related health organizations. Moreover, the Chapter 2 and Chapter 4 manuscripts were submitted to appropriate peer-reviewed journals, and the Chapter 3 manuscript is currently being prepared for publication.

In sum, the findings of this study improve and enhance the use of the KDE method in health research, provide better awareness and understanding of disease mapping methods, and offer an alternative method to identify populations at risk in areas with limited health data. Overall, these findings will benefit health society as well as enhance disease surveillance systems.

5.4 References

- Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(18). doi:10.1186/1476-072X-3-18
- Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43(1), 1-20.
- Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-59. doi:10.1191/0962280205sm388oa

- Beyer, K. M., Tiwari, C., & Rushton, G. (2012). Five essential properties of disease maps. *Annals of the Association of American Geographers*, *102*(5), 1067-1075. doi:10.1080/00045608.2012.659940
- Bithell, J. F. (2000). A classification of disease mapping methods. *Statist. Med.*, *19*, 2203-2215.
- Cai, Q. (2007). *Mapping disease risk using spatial filtering methods* (Unpublished doctoral dissertation). The University of Iowa, Iowa City, Iowa.
- Carlos, H. A., Shi, X., Sargent, J., Tanski, S., & Berke, E. M. (2010). Density estimation and adaptive bandwidths: a primer for public health practitioners. *International Journal of Health Geographics*, *9*(39). doi:10.1186/1476-072X-9-39
- Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2010). *Establishing a holistic framework to reduce inequities in HIV, viral hepatitis, STDs, and tuberculosis in the United States*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved from <https://www.cdc.gov/socialdeterminants/docs/SDH-White-Paper-2010.pdf>
- Chi, W., Wang, J., Li, X., Zheng, X., & Liao, Y. (2007). Application of GIS-based spatial filtering method for neural tube defects disease mapping. *Wuhan University Journal of Natural Sciences*, *12*(6), 1125-1130. doi:10.1007/s11859-007-0097-6
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*(3), 671-681.
- Cromley, E. K., & McLafferty, S. L. (2012). *GIS and public health* (2nd ed.). New York, NY: The Guilford Press.
- Ingram, D. D., & Franco, S. J. (2014). 2013 NCHS urban-rural classification scheme for counties. *National Center for Health Statistics. Vital Health Stat*, *2*(166). Retrieved September 29, 2016, from https://www.cdc.gov/nchs/data/series/sr_02/sr02_166.pdf
- McLafferty, S. (2015). Disease cluster detection methods: recent developments and public health implications. *Annals of GIS*, *21*(2), 127-133. doi:10.1080/19475683.2015.1008572
- Moraga, P., & Lawson, A. B. (2012). Gaussian component mixtures and CAR models in Bayesian disease mapping. *Computational Statistics and Data Analysis*, *56*, 1417-1433. doi:10.1016/j.csda.2011.11.011
- Oppong, J. R., Tiwari, C., Ruckthongsook, W., Huddleston, J., & Arbona, S. (2012). Mapping late testers for HIV in Texas. *Health & Place*, *18*, 568-575. doi:10.1016/j.healthplace.2012.01.008
- Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, *15*, 717-726.

- Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over in homogeneous backgrounds. *International Journal of Geographical Information Science*, 24(5), 643-660. doi:10.1080/13658810902950625
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York, NY: Chapman and Hall.
- Talbot, T. O., Kulldorff, M., Forand, S. P., & Haley, V. B. (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine*, 19, 2399-2408.
- Tiwari, C., & Rushton, G. (2005). Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In P. Fisher, *Developments in spatial data handling* (pp. 665-676). Berlin: Springer-Verlag.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Boca Raton, FL: Chapman & Hall/CRC.
- Washington State Department of Health. (2012, October). *Guidelines for working with small numbers*. Retrieved from <http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf>

APPENDIX A
SUPPLEMENTARY DATA

Table A.1 A list of the International Classification of Diseases (ICD) for heart diseases used in this study

Cause of death	ICD-10
Acute rheumatic fever	I00-I02
Chronic rheumatic heart diseases	I05-I09
Hypertensive heart disease	I11
Hypertensive heart and renal disease	I13
Ischemic heart diseases	I20-I25
Pulmonary heart disease and diseases of pulmonary circulation	I26-I28
Other forms of heart disease	I30-I51

Source

Heron, M. 2015. Deaths: Leading causes for 2011. *Nation Viatal Statistics Reports* 64(7). Hyattsville, MD: National Center for Health Statistics. Available at http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_07.pdf [last accessed 5 November 2015].

Table A.2 A list of social determinants of health variables used in this study

Social environment	Physical environment	Health services
Race/ethnicity	Households by size	Health insurance coverage
White	Number of rooms in household 2-4 rooms 5-7 rooms 8 or more rooms	Health insurance coverage by age Under 18 years 18-34 years 35-64 years 65 years and older
Black		
Hispanic (non-White)		
Educational attainment Less than 9 th grade 9 th -12 th grade High school completion Some college or associate's degree Bachelor's degree Master's degree or higher	Housing units by structure Single-family detached homes Mobile home	Household language Total household of limited English Spanish and limited English speaking household Other languages and limited English speaking household
	Housing units by year built Before 1960 1960-1979 1980-1999 2000 or after	
	Median value of home	
Ratio of income to poverty level Less than 0.50 0.50-0.99 1.00-1.24 1.25-1.49 1.50-1.99 2.00 or greater	Values of homes Less than \$50,000 \$50,000-\$99,999 \$100,000-\$149,999 \$150,000-\$199,999 \$200,000-\$299,999 \$300,000 and higher	
	Transportation to work By car, truck, or van By public transportation By bicycle or walked Worked at home	
Household income Less than \$15,000 \$15,000-\$24,999 \$25,000-\$34,999 \$35,000-\$49,999 \$50,000-\$74,999 \$75,000-\$99,999 \$100,000-\$149,999 \$150,000 or greater		
Median household income		
Median household income by race and Hispanic origin White (not Hispanic) White Hispanic		
Tenure by vehicles available No vehicle 1 vehicle 2 vehicles 3 vehicles or greater		
Employment status		

(table continues)

Table A.2 (continued)

Social environment	Physical environment	Health services
Occupation for civilian employed Management Services Sales and office Construction Production, transportation, and material moving		

APPENDIX B
SUPPLEMENTARY RESULTS

Table B.1 Summation of the average, standard deviation and RMSE of estimated disease rates from 100 replications.

	Aged 35-44			Aged 45-54			Aged 55-64			Aged 65+			Aged 35+ (overall)		
	Avg	SD	RMSE	Avg	SD	RMSE	Avg	SD	RMSE	Avg	SD	RMSE	Avg	SD	RMSE
Simulated baseline rate	33.9			115.2			297.6			1245.4			351.1		
Crude rate	35.8	299.7	9.2	115.8	476.5	12.3	298.0	822.2	20.8	1243.0	1614.5	41.3	409.3	614.9	60.7
Arbitrary choices															
<i>50</i>	34.4	126.2	3.7	115.1	222.0	7.3	298.0	374.1	10.5	1244.1	746.3	19.9	407.8	351.9	57.4
<i>100</i>	34.1	98.7	3.1	115.3	178.2	6.0	298.1	304.3	8.4	1243.0	615.3	17.7	407.1	292.4	56.6
<i>500</i>	34.1	58.4	2.3	115.4	107.2	3.8	298.4	181.2	5.8	1244.4	371.2	10.8	401.6	188.7	50.7
<i>1000</i>	34.1	46.2	2.0	115.4	85.1	3.3	298.1	142.8	5.2	1244.6	290.8	9.9	398.7	158.9	47.7
<i>5000</i>	34.1	24.2	1.6	115.4	44.4	2.4	297.6	72.4	3.5	1244.7	147.1	6.8	390.9	109.9	39.9
<i>10000</i>	34.0	17.6	1.5	115.5	32.5	2.1	297.9	52.3	3.1	1244.7	106.7	6.0	387.4	91.7	36.4
Calculated threshold															
h_{pi}	34.3	123.4	3.6	115.2	214.3	7.0	298.1	388.8	10.8	1242.9	785.5	20.9	404.9	243.2	54.2
h_{ns}	34.2	71.5	2.6	115.6	133.3	4.6	298.6	253.8	6.8	1244.0	535.1	14.7	399.3	165.5	48.4
<i>median</i>	34.2	67.2	2.4	115.5	115.7	4.1	298.8	206.0	6.5	1244.7	428.3	11.7	396.5	145.3	45.6

Table B.2 Summation of the average, standard deviation and RMSE of estimated state rates from 100 replications (global difference).

<i>Threshold</i>	Age 35-44 Avg \pm SD (RMSE)		Age 45-54 Avg \pm SD (RMSE)		Age 55-64 Avg \pm SD (RMSE)		Age 65 years and older Avg \pm SD (RMSE)		Overall Avg \pm SD (RMSE)	
	<i>KDE</i>	<i>SEB</i>	<i>KDE</i>	<i>SEB</i>	<i>KDE</i>	<i>SEB</i>	<i>KDE</i>	<i>SEB</i>	<i>KDE</i>	<i>SEB</i>
<i>50</i>	34.4 \pm 126.2 (0.002275)	34.4 \pm 127.7	115.1 \pm 222.0 (0.001141)	115.1 \pm 223.1	298.0 \pm 374.1 (0.001965)	298.0 \pm 374.8	1,244.1 \pm 746.3 (0.003566)	1,244.1 \pm 746.8	407.8 \pm 351.9 (0.002343)	407.8 \pm 352.4
<i>100</i>	34.1 \pm 98.7 (0.002280)	34.1 \pm 99.3	115.3 \pm 178.2 (0.001310)	115.3 \pm 178.7	298.1 \pm 304.3 (0.002148)	298.1 \pm 304.7	1,243.0 \pm 615.3 (0.005295)	1,243.0 \pm 615.6	407.1 \pm 292.4 (0.002553)	407.1 \pm 292.8
<i>500</i>	34.1 \pm 58.4 (0.002802)	34.1 \pm 58.5	115.4 \pm 107.2 (0.000977)	115.4 \pm 107.3	298.4 \pm 181.2 (0.166347)	298.3 \pm 181.2	1,244.4 \pm 371.2 (0.003041)	1,244.4 \pm 371.4	401.6 \pm 188.7 (0.002632)	401.6 \pm 188.8
<i>1000</i>	34.1 \pm 46.2 (0.004345)	34.1 \pm 46.3	115.4 \pm 85.1 (0.001318)	115.4 \pm 85.1	298.1 \pm 142.8 (0.001082)	298.1 \pm 142.8	1,244.6 \pm 290.8 (0.001729)	1,244.6 \pm 291.0	398.7 \pm 158.9 (0.088989)	398.6 \pm 158.8
<i>5000</i>	34.1 \pm 24.4 (0.010697)	34.1 \pm 24.3	115.4 \pm 44.4 (0.000170)	115.4 \pm 44.5	297.6 \pm 72.4 (0.000196)	297.6 \pm 72.4	1,244.7 \pm 147.1 (0.124612)	1,244.5 \pm 147.2	390.9 \pm 109.9 (0.259353)	390.7 \pm 109.2
<i>10000</i>	34.0 \pm 17.6 (0.006893)	34.0 \pm 17.7	115.5 \pm 32.5 (0.023784)	115.5 \pm 32.5	297.9 \pm 52.3 (0.008555)	297.9 \pm 52.4	1,244.7 \pm 106.7 (0.175745)	1,244.5 \pm 106.5	387.4 \pm 91.7 (0.201128)	387.2 \pm 91.3
<i>h_{pi}</i>	34.3 \pm 123.4 (0.002270)	34.3 \pm 124.7	115.2 \pm 214.3 (0.001208)	115.2 \pm 215.2	298.1 \pm 388.8 (0.001930)	298.1 \pm 389.6	1,242.9 \pm 785.5 (0.003700)	1,242.9 \pm 786.0	404.9 \pm 243.2 (0.002418)	404.9 \pm 243.4
<i>h_{ns}</i>	34.2 \pm 71.5 (0.002603)	34.2 \pm 71.7	115.6 \pm 133.3 (0.054586)	115.5 \pm 133.4	298.6 \pm 253.8 (0.001963)	298.6 \pm 253.9	1,244.0 \pm 535.1 (0.003172)	1,244.0 \pm 535.3	399.3 \pm 165.5 (0.002480)	399.3 \pm 165.6
<i>Median</i>	34.2 \pm 67.2 (0.005161)	34.2 \pm 67.2	115.5 \pm 115.7 (0.001738)	115.5 \pm 115.7	298.8 \pm 206.0 (0.001741)	298.8 \pm 206.0	1,244.7 \pm 428.3 (0.200337)	1,244.6 \pm 428.1	396.5 \pm 145.3 (0.002995)	396.5 \pm 145.3
<i>Crude rate</i>	35.8 \pm 299.7		115.8 \pm 476.5		298.0 \pm 822.2		1,243.0 \pm 1,614.5		409.3 \pm 614.9	
<i>Simulated baseline rate</i>	33.9		115.2		297.6		1,245.4		351.1	

Note: KDE – the adaptive KDE method; SEB – the spatial empirical Bayes method.

Table B.3 Summary of RMSE at ZCTA level (local difference)

Threshold	Age 35-44 Median RMSE [min, max]	Age 45-54 Median RMSE [min, max]	Age 55-64 Median RMSE [min, max]	Age 65 years and older Median RMSE [min, max]	Overall Median RMSE [min, max]
<i>50</i>	0.0089 [0, 0.96]	0.0150 [0, 0.19]	0.0325 [0.01, 0.23]	0.0656 [0.02, 0.42]	0.0133 [0, 0.96]
<i>100</i>	0.0101 [0, 0.96]	0.0210 [0, 0.16]	0.0407 [0.01, 0.28]	0.0739 [0.03, 1.15]	0.0177 [0, 0.96]
<i>500</i>	0.0120 [0, 0.96]	0.0240 [0.01, 0.16]	4.5180 [0.02, 22.59]	0.0713 [0.03, 0.66]	0.0239 [0.01, 0.96]
<i>1,000</i>	0.0126 [0, 1.06]	0.0192 [0.01, 0.43]	0.0332 [0.01, 0.22]	0.0749 [0.03, 0.23]	3.4843 [0.01, 8.71]
<i>5,000</i>	0.0173 [0, 1.67]	0.0112 [0.01, 0.03]	0.0140 [0.01, 0.03]	4.6476 [1.75, 7.95]	11.0038 [4.40, 16.73]
<i>10,000</i>	0.1878 [0, 0.75]	0.7070 [0.28, 1.72]	0.2256 [0, 1.13]	7.5849 [5.07, 10.96]	6.9595 [4.25, 9.73]
<i>h_{pi}</i>	0.0092 [0, 0.96]	0.0175 [0, 0.22]	0.0307 [0.01, 0.39]	0.0725 [0.02, 0.59]	0.0190 [0.01, 0.96]
<i>h_{ms}</i>	0.0111 [0, 0.96]	0.0258 [0.01, 13.83]	0.0342 [0.01, 0.21]	0.0807 [0.03, 0.42]	0.0214 [0.01, 0.96]
<i>Median</i>	0.0118 [0, 2.03]	0.0257 [0, 0.48]	0.0401 [0.01, 0.40]	6.0184 [0.04, 24.07]	0.0224 [0.01, 0.96]

Table B.4 Factor loadings of all 26 variables included in the factor analysis (initial results).

Variables	Public Health Region 3						Public Health Region 4					
	Communalities	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communalities	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
<i>Percent of variance</i>		26.1	15.6	12.4	11.5	5.7		17.1	15.9	12.9	8.8	7.1
<i>Race/ethnicity</i>												
White	0.782	-0.650		0.467		0.319	0.792		0.555	-0.510		0.372
Black	0.672			-0.712			0.687		-0.613			
Hispanic	0.857	0.903					0.833			0.895		
<i>Education attainment</i>												
Less than high school	0.887	0.871					0.755			0.701	-0.320	
High school degree	0.850		-0.830				0.628	-0.410		-0.302	-0.410	0.413
Some college and associate degree	0.697	-0.502	-0.660				0.482			-0.334	0.495	
Bachelor degree and higher	0.831	-0.515	0.517		0.446		0.779	0.715				-0.307
<i>Income to poverty ratio</i>												
Less than 0.50	0.500			-0.643			0.411		-0.617			
Between 0.50 and 0.99	0.525	0.506		-0.449			0.443		-0.535			
Equal or greater than 2.00	0.873	-0.634		0.546	0.367		0.818	0.391	0.733	-0.321		
<i>Household incomes</i>												
\$50,000-99,999	0.630		-0.500	0.562			0.607		0.757			
\$100,000-149,999	0.636	-0.379		0.327	0.614		0.415	0.337	0.495			
\$150,000-199,999	0.592	-0.362	0.404		0.518		0.326	0.543				
Equal or greater than \$200,000	0.839	-0.423	0.795				0.620	0.775				
<i>Occupation</i>												
Management	0.831	-0.640	0.419		0.403		0.632	0.594			0.325	
Construction	0.594	0.677					0.649					0.767
Production	0.444	0.448			-0.350		0.640	-0.365		0.337	-0.546	

(table continues)

Table B.4 (continued)

Variables	Public Health Region 3						Public Health Region 4					
	Communalities	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communalities	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
<i>Value of house</i>												
Less than \$100,000	0.777	0.494		-0.363	-0.501	0.332	0.809	-0.589	-0.372		-0.543	
Between \$100,000 and 149,999	0.688		-0.480			-0.549	0.643				0.749	
Between \$150,000 and 299,999	0.888				0.901		0.534	0.528	0.321		0.326	
Equal or greater than \$300,000	0.840	-0.403	0.808				0.629	0.764				
Living in mobile home						0.759	0.552				-0.346	0.575
Housing with ≥ 8 rooms	0.671	-0.450	0.532		0.371		0.576	0.659	0.306			
No vehicle own	0.570			-0.724			0.551		-0.693			
No health insurance	0.745	0.770					0.544	-0.312	-0.311	0.555		
Limited speaking English in the household	0.716	0.827					0.716			0.839		

Note: All variables are in percentage unit; **Boldfaced** values indicated variables that contribute in the composite factor (factor loadings $\geq |0.45|$); Factor loadings $< |0.3|$ are suppressed.