Facilitators and barriers to safely managed water and sanitation: A spatio-temporal investigation of the association between socioeconomic factors and shigellosis incidence

by

Sabrina Li

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Geography

Waterloo, Ontario, Canada, 2017

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

The lack of access to safe drinking water and sanitation worldwide has contributed to the occurrence of shigellosis, a waterborne infectious disease. Previous research has shown that shigellosis can be prevented by access to safe drinking water and adequate sanitation, however access is facilitated or hindered by socioeconomic conditions. The effects of socioeconomic conditions on shigellosis incidence are unclear in the context of rural China. This thesis explored the spatial patterns of shigellosis over time and the spatio-temporal association between shigellosis incidence and socioeconomic conditions of the rural population in Jiangsu province.

Choropleth maps were created to understand the geographic distribution of shigellosis incidence at the county level. Spatial analysis methods such as spatial autocorrelation, Local Moran's I, and the Getis Ord Gi were used to identify disease clusters, outliers, and hotspots. Based on the findings from the hot spot analysis and evidence from literature, a field visit to the northwestern county of Suining was conducted to further investigate the relationships between water and sanitation access and shigellosis incidence in the rural context. The temporal variability of the association between socioeconomic factors and shigellosis at the county level was investigated using negative binomial and quasi-Poisson regression models. The spatial relationship between socioeconomic factors and shigellosis at the county level was explored using a Bayesian spatial model.

Results showed that shigellosis morbidity was characterized by significant declines in most regions from 2011 to 2015; however, high morbidity rates were still evident in the northwestern region of Jiangsu. At the county level, the factors such as the number of hospital beds per capita and the percentage of rural households has shown to be significantly associated with shigellosis incidence for years 2011, 2012, and 2014, respectively. The percentage of rural households was negatively correlated with shigellosis incidence; this relationship was further confirmed by results from the Bayesian spatial model. In addition, results showed that rural employment and the number of hospital beds per capita, respectively, were correlated with a decrease in shigellosis incidence. In contrast, the number of hospitals per capita was positively correlated with an increase in shigellosis incidence. Underreporting of shigellosis in rural areas was suspected to be the cause of the low rate of shigellosis in rural areas. The quality of the rural healthcare system and living conditions may have influenced this underreporting. Thus, a more active surveillance method should be adopted to gauge the real occurrence of shigellosis amongst the rural population.

ACKNOWLEDGEMENTS

I'd like to acknowledge and thank the funding I've received from the MITACS Globalink Award for supporting my thesis research in China. This thesis was completed with the support of many individuals. I would like to extend my sincerest thanks to all of them.

Firstly, I would like to thank my supervisor Dr. Susan Elliott for supporting my interests and for dedicating her time to this thesis since the conception of my research topic. I feel like I became a more experienced researcher under your invaluable mentorship.

Next, I'd like to thank my thesis committee members, Dr. Peter Deadman, Dr. Peter Johnson, and Dr. Lisa Guppy. Thanks for dedicating your time to reviewing and sharing your feedback on the content of this thesis. I would also like to extend my thanks to Jiang Han (Department of Geography and Environmental Management) for helping to translate data collected from China, Dr. Derek Robinson (Geography and Environmental Management) for suggesting tools in ArcGIS for spatial analysis, and Dr. Changbao Wu (Department of Statistics and Actuarial Science) and Dr. Jane Law (School of Planning) for verifying the procedure taken for the generalized linear regression analysis.

This thesis would not have been possible without the support I have received in China. Thanks to Dr. Jinhe Zhang from Nanjing University for giving me the opportunity to connect with students and obtain the resources I needed for this thesis, and Dr. Yongyue Wei from Nanjing Medical University for his suggestions on epidemiological data analysis. A special thanks goes out to Victor Xu Qi, and Kang Jing Yao for helping me with finding GIS data, and Hu Huan for helping me to coordinate the paper work needed for my field visits. Furthermore, a BIG thank you goes to Di Meng and her friends at Suining County for connecting me to all the resources in Suining County. Thank you for showing me your hometown and providing me incredible support throughout.

Lastly, thanks to my friends and family for their endless support. To my grad school friends, thanks for all the (much-needed) social outings, potlucks, zumba hangouts, fun side projects (Polygone and WAMS), and board game nights that have made grad school enjoyable. A special shout out to Sabrina Bedjera, Hongjing Chen, Sarah Irvine, Kelvin Liew, and Sabrina Touchette for all your kind words and hugs, especially when I needed them the most. To my friends Elaine Wong and Mark Vaz, thanks for always being there for me from the start to finish. To Francis Li, thanks for your optimism and supporting me in all my interests and endeavors. Finally, thanks to my parents for all your love and encouragement during this journey. Many hugs and kisses.

DEDICATION

This thesis is dedicated to my loving grandparents, lao lao and lao ye. Thank you both for fostering my creativity and always believing in me.

TABLE OF CONTENTS

AUTHOR'S DECLARATION ..... **ii**

ABSTRACT ..... **iii**

ACKNOWLEDGEMENTS ..... **iv**

DEDICATION ..... **v**

LIST OF FIGURES ..... **viii**

LIST OF TABLES ..... **ix**

CHAPTER 1: INTRODUCTION **1**

1.1 Research Context 4
1.2 Research Objectives 6
1.3 Research Contributions 8
1.4 Chapter Outline 9

CHAPTER 2: LITERATURE REVIEW **10**

2.1 Introduction 10
2.2 Shigellosis: Causes, transmission pathways, and prevention 10
2.3 Water and Sanitation in Rural China 16
2.4 Theoretical Framework 19
2.5 Conceptual Framework 22
2.6 Methodological Literature 27
2.6.1 Spatial Data Visualization 27
2.6.2 Spatial Analysis of Infectious Disease 28
2.6.3 Multivariate Regression Models for Small area Incidence 31
2.7 Chapter Summary 41

CHAPTER 3: METHODOLOGY **42**

3.1 Introduction 42
3.2 Study Area 42
3.3 Data 47
3.3.1 Data Sources 47
3.3.2 Data Preprocessing and Geocoding 51
3.3.3 Missing Data Imputation 53
3.4 Generalized Linear Model Regression 54
3.4.1 Analysis Workflow 54
3.4.2 Generalized Linear Regression: Poisson Model 57
3.4.3 Overdispersion: Quasi-Poisson and Negative Binomial Model 59
3.4.4 Bayesian Spatial Regression Model 64
3.5 Spatial Data Visualization and Analysis 65
3.5.1 Analysis Workflow 65
3.5.2 Choropleth Maps 67
3.5.3 Spatial Autocorrelation: Global Moran's I 69
3.5.4 Hot Spot Analysis: Local Moran's I and Getis Ord Gi* 72
3.6 Chapter summary 74

CHAPTER 4: RESULTS **75**

4.1 Introduction 75
4.2 Data Visualization and Analysis 75
4.2.1 Spatiotemporal distribution of shigellosis incidence 75

4.2.2 Spatial distribution and correlation of socioeconomic determinants....84
4.3 Generalized Linear Model Analysis....89
4.3.1 Resulting Analysis Workflow....89
4.3.2 Exploratory Analysis of Regression Variables....90
4.3.3 Poisson Regression....92
4.3.4 Adjustment for Overdispersion....93
4.3.5 Negative Binomial Regression Output....97
4.3.6 Bayesian Spatial Regression Output....98
4.4 Chapter Summary....99

CHAPTER 5: DISCUSSION AND CONCLUSIONS....101
5.1 Introduction....101
5.2 Summary of Key Findings....101
5.3 Limitations....109
5.4 Challenges with Secondary Data in Data Deprived Areas....112
5.5 Contributions....115
5.6 Directions for Future Research....117

REFERENCES....120

APPENDIX A: REGRESSION RESULTS....135

APPENDIX B: SCATTER PLOTS....137

APPENDIX C: EXPLORATORY PLOTS....139

APPENDIX D: SOCIOECONOMIC DATA....141

LIST OF FIGURES

Figure 2.1 The classic F diagram adapted from Wagner & Lanoix (1958) 12
Figure 2.2 Proposed Conceptual Framework 22
Figure 3.1 Thirteen prefecture level cities in Jiangsu province with their county and district boundaries 43
Figure 3.2 One of the few hand pumps left in Wangji Township 44
Figure 3.3 (a) Jet pump draws water from underground (b) Discharge faucet 45
Figure 3.4 (a) Typical location of sanitation facility (b) Pour flush using buck pit latrine (c) Manure from human feces 46
Figure 3.5 Qingan reservoir 46
Figure 3.6 Workflow of generalized linear regression analysis 56
Figure 3.7 Histograms for shigellosis counts in Jiangsu province 58
Figure 3.8 Spatial analysis process for shigellosis incidence 66
Figure 3.9 Spatial analysis process for socioeconomic indicators 67
Figure 4.1 Choropleth maps illustrating the incidence rate of shigellosis by county in Jiangsu province from 2011 to 2015 79
Figure 4.2 Clusters and outliers determined by the Local Moran's I 82
Figure 4.3 Shigellosis hot spots and cold spots determined using local Getis Ord Gi* 83
Figure 4.4 Choropleth maps of socioeconomic indicators 86
Figure 4.5 Spearman Ranked Coefficient Correlograms 88
Figure 4.6 Workflow of resulting analysis 89
Figure 4.7 Scatterplots between predictors and response variable 91
Figure 4.8 Mean-variance plots 96

LIST OF TABLES

Table 2.1 Improved and Unimproved water and sanitation interventions (Adapted from JMP, 2015) 15
Table 2.2 Facilitators and Barriers 23
Table 3.1 Socioeconomic determinants of health framework 47
Table 3.2. Socioeconomic, demographic, and water and sanitation data 50
Table 3.3 Sample size of socioeconomic indicators by administrative division type 52
Table 3.4 Missing Data 54
Table 4.1 Distribution of shigellosis incidence rates per 100, 000 persons by county in Jiangsu province from 2011-2015 76
Table 4.2 Incremental Spatial Autocorrelation Peak Distances 77
Table 4.3 Spatial autocorrelation in shigellosis incidence rates of counties 80
Table 4.4. Summary statistics of socioeconomic factors investigated for association with shigellosis incidence in counties of Jiangsu province 87
Table 4.5 Shigellosis incidence counts used in regression 90
Table 4.6 Shigellosis incidence rates (per 100,000 persons) used in regression 90
Table 4.7 Goodness-of-Fit Test 92
Table 4.8 Comparison of Poisson, quassi-Poisson, and Negative Binomial model outputs 94
Table 4.9 Goodness of Fit Test for quasi-Poisson and Negative Binomial 95
Table 4.10 Association between shigellosis incidence and significant socioeconomic determinants 97
Table 4.11 Coefficient Estimates of socioeconomic determinants using Bayesian Spatial Model 98

CHAPTER 1: INTRODUCTION

In 2000, the United Nations established eight measurable goals known as the Millennium Development Goals (MDGs) to reduce poverty by improving health and social development (WHO, 2015). As a part of Goal 7, Target 7c aimed to "reduce the proportion of the population without sustainable access to safe drinking water and basic sanitation by half" by 2015 (WHO & UNICEF, 2000). In 2015, the United Nations established the successors to the MDGs, known as the Sustainable Development Goals (SDGs), which aim to be achieved by 2030. Goal 6 of the SDGs aims to "ensure availability and sustainable management of water and sanitation for all".

While the MDGs have made unprecedented progress worldwide, around 884 million people still lack access to a basic drinking water service while 2.3 billion people do not have access to a basic sanitation service (WHO/UNICEF, 2017). As defined by WHO/UNICEF JMP, a basic drinking water service ensures access to an improved drinking water source, which has the potential to deliver safe water based on its construction and design. A basic drinking water service is protected from outside contamination, provided that collection time is no more than 30 minutes. In addition, a basic sanitation service refers to having access to an improved sanitation facility, which is defined as a facility where human excreta is hygienically separated from human contact and is not shared with other households.

Targets 6.1 and 6.2 of the SDGs aim to move beyond basic services and achieve universal access to safely managed water and sanitation services by 2030. Based on the definitions provided by WHO/UNICEF (2017), a safely managed drinking water service is defined as "drinking water from an improved water source that is accessible on premises, available when needed, and free from fecal and priority chemical contamination" while a

safely managed sanitation facility is defined as “improved facilities that are not shared with other households, and where excreta are safely disposed of in situ, or transported and treated offsite”.

Exposure to contaminated drinking water and lack of sanitation can lead to waterborne diseases that affect human health (WHO, 2017a). Open defecation or the lack of a toilet facility, along with improper handling of human excreta, can promote the spread of fecal matter through groundwater and surface water. This leads to fecal-oral transmission of waterborne diseases upon ingestion of water or food contaminated with fecal matter. Waterborne diseases, which are mainly caused by microorganisms such as bacteria, can cause intestinal infections that lead to diarrhea.

Globally, the number of young children dying from diarrhea is more than the combined number of deaths from AIDS and malaria (Boschi-Pinto et al., 2008; L. Liu et al., 2017). Diarrhea is the second leading cause of death for children under five years of age (WHO, 2017b). The diseases caused by poor water and sanitation can trigger malnutrition amongst young children, which makes them more vulnerable to major childhood diseases such as measles and pneumonia (Bartram & Cairncross, 2010; Carlton et al., 2012). Due to the lack of water supply and a sanitation facility, children are compelled to spend more time collecting water and seeking a place to defecate. These conditions may delay children’s entry into school and deter girls from attending school once they reach menarche (Bartram & Cairncross, 2010; Pearson & McPhedran, 2008).

The lack of access to a safely managed water and sanitation can significantly contribute to the existing water issues in China. China’s water resources, which are only less than one

quarter of the world average supply, are required to sustain 20% of the world's population (Jiang et al., 2015; Junfeng Zhang et al., 2010; Xie et al., 2009). In addition to limited water availability, the majority of the water that comes from lakes and major rivers in China is severely polluted (Junfeng Zhang et al., 2010). In some places, water from reservoirs is unsuitable for drinking even after wastewater treatment (H. Yang et al., 2012; Junfeng Zhang et al., 2010) due to contamination and lack of adequate treatment prior to reaching the point of water delivery (Shaheed et al., 2014). For instance, piped water at the point-of-use may be of suboptimal quality as a result of piping minimally treated water at the source that does not meet drinking water standards for microbial safety (Carlton et al., 2012; Prüss-Üstün et al., 2014; Shaheed et al., 2014). As a result of China's water scarcity and contamination problems, 4% of the population (54.8 million) still lacks access to a basic drinking water service while 30% of the population (411.3 million) still lacks access to safely managed sanitation facility (WHO/UNICEF, 2017).

To improve access to safe drinking water, China has implemented a five-year plan worth 410 billion RMB ($66 billion USD) on the expansion of its public water infrastructure to 54% of the population living in cities and towns by 2015 (Tao & Xin, 2014). Despite this expansion, regional disparities continue to affect water access for the rural population, which makes up 43% of the total population in China (World Bank, 2016). Currently, 38% of the rural population still lacks access to piped water (WHO/UNICEF, 2017). Moreover, it has been found that the lack of access to treated water has been underestimated by the WHO/UNICEF in the past (Zhang and Xu, 2016), so it is likely that the proportion with access to untreated drinking water is much greater in reality (Yang et al., 2012). Results from a survey on 31 provinces conducted in 2006 indicated that half of the sampled 60 000 rural households across China

relied on untreated water from hand pumps, wells, or surface water from nearby rivers (Rong et al., 2009). This study is the most comprehensive study on rural water and sanitation in China published to date, and has shown that more than half of the water samples collected were unsafe for drinking, primarily due to contamination with untreated sewage. Thus, the lack of access to safely managed drinking water renders the rural population in China particularly vulnerable to waterborne diseases.

1.1 Research Context

It is estimated that 700, 000 deaths each year are caused by shigellosis, a water-borne diarrheal disease caused by a single type, gram-negative bacteria called *Shigella* (WHO, 2005). It is the third most commonly reported infectious disease in China (Z. Li et al., 2015; X. Liu et al., 2017; Jianmin Zhang et al., 2014; H. Zhang et al., 2016), after tuberculosis and hepatitis B (Z. Li et al., 2015; Xiao et al., 2014). Shigellosis continues to pose a considerable disease burden especially amongst the elderly and children (Ma et al., 2015; Li et a., 2016; X. Wang et al., 2006; L. Zhao et al., 2017). In China, *Shigella* resistance to several first-line antibiotics makes effective antimicrobial treatment a significant challenge (Seidlein et al., 2006; X. Wang et al., 2006; Jianmin Zhang et al., 2014). Qu et al. (2014) found that over 90% of *Shigella* isolates were resistant to at least three different kinds of antibiotics.

Shigellosis is monitored by the Chinese Centre for Disease Control and Prevention (CDC), which requires cases of shigellosis to be reported by all clinics and hospitals. Transmission occurs through oral contact and can be diagnosed through the stool of a person infected with *Shigella*. Transmission may occur due to inadequate hand hygiene, contact with food that has been processed or washed with contaminated water, drinking water that has

contaminated with fecal matter, or exposure to feces through sexual contact. These transmission processes are exacerbated in regions of China where people rely heavily on polluted water due to lack of access to safe managed water.

The morbidity rate of shigellosis varies by region. In particular, high morbidity caused by shigellosis has been reported in the eastern province of Jiangsu (X. Wang et al., 2006), one of the most densely populated provinces in China. From 2001 to 2011, Tang et al. (2014) analyzed the spatiotemporal trends of shigellosis incidence trends in Jiangsu, and found that shigellosis incidence rates followed a decreasing trend over time but peaked in 2004, 2006, and 2011. Clusters of high shigellosis incidence were found in the southwestern and northwestern regions. Shigellosis incidence varies geographically as it is driven by high temperatures and relative humidity (Tang et al., 2014). Due to these meteorological drivers, shigellosis is more prevalent in high temperature environments as the majority of shigellosis incidents occur during the summer and autumn months (July - October) (Ma et al., 2015; Tang et al., 2014; Xu et al., 2014; Y. Zhang et al., 2007).

Despite the lack of effective antimicrobial treatment, the prevalence of shigellosis can be easily controlled and reduced by improved water and sanitation in rural communities (Nie et al. 2014). Bartram & Cairncross (2010) claims that improved water and sanitation could reduce diarrhea prevalence by one third. However, there exists a disparity in access to improved water and sanitation in rural communities, which is influenced by socioeconomic factors such as education attainment, the existence of a social support network, and culture (e.g. values of a place that contribute to perpetuation of marginalization and stigmatization). These factors are defined as a subset of the social determinants of health (Public Health Agency of Canada,

2011), which are the living conditions that facilitate or hinder one's access to resources to good health, including safe managed water and sanitation (Q. Wang & Yang, 2016).

While present literature has investigated the spatial and temporal distribution of shigellosis over the years from 2001 to 2011 (Ma et al., 2015; Tang et al., 2014; Xu et al., 2014), recent spatiotemporal patterns after 2011 have not been explored in Jiangsu, a province with regions that experience high shigellosis morbidity. Furthermore, very little research has focused on how the incidence of shigellosis clusters is influenced by access to safely managed water and sanitation, particularly in rural areas of China (H. Zhang et al., 2016). Tang et al. (2014) have identified inadequate water and sanitation and low family income as risk factors of shigellosis in Jiangsu province, however it is uncertain how some of these factors that are part of the social determinants of health (herein referred to as facilitators and barriers) that determine access to safely managed water and sanitation. Therefore, it is hypothesized in this thesis that the facilitators and barriers mediate the relationship between water and sanitation access and shigellosis incidence. It is important to examine this relationship in recent years to better understand the current water and sanitation conditions in rural China and how this is linked with shigellosis incidence during the period of the SDGs.

1.2 Research Objectives

As China moves from the period of the MDGs to the SDGs, it is critical to gain an updated understanding of shigellosis prevalence in the highly populated province of Jiangsu, which has been previously identified as a region with high shigellosis morbidity. The disparity in access to safely managed water and sanitation in rural areas is largely determined by a subset of social determinants acting as facilitators and barriers, but very little information is

available on how these factors are linked with recent incidence and distribution of shigellosis across time and space (Schmidt, 2014; H. Zhang et al., 2017). Thus, the province of Jiangsu was selected as an area of interest in this thesis due to its high shigellosis morbidity and limited understanding of socioeconomic determinants that act as risk factors in that region. This study area was also chosen to serve as a case study so that findings can be transferred to other areas in eastern China. To address these knowledge gaps, this thesis will explore the following objectives:

1) To examine spatiotemporal variation of shigellosis incidence across Jiangsu province
2) To identify the facilitators and barriers to safely managed water and sanitation
3) To investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province

The research design and methods of this thesis are based on a post-positivist approach. This approach is adopted to explore the research objectives outlined above from an objective perspective using quantitative methods, but recognizes that all observations can be challenged and that data are susceptible to error and bias. Objective one was met by conducting spatial analyses to investigate the spatiotemporal distribution of shigellosis incidence from 2011 to 2015. Objective two was met by developing a conceptual framework based on existing literature. Objective three was met by conducting an exploratory analysis using survey data on rural water and sanitation and shigellosis incidence in a rural county located in Jiangsu province. A rural county was chosen because counties are the smallest administrative unit that has the most comprehensive data on a household access to water and sanitation access. Field observations of water and sanitation facilities were also used to complement and validate the findings. Lastly, objective four was met via a multivariate regression analysis that investigated

the association between shigellosis incidence and several socioeconomic factors that are part of the social determinants of health framework.

1.3 Research Contributions

This thesis presents new findings on water and sanitation in rural China. While much of the research on water and sanitation has been conducted in Southeast Asia and Africa, very little is known about the conditions of water sources and sanitation facilities in rural areas in China. Despite rapid economic development, China's large urban-rural disparities continue to affect access to clean water and improved sanitation facilities, and as a result, water and sanitation problems are still prevalent in rural areas.

This thesis makes two substantive contributions. First, this thesis unveils insights into the socioeconomic indicators that act as facilitators and barriers to safely managed water and sanitation, which may be associated with shigellosis incidence in rural areas of Jiangsu province. This would not be possible using data reports aggregated at the provincial and national level. Secondly, this thesis contributes to the current literature on water and sanitation in rural areas of low-middle income countries. The findings of this thesis could be used to provide up-to-date information to local stakeholders to support rural water management and diarrhea surveillance in Jiangsu province. The findings of this thesis could be used to educate and help the government to prioritize the improvement of certain social and living conditions to bring effective change to water and sanitation services that will benefit, the health of vulnerable areas.

1.4 Chapter Outline

This thesis consists of five chapters and is organized as follows. Chapter 2 provides a review of the current literature on the epidemiology of shigellosis (Chapter 2.2) and the conditions of water and sanitation in rural China (Chapter 2.3). In addition, this chapter also includes elaboration on the proposed theoretical (Chapter 2.4) and conceptual (Chapter 2.5) frameworks, as well as the methodological literature (Chapter 2.6) adopted and considered for this thesis.

The methodology is outlined in Chapter 3 and is separated into Chapter 3.2, which covers all information on data, and Chapter 3.3, which introduces the study area. The methodology is broken down into two main sections; Chapter 3.4 discusses generalized linear regression model, Chapter 3.5 discusses spatial data visualization and analysis techniques, and Chapter 3.6 discusses the descriptive analysis of rural water and sanitation conditions.

Results are presented in Chapter 4. This section includes results for spatial data visualization and analysis (Chapter 4.2), generalized linear regression analysis (Chapter 4.3), and a descriptive analysis of survey observations on water and sanitation conditions in a rural county (Chapter 4.4) supported by field images.

Lastly, discussion and conclusions presented in Chapter 5 will summarize this thesis' findings, limitations, contributions, policy implications, and provide recommendations for future work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter explores the theoretical, conceptual, and methodological frameworks used to construct this thesis and its objectives:

1) To examine spatiotemporal variation of shigellosis incidence across Jiangsu province
2) To identify the facilitators and barriers to safely managed water and sanitation
3) To investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province

This chapter discusses the disease characteristics of shigellosis, the role of water and sanitation for shigellosis prevention and control, and the changes in water and sanitation over time in rural China. The theoretical and conceptual frameworks used to address these objectives are also outlined and justified. Lastly, this chapter examines literature on applicable methodological approaches for addressing the objectives of this thesis.

2.2 Shigellosis: Causes, transmission pathways, and prevention

Many scientific studies based in microbiology, vaccines, and infectious diseases have shed light on shigellosis' causes and effects. Shigellosis is known as bacillary dysentery (Public Health Agency of Canada, 2011b), an enteric infectious disease that causes diarrhea. Infectious diseases such as shigellosis are transmissible from person to person and occur when human body tissues are invaded by a disease-causing agent such as bacteria. There are four species of *Shigella* bacteria: *S. dysenteriae* (group A), *S. flexneri* (group B), *S. boydii* (group C) and *S. sonnei* (group D). *S.flexneri* and *S.dysenteriae* (bacillary dysentery) accounts for the

majority of deaths from shigellosis in developing countries (Bardhan et al., 2010; Kotloff et al., 1999; Nyogi, 2005; Von Seidlein et al., 2006).

Symptoms of shigellosis include watery diarrhea (that may contain mucus and blood), painful bowel movements, abdominal pain, nausea and vomiting, rapid dehydration and weight loss. Symptoms may occur between 12 to 50 hours and last one to seven days (Public Health Agency of Canada, 2011b). Acute diarrhea may develop within one to two days. Patients can recover completely from shigellosis, however it will take several months for their bowel movements to become normal (CDC, 2017). Once a person gets infected with a specific strain of *Shigella*, they will not get infected with that strain for at least several years. However, they can still get infected with other strains of *Shigella* (CDC, 2017).

Globally, shigellosis results in 700,000 deaths a year, with most cases occurring in the developing world (WHO, 2005). It is estimated that shigellosis has caused the death of 34, 400 children under the age of five (Mani et al., 2016). In addition, it is estimated that shigellosis has resulted in the deaths of 40,500 persons over the age of five in 2013 (Peterson et al., 2015). Travelers and military service members frequently contract shigellosis when visiting *Shigella* endemic areas (Mani et al., 2016).

Shigellosis outbreaks are common in areas suffering from lack of access to safely managed drinking water, overcrowding, and poor sanitation (Nelson & Williams, 2007). Children under the age of five are highly susceptible to contracting diarrheal diseases such as shigellosis (Prüss-Üstün, 2008; WHO, 2017b). This is particularly pertinent in the context of rural China, where many private wells that supply water to schools are built in close proximity to toilets, septic tanks, sewer ditches, lakes and ponds. This allows the exposure to high

concentrations of untreated sewage (T. Chen et al., 2014) cultivated with bacteria like *Shigella* to spread through fecal-oral pathways. According to Wagner and Lanoix (1958), there are numerous pathways for fecal-oral transmission (Figure 2.1)

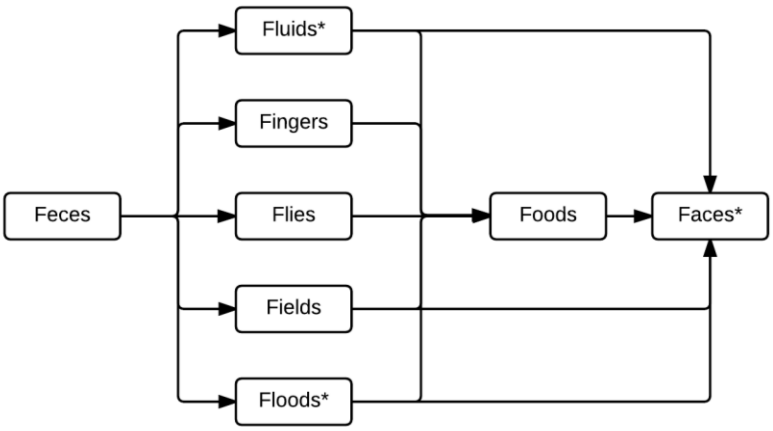

Figure 2.1 The classic F diagram adapted from Wagner & Lanoix (1958) illustrates the transmission of shigellosis from human feces to human host. *Fluids refer mainly to drinking water. Floods refer to surface water and ground water. Faces refer specifically to the mouth.

As seen in Figure 2.1, some pathways are direct while some are indirect. For instance, the indirect transfer of shigellosis through fecal-oral pathways can lead to person-to-person transmission and contamination of food and water (Stauber & Casanova, 2015; T. Chen et al., 2014). In addition, humans are not the only vectors of *Shigella*; houseflies and other anthropods also have the capability to transmit *Shigella* (WHO, 2005). T. Chen (2014) found that the two primary transmission routes for *Shigella* were person-to-person and person-to-water-to-person.

Treating shigellosis continues to be a challenge in the developing world. No vaccines exist and a successful preventative treatment is currently unavailable (Public Health Agency of Canada, 2011b; Tang et al., 2014). Moreover, several species of *Shigella* have acquired increasing resistance to antimicrobial drugs (Pazhani et al., 2005; Sack et al., 1997). It has been

found by several studies that persons identified as HIV positive experienced reoccurring *Shigella* infections despite the use of antimicrobial drugs (Baer et al., 1999; Kotloff et al., 1999; Mayer & Wanke, 1994; Sanchez et al., 2005).

Compared to urban areas, rates of shigellosis may appear lower in rural areas of China. A recent study by Cheng et al. (2017) found that in Hefei province, the neighbouring province located west of Jiangsu, urban areas were more susceptible to shigellosis than rural areas due to higher population densities that may facilitate transmission. Wang et al. (2005) investigated shigellosis reporting amongst four rural townships in Zhengding County, Hebei province, and found that the incidence rate observed for shigellosis was almost 10 times higher than what was reported by the National Noticeable Infectious Disease Report System (which is equivalent to the current CDC). Rural areas had more shigellosis infections, but cases were often underreported.

Shigellosis can be prevented by interventions that provide safe water supply and sanitation. This includes ensuring people have access to a safely managed improved water source, safely managed improved sanitation facility, and hygiene education (Fuller et al., 2014; Jing Zhang et al., 2012; Prüss-Üstün et al., 2014; Qu et al., 2012; Waddington & Snilstveit, 2009; Wolf et al., 2014). A safely managed drinking water service, as previously defined, is "drinking water from an improved water source that is accessible on premises, available when needed, and free from fecal and priority chemical contamination", while a safely managed sanitation facility is defined as, "improved facilities that are not shared with other households, and where excreta are safely disposed of in situ, or transported and treated offsite".

Accessibility to the water source is based on the type of source on site. Examples of improved and unimproved water sources are summarized in Table 2.1. Improved drinking water sources, such as piped water from a public tap, borehole, or a protected spring, are meant to provide safe water based on their construction and design (Wolf et al., 2014; WHO/UNICEF, 2017). Despite water sources being classified as "improved", some sources can provide water that is still unsafe for consumption (Bain et al., 2014, as cited in Prüss-Üstün et al., 2014). This is because the definition of "improved" water does not precisely predict microbial safety. The exposure of water to bacteria from improved water sources can result from open water storage, lack of water treatment before entry into the pipe system, and inconsistent usage of improved sources due to household water management (Shaheed et al., 2014). Thus, it is important that these facilities are considered safely managed to prevent contamination and contact with human excreta.

The fecal-oral transmission of *Shigella* can be effectively prevented by gaining access to a safely managed improved sanitation facility such as a toilet (can be supported by a piped sewer system or septic tank), which is a private facility separated from the kitchen or other living spaces. While a shared sanitation facility between two or more households can be accepted as an improved sanitation facility in certain situations (Rheinlände et al., 2015), a public toilet is not (WHO, 2012). According to the JMP, only access to a private sanitation facility would be considered having access to "improved" sanitation. Examples of improved and unimproved facilities are summarized in Table 2.1. The lack of a toilet facility is considered unimproved as it can lead to open defecation. Inappropriate disposal and lack of treatment of fecal matter can cause the contamination of water and create a breeding ground for waterborne diseases such as shigellosis.

Table 2.1 Improved and Unimproved water sources and sanitation facilities (Adapted from JMP, 2015)

	Improved	Unimproved
Water Source	• Piped water into dwelling • Piped water to yard • Public tap or standpipe • Tubewell • Borehole • Protected dug well • Protected spring • Rainwater	• Unprotected spring • Unprotected dug well • Cart with small tank/drum • Tanker-truck • Surface water • Bottled water
Sanitation Facility	• Flush toilet • Piped sewer system • Septic tank • Flush/pour flush to pit latrine • Ventilated improved pit latrine • Pit latrine with slab • Composting toilet	• Flush/pour flush to elsewhere • Pit latrine without slab • Bucket • Hanging toilet or latrine • Shared sanitation • No facilities or bush or field

The distribution of water can also influence the safety of drinking water, as some sources do not always provide water when needed. Distribution can be categorized into networked systems and un-networked systems (Marks & Kellogg, 2015). Networked systems distribute water through pumped, gravity fed systems that deliver water to public kiosks, yard taps, and household taps. Un-networked sources provide water at the location of source. In a systematic review and meta-regression analysis by Wolf et al. (2014), it was found that providing continuous high quality water through robust water infrastructure such as pipes and sewer connections was associated with greater reductions of diarrhea compared to other sources, such as those based on non-continuous water supply that may require water storage. These types of water supply are prone to microbial risks through infiltration into non-pressurized tap distribution systems and recontamination during household storage.

To reduce the contamination of water, regular usage of household water treatment (HWT) can be adopted to provide safe water for consumption (Luoto et al., 2014; Montgomery & Elimelech, 2007; Rosa & Clasen, 2010; Sobsey et al., 2008). Adopting HWT has considered as

one of the most cost effective interventions that advanced the MDGs (Clasen et al., 2007; C.Yang et al., 2009; Hutton & Haller, 2004). In rural China, the most common HWT is boiling water before consumption (H. Yang, 2012; Junfeng Zhang et al., 2010; Tao & Xin, 2014). However, WHO/UNICEF (2008) indicates that rural populations may not treat their water adequately. This claim requires further investigation as awareness of proper treatment methods vary by country as well as between the urban and rural population.

2.3 Water and Sanitation in Rural China

Before 1980, rural residents in China relied on untreated water from wells, rivers, and lakes. According to the China Health and Nutrition Survey (Jing Zhang & Xu, 2016), 70% of rural residents had access to only untreated water in 1989. Open defecation amongst humans and livestock were common in villages, which led to poor sanitation conditions.

Water and sanitation infrastructure in China has drastically improved in the last three decades as a result of increasing national level investment and regulatory action (Carlton et al., 2012; Qiu 2011; Tao & Xin, 2014). Since 1980, China's rural development projects have made access to improved water and sanitation a national priority (Carlton et al., 2012). The rural drinking water program from early 1980s have incurred a cost of $8.8 billion USD by 2002, and have covered 300 million people by 2008 (Jing Zhang & Xu, 2016). The program aimed to build water treatment plants to eliminate chemical and fecal contaminants and pipelines to provide rural residents with access to safe drinking water. Every five years, the Ministry of Health, the Ministry of Construction, the Bureau of Environmental Protection and the Ministry of Agriculture would hold a steering meeting to set goals and strategies for improving water

and sanitation over the upcoming five year period. This created the Five Year Plan initiative, which has been effective since 1986.

In 2012, China was one of the first developing countries to implement strict regulations on national drinking water quality for both urban and rural areas (Qu et al., 2012). In the more economically developed areas such as southern China, household purification units such as drinking water filters are increasingly being adopted in rural households (Ying, 2005, as cited in Yang & Wright, 2012). However, access to safely managed water services in rural areas is still difficult due to the lack of technology, finances, and human resources (Waters & Kellogg, 2015). Boiling water is still the most commonly used water treatment method used in rural households in China due to its convenience and affordability (H. Yang et al., 2012).

According to the 2014 National Health and Family Planning Commission (as cited in Tong et al., 2016), a sanitation project called the "Toilet Revolution" was created to improve sanitation conditions in rural areas. Since 1995, access to toilets in rural regions increased from 7.5% to 76.1% in 2014. Through a program called the "China Women's Development program", the Chinese government planned to ensure that 85% of the population in rural areas has access to toilets by 2020 (Tong et al., 2016). Currently, the most common toilet in rural China is the pit latrine. Pit latrines generally do not prevent stored feces from leaching into water sources and are not effective at reducing the spread of fecal contaminants in the environment (Tong et al., 2016). The reduction of fecal contaminants in the environment is important for reducing public health risks and future management strategies (Fuhrmeister et al., 2015).

Despite ongoing plans set by the government, access to improved water and sanitation infrastructure still remains a barrier for the rural population in China. More than 95% of untreated wastewater in rural areas of China is still drained into rivers and lakes (Sun et al., 2008, as cited in Yang & Wright, 2012). Small rural industries contribute significant water pollution, which includes agricultural runoff. Only 1% of villages and towns (approximately 25,000) have wastewater treatment facilities despite significant water pollution from small rural industries (Jiang, 2015).

H. Li (2016) conducted a meta-analysis of literature published since 1980 to determine effective water and sanitation interventions for controlling diarrheal diseases in rural China. Results showed that the four effective water and sanitation interventions for reducing diarrhea in rural China are: (1) improving water supply, (2) building sanitary latrines, (3) implementing multiple interventions at once and (4) promoting health education and behaviour. Firstly, rural water supply has been established as a government goal in the 12th Five Year Plan (2010-2015), which continues to assist with the provision of tap water in rural areas. Secondly, the provision of toilets in rural areas has resulted in an increase in the proportion of rural population with access to hygienic and sanitary latrines; the rate rose from 7.5% in 1993 to 74.1% in 2012, which resulted in a decrease of fecal-oral transmitted diseases. Thirdly, in 2010, the Chinese government established an "Urban-Rural Environment Clean Action Plan" to implement multiple interventions at once, including providing piped water, constructing sanitary latrines, and constructing waste and water treatment facilities in rural areas. Lastly, the Chinese government has launched a series of educational activities to promote awareness of basic health and hygiene to rural residents. The study concludes that multiple water and sanitation interventions must take place concurrently to effectively reduce diarrhea.

M. Wang et al. (2008) claims that the control of rural water pollution is not entirely based on introducing new regulations, policy or acquiring more funding, but lies in revising the current development model used for rural areas. This includes establishing consistent monitoring through a centralized governing body and raising environmental awareness by engaging media and citizens in local government affairs. H. Li et al. (2016) has also suggested increasing population engagement to build a more comprehensive water and sanitation intervention system for preventing diarrhea.

2.4 Theoretical Framework

The post-positivist paradigm is adopted in this thesis to explore several research objectives using observations and measurements. The ontology of this approach states that a truth exists, but will not exist independently of what individuals perceive. This truth can be altered by biases and mechanisms of the social and physical world, making all observations susceptible to fallacy. In health geography, post-positivism seeks to uncover causes, but typically the best that can be achieved is the finding of a strong association between factors.

A post-positivist approach is suitable for this research for several reasons. Firstly, data used in a post-positivist approach is high in repeatability and covers both a wide and uniform representation of the population. This is especially beneficial to study the variation of infectious disease incidence across geographies that cover various regional populations. For instance, Y. Chen et al. (2015) utilized disease incidence data collected from the CDC to analyze shistosomiasis infections across counties in the province of Hubei, China from 2009 to 2013. This incidence data is representative of a population as it is monitored and collected on a constant basis (e.g. daily, weekly, monthly etc.) from each county. Its repeatability allows the

data to be aggregated and studied to assess spatial and temporal trends. Secondly, the post-positivist approach aims to be objective and consistent in primary data collection by adopting a structured and controlled communication medium. Data can be easily sorted because they are often represented using simple categorized value sets such as a numerical value, a rating scale (e.g. 1-10) or a Likert scale (e.g. “Strongly Agree – Strongly Disagree”), and a Boolean response (“yes” or “no”). For instance, categorical data has been effectively used to determine the percentage of rural households that are using household water treatment in villages of China (Junfeng Zhang et al., 2010; H. Yang, 2012).

In health geography, spatial analysis is often used as a post-positivist approach to study how location can affect the health of a population. Maantay (2007) used Geographic Information Systems (GIS) to study the association between air pollution and people who have been hospitalized for asthma in the Bronx area of New York City and how this association is influenced by location. They concluded that people living near sources of air pollution in the city were 66% more likely to be hospitalized for asthma, 33% more likely to be poor, and 13% more likely to be a minority, compared to people living outside the designated buffer zones.

Spatial analysis can also be utilized to illustrate how a disease varies across time, space, and different spatial scales, which can reveal information on the extent of an outbreak, transmission pathways, and associated risk factors (Gondhalekar et al., 2013; Sarkar et al., 2007). This helps to illustrate regional health disparities caused by environment related diseases, which has been poorly documented and mainly concealed at the national level in China (Carlton et al., 2012; Junfeng Zhang et al., 2010). Sarkar et al. (2007) employed spatial cluster analysis of household disease data to investigate a diarrheal outbreak in a small village in Southern India. They found that the disease was distributed uniformly over the village

without any clustering. By overlaying maps containing household information with the location of the water supply system, sewage channels, and areas with observed fecal contamination, it was found that water was distributed via pipes that were placed within soil contaminated with fecal matter.

A concern of the post-positivist approach is that people only appear as “numbers on a map”. These numbers do not provide a personal account of how an individual’s lifestyle has changed as a result of exposure to waterborne diseases, or how rural areas cope with the lack of safely managed water and sanitation. Despite the fact that a post-positivist approach cannot be used to unveil answers to these questions, it can still be used to scope answers to important health issues that occur on a population level. For instance, information on the social determinants of health such as family income, level of education, social support networks, and cultural aspects of a population’s lifestyle can be collected through surveys and national surveys to gauge how socioeconomic conditions facilitate or hinder exposure to a disease on a population level (Arku et al., 2016; Ma et al., 2015; Nie et al., 2014; Odone et al., 2013; J. Zhao et al., 2016). This data can be used to identify knowledge gaps that may require application of other research approaches to understand health factors at the individual and community level.

A significant drawback of a post-positivist approach is that studying social phenomena may not always draw valid conclusions. In general, it is difficult to measure behavior as it varies from person to person. Social behavior is part of a complex system that is often hard to predict and understand. For instance, based on previous observations in China, open defecation for children is common practice in both urban and rural areas, even when toilets and pit latrines are available and can be easily accessed. The limitations of this approach will be further

discussed in Chapter 5.

2.5 Conceptual Framework

A conceptual framework is a system of concepts that supports, verifies, and outlines the process of a research protocol. In this thesis, the conceptual framework is built on the relationships between socioeconomic determinants and environmental policy, which can act as facilitators and barriers to safely managed water and sanitation, as shown in Figure 2.2. This conceptual framework was developed to meet objective two of this thesis, which was to identify the facilitators and barriers to safely managed water and sanitation in the province of Jiangsu, China. This framework is further used to select the socioeconomic indicators required to explore objective four, which is the relationship between socioeconomic determinants and shigellosis incidence.

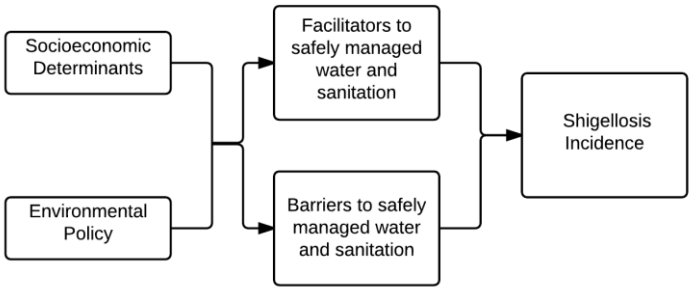

Figure 2.2 Proposed Conceptual Framework

Facilitators are factors that assist with access to safely managed water and sanitation, while barriers refer to factors that hinder that access. This thesis hypothesizes that facilitators and barriers are may be linked to shigellosis incidence. This thesis proposes that facilitators and barriers within this framework can be further categorized into the following themes based on the social determinants of health framework: socioeconomic determinants and

environmental policy implementation (Table 2.2). These two categories define the underlying physical, social, and political structures that affect access to safely managed water and sanitation, which is linked to shigellosis incidence.

Table 2.2 Facilitators and Barriers

	Socioeconomic Determinants	Environmental Policy
Facilitators	• High Income • Employment • Access to Health Services • High Education Attainment • High Rate of School Enrollment • Social Capital	• Effective Rural Water Resource Management • High Political Integrity and Transparency • Regulated Waste Disposal
Barriers	• Low Income • Unemployment • Lack of Access to Health Services • Low Education Attainment • Low Rate of School Enrollment • Lack of Social Capital (i.e. left over children phenomenon)	• Weak Rural Water Resource Management • Poor Political Integrity and Transparency • Poor/Lack of Waste Disposal Regulations

It is recognized that exposure to diarrheal diseases such as shigellosis is influenced by social and economic factors (Adams et al., 2016; Bartram & Cairncross, 2010; Evans & Kantrowitz, 2002; Fotso & Kuate-Defo, 2005; Mock et al., 1993). For instance, it has been found that individuals of high socioeconomic status (SES) are better prevented from contracting diarrhea due to better access to water and sanitation, as opposed to those of low SES (Larson et al., 2007; Woldemicael et al., 2001). This is further emphasized in rural areas of China, where family and village income could dictate a family's access to safely managed water and sanitation (C.Yang et al., 2009; Yi-Xin and Manderson et al., 2005). For instance, Q. Wang and Yang (2016) found that low-income households are particularly susceptible to water pollution due to poor living conditions and the inability to access or afford safe drinking water.

Income may be inversely related to the time and distance taken to access the water source. For instance, wealthy households have private water connections in their homes while less

wealthy household may be required to collect water if they do not have access to piped water within their home. This may involve traveling long distances to the nearest water source. This is a barrier as it is a time consuming task that prevents the fetcher from doing other socially valuable tasks such as working and going to school (Dreibelbis et al., 2011; Graham et al., 2016; Hemson, 2007; Sorenson et al., 2011).

In developing counties, access to water and sanitation is closely linked with the level of education attainment, particularly amongst children (Ortiz-Correa et al., 2016). Hunter et al. (2014) found a strong association between the provision of safely managed drinking water and reduced absenteeism in school. Therefore, it is important to consider how the lack of safely managed water and sanitation could lead to *Shigella* exposure, which could hinder children from attaining an education and maintaining school enrollment.

Education attainment here is defined as the number of years of education a person chooses to complete, and has shown to result in huge influence on a child's future income and welfare (Jing Zhang & Xu, 2016). The level of education is relative to water and sanitation access. For instance, the children of a family may be responsible for fetching water, which is a task that competes with schooling and may lead to an overall reduction in the amount of time the child spends in school.(Jing Zhang & Xu, 2016; Hemson, 2007; Nauges & Strand, 2014). In addition, children fending for themselves in the absence of adults may be responsible for non school-related responsibilities (e.g. such as caring for sick relatives) and thus may struggle with obtaining clean water and finding a place to defecate. These situations create a feedback loop that links lower education attainment with poor access to water and sanitation. When children are not educated on sanitation and hygiene, there is a higher likelihood for them to contract waterborne diseases since they are not aware of water consumption safety and

hygienic practices.

Water access is not only related to children's education attainment, but also that of teens. A study on water access and quality in rural China found that a significant decline in school enrollment is witnessed amongst girls from households lacking in access to sufficient water after they have reached menarche (Maimaiti & Siebert, 2009). Jing Zhang and Xu (2016) also found that an increase in access to treated water could greatly benefit girls in terms of school attainment and eliminate the gender disparity in school attainment in rural areas of China. Their study found that young people with access to treated water had better education than those without. Given access to treated water, they found that the completed grade of education amongst teens increased by 1.1 years on average.

For certain individuals such as children and the elderly, gaining access to safely managed water and sanitation can be difficult. Many of them lack a social support network (Murphy et al., 2016; Lu et al., 2016; X. Zhao et al., 2014) that hinders them from accessing safely managed water and sanitation. This is often attributed to the migration of working members of the family to the city, leaving behind the elderly and children to fend for themselves (Biao, 2007). This includes "left-behind children", which is a cultural phenomenon where children are left behind in rural areas to live with their relatives. Often times, these children live with their grandparents that have little to no education, financial support, or the physical ability to look after themselves and the child (Lu et al., 2016).

A study conducted by Wen & Lin (2012) found that the lack of social capital in the family of left-behind children have left the children with weak social ties to the people in their community and thus can negatively impact school engagement and performance. According to

the CPC town secretary in Wangji, approximately 1/3 of the households have left-behind children. While Murphy et al. (2016) have found that left-behind children is associated with barriers to health and wellbeing, it is unclear whether left-behind children are associated with a lower education attainment that contributes to lack of awareness of safely managed water or sanitation.

Children can also be exposed to *Shigella* from oral contact with garbage and contaminated food scraps. Tang et al. (2014) claims that unregulated waste disposal is a factor that exacerbates *Shigella* transmissions in Jiangsu. Rego et al. (2005) also found that children exposed to garbage in their surrounding environment were four times more likely to contract and develop diarrhea. Children playing with garbage may contract diarrhea through oral contact due the lack of hygiene awareness and practices

For sustainable and long-term prevention of diarrhea in rural China, water quality must be regulated and monitored. In China, environmental legislations are still undermined by the low bureaucratic status of environmental regulatory bodies. Rural water resource management in China is dependent on local water quality monitoring agencies which face difficulty enforcing the "polluter pays" principle and securing active participation and cooperation among other governmental agencies and industries (Jiang, 2015; Jahiel, 1998; M. Wang et al., 2008). In addition to weak authoritative influence, environmental regulators also lack human and financial resources to implement effective water resource management strategies; this includes the lack of technical support, training resources, and funding to purchase new technologies (Xiaoman Yu et al., 2015).

The socioeconomic determinants that act as facilitators and barriers presented in this

section lead the research direction of this thesis. It is important to understand how these factors affect access to safely managed water and sanitation. The next section will explore how these factors can be explored and utilized to meet the main research objectives of this thesis.

2.6 Methodological Literature

In order to study the spatial distribution of shigellosis and analyze the socioeconomic determinants that act as facilitators and barriers, quantitative methods are adopted. In health geography, quantitative methods involve numerical measurements and often consist of three stages: (1) mapping for visualization of data, (2) graphical and statistical analysis for exploratory spatial data analysis, and (3) modeling to explore relationships (Gatrell and Elliott, 2015).

2.6.1 Spatial Data Visualization

Within a predefined geographic region, observations are often collected in the form of arbitrary geographic units also known as small areas (Lawson & Corberán-Vallet, 2016). These areal units can be in the form of postal zones, census tracts, or larger units such as municipalities, counties, province, or country. Health studies collect data on disease incidence often in the form of counts.

Mapping the geographical distribution of an infectious disease is the first stage, and can allow visualization of the extent and magnitude of an infectious disease outbreak (Hay et al., 2013). This can be done through choropleth maps that can be produced using standard desktop mapping and GIS software (Fotheringham, 2007; Fradelos et al., 2014). Choropleth maps are commonly used for illustrating the spatial distribution of aggregated disease data. They can effectively visualize aggregated disease counts across space (Pfeiffer, 2008). Despite the usage

of choropleth maps in this study, it should be noted that choropleth maps have three inherent limitations. First, large feature polygons of the study area tend to dominate and introduce bias in interpretation (Monmonier and De Blij, 1996). Second, the modifiable areal unit problem (MAUP) is evident as analysis may change with the shape and size of the aggregation unit (Openshaw, 1984). Last but not least, the highly skewed distribution of infectious disease count data is difficult to effectively visualize using a finite number of colour shade categories. However,

2.6.2 Spatial Analysis of Infectious Disease

Mapped data can be assessed using spatial and statistical analysis. According to Tobler (1970), "everything is related to everything else, but near things are more related than distant things". In studies where data corresponds to areal units, the assumption of independence is not likely satisfied. According to O'Sullivan and Unwin (2010), a test for autocorrelation should always be carried out before any theories are developed for the patterns that are observed in map. Thus, Tests for spatial autocorrelation should be conducted to identify spatial dependence between disease incidences at different locations.

The Global Moran's I is a measure of overall spatial autocorrelation that has been commonly applied in infectious disease literature (Gu et al., 2017; Guo et al., 2017; Liao et al., 2016; Y. Chen et al., 2015). The Global Moran's I is used to identify whether an infectious disease is clustered or dispersed. For example, both Gu et al. (2017) and Guo et al. (2017) adopted Global Moran's I as a preliminary step of spatial analysis to investigate whether there is any spatial clustering of paratyphoid fevers and tuberculosis in the first place, such that they have been aggregated in a certain part of an area.

Since the Global Moran's I can only provide a summary of spatial clustering and cannot identify local spatial dependence, the Anselin's Local Indicator of Spatial Autocorrelation (LISA) and Getis Ord Gi* statistics can be employed to further investigate the local relationship between a feature and its neighbours (Cromely & McLafferty, 2011; Pffeifer et al., 2008). Anselin (1995) stated the Local Moran's I is proportional to the global statistic of spatial association and can make important inferences on the local instability as a measure of the global statistic. Getis and Ord (1992) recommended applying both Moran's I and Getis Ord Gi* as they take different measure (global versus local) and may point to different drivers that contribute to the spatial distribution of disease.

While both the Local Moran's I and Getis Ord Gi* are local indicators of spatial association (LISA), these two statistics measure different concepts of spatial association (Anselin, 1995). These two measures are cluster detection methods; Anselin's LISA is used to detect spatial clusters of similar and dissimilar values, while Getis Ord Gi is used to identify whether clusters are "hot spots" or "cold spots". It is important to investigate spatial clustering in both time and space as clustering over time may be due to the infectiousness of the disease or environmental hazards (Marshall et al., 2001).

Spatial autocorrelation methods such as the Global Moran's I, Anselin's local indicator of spatial autocorrelation (LISA), and Local Getis Ord Gi statistics have been commonly applied in conjunction to identify infectious disease clusters and hot spots. Y. Chen et al. (2015) used Global Moran's I and Local Getis-Ord statistics to determine whether shistosomiasis cases were clustered, dispersed, or randomly distributed, in Hubei, China. Phung et al. (2015) used Anselin's local indicators of spatial autocorrelation (LISA) in addition to Global Moran's I to determine spatial clusters of diarrhea in the Mekong Delta area of Vietnam. Bayles and

Allan (2014) also used both Anselin's LISA and Global Moran's I to identify high incidence clusters of tick borne disease ehrlichiosis in Missouri, United States. Thus, both methods of spatial autocorrelation and cluster detection should be applied after initial data visualization to explore spatial patterns.

The spatial distribution of shigellosis in China during years before 2011 has been widely studied using spatial analysis methods. The Global Moran's I and LISA (Anselin and Getis Ord Gi) have been the most common indicators used to identify spatial clustering of shigellosis in Jiangsu (Ma et al., 2015; Nie et al., 2014; Tang et al., 2014). These indicators were used to detect spatial clustering of shigellosis cases at the county level. However, no studies have explored the spatial patterns of shigellosis incidence in Jiangsu after 2011.

To a lesser extent, few studies have conducted spatial analysis to understand the association between shigellosis and socioeconomic determinants in China at a small area level. Both Nie et al. (2014) and Ma et al. (2015) studied the correlation of shigellosis incidence rates and socioeconomic factors using the spatial autocorrelation indicator Moran's I. Nie et al. (2014) used Global Moran's I for multivariate spatial autocorrelation while Ma et al. (2015) used Bivariate Moran's I. Both studies analyzed similar socioeconomic indicators such as the proportion of working industries (primary, secondary, tertiary), GDP, percentage of illiterate population in total population under aged 15, popularization rate of tap water in rural area, access to sanitation toilets in rural area at the county level. The Bivariate Moran's I between shigellosis incidence and each variable was done by (1) establishing a corresponding weight matrix, followed by (2) a visual analysis to obtain information on the distribution, and (3) a measurement of the spatial correlation between shigellosis incidence rate and socioeconomic variables using the GeoDa software (Version 1.3.28). However these studies did not explore

other rural indicators like the percentage of rural households and rural employment.

2.6.3 Multivariate Regression Models for Small area Incidence

Modeling small area incidence

In health geography, relationships between various factors can be explored quantitatively through a model. To explore the relationship between shigellosis incidence and socioeconomic determinants, multivariate regression models can be adopted to identify significant factors associated with the infectious disease. To study the associations between exposure to infectious diseases and socioeconomic determinants on a population across space, studies have adopted spatial models, such as the Bayesian spatial (Chitunhu & Musenge, 2016; Wilking et al., 2012) and spatiotemporal models (Arku et al., 2016; Ma et al., 2015; Waller et al., 2012) and geographically weighted regression models (GWR) (Delmelle et al., 2016); and non-spatial models, such as logistic regression models (Fuller et al., 2014; Roka et al., 2012; J. Zhao et al., 2016) and generalized linear regression models such as Poisson (Weinberger et al., 2013; Thompson et al., 2015).

At small area levels, disease incidence reported as positive discrete values per unit area makes disease incidence a count data variable based on the Poisson distribution (Cameron & Trivedi, 2013; Myers et al., 2010). The Poisson distribution has been used to describe the distribution of infectious disease counts (Qian et al., 2010; Kleinmen et al., 2004 as cited in Unkel et al., 2012) and is represented mathematically as

$$F(y_i|\mathbf{x_i}) = \Pr = \frac{{u_i}^{y_i} e^{-u_i}}{y_i!}, \qquad y_i = 0,1,2,\ldots \tag{2.1}$$

$$\mathrm{E}(\mathrm{y_i}|\mathbf{x_i}) = u_i \tag{2.2}$$

$$V(y_i|\mathbf{x_i}) = u_i \quad (2.3)$$

where a discrete random variable y_i follows a probabilistic distribution with a mean and variance of u_i. The essential criteria of the regression model is based on the following assumptions of the Poisson distribution:

1. Homogeneity: the mean rate at which events occur is the same.
2. Independence: one case does not affect the probability of a second case.
3. Fixed time period.

The Poisson distribution is notable for the first criterion, which is a restrictive property that requires the mean and variance to be the same. This criterion is rarely met when modeling infectious diseases. Criterion two is not met in non-spatial models because of the spatial clustering characteristic of infectious disease; the incidence of infectious disease will be similar for small areas close to each other. Lastly, criterion three is satisfied when the time is fixed in the model.

Poisson Model

The Poisson distribution has been used to model the distribution of infectious diseases at small area level. Shekar et al. (2016) assumed malaria incidence in the small area of Kalaburagi, India (consisted of 139 spatial units - 138 villages and one city) followed a Poisson process. Azage et al (2015) applied the Poisson distribution to study childhood diarrhea incidence across the 33 districts of the Awi, East Gojjam, and West Gojjam areas of Amhara state Ethiopia. Jones et al. (2012) assumed a Poisson distribution for studying the distribution of tick-borne and mosquito-borne diseases across 95 counties within Tennessee, United States. All three studies performed their analysis using the SaTScan software, which "uses a Poisson-

based model where the number of events in a geographical area is Poisson distributed" (SaTScan, 2005).

In addition, Poisson regression models have been frequently applied to study small area infectious disease incidence across time and space. Yeshiwondim et al. (2009) used Poisson regression to model the spatiotemporal transmission of daily individual malaria incidence from 2002 to 2006 using demographic indicators across 543 villages in East Shoa, central Ethiopia. Weinberger et al. (2013) applied a multivariate Poisson regression model to investigate the effects demographic characteristics including the degree of urbanization on annual incidence of non-typhoid *Salmonella* (NTS) reported per sub-district from 1996 to 2007 across 15 administrative sub-districts in Israel. Thompson et al. (2015) developed a multivariate Poisson regression model to study the effects of climate conditions on monthly diarrhea incidence rate reported per district from 2005 to 2010 across the 24 districts of Ho Chi Minh City in Vietnam. In addition, Xu et al. (2016) applied Poisson regression to explore the relationship between temperature variability and daily incidence count per district of infectious child hand, foot, and mouth disease from 2012 to 2014 across six districts in Huainan city China. Overall, Poisson regressions have been effectively applied to study small area incidence reported at various administrative levels.

Two studies have used Poisson regression to investigate the relations between socioeconomic indicators and diarrhea. Simonsen et al., (2008) used longitudinal data in a Poisson regression model to predict incidence rate ratios of various diarrheal diseases (including *Salmonella* and *Shigella*) for different socioeconomic groups in Denmark, while Colombara et al. (2013) used hospital diarrhea surveillance records in a multivariate Poisson regression model to study the association between various socio-demographic determinants

and cholera burden among children under the age of five in rural Bangladesh. Wilking et al. (2012) used a spatial Bayesian Poisson regression model to study the association between rotavirus incidence and socio-demographic and economic variables in 447 neighbourhoods of Berlin, Germany. However, the association between socioeconomic determinants and infectious diseases in rural settings and how they vary across geography at small area levels are still less understood (Odone et al., 2013).

Quasi-Poisson Model

The application of the Poisson regression model to study spatial infectious disease patterns is limited due to the spatial clustering characteristic of infectious diseases, which leads to variability in the number of cases across geography. The Poisson model neglects overdispersion, which would underestimate the standard error and increases the probability of obtaining a false positive result (Hinde and Demétrio, 1998 as cited in Liao et al., 2016). Thus the assumption of the Poisson model, which states that the mean is equal to the variance, is rarely satisfied. This results in an extra-Poisson behaviour known as overdispersion, which is common amongst infectious disease counts as they have a variance above the expected value. This can be adjusted through a dispersion parameter that adjusts the regression variance or standard errors. Farrington et al. (1996) established the log-linear model regression model known as the quasi-Poisson model, which consists of a dispersion parameter that accounts for overdispersion, which is discussed in more detail in Chapter 4. The quasi-Poisson model by Farrington et al. is now used routinely by the Health Protection Agency to detect infectious disease outbreaks using laboratory based surveillance data in the United Kingdom (Unkel et al., 2012).

Studies have referenced Farrington's quasi-Poisson regression model when dealing with

infectious disease counts. Green et al. (2006) accounted for overdispersion by adjusting the variance via a dispersion parameter when modeling the association between socio-demographic, landscape characteristics and campylobacteriosis (an infectious foodborne disease that is characterized by bloody diarrhea) across 498 neighbourhoods in the Canadian province of Manitoba from 1996 to 2004. Yupiana et al. (2010) also applied a dispersion parameter to the variance when studying the risk factors of poultry outbreaks and H5N1 avian influenza at the small area level 25 districts of West Java Province in Indonesia. Y. Li et al. (2013) applied a dispersion parameter to the standard errors of the regression coefficients when modeling the association between risk factors and overdispersed human brucellosis incidence at the province level for four provinces with high incidence. In summary, the application of the dispersion parameter to the variance or standard deviation has been very commonly adopted when applying the quasi-Poisson model.

Negative binomial model

Overdispersion can also be adjusted via a negative binomial model, which according to Held et al (2005), is a more flexible model that allows for "overdispersion caused by the influence of unobserved covariates that affect the disease incidence". The negative binomial model is also known as the "gamma-Poisson" model, as it assumes that the number if events occur follows a Poisson distribution, but the mean number of events is not the same. Instead, the mean number of events that occur follows a gamma distribution in the population (Land et al., 1996). The negative binomial is advantageous due to its ability to model count data with varying degrees of overdispersion (Lloyd-Smth, 2007).

The application of negative binomial regression to adjust for overdispersion at small area incidence has been explored in the following three studies. A study by Coelho et al. (2016)

adopted a negative binomial regression model after observing overdispersion using a Poisson regression to model Zika and dengue incidence (aggregated by age classes) in Rio De Janeiro, Brazil. In addition, Fornace et al. (2016) adopted the negative binomial model at the village level (n=405) to study the association between landscape factors and spatial patterns of *Plasmodium knowlesi*, the main cause of human malaria in Malaysian Borneo. Both studies started with the fitting of the Poisson model and detection for overdispersion prior to fitting using the negative binomial model.

Last but not least, Hughes and Gorton (2015) experimented with three GLM (Poisson, Zero-Inflated Poisson, and Negative Binomial) and was the only study to experiment with all three models for small area incidence. Their study explored the association between the Multiple Index of Deprivation and incidence of 21 infectious diseases in small areas of North East England from 2007 to 2011. The best fitting model was selected from a hierarchical approach that started with 1) fit of a Poisson model, 2) fit of zero-inflated Poisson (ZIP) model, 3) fit of negative binomial model, and 4) zero-inflated negative binomial model. From their experiments, the best fitting model was selected for each disease by assessing the Akaike's Information Criteria (AIC) where an AIC > 2.5 was considered a significant improvement in model fit. Negative Binomial model was found to be the suitable model for modeling 13 out of the 21 infectious diseases. Of the remaining eight diseases, three were modeled using the Poisson model while five were modeled using the zero-inflated Poisson model.

Bayesian spatial models

When working with spatial data, it is critical to incorporate the spatial structure of the data into the regression model. Recent literature has shown that a geographic weighted regression (GWR) model can also be applied to Poisson and negative binomial so that spatial

relationships between variables over space are considered. A GWR is a local regression method that identifies the non-stationary relationships of variables for each feature via a moving window or kernel (Nakaya et al. 2005). For each feature in a dataset, a different regression model with different coefficients is fitted into the dataset.

Weisent et al. (2012) used a Poisson and negative binomial GWR to study how socioeconomic factors can be used to determine the risk of campylobacteriosis, a gastroenteritis disease. Delmelle et al. (2016) also used a GWR model to study the associations between socioeconomic and environmental determinants and dengue fever. However, in order to apply geographic weighted regression, there must be sufficient geocoded data so that spatial weights can be computed and applied appropriately. In both studies conducted by Weisent et al. (2012) and Delmelle et al. (2016), the number of cases were 3,756 and 9,287, respectively, and were analyzed at the neighbourhood level. Due to the small sample size used in this thesis, the GWR method was not considered a viable option.

Another method to incorporate spatial relationships between neighbouring areal units is through Bayesian methods, which can be combined with generalized linear models. Bayesian methods, unlike frequentist methods, are based on the idea that only one form of uncertainty exists and that this uncertainty can be described by a probability distribution. The uncertainty prior to the introduction of new information is described by a *prior* distribution. The process of making an inference using this model consists of combining the prior and current data model to derive the posterior distribution, which provides information on the hypothesis. According to Blangiardo et al. (2013), there are two advantages of this approach. Firstly, this method will be inclusive of previous information on the model (prior distribution). Secondly, the

likelihood that a predictor variable will exceed its threshold under certain conditions can be easily derived from the posterior distribution and can be interpreted in a relative risk context, as opposed to a p-value used in the frequentist approach.

The Bayesian generalized linear regression model is effective for small area disease modeling because it acknowledges the spatial effects between areas. The Bayesian model consists of a spatial unstructured residual that accounts for the clustering nature of infectious diseases, and a spatial structured random residual that accounts for spatial dependence between areas. This has been selected as a model of choice by several studies that focused on small areas and the association between socioeconomic variables and infectious disease incidence. For instance, Wilking et al. (2012) applied a spatial Bayesian Poisson regression model to study rotavirus incidence = across 447 neighbourhoods in Berlin, Germany from 2007 to 2009. Chitunhu and Musenge (2016) applied a similar Bayesian Poisson spatial model to study malaria incidence in 140 areal clusters across Malawi in 2012, while Wijayanti et al. (2016) applied a Bayesian Poisson model to study dengue incidence across 329 villages in Java, Indonesia. This model has also been adopted similarly by Kara-giannis-Voules et al. (2013) and Bessell et al. (2010) to study small area incidence. Overall, the Bayesian Poisson model has been the most common Bayesian model used to account for spatial effects.

However, very few studies have adopted a Bayesian negative binomial model despite the common occurrence of overdispersion in infectious disease incidence. The only study found in this literature search that adopted a Bayesian spatial-temporal negative binomial model was Liao et al. (2016). They adopted both a Bayesian Poisson spatio-temporal model and a Bayesian negative binomial model, which consisted of an additional time parameter, to study hand, foot, and mouth disease incidence across the 135 counties in Sichuan province of China

from 2009 to 2013. They found that the Poisson model estimated a wider 95% coverage interval for posterior means estimation of the parameters and was extremely sensitive to changes in the hyper criteria of the priors used to model the random structured and unstructured effects. In general, the authors claim that the Bayes spatio-temporal model is more flexible than maximum likelihood estimation methods, which is sensitive to population size and spatial effects. In addition, using Bayesian methods allows the consideration of spatial trends between geographically close areas and prior disease rates, which is beneficial for small samples of spatially correlated data. The adoption of a Bayesian negative binomial model to study overdispersed incidence has been proven to be better than the Poisson model, but should be further explored to understand the extent of its application.

Model Limitations

Using negative binomial regression on small sample sizes may underestimate the degree of overdispersion in the data, as reported in previous studies (Lloyd-Smith, 2007; Saha & Paul, 2005). According to Lolyd-Smith (2007), the sample sizes of datasets should be N=100 or more to allow accurate maximum likelihood estimates of the dispersion parameter, while a sample size of N=30 will result in minimal bias with a sampling distribution that tends to skew towards to high values. A sample size N=10 would be the least feasible as it would yield unreliable estimates. In this thesis, the number of counties of interest is greater than N=30, which makes this model viable.

According to Marshall et al. (1991), there are several issues with the statistical analysis of spatial disease patterns. In particular, this occurs during the stage of modeling and interpreting the association between the predictor and response variables. Firstly, there is the

issue of ecological fallacy. Coefficient parameters in the regression model can change drastically based on the geographic scale of aggregation, thus coefficients are subject to spatial bias. Secondly, the assumption of spatial independence is often challenged due to the proximity of small areas next to each other. During the fitting of a regression model, this may result in a spatial correlation between residuals. Thus, spatial autocorrelation should be tested and considered when interpreting the results of the regression model.

Despite these limitations, the value of assessing the association between factors for spatial based studies lies in understanding the impact of processes and structures of social organizations in determining health outcomes. The availability of surveillance data means they can be quickly used and are easily accessible, in particular for research on small areas as surveillance data is often available at various administrative divisions (e.g. census tracts). In addition, applying spatial or non-spatial regression is effective at identifying the effects underlying social and economic factors of a particular space. From a health service planning perspective, spatial analysis is useful for identifying areas of poor health.

In summary, due to the clustering nature of infectious disease incidence across geography, an overdispersed Poisson regression model such as the quasi-Poisson or the negative binomial model should be adopted. It is important to acknowledge the limitations of non-spatial regression models due to the spatial autocorrelation of infectious disease data. Thus, a spatial Bayesian model should be adopted to incorporate the spatial relationships between areas. There have only been a few studies that have adopted a non-spatial or spatial negative binomial regression model to study the association between socioeconomic factors and infectious disease incidence at a small area level. This thesis will contribute to existing

literature on the application of negative binomial regression models for modeling small area incidence across time and space. This contribution will be further discussed in Chapter 5.

2.7 Chapter Summary

This chapter examined the theoretical, conceptual, and methodological frameworks that were used to construct this thesis. The disease characteristics of shigellosis and prevention of shigellosis through water supply and sanitation were described. In addition, a review of water and sanitation in rural China was provided. Next, the post-positivist theory and the facilitators and barriers to safely managed water and sanitation, which falls within the social determinants of health framework, were described and examined in the context of this thesis. Lastly, quantitative methodological literature on spatial analysis and regression analyses for infectious diseases were examined and reviewed. This literature review identified (1) the socioeconomic determinants that act as facilitators and barriers to safely managed water and sanitation (2) spatial approaches that can be used to identify spatiotemporal disease patterns and (3) quantitative approaches that can be applied to study the association between socioeconomic determinants and shigellosis incidence. Facilitators and barriers identified through the conceptual framework set the basis for the data used in quantitative analysis to satisfy objective four. The next chapter will discuss how quantitative approaches are employed to meet the objectives of this thesis.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter describes the methodology used to address the objectives introduced in chapter 1:

1) To examine spatiotemporal variation of shigellosis incidence across Jiangsu province
2) To identify the facilitators and barriers to safely managed water and sanitation
3) To investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province

As summarized in Chapter 2, quantitative methods based on the post-positivist approach have been adopted to address all three objectives in this thesis. The facilitators and barriers identified in Chapter 2 were also use to set the basis for quantitative analysis of socioeconomic determinants and shigellosis incidence. These methods are described in the following sections of this chapter.

3.2 Study Area

Jiangsu is a province in eastern China located between 116°18' -121°57' E and 30°45' – 35°20' N with an estimated 2014 population of 79.8 million people and an area of 107 200 km^2 (Jiangsu Statistical Yearbook, 2015). Jiangsu is one of the densest provinces in China with a population density of 742 people per km^2. It was chosen the study area of interest as it has been previously identified as a province with regions of high shigellosis morbidity (Tang et al., 2014). The regions in Jiangsu province are organized into thirteen prefecture level cities as shown in Figure 3.1. At the county level[1], these cities are further divided into 96 administrative divisions that correspond to 21 county cities (i.e. the merge of a "city" and a "county" into one

[1] The county level refers to the third administrative division after prefecture (second) and province (first).

unified administrative division under a prefecture), 20 counties, and 55 districts (i.e. subdivisions of a municipality or prefecture level city). Counties and districts are the finest scale at which the province reports its socioeconomic and demographic information.

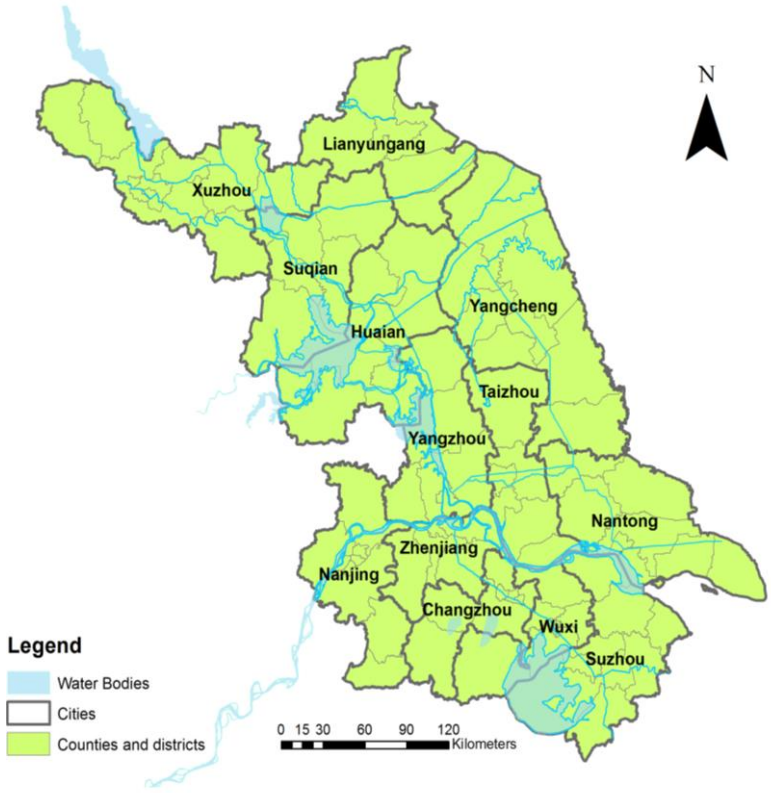

Figure 3.1 Thirteen prefecture level cities in Jiangsu province with their county and district boundaries

Jiangsu has distinct seasons characterized by warm climate and moderate rainfall. The climate is influenced by subtropical monsoons. The topography of Jiangsu consists of low and flat terrains surrounded by lakes and rivers. As part of the Yangtze River Delta, the Yangtze River flows through the province from west to east for more than 400 kilometers.

To observe and understand water and sanitation conditions in the field of a rural county IN Jiangsu, a field visit to two townships in Suining County, Wangji and Qingan, was

conducted on August 16. 2017. Suining (33.9 N, 117.9 E) is a rural county within Xuzhou city located in northwestern Jiangsu. It has a total 2014 population of 1.44 million people and a land area of 1 767 km^2. Suining was previously identified as a county with relatively low access to improved water sources and sanitation facilities (Tang et al, 2014). According to the township CPC secretary (magistrate), the township of Wangji mainly depended on hand pump wells for drinking water in the past (Figure 3.2). Today, many households have a jet pump well, which draws water from 20 m under the ground (Figure 3.3 a). A jet pump is powered by electricity and consists of a centrifugal pump and an injector. The injector creates a vacuum, which draws water from a nearby well and discharges the water via a faucet (Figure 3.3 b).

Figure 3.2 One of the few hand pumps left in Wangji Township

(a) (b)

Figure 3.3 (a) Jet pump draws water from underground (b) Discharge faucet

Wangji was in the process of installing a centralized tap water system that will transport water from the county of Suining. Since 2014, 56% of residents have access to tap water (Suining Statistics Bureau, 2016).

In Wangji, each household had a home toilet facility outside of their home (Figure 3.4 a). The typical toilet was a pour flush pit latrine (Figure 3.4 b). Outside the toilet facility, human wastes from the pit latrine would be compressed into manure and later used as a fertilizer (Figure 3.4 c). Human feces were not treated prior to being converted into manure. When asked why a flush toilet was not employed, the township secretary said flush toilets were too costly to be installed. Aside from an outdoor toilet, wealthier households also had a shower facility indoors that was powered by solar energy.

Figure 3.4 (a) Typical location of sanitation facility (b) Pour flush using buck pit latrine (c) Manure from human feces

Approximately 20 km East of Wangji is the township of Qingan. The township of Qingan has already adopted a centralized tap water system. Since 2014, 99% of residents in Qingan have access to tap water (Suining Statistics Bureau, 2016).

Figure 3.5 Qingan reservoir

While drinking water was still sourced from groundwater, a portion of the drinking water has been sourced from the surface water stored in the Qingan Reservoir since 2015 (Figure 3.5). The local government has prohibited residents from drinking water from the reservoir or using it for recreational activities.

3.3 Data

3.3.1 Data Sources

The facilitators and barriers to safely managed water and sanitation identified through the conceptual framework in Chapter 2 will set the basis for the data used in quantitative analysis to satisfy objective four, which aims to investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province. Due to data availability, the socioeconomic determinants that will be investigated in this thesis are highlighted in Table 3.1.

Table 3.1 Socioeconomic determinants of health framework

Socioeconomic Determinants	Environmental Policy
• **Income** • **Employment** • **Access to Health Services** • Education Attainment • School Enrollment • Social Capital	• Effective Rural Water Resource Management • High Political Integrity and Transparency • Regulated Waste Disposal

To collect data on the highlighted factors, this research employed four sources of data, namely (1) number of shigellosis cases at county and district level, (2) socioeconomic datasets from the 2011-2014 provincial statistical yearbooks and county statistical yearbooks, and (3) data on water and sanitation conditions from the 2011-2014 county level statistical yearbooks and (4) field observations of rural water and sanitation facilities.

Spatial Data Sources

Cartographic boundary files of cities, counties, and districts in Jiangsu were obtained from the School of Geographic and Oceanic Sciences at Nanjing University. Cartographic information on geographic features such as the spatial location and boundaries of primary rivers, secondary rivers, and water bodies were also obtained.

Shigellosis and Socioeconomic Factors Data Sources

Shigellosis data was collected from the Jiangsu Provincial Center for Disease Control and Prevention (CDC). The CDC mandates clinical and hospital doctors to report all shigellosis cases. All cases of shigellosis are diagnosed in the laboratory by the extraction of the *Shigella* strain from the stool specimen or rectal swab. The collected dataset covered the period from 2011 to 2015. The number of new shigellosis cases per year (incidence) and the number of new cases per 100, 000 persons (incidence rate) were available in aggregated count per year for every prefecture level city, city county, county, and district. Age and gender were not disclosed in the dataset. This poses two limitations: (1) these factors were not considered in the analysis and (2) available disease data was not standardized by age or gender.

Socioeconomic data was collected from the provincial Jiangsu Statistical Yearbook and city statistical yearbooks. Data on demographic and socioeconomic indicators for each city, city county, and county were collected from the provincial statistical yearbook. Due to missing and overlapping data in the provincial statistical yearbook, additional data was collected from individual city statistical yearbooks to cover additional socioeconomic indicators and information not covered in the provincial statistical yearbook. Data on rural households and rural income was obtained from the city level statistical yearbooks.

Collected socioeconomic data was available as aggregated counts per city, city county, or county, as shown in Table 3.2. In this context, the spatial unit of cities only represents the socioeconomic data on city districts, since data on each individual district was not available. In addition, socioeconomic data for Nanjing was collected at the prefecture city level since data was not available for its city districts. Despite these discrepancies, the aggregated form of the collected data is still representative of socioeconomic conditions in those regions and can reasonably illustrate its spatial distribution. Each socioeconomic indicator is used as a predictor variable for the regression analysis that explores the association between socioeconomic determinants and shigellosis incidence.

Selected demographic indicators included permanent population, registered population, population density, total households by year-end, and total rural households. These indicators were included because they provide information on the general population (used to calculate the proportion of rural households), rural population, and the number of households, which will be used to determine the proportion of the rural population.

Socioeconomic indicators were selected based on the key socioeconomic determinants of health (Public Health Agency of Canada, 2011a) hypothesized to act as facilitators and barriers to safely managed water and sanitation. This included income (rural income per capita), rural employment (rural employed persons), and access to health care (number of health institutions and beds). While data on the level of education amongst the population was unavailable in the provincial statistical yearbooks, data on school enrollment was available. However this data was not available in a format that allowed meaningful comparisons between counties to be made. Thus, it was omitted from the analysis.

Table 3.2. Socioeconomic, demographic, and water and sanitation data

Data	Definition	Year	Reported Unit for County
Registered Population	Number of people registered to live during a certain time within a given area.	2011-2014	Count
Permanent Population	Number of people permanently living within a given area. Represents total population in a given area.	2011-2014	Count
Population Density	Number of persons per city county/county/district.	2011-2014	Persons/ Km^2
Total Households by Year End	Total number of households in counties by the end of the year.	2011-2014	Count
Total Rural Households	Total number of households in townships under the jurisdiction of counties.	2011-2014	Count
Rural Employed Persons	Total rural labour force during a certain period of time within a county that receive remuneration.	2011-2014	Count
Rural Income per Capita	Total income of permanent residents living in rural households after the deduction of all expenses including tax.	2011-2014	Yuan/Capita
No. of Health Institutions	Number of hospitals (city and county), neighbourhood medical services stations, clinics (urban and rural), first aid centres, disease prevention and control centres, and health education centres.	2011-2014	Count
No. of Beds in Health Institutions	Number of beds available in applicable health institutions listed above.	2011-2014	Count
Households with access to tap water	Number of households in township with access to piped water	2011-2014	Count
Private vs Public Water Access	Number of households with access to 1) private water source and 2) shared water source, respectively.	2011-2013	Count (out of 100 samples)
Type of sanitation facilities	Toilet and shower and unimproved sanitation: no sanitation facility, publicly shared toilet)	2011-2013	Count (out of 100 samples)

According to the 2014 Jiangsu Statistical Yearbook, the Rural Population indicator refers to the population not residing in cities or towns, which is calculated as part of the urban population. Rural Households refer to households residing at a rural address under the township administration for more than one year. Families that are part of a rural household but

have moved away are not included in the calculation. This indicator was collected to gauge the number of people living in rural areas. Rural Employed Persons refer to persons that receive remuneration for their profession. This includes persons who work as employees, employers, self-employed workers, and teachers for religious services. This indicator provides information on the total rural labour force of China. Rural Income is reported from two indicators: Net Rural Income and Rural Disposable Income. Net income refers to the total income of the rural household after subtracting all expenses including taxes and household operation costs. Rural Disposal Income refers to the income at the disposal of the rural person, whether for personal consumption, expenditure, or savings. Health institutions include hospitals (city and county), neighbourhood medical services stations, clinics (urban and rural), first aid centres, disease prevention and control centres, and health education centres.

3.3.2 Data Preprocessing and Geocoding

Spatial Visualization

Initial geospatial data exploration and preprocessing were performed in Microsoft Excel (Vers. 14.5.5) and ArcMap (Vers. 10.4.1). Shigellosis incidence rates for each year (n=96) from 2011 to 2015 were geocoded to their respective counties (n=41) and districts (n=55) based on division codes. Rates that were not geocoded were either missing or duplicates. Missing cases and rates were not left blank; instead, any missing data was estimated using multiple missing data imputation.

Multivariate Generalized Linear Regression Variables

Data on socioeconomic indicators and water and sanitation indicators were preprocessed in Excel and RStudio (Vers.1.0.136). Each socioeconomic indicator was geocoded to its smallest available administrative division, such as city, city-county, county, or district by

administrative division codes (Table 3.3). This discrepancy in geographic scale was caused by the lack of available data on individual counties and districts within certain cities. Thus, certain districts were aggregated to the city level due the lack of data available for individual districts. In order to ensure the rate of shigellosis is accurately portrayed per geographic area, an offset value equal to the population of each geographic area was incorporated.

Table 3.3 Sample size of socioeconomic indicators by administrative division type

Year	Total (N)	City(n)	City-county(n)	County(n)	District(n)
2011	60	13	20	21	6
2012	59	13	20	21	5
2013	58	13	20	21	4
2014	57	13	20	21	3

For every city, city-county, and county in the dataset, shigellosis incidence rate was matched with the data on each socioeconomic indicator. This resulted in N = 60 (2011), N=59 (2012), N=58 (2013), and N= 57 (2014) for the multivariate generalized linear regression analysis. For the Bayesian spatial model, values were taken by averaging the dataset of each variable from all four years.

Variable Unit Conversions

Unit conversions were performed on socioeconomic and water and sanitation data so that data trends can be more effectively understood and illustrated during analysis. To analyze the distribution of the rural population, the percentage of rural households was calculated by dividing total rural households from total households by year-end. In addition, the percentage of rural employed was calculated by dividing the total number of persons employed by the permanent population. For comparison purposes, data on healthcare was converted to per capita values. The number of health institutions was converted to number of health institutions per 10,000 people by dividing the variable by its corresponding permanent population then

multiplying the number by 10,000. The number of hospital beds was converted to number of hospital beds per 1,000 people by dividing the variable by its corresponding permanent population then multiplying the number by 1,000. Missing data was computed using multiple missing data imputation. Tap water access, which was initially reported as a count, was converted to a rate by dividing the number of households with access to tap water from the total number of households in each respective town. Private and shared water access, which was presented as a rate (count per 100 households) respectively, was converted to a count based on the total number of households. This was calculated by multiplying the rate by total number of households. Similarly, this conversion was also done for improved and unimproved sanitation facilities, which were initially reported as rates. Due to missing data, water access and sanitation facilities in 2014 were omitted from the analysis.

3.3.3 Missing Data Imputation

Missing values for predictor variables was dealt based on the multiple imputation strategy. This strategy was selected because it has been recognized by Haan (2013) for producing more accurate and consistent results in social science studies. An advantage of the multiple imputation method over single imputation methods is that it retains a level of uncertainty, which helps to preserve the integrity and accuracy of the standard errors and model fit coefficient estimates.

This method was applied to both predictor and response variable data sets with missing values. Given a data set with missing cases, five random draws were taken from the group of valid cases in the data set. This was used to create a data set of five random values. From this data set, an average was taken and adopted as the value for the missing data point (Table 3.4).

Table 3.4 Missing Data

Type of Data	Number of missing cases	% Total	Notes
Shigellosis incidence	17	3.5%	Total number of missing cases out of 2011-2014 datasets
Socioeconomic data	2	3.4%	Rural Income for 2013 dataset

The method of multiple imputation also has several disadvantages. It may be computational intensive and time consuming to apply multiple imputation for large data sets. Every time when multiple imputation is computed, different estimates are produced, which produces a different result when averaged every time. Since the data sets used in this study were relatively small, these limitations were avoided.

3.4 Generalized Linear Model Regression

3.4.1 Analysis Workflow

The flow of the non-spatial generalized linear regression analysis is illustrated in Figure 3.6. Data on shigellosis incidence and socioeconomic data will be preprocessed and geocoded based on their administrative division codes. After, an exploratory analysis on regression variables will be conducted. The Poisson regression model will be used to study the association between the predictors being the socioeconomic indicators and the response, being shigellosis incidence. Overdispersion within the data will be tested via a Goodness-Of-Fit test. If overdispersion is evident, quasi-Poisson and negative binomial models will be considered and tested to determine which model is the most suitable by conducting two tests, namely (1) goodness-of-fit test and (2) mean-variance plots. If the negative binomial model is chosen as the most suitable model, then a negative binomial Bayesian spatial model will be adopted. Alternatively, if the quasi-Poisson model was found to avail a better fit, then a quasi-Poisson

Bayesian spatial model would be adopted. Model outputs between the non-spatial and spatial models are compared. Moreover, if overdispersion was not evident, then a Poisson regression model would be adopted and a Poisson Bayesian spatial model would be chosen.

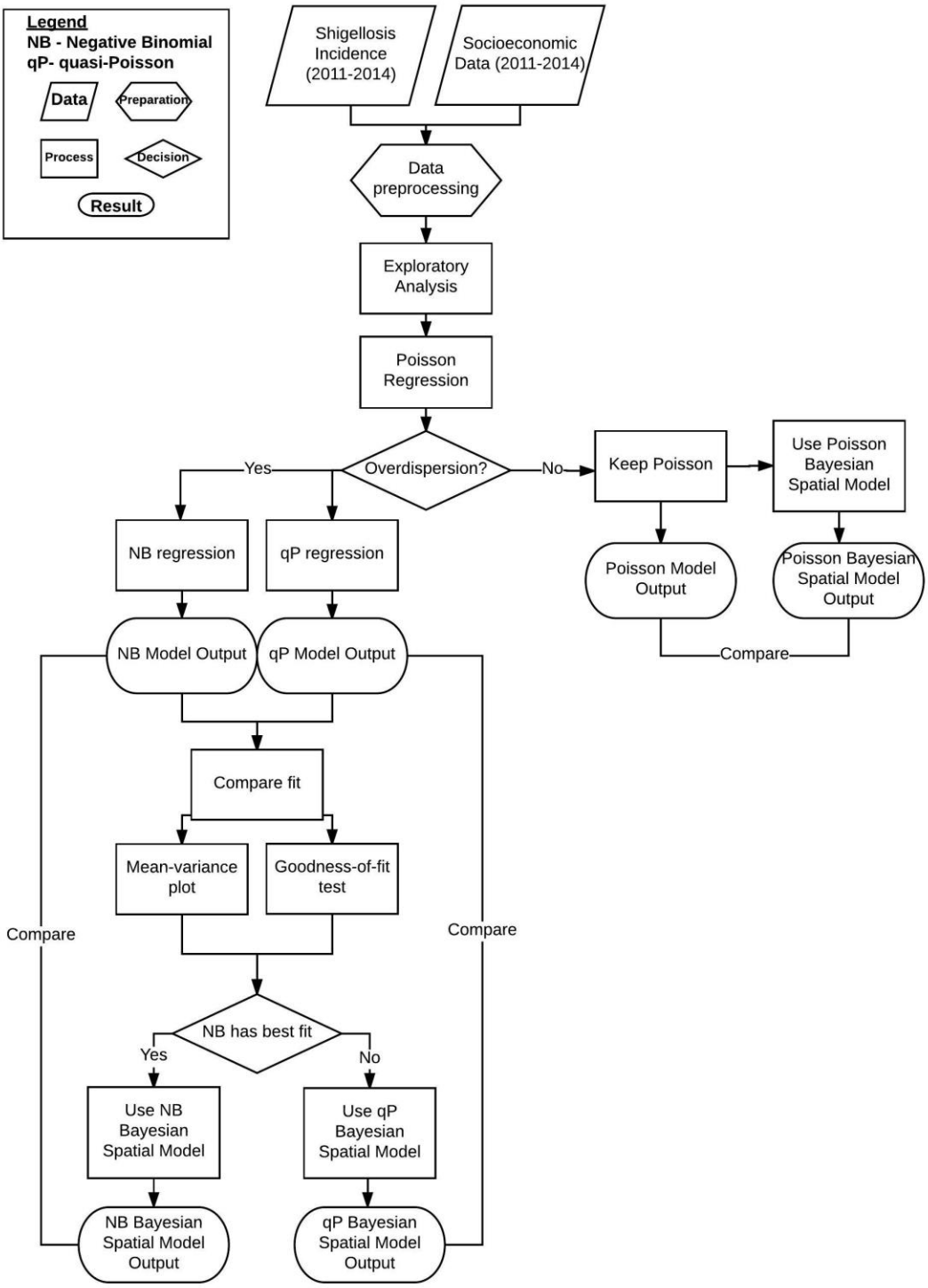

Figure 3.6 Workflow of generalized linear regression analysis

3.4.2 Generalized Linear Regression: Poisson Model

To study the association between socioeconomic determinants and shigellosis incidence, a generalized linear regression model was adopted. Generalized linear models (GLMs) refer to a group of linear regression models that are used to study a response variable reported in the form of a count, binary value, or proportion (Hilbe, 1994). Such response variables follow a distribution part of the exponential family, which includes many distributions such as the normal, binomial, gamma, Poisson, negative binomial, Weibull, and more.

An exploratory analysis was conducted prior to the regression to assess whether each dataset satisfied the conditions of a generalized linear regression. Histograms were created to understand the distribution of the response variable. Scatterplots were created to gauge linearity and patterns between each independent predictor and the response variable. Collinearity between independent variables was assessed using a Spearman ranked correlogram to identify whether there was any linear relationship between independent variables as this may affect the independent impact of each predictor variable on the response.

The response variable in a Poisson regression model follows a Poisson distribution. At small area levels, the distribution of infectious disease counts can be assumed to follow a Poisson distribution (Herrador et al., 2015; Shekhar et al, 2017). In this thesis, shigellosis incidence is presented as a count per county. Histograms of shigellosis incidence created for each year (Figure 3.7) shows that the incidence is highly skewed. Very few counties have high shigellosis incidence (greater than 400) while the majority of counties have a shigellosis incidence count less than 200. Given this condition, the GLM method based on the Poisson distribution was adopted.

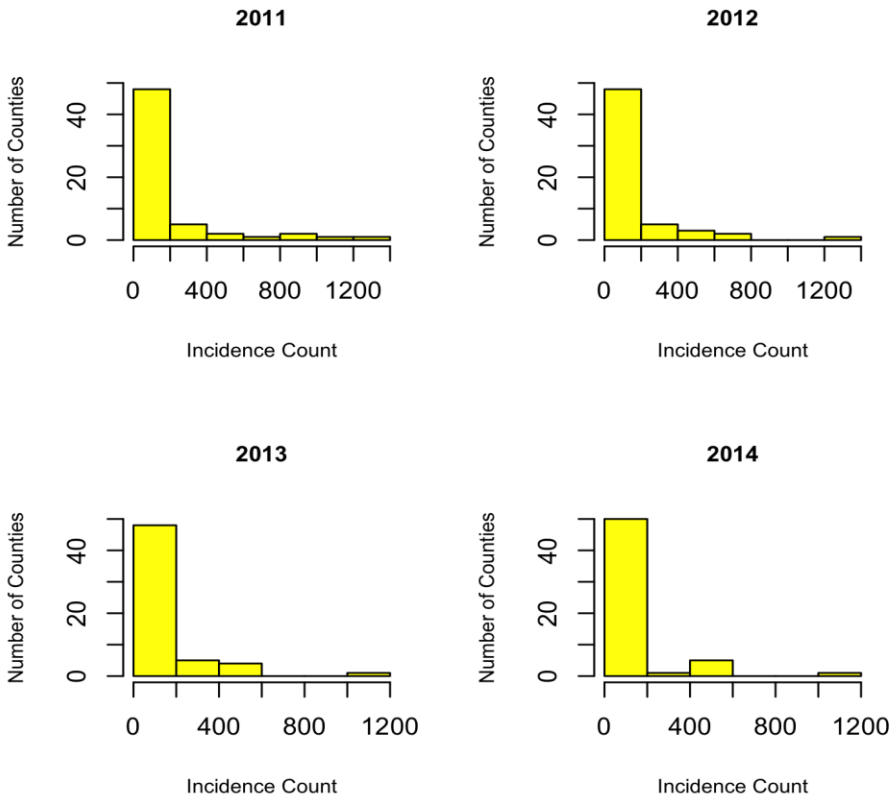

Figure 3.7 Histograms for shigellosis counts in Jiangsu province

The count variable is often assumed to follow a Poisson distribution (Qian, 2010), where a discrete random variable y_i follows a probabilistic distribution with a mean and variance of u_i:

$$F(y_i|\mathbf{x_i}) = \Pr = \frac{{u_i}^{y_i} e^{-u_i}}{y_i!}, \qquad y_i = 0,1,2,\ldots \tag{3.1}$$

The mean and variance of the function are the same as shown below:

$$\mathrm{E}(\mathrm{y_i}|\mathbf{x_i}) = u_i \tag{3.2}$$

$$\mathrm{V}(\mathrm{y_i}|\mathbf{x_i}) = u_i \tag{3.3}$$

The Poisson distribution is part of an exponential family with a log link function $f(u) = \log(u)$. This link function connects the mean parameter u_i to a function of predictors $\mathbf{x_i}$, which is expressed as

$$\log(E[y_i|x_i]) = \log(u_i) = \mathbf{x_i'}\beta \qquad (3.4)$$

The above function indicates that the logarithmic expected number of shigellosis incidents y_i can be modeled by a GLM of predictors $\mathbf{x_i}$ with coefficients β. Therefore, the log-likelihood function of a Poisson regression for a rate is defined by

$$\log[E(y_i)] = \log(n_i) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k \quad i = 1, \dots, n \qquad (3.5)$$

where shigellosis incidence is the expected count [E(yi)] for county *i* modeled as a function of a series of β explanatory variable and an offset variable $\log(n_i)$, which adjusts count based on an exposure designated by the population size of county *i*, and an error term ε_i that represents the lognormal measure. Here, the offset variable was determined by the "Permanent Population" of each county. Permanent Population was chosen as the offset variable to ensure consistency with the calculations performed by the Chinese CDC.

3.4.3 Overdispersion: Quasi-Poisson and Negative Binomial Model

Unlike non-infectious diseases, infectious disease counts display a behavior of overdispersion (Imai et al., 2015), which occurs when the variance of observed counts is greater than the mean (Cameron & Trivedi, 2013). This is evident in the histograms shown in Figure 3.2. When the data is overdispersed, the Poisson regression model will underestimate the uncertainty of regression coefficients. The most common GLM models that allow for

overdispersion are the quasi-Poisson and negative binomial regression models, which are both based on the Poisson regression model. Both methods may yield similar regression coefficients, however this may not be the case when overdispersion is high. Thus, both models were adopted in this study to adjust for overdispersion and assessed for best fit.

The most common method for adjusting for overdispersion for both the quasi-Poisson and negative binomial model is the quasilikelihood approach, which bases inference on robust standard errors (Lee et al. 2012; Myers et al., 2010; Unkel et al., 2012; Ver Hoef & Boveng, 2007). This employs an overdispersion parameter (σ^2) that can be estimated with a Goodness-of-Fit test. For the test, both the Pearson and deviance statistics were used as the test statistic, respectively, and compared to the 5% critical value of $\chi^2_{0.95}$ to determine the p-value. The Pearson statistic can be derived from the Pearson residual (z_i), which was calculated by

$$z_i = \frac{y_i - \exp(\mathbf{x}_i'\hat{\beta})}{\sqrt{\exp(\mathbf{x}_i'\hat{\beta})}} = \frac{y_i - \widehat{y_i}}{\sqrt{\widehat{y_i}}} \tag{3.6}$$

Where $\widehat{y_i}$ is the model predicted values of the response variable. From this, the Pearson statistic was calculated as

$$P_P = \sum_{i=1}^{n} {z_i}^2 \tag{3.7}$$

P_P follows a χ^2 distribution with a degree of freedom(d.f.) of $n - k$, where n is the number of observations and k is the number of regression coefficients, respectively. This distribution also has a mean of $n - k$. Therefore, if there were no overdispersion, the value of P_P would be close to $n - k$. If there were overdispersion, P_P would be larger than the expected value.

Therefore, a ratio of P_P and $n-k$ was used to determine the overdispersion parameter, which was estimated by

$$\sigma^2 = \frac{1}{n-k}\sum_{i=1}^{n} {z_i}^2 = \frac{{\chi^2}_{predicted}}{d.f.} \tag{3.8}$$

where σ^2 is the overdisperion parameter, ${\chi^2}_{predicted}$ is the Pearson goodness-of-fit statistic, and d.f. is the degree of freedom (McCullagh & Nelder, 1989). This equation was also applied for the deviance goodness-of-fit statistic, which is exclusive to a GLM. The deviance statistic (D_P), also known as the G-squared statistic, was calculated by

$$\begin{aligned} D_P &= 2\ \textstyle\sum_{i=1}^{n}[y_i \log(\frac{y_i}{\exp(\mathbf{x}_i'\hat{\beta})}) - (y_i - \exp\{\mathbf{x}_i'\hat{\beta}\})] \\ &= 2\ \sum_{i=1}^{n}[y_i \log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)] \end{aligned} \tag{3.9}$$

Likewise, the deviance statistic also follows an approximate χ^2 distribution with $n-k$ degrees of freedom. When the Goodness-of-fit test is rejected due to the overdispersion parameter and p-value, there is an indication of a lack of fit (Qian, 2010).

Once the overdispersion parameter was calculated, the quasilikelihood method was applied. The quasilikelihood requires the standard errors generated from the Poisson model be multiplied by the factor ω, which is the square root of the overdispersion parameter:

$$\omega = \sqrt{\sigma^2} \tag{3.10}$$

This generated the "quasi-Poisson" regression model, which has the same mean function as the Poisson regression model but with a variance that is ω times the mean (Land et al., 1996; Lee et al., 2012; Ver Hoef & Boveng, 2007) as represented by

$$\mathrm{V}(\mathrm{y_i}|\mathrm{x_i}) = \omega u \tag{3.11}$$

The resulting regression coefficients would be the same as the estimated regression coefficients from the Poisson model, but only the standard errors of the coefficients would be larger. To adjust for overdispersion, ω was multiplied to the estimated coefficient standard errors of the Poisson model. In RStudio, this was done using the quasi-Poisson GLM.

In the presence of an overdispersed quasi-Poisson model, the negative binomial distribution can be adopted to avail a better model (Lee et al., 2012; Lloyd-Smith, 2007; Neyen et al., 2012). The negative binomial distribution model holds the same structure as the Poisson regression model and is often used for overdispersed count data. The distribution is based on X number of failures before the r^{th} success in independent trials, with a probability of success equal to p in each trial ($r \geq 0$ and $0 \leq p \leq 1$). The probability mass function is expressed as

$$P(X = x|r, p) = \frac{\Gamma(x + r)}{x!\,\Gamma(\mathrm{r})}\, p^r (1 - p)^x \tag{3.12}$$

The regression model was expressed as

$$Y \sim NB(r, p) \tag{3.13}$$

where the mean and variance are represented as

$$\mathrm{E}(\mathrm{y_i}|\mathbf{x_i}) = u_i = \frac{pr}{(1 - p)} \tag{3.14}$$

$$\mathrm{V}(\mathrm{y_i}|\mathbf{x_i}) = \frac{pr}{(1-p)^2} = u_i + \frac{1}{r}{u_i}^2 = u_i + k{u_i}^2 \qquad (3.15)$$

If $r \to \infty$, the negative binomial statistics simplifies to the Poisson distribution. The parameter k represents the negative binomial dispersion parameter, which can be estimated by maximum likelihood. The Goodness-of-Fit test was conducted using both the Pearson and deviance statistics to assess the goodness of fit and whether overdispersion exists. The negative binomial regression model is equivalent to Eqn. 3.3 and is expressed as

$$\log[E(y_i)] = \log(n_i) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k \;\; i = 1, \ldots, n \qquad (3.16)$$

In RStudio, the negative binomial regression model was computed using GLM.NB.

Following the regression, additional diagnostic test and plots were made to assess the model fit of each GLM. Firstly, a goodness-of-fit test was conducted to compare which model was more effective at adjusting for overdispersion (Potts and Elith, 2006; Ver Hoef and Boveng, 2007). Next, mean-variance plots were created to validate the fit of the data using quasi-Poisson and negative binomial (Ver Hoef & Boveng, 2007). Afterwards, diagnostic plots such as the deviance residual plot and the deviance QQ plot were produced. Based on the recommendation of McCullagh and Nelder (1989), Myers et al. (2010) recommended plotting the deviance residuals against fitted values because they are nearly the same as those generated by the best possible normalizing transformation. Lastly, a normal probability QQ plot of deviance residuals was also created upon recommendations in literature.

3.4.4 Bayesian Spatial Regression Model

The Bayesian spatial Regression model is applied to account for spatial dependence between areas in the dataset. The Bayesian spatial model adopts a distribution that fits the data. Due to the overdispersed nature of shigellosis incidence in this thesis, the Bayesian negative binomial model was adopted and will be further discussed in this section. For the *i*th area, the number of shigellosis cases (y_i) is modeled as

$$Y \sim NB(r, p) \tag{3.17}$$

$$\log[E(y_i)] = \log(n_i) + \alpha + \beta_1 x_{1i} + B_2 x_{2i} \ldots B_n x_{ni} + \nu_i + \upsilon_i \tag{3.18}$$

Where $\log(n_i)$ refers to the offset, α refers to the intercept quantifying the average shigellosis count, and B_n is the regression coefficient of the predictor variable. Variables ν_i refer to the unstructured non-spatial residual while variables υ_i refer to the unstructured spatial residual. This model assumes a Besag – York – Mollie (BYM) specification for identifying the residuals as described in Besag et al. (1991). The BYM model formulation is commonly applied for disease mapping (Riebler et al., 2016). The spatial structured residual is modeled using an intrinsic conditional autoregressive structure (iCAR)

$$\upsilon_i | \upsilon_{j \neq i} \sim Normal\left(m_i, s_i^2\right), m_i = \frac{\sum_{j \epsilon N(i)} \upsilon_j}{\#N(i)} \; and \; s^2{}_i = \frac{\sigma_\upsilon^2}{\#N(i)} \tag{3.19}$$

Where the #N(i) is the number of areas that share boundaries with its neighbours. The non-spatial unstructured residual is modeled as $\nu_i \sim Normal(0, \sigma_\nu^2)$.

The models were fitted using the Integrated Nested Laplace Approximation (INLA) method and is implemented in RStudio (Vers 3.4.1). The coded program determines the relationship between neighbours using the spatial structured and unstructured residuals calculated using the BYM specification. Afterwards, incidence and socioeconomic data are matched with each respective area prior to running the Bayesian regression model.

3.5 Spatial Data Visualization and Analysis

3.5.1 Analysis Workflow

The flow of the analysis for exploring the spatiotemporal patterns of shigellosis incidence is shown in Figure 3.8. Firstly, shigellosis incidence data is preprocessed and entered into ArcGIS for data classification to create choropleth maps. Secondly, the classified data undergoes incremental spatial autocorrelation to determine a threshold distance for Global Moran's I analysis. Next, if the Moran's I analysis determines the data to be clustered, Local Indicators of Spatial Autocorrelation (LISA) are adopted for further exploration. This is done through two LISA methods, namely the Local Moran's I and the Getis Ord Gi. The Local Moran's I determine whether there is clustering of similar data or dissimilar data, while the Getis Ord Gi determines whether these clusters are hotspots or cold-spots. If there is no spatial clustering, then the process is terminated.

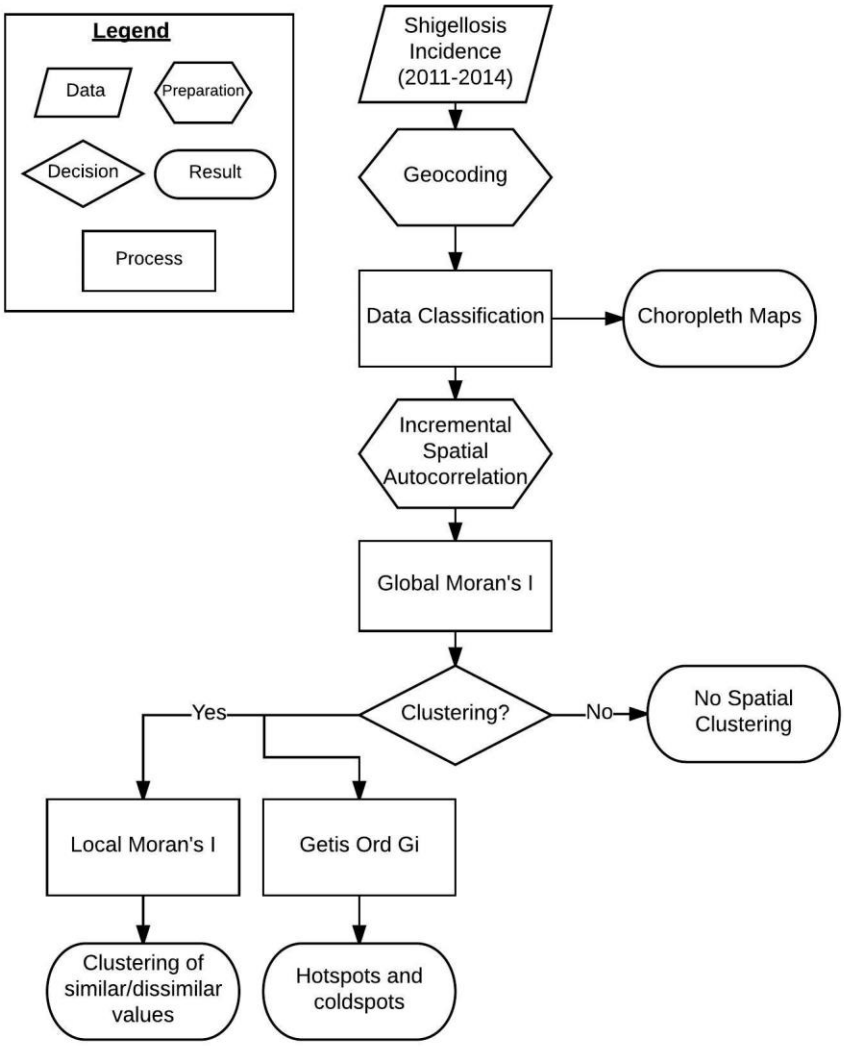

Figure 3.8 Spatial analysis process for shigellosis incidence

The flow of the analysis for exploring the spatial patterns of socioeconomic indicators is shown in Figure 3.9. Data on each socioeconomic indicator was imported into ArcMap and classified for data visualization.

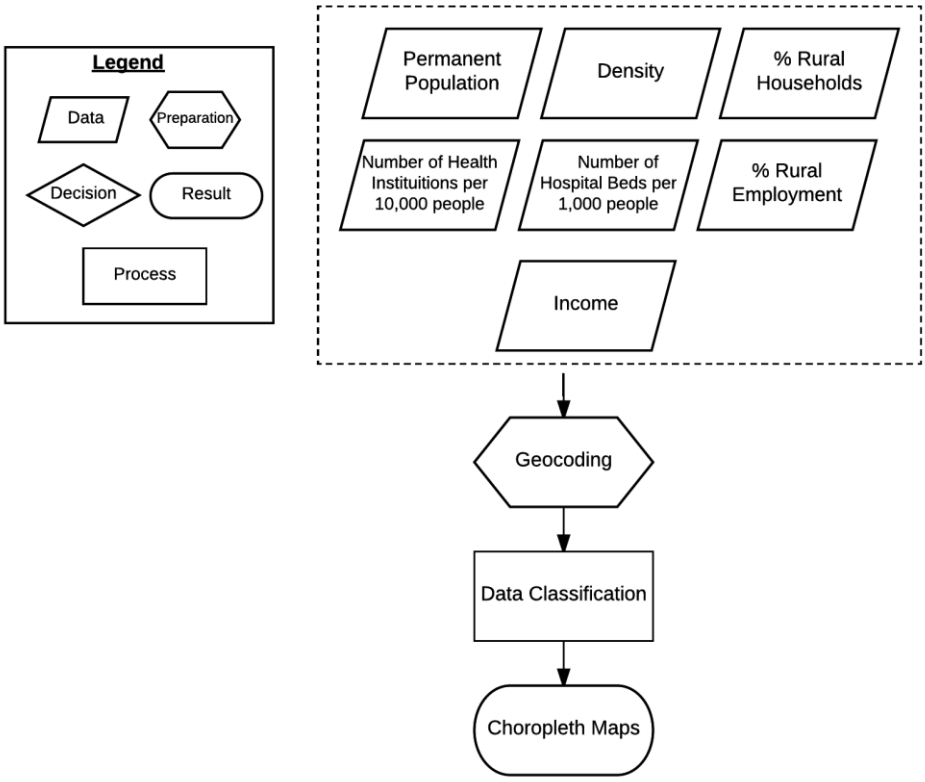

Figure 3.9 Spatial analysis process for socioeconomic indicators

3.5.2 Choropleth Maps

Choropleth maps were created to illustrate the spatial distribution of shigellosis incidence from for 2011 to 2015 in ArcMap. In this thesis, the Equal interval classification method was adopted to illustrate the spatial distribution of shigellosis incidence. The Equal Interval method was chosen for shigellosis incidence so data can be compared by year. This allows the identification of temporal patterns.

The Jenks natural breaks classification method was adopted to illustrate the spatial distribution of socioeconomic indicators. The Jenks natural breaks approach is most suitable for data with high variance and was mainly chosen due to the spatial clustering of some

socioeconomic indicators. It uses an iterative approach to determine the best classification arrangement so that trends in the data can be accurately depicted. By reducing the variance within classes and maximizing the variance between classes, it arranges groups of values so there is less variation in each class. According to Jenks (1967), the natural breaks method is based on repeated calculations using different breaks in the dataset to see which break yields the smallest in-class variance. The first step involves ordering the data and breaking them into arbitrary groups. Then, the following steps are conducted:

1. Calculate the sum of squared deviations between classes (SDBC). This represents the variance between classes.
2. Calculate the sum of squared deviations from the array mean (SDAM).
3. Subtract the sum of squared deviations between classes from the sum of squared deviations from the group mean. This is equal to the sum of the squared deviations from the class means (SDCM), which represents the variance within classes.
4. Inspect each squared deviation between classes. The class with the highest squared deviation between classes moves one unit towards the class with the lowest squared deviation between classes.

Given the new class deviations, repeat steps 1 to 4 until the sum of the squared deviations from the class mean reaches a minimum. Lastly, the goodness-of-variance-fit statistic is calculated to determine the fit. This process stops when the goodness of variance fit statistic (SDAM – SDCM/ SDAM) can no longer be increased. This process has been streamlined in ArcMap under “Layer Properties -> Symbology”, where the “Nautral Breaks (Jenk)” option for was selected for classification method. ArcMap was used to indicate the classification method.

3.5.3 Spatial Autocorrelation: Global Moran's I

The Global Moran's Index (I), a global measure of spatial autocorrelation based on feature locations and attributes that evaluates whether a spatial pattern is clustered, dispersed, or random was adopted to study overall spatial autocorrelation. The Global Moran's I is given as

$$I = \frac{n}{S_o} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{3.20}$$

Where z_i is the deviation of an attribute for feature i from the mean, w_{ij} is the spatial weight between feature i and j, S_o is the aggregate of spatial weights, and n is the total number of features. If the values in a data set cluster spatially (i.e. areal units with high values are near other areal units with high values), the Global Moran's Index would be positive. Contrariwise, if high values were near low values, the index would be negative. If the values balance, the index would be near zero. Most index values will fall between -1 and +1.

The Global Moran's I is an inferential statistic where the result of the analysis will be interpreted relative to the null hypothesis. The null hypothesis of the Global Moran's I statistic states that the feature of interest is randomly distributed among other features in a given study area. If the p-value is determined be statistically significant, the null hypothesis is rejected.

In ArcMap (Vers. 10.4.1), the Global Moran's I was computed using the Spatial Autocorrelation tool. Inputs required for the tool included 1) Conceptualization of Spatial Relationships and 2) Distance Method and 3) Distance Band or Threshold Distance. The conceptualization parameter input corresponds to the spatial weight in Equation 2.16 and was used to define the spatial relationship between features. The conceptualization parameters

listed in the tool include Inverse Distance (Normal and Squared), Fixed Distance, Zone of Indifference, and Contiguity (Edges and Corners).

The Inverse distance method defines that nearby neighbouring features will have a greater spatial influence than features that are farther away. Inverse distance squared is similar, except influence drops more quickly as distance increases from the target feature. The fixed distance band method analyzes features based on its neighbouring features that are within a critical distance; features within the critical distance receive a weight of 1 while features outside of the critical distance receive a weight of 0. Similar to fixed distance, zone of indifference also determines the spatial influence of features within a critical distance. However, features outside of the critical distance can still influence the target feature albeit at a very reduced capacity. Contiguity edges only calculates spatial influence from neighbouring polygon features that share a boundary or overlap with the target polygon, while contiguity corners calculated influence from neighbouring polygon features that share a boundary, or a node, or overlap with the target feature.

Inverse Distance is best for continuous data and the Contiguity methods are best when feature polygons are similar in size and distribution. None of these conditions satisfy the aggregated count data that have been applied to county polygons of various sizes in this study. The Fixed Distance method works well for polygon features that vary in size. The Zone of Indifference also works well for polygons, but it considers every feature to be a neighbour of every other feature since features outside of the critical distance are still analyzed. This is not suitable for administrative areal units near the province boundaries. Due to these reasons, the Fixed Distance Band method was chosen for the conceptualization parameter.

To determine how the distance between two neighbouring features is calculated, the input for the Distance Method must be specified. The two options listed for the Distance Method are “Euclidean” and “Manhattan”. The Euclidean distance refers to a straight line distance between two points, while the Manhattan distance refers to the distance measured along x and y axes at right angles, then calculated using the difference between the x and y coordinates. The Euclidean method was selected since Fixed Distance method is only interested in the Euclidean distance between feature i and j.

Using the Fixed Distance Band method requires specifying a cut-off distance so that only features within the specified distance are analyzed. To find an appropriate distance band, the incremental spatial autocorrelation tool was utilized in ArcMap. This tool measures the spatial autocorrelation of a series of distances and creates a line graph with z-values that correspond to each distance. The z-value reflects the intensity of spatial clustering. Therefore, peak z-scores corresponds to peak distances that yield a high level of spatial clustering. These peak distances were applied to the Distance Band parameter.

To employ this tool, the Beginning Distance, the Distance Method, and the Number of Distance Bands were specified. During the first run, no value was specified for the Beginning Distance. In this case, the minimum distance between each feature and one neighbour was applied. After the initial run has been completed, the beginning distance was reported in the output. In the second run, the Beginning Distance was set to the beginning distance from the previous output. The Number of Distance Bands was also specified. This input is used to indicate the number of times the neighbourhood size should be incremented, with the beginning distance being the starting point.

To select an appropriate number of distance bands, a sensitivity analysis was conducted. This was done to identify the peak distance that corresponded to the highest level of spatial clustering. The minimum number of distance bands that can be chosen was 2, while the maximum number of distance bands was 30. To start, the initial number of distance bands was 6, then was increased to 8,10, 12, and 14. These numbers were chosen to optimize the number of neighbours without exceeding the maximum distance where all features are neighbours with each other. After the peak distance was selected, it was applied to the distance band parameter to compute the Global Moran's I for each year.

3.5.4 Hot Spot Analysis: Local Moran's I and Getis Ord Gi*

Since the Global Moran's I can only provide a summary of spatial clustering, the Anselin Local Moran's I and Getis Ord Gi* statistics were employed to study the local relationship between a feature and its neighbours. A positive value of the Local Moran's I indicates a cluster of similar values (high or low), while a negative value indicates a cluster of dissimilar values (e.g. high values surrounded by neighbours of low values). For the Gi* statistic, a positive value refers to the clustering of high values, while a negative value refers to the clustering of low values.

The Local Moran's I statistic of spatial association is presented as:

$$I_i = \frac{x_i - \bar{X}}{{S_i}^2} \sum_{j=1, j \neq i}^{n} w_{ij}(x_j - \bar{X}) \qquad (3.21)$$

Where x is the attribute for feature I, $\bar{X}$ is the mean of the corresponding attribute, w_{ij} is the spatial weight between feature i and j, and ${S_i}^2$ is the aggregate of spatial weights given as

$$S_i^2 = \frac{\sum_{j=1,j\neq i}^{n}(x_j - \bar{X})^2}{n-1} \tag{3.22}$$

where n is the total number of features.

The Gi* statistics compares local averages to the global average (e.g. Moran's I) and is calculated by

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2\right]}{n-1}}} \quad for\ all\ j, x_j \neq 0 \tag{3.23}$$

Where x_j is the attribute value of feature j, $\bar{X}$ is the mean, and S is the standard deviation. The resulting Gi* statistic is a z-score which not only includes the values of the nearest neighbours but also the value of the feature of interest.

In ArcMap, the Local Moran's I was calculated and mapped for each year using the Cluster and Outlier Analysis tool, while the Getis Ord Gi* was calculated using the Hot Spot Analysis tool. To apply both tools, the Conceptualization of Spatial Relationships parameter must be specified. Similar to the Global Moran's I analysis, the "Fixed_Distance_Band" was selected. Distance method was set to "Euclidean_Distance" and Threshold Distance was set to the beginning distance calculated using the incremental spatial autocorrelation tool.

The Local Moran's I output in ArcMap consists of a local Moran's I index, z-score, p-value, and COType. The cluster/outlier type (COType) identified whether the cluster is statistically significant cluster of high values (High-High), a cluster of low values (Low-Low), an outlier in which a high value is surrounded primarily by low values (High-Low), or an outlier in which a low value is surrounded by high values (Low-High).

Gi* statistic results for each feature are reported with a z-score, p-value, and confidence level bin (Gi_Bin). The confidence level bins are sorted into 4 groups: bin 0, +/- 3 bins, +/- 2 bins, and +/- 1 bins. +/- 3 bins correspond to feature clustering at a statistical significance with 99% confidence level, +/- 2 bins correspond to feature clustering at a statistical significance of 95% confidence, while +/- 1 bins correspond feature clustering at a statistical significance of 90% confidence. Features in bin 0 indicate that clustering of features is not significant.

3.6 Chapter summary

This chapter described the methodology adopted for this thesis. It introduced the study area, summarized the data sources, and described the data preprocessing procedure taken. An overview of the Poisson and negative binomial regression models used to investigate the association between socioeconomic determinants and shigellosis incidence was provided. Finally, spatial analyses used to test for spatial autocorrelation and hotspots amongst county-level shigellosis incidence in Jiangsu province were described. The results of these analyses are presented in the following chapter.

CHAPTER 4: RESULTS

4.1 Introduction

This chapter presents the results from 1) spatial analysis, 2) generalized linear model regression (GLM) analysis, and 3) exploratory survey data analysis of water and sanitation in Suining county. Field observations collected in the rural areas of two towns in Suining are used to support and contextualize the results presented in this thesis. These three analyses are used to address the following research objectives:

1) To examine spatiotemporal variation of shigellosis incidence across Jiangsu province
2) To identify the facilitators and barriers to safely managed water and sanitation
3) To investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province

Maps, tables, graphs, and field photos are used to support the results presented.

4.2 Data Visualization and Analysis

4.2.1 Spatiotemporal distribution of shigellosis incidence

The shigellosis incidence rate is the number of new shigellosis cases per 100,000 persons. Summary statistics were produced to describe the distribution of shigellosis incidence rates at the county level, which consists of counties and districts[2] (N=96) in Jiangsu province from 2011 to 2015 (Table 4.1). In 2011, the average number of shigellosis cases per 100,000 persons was 11.2. From 2011 to 2015, the average incidence rate declined to 5.69.

[2] As defined in Chapter 3.2.

Table 4.1 Distribution of shigellosis incidence rates per 100, 000 persons by county in Jiangsu province from 2011-2015

Year	N	Min	Median	Mean	Max
2011	96	0.240	7.13	11.2	68.4
2012	96	0.101	4.67	8.54	38.0
2013	96	0.102	3.32	6.82	37.8
2014	96	0.116	3.01	5.93	45.1
2015	96	0.116	2.81	5.69	40.6

Shigellosis incidence data was then geocoded in ArcMap to create choropleth maps that llustrateed the geographic variability in shigellosis incidence at the county level (N=97) of Jiangsu province from 2011 to 2015 (Figure 4.1) [3]. It can be seen in Figure 4.1 that relatively high incidence rates were reported near the northwestern and southwestern regions in 2011. From 2011 to 2015, the number of cases in the southwesstern region diminished gradually from 10 to 30 cases per 100,000 persons a year to less than 10 cases per 100,000 persons in a year in some counties. From 2012 to 2015, counties in the northwestern region experienced the highest number of shigellosis cases per year compared to the rest of the province. Incidence rates for the two counties, one near the southwest coast, and the other near the northeast coast, were not determined due to missing data from the Chinese CDC. Despite missing data, it could be assumed that rates in those areas are similar to neighbouring counties under normal circumstances (i.e. they are not outliers). In summary, while shigellosis is mainly concentrated in the northwestern and southern regions, it has shown to persist predominantly in the northwestern region from 2011 to 2015.

In order to investigate the spatial autocorrelation of shigellosis distribution and to identify shigellosis hot spots in ArcMap using the Global Moran's I, Local Moran's I, and Getis Ord Gi*, a threshold distance setting must be determined. This was done by selecting an

[3] These maps adopted a Transverse Mercator projection and were projected using the WGS 1986 UTM Zone 50N coordinated system.

appropriate peak distance using the "Incremental Spatial Autocorrelation tool" in ArcMap, which is based on the Global Moran's I, the distance corresponding to the highest level of spatial clustering of shigellosis cases. To determine the peak distance, a sensitivity analysis with the "Beginning Distance" set to 44, 767 m was conducted. The following peak distances in Table 4.2 were determined.

Table 4.2 Incremental Spatial Autocorrelation Peak Distances

Distance Band	# of Peaks	2011 (m) (z-score)	2012 (m) (z-score)	2013 (m) (z-score)	2014 (m) (z-score)	2015 (m) (z-score)
6	1	67171.60 (5.441)	67171.60 (5.999)	89576.19 (5.327)	89576.19 (4.694)	89576.19 (3.993)
8	1	67171.60 (5.441)	67171.60 (5.999)	89576.19 (5.327)	89576.19 (4.694)	89576.19 (3.993)
10	1	67171.60 (5.441)	67171.60 (5.999)	89576.19 (5.327)	89576.19 (4.694)	89576.19 (3.993)
12	1	67171.60 (5.441)	67171.60 (5.999)	89576.19 (5.327)	89576.19 (4.694)	89576.19 (3.993)
14	1	67171.60 (5.441)	67171.60 (5.999)	89576.19 (5.327)	89576.19 (4.694)	89576.19 (3.993)

The most commonly identified peak distances are 67,171.60 m and 89,576.19 m, respectively. Both distances were tested using the Local Moran's I and Getis Ord Gi* tools in ArcMap. It was found that the peak distance of 89,576.19 m was more effective at capturing spatial clustering, thus, this distance was adopted as the threshold distance for Global Moran's I, Local Moran's I, and Getis Ord Gi*.

Spatial autocorrelation, or the clustering of shigellosis cases throughout the province of Jiangsu, was investigated using the Global Moran's I tool from the Spatial Statistics toolbox in ArcMap. Table 4.3 shows the value, variance, and significance of Global Moran's I calculated for each year. The null hypothesis of Moran's I state that there is no spatial autocorrelation in incidence rates of shigellosis (Moran's I = 0). Positive Moran's I values indicated that there

was significant positive clustering of similar incidence rates throughout counties in Jiangsu ($p<<0.01$). Since spatial autocorrelation was evident, the null hypothesis was rejected.

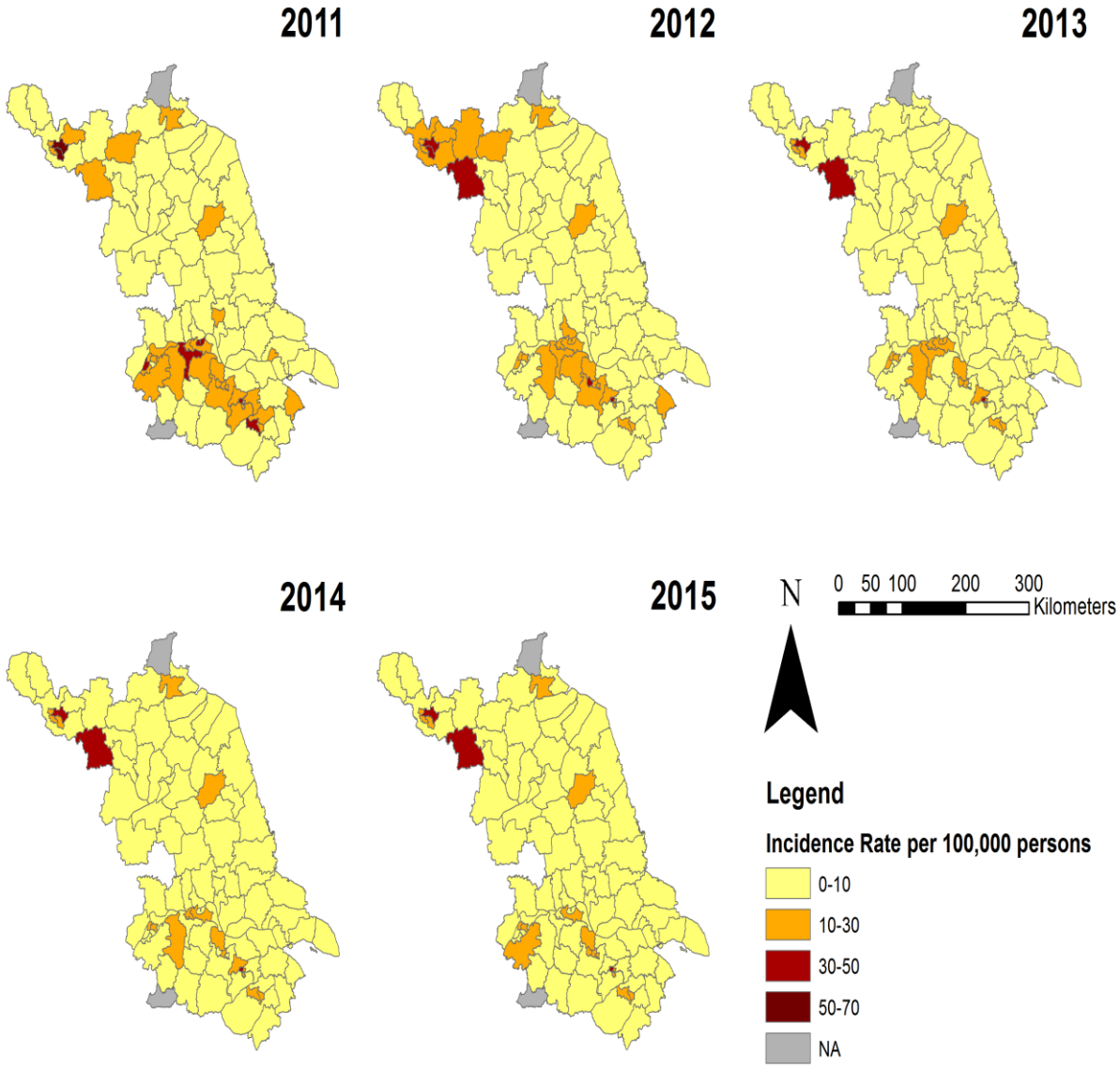

Figure 4.1 Choropleth maps illustrating the incidence rate of shigellosis by county in Jiangsu province from 2011 to 2015

Table 4.3 Spatial autocorrelation in shigellosis incidence rates of counties

Year	Moran's I	Variance	p-Value
2011	0.106	0.000676	0.000008
2012	0.118	0.000704	0.000001
2013	0.109	0.000686	0.000005
2014	0.0866	0.000660	0.000166
2015	0.0664	0.000663	0.002925

The Local Moran's I was employed to identify shigellosis clusters and outliers (Figure 4.2). As explained in Chapter 3, the distance method and conceptualization parameter, which were used to determine the method for identifying clusters and outliers, were set as "Euclidean Distance" and "Inverse_Distance" respectively. The Distance Band or Threshold Distance was set to 89, 576.9 m to reflect the intensity of spatial clustering. The significance level was set to a 95% confidence level. From 2011 to 2015, High-High clusters were predominantly found in the northwestern and southern regions, while Low-Low clusters were found predominantly in central and eastern coastal regions. From 2011 to 2015, a few Low-High outlier clusters were found near the High-High clusters, while High-Low clusters were found near the Low-Low outlier clusters. Thus, clusters of high shigellosis incidence were mainly found in the northwestern and southern regions of Jiangsu province.

In order to investigate the location, size, and intensity of incidence clusters, the local Getis Ord Gi* statistic was performed. This statistic was conducted using the Hot Spot Analysis tool in ArcMap's Spatial Statistics toolbox. Similarly, a distance band of 89 579.9 m was also set to define the intensity of spatial clustering. Figure 4.3 illustrates the spatial variation of hot spot and cold spot clusters of shigellosis incidence by year. During the 2011-2015 time frame, the majority of hot spots concentrated in the northwestern and southern counties of Jiangsu were considered statistically significant. This is because features in these regions have high values surrounded by other high values. While hot spots were identified

mainly in the southern region for all five years, the number of hot spot counties in the region decreased from 2011 to 2015. In 2014, only five counties in the southwestern region were identified as hotspots, but only with 90% to 95% of confidence. In contrast, counties within the central region and coastal areas were predominantly identified as cold spots. From 2013 to 2015, the number of cold spot counties decreased. By 2015, cold spots were identified to be the central and western regions with only 90% to 95% confidence. In summary, from 2011 to 2015, significant shigellosis hot spots were mainly found in the northwestern and southern regions, while significant shigellosis cold spots were identified in the central regions and coastal areas. This clustering of shigellosis incidence shows that more attention paid to areas with clusters of high shigellosis incidence.

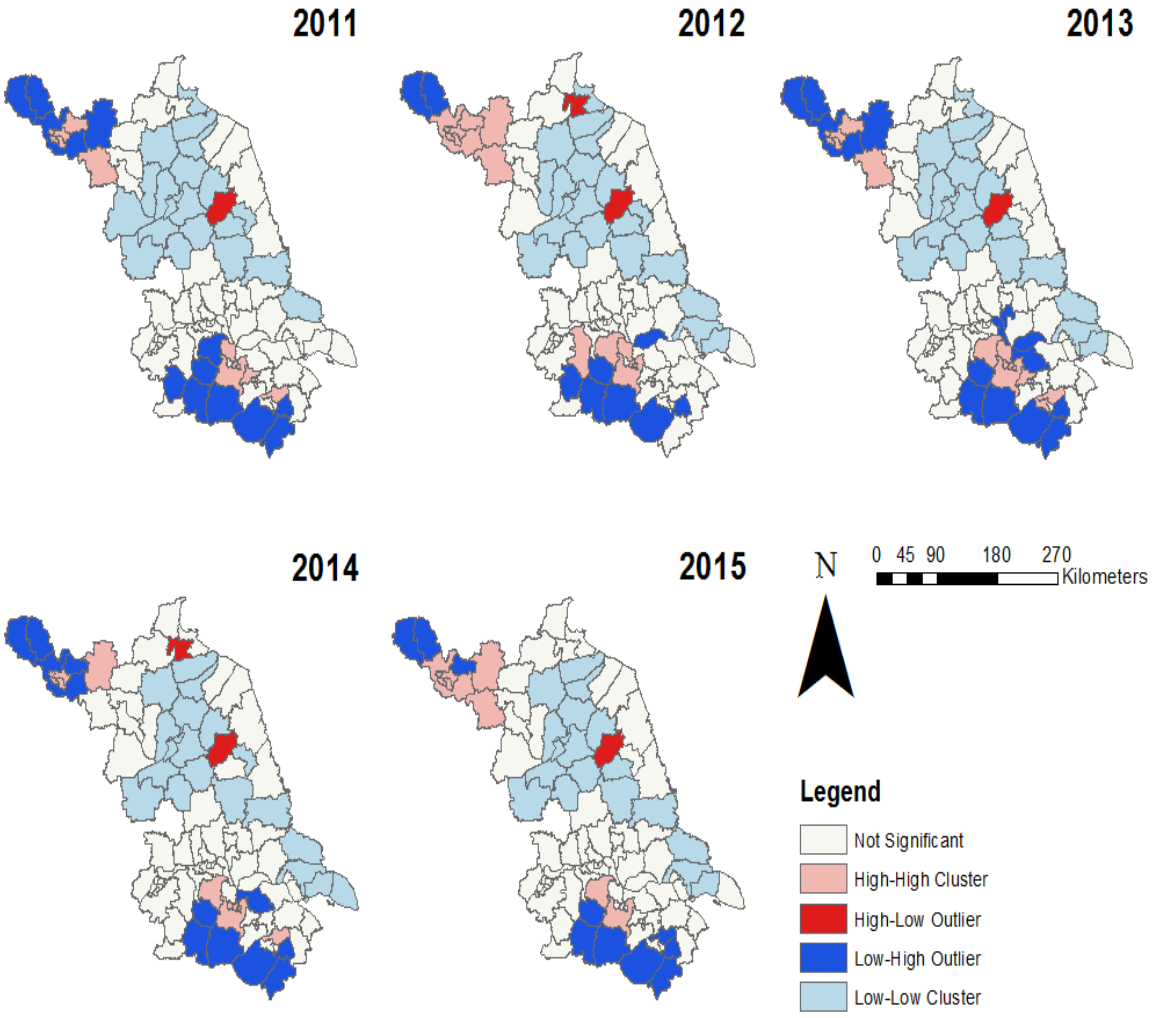

Figure 4.2 Clusters and outliers determined by the Local Moran's I

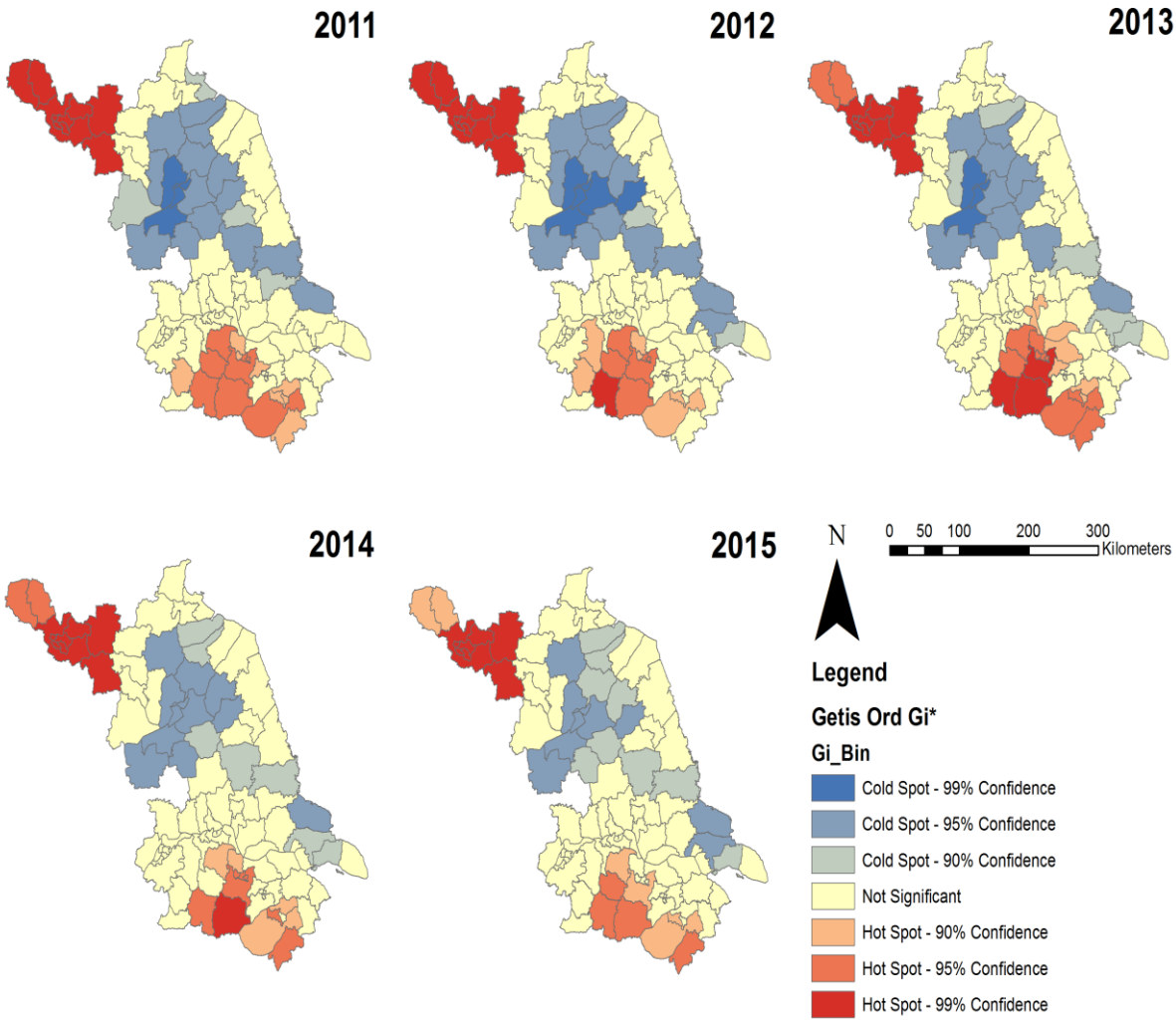

Figure 4.3 Shigellosis hot spots and cold spots determined using local Getis Ord Gi*

4.2.2 Spatial distribution and correlation of socioeconomic determinants

Data on socioeconomic indicators were organized by year from 2011-2014. The socioeconomic indicators selected for analysis included the following: 1) Percentage of Rural Households, 2) Number of Health Institutions per 10,000 people, 3) Number of Hospital Beds per 1,000 people, 4) Percentage of Rural Employees, and 5) Rural Net Income (Yuan Per Capita). Demographic indicators such as population and density were also included for visual exploratory analysis.

An exploratory analysis using choropleth maps was conducted to understand the recent spatial distribution of demographic and socioeconomic conditions in the Jiangsu province averaged between 2011- 2014 (Figure 4.4). The geographic divisions used in these maps include cities, city-counties, and counties. In these maps, "cities" only represent the socioeconomic data on city districts, since data on each individual district was not available. In addition, Nanjing was the only city presented as a prefecture city in these maps due to the lack of data available for its city districts. The choropleth maps were created by merging these geographic units together. Despite the discrepancy in geographic scale of these maps, the data used is still representative of each respective region and efforts have been taken via data preprocessing to ensure that values are representative of all districts in the city.

Choropleth maps were used to identify the spatial distribution in the population. It was also of interest to identify how the distribution in population corresponded to the distribution in access to healthcare. It can be seen that the majority of the population resided in the southern region of the province, but a relatively significant population also resided in some areas in the north. Most counties in the north had a relatively high number of health institutions per 10,000

people. Despite this, the majority of the central and northwestern regions had the least number of hospital beds per 1,000 people relative to other areas of the province.

In addition, choropleth maps were also used to explore the spatial distribution of the rural population. A high percentage of rural households and rural employees are predominantly concentrated in the central and southeast regions, and some areas in the north. Rural income for those regions were in general, relatively low, with the exception of a few areas in the southeastern region.

Summary statistics of socioeconomic indicators were produced in order to describe the variation in socioeconomic determinants that were investigated in the regression (Table 4.4). The spatial unit for analysis was mostly city-counties and counties, with the exception of Nanjing and a few "cities" that were the representation of city districts within the city. It should be noted again that socioeconomic data for individual city districts were not published by the provincial and city statistical yearbooks. Despite these discrepancies, the data adopted for this analysis is still representative of the conditions at the county level in Jiangsu.

As shown, the province was predominantly rural from 2011 to 2014. In one county, 84% of the total households were rural. The average number of health institutions (hospitals, clinics, and medical stations, etc.), number of hospital beds, and rural net income per capita increased from 2011 to 2014. The percentage of rural employment fluctuated over the years. In 2012, the maximum percentage of rural persons employed was 60%, but dropped to 49% in 2014.

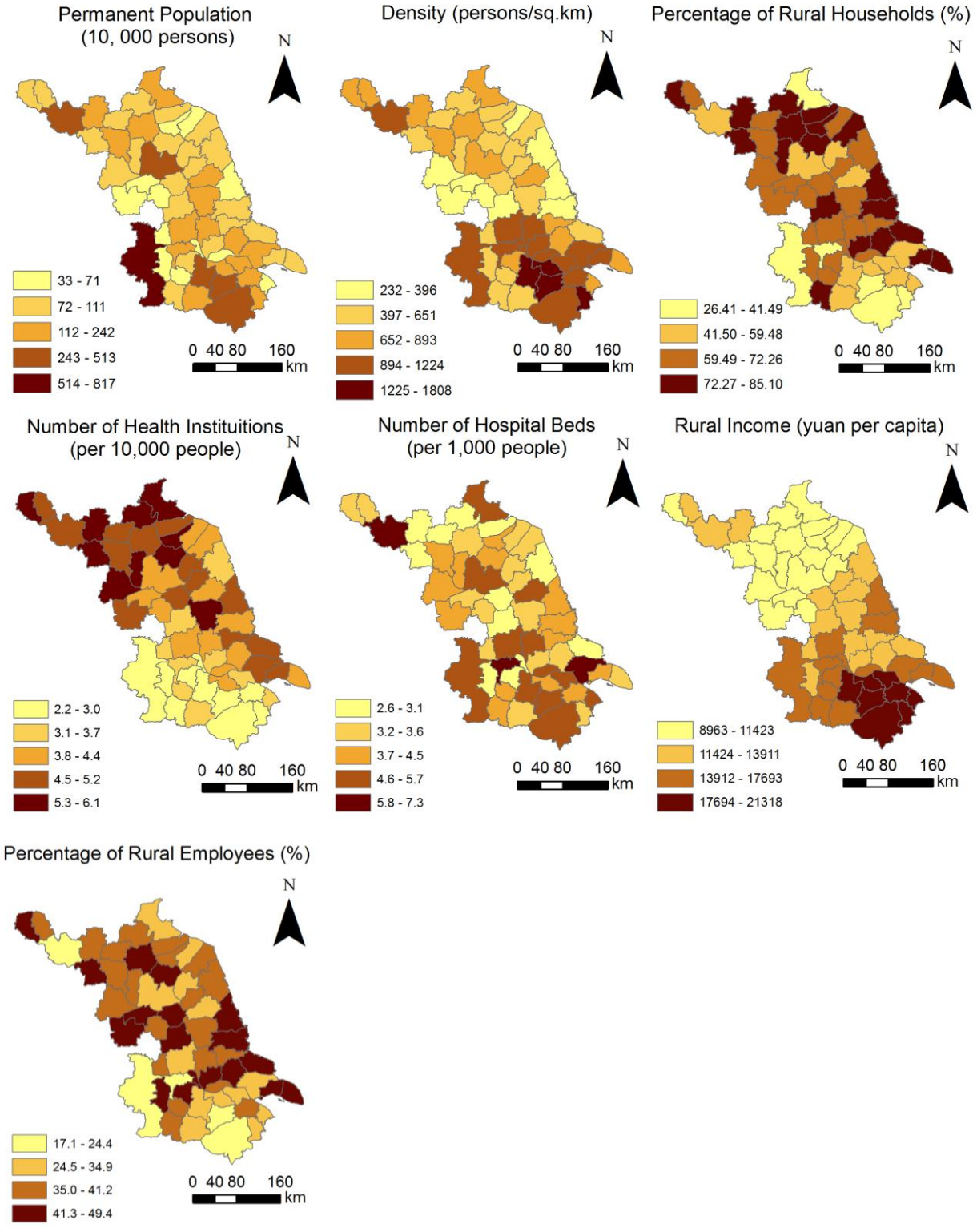

Figure 4.4 Choropleth maps of socioeconomic indicators

Table 4.4. Summary statistics of socioeconomic factors investigated for association with shigellosis incidence in counties of Jiangsu province

Year		% Rural Households	# Health Institutions per 10,000 people	# Hospital Beds per 1,000 people	Rural Net Income per capita	% Rural Employees
2011 (N=60)	Min	26.3	2.13	1.78	7451	9.90
	Median	68.1	4.22	3.05	10 810	42.4
	Mean	65.5	4.3	3.33	11 490	39.9
	Max	85.7	7.87	6.27	17 460	59.0
	SD	14.1	1.28	1.03	2899.63	12.5
2012 (N=59)	Min	27.0	2.17	2.45	8472	3.95
	Median	68.1	4.25	3.53	12 230	42.2
	Mean	64.9	4.22	3.8	12 970	39.7
	Max	85.2	7.39	6.98	19 660	60.2
	SD	13.9	1.07	1.03	3196.30	12.9
2013 (N=58)	Min	25.8	2.08	2.72	9488	11.2
	Median	68.1	4.29	3.91	13 470	42.0
	Mean	64.2	4.18	4.2	14 550	39.6
	Max	85.0	7.3	7.78	23 640	59.3
	SD	14.5	1.08	1.06	3647.94	12.8
2014 (N=57)	Min	24.7	2.25	2.84	10 440	8.60
	Median	68.5	4.45	4.46	14 850	41.7
	Mean	64.3	4.36	4.56	15 800	39.5
	Max	85.3	8.7	9.7	26 370	58.7
	SD	14.3	1.21	1.25	4050.47	12.7

Correlation matrices were created to identify the dependence between socioeconomic variables. Figure 4.5 shows a Spearman ranked correlogram for each year, with the coloured dot representing a statistically significant correlation ($p<0.01$). The percentage of rural households experienced a strong positive correlation with rural employment, showing the parallel connection between rural household and employment. Furthermore, there also exists a significant negative correlation between rural employment and income, showing that counties with higher rural employment tended to have lower income. In contrast, there is a statistically significant negative correlation between the number of health institutions and rural income, showing that counties with more health institutions tended to have lower income.

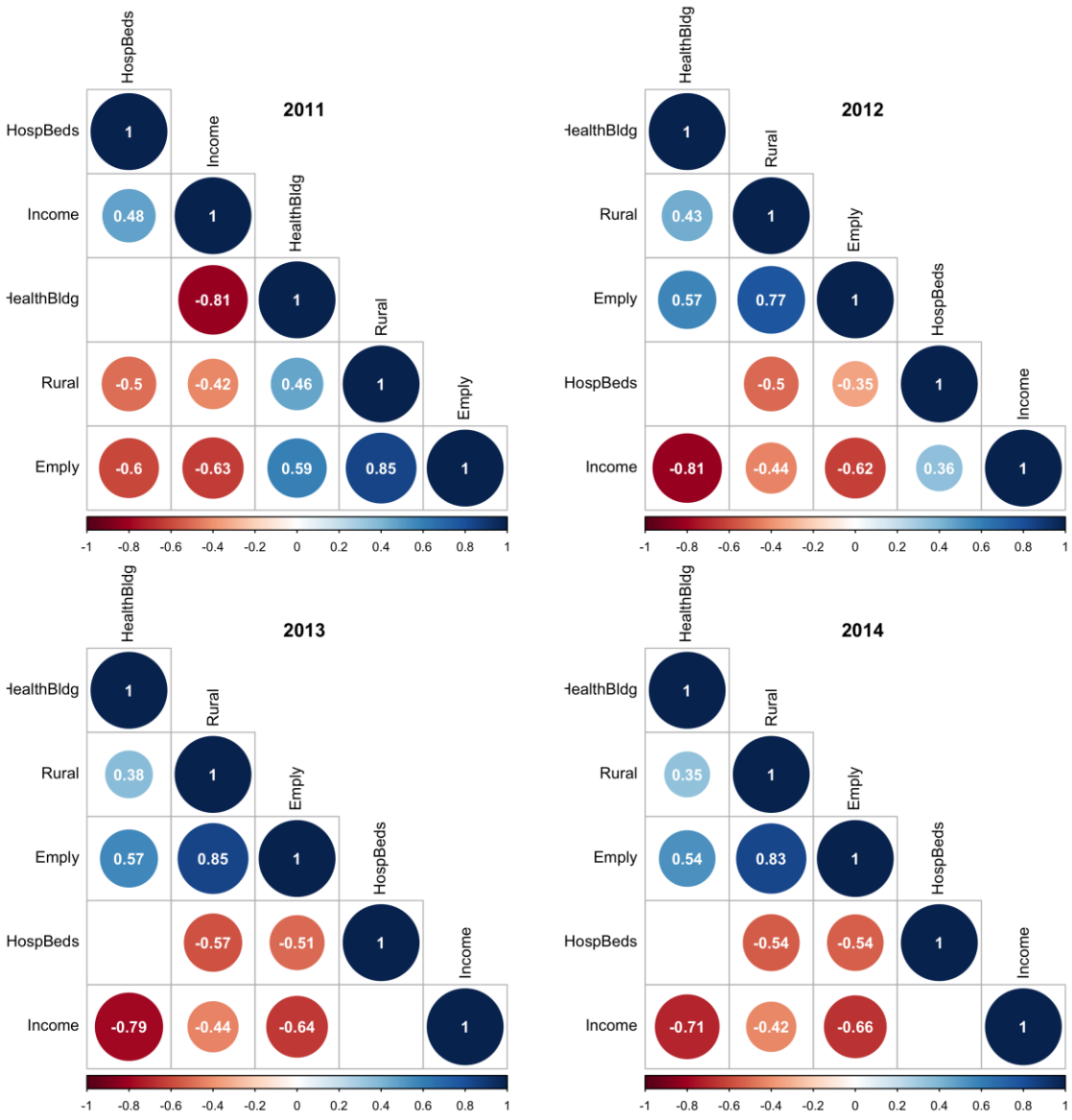

Figure 4.5 Spearman Ranked Coefficient Correlograms

4.3 Generalized Linear Model Analysis

4.3.1 Resulting Analysis Workflow

The resulting workflow of this analysis is illustrated and highlighted in Figure 4.6.

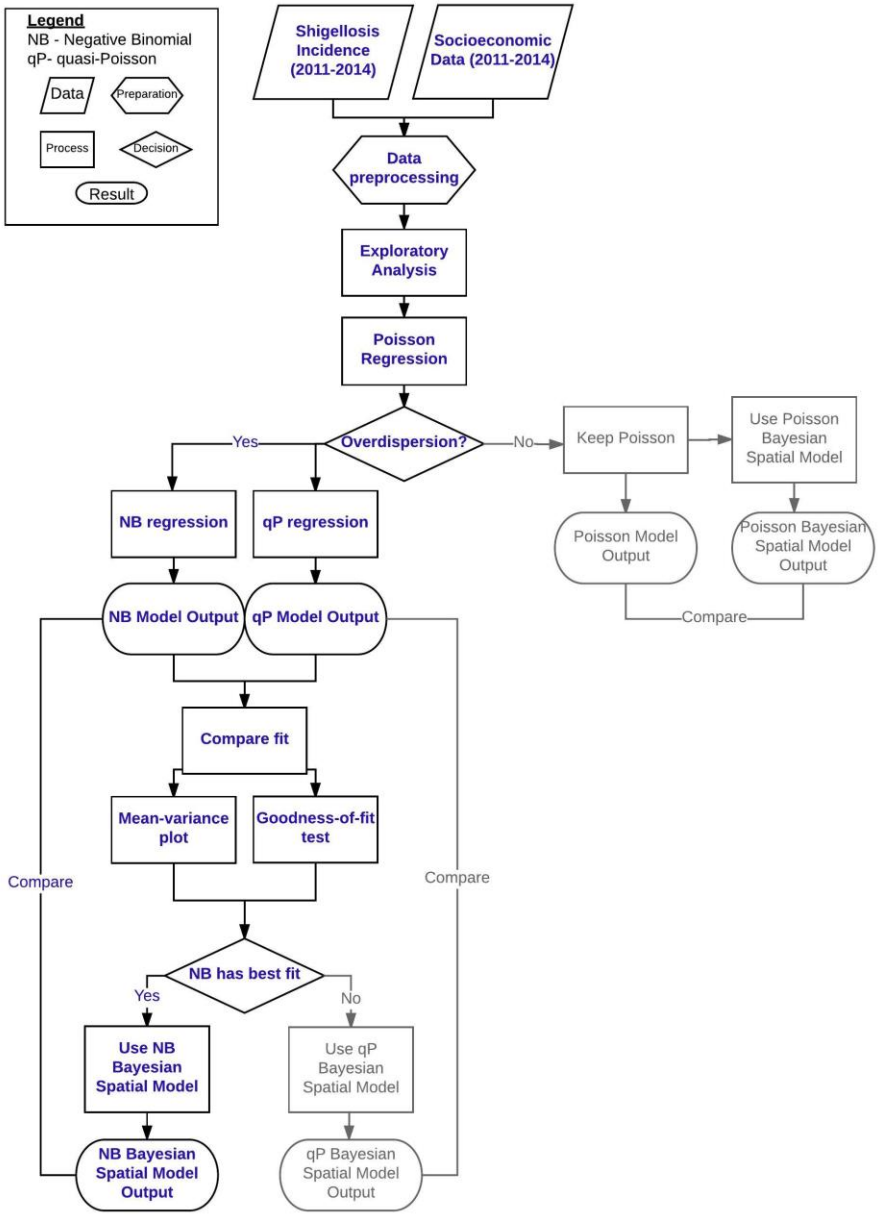

Figure 4.6 Workflow of resulting analysis

After data preprocessing, exploratory analysis of the regression variables was conducted. The Poisson regression was employed and overdispersion was found. As a result, both the negative binomial and quasi-Poisson regression models were adopted and tested for best model fit. It was found that the negative binomial model availed a better fit and thus, a negative binomial Bayesian spatial model was adopted to incorporate spatial dependencies. The results of this analysis are discussed below.

4.3.2 Exploratory Analysis of Regression Variables

Prior to the regression analysis, an exploratory analysis was conducted for the data. Table 4.5 and Table 4.6 summarize the distribution in incidence count and rate of shigellosis used in the regression. In 2011, the average number of cases per 100,000 persons in one county was 166 and decreased to 95 in 2015. These values decreased from 10.7 cases per 100,000 persons and 5.23 cases per 100,000 persons in 2015, almost half of the rate in 2011.

Table 4.5 Shigellosis incidence counts used in regression

Year	n	Min	Median	Mean	Max	Variance
2011	60	2	54	166	1348	80938
2012	59	2	41	133	1349	56743
2013	58	1	22	106	1056	36927
2014	57	1	19	95	1080	34440

Table 4.6 Shigellosis incidence rates (per 100,000 persons) used in regression

Year	n	Min	Median	Mean	Max	Variance
2011	60	0.560	5.43	10.7	51.4	168.9
2012	59	0.246	4.34	7.79	43.0	87.9
2013	58	0.237	2.78	5.98	38.3	65.0
2014	57	0.0880	2.07	5.23	32.6	57.7

As shown in Table 4.6, the variance and mean of each dataset does not equal each other, violating an assumption of the Poisson distribution. This is a common occurrence of overdispersed data sets (Cameron & Trivedi, 2013).

Next, each predictor was plotted against the response for each year to explore the relationship between each predictor variable and response (Appendix B). Figure 4.7 shows the scatterplots for the average shigellosis incidence from 2011 to 2014. There appears to be a negative linear relationship between the percentage of rural households and shigellosis counts. Additionally, the scatterplot of the percentage of rural employment versus shigellosis counts also depicted a negative linear relationship. This showed that as the percentage of rural employment and rural households increased, shigellosis counts decreased. On the other hand, there exists a positive linear relationship between the number of hospital beds per 1,000 persons and shigellosis counts. The relationships between other predictors and the response illustrated non-linear behaviour.

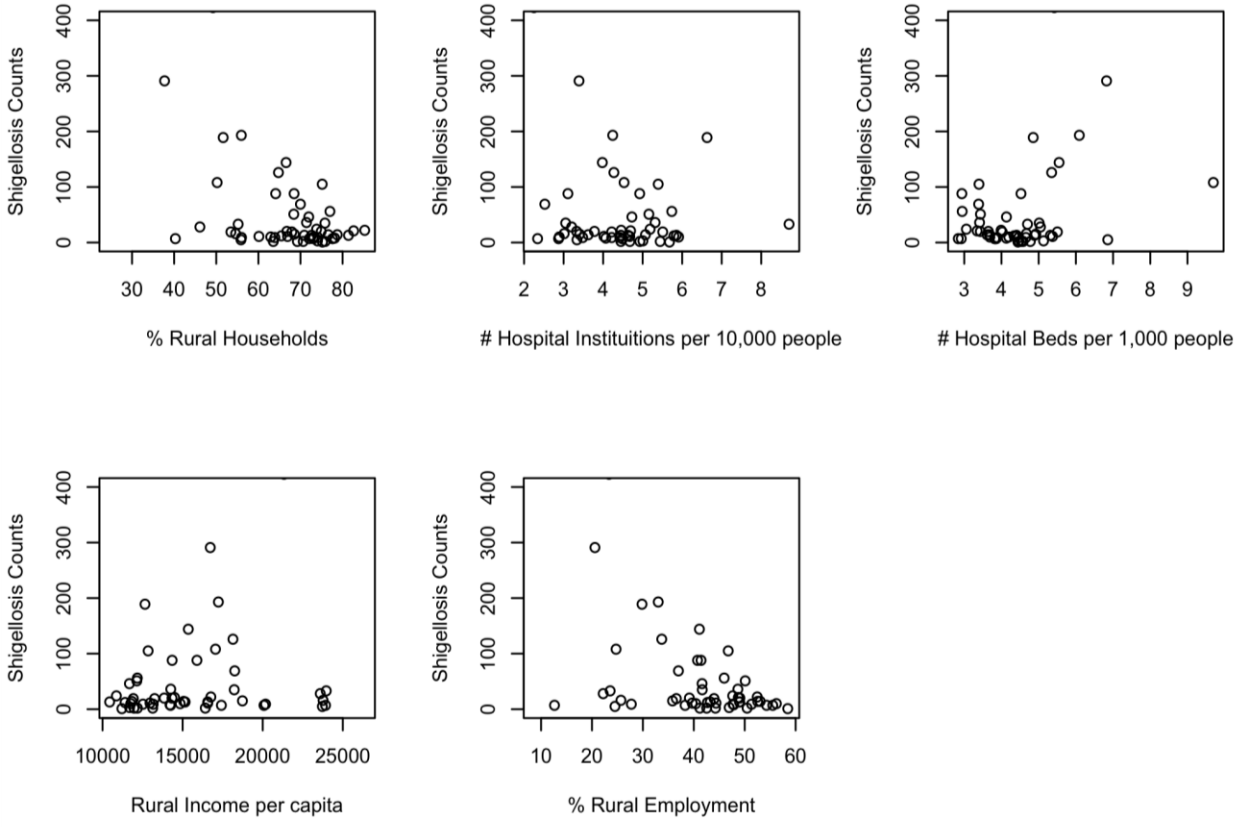

Figure 4.7 Scatterplots between predictors and response variable

4.3.3 Poisson Regression

Using RStudio (Vers. 1.0.136), regression outputs for each year were generated using a multivariate Poisson regression model (Appendix A – Table 1). To assess the model fit, a Goodness-of-Fit test was performed using the Pearson statistic (χ^2) and Deviance statistic (D) (Table 4.7). Both tests are chi-square distributed with *n-p* (number of observations – number of predictors) degrees of freedom. For example, using the dataset from 2011 and assuming α=0.05, the critical value for a Pearson χ^2 at 52 degrees of freedom was 70. In comparison, the Deviance and Pearson chi-squared statistic were 4370 and 4910 respectively, almost 70 times greater than the critical value. The probability that any randomly draw number from this distribution was as large or larger than 4370 or 4910 was 0.

Table 4.7 Goodness-of-Fit Test

Year	Test	DF (n-p)	χ^2 predicted	χ^2 critical	p-value	Overdispersion parameter (σ^2)
2011	Deviance	52	4370	70	0	94
	Pearson	52	4910			
2012	Deviance	51	5064	69	0	128
	Pearson	51	6519			
2013	Deviance	50	3875	68	0	120
	Pearson	50	5980			
2014	Deviance	49	4657	66	0	186
	Pearson	49	9132			

The lack of fit in a Poisson model required a test for overdispersion. As shown in Table 11 above, the overdispersion scale parameter (σ^2) was calculated for each dataset. Due to the presence of significant overdispersion observed in all four datasets, quasi-Poisson and negative binomial regression models were adopted to adjust for dispersion.

4.3.4 Adjustment for Overdispersion

To adjust for overdispersion, the association between shigellosis incidence and several socioeconomic determinants was modeled using multivariate quasi-Poisson regression model and negative binomial regression model. Table 4.8 summarizes coefficient estimates and standard errors for Poisson, quasi-Poisson, and negative binomial.

Overdispersion was evident by analyzing the reported standard errors. Therefore, while the Poisson and quasi-Poisson model have the same estimated regression coefficients, quasi-Poisson's coefficient standard errors were larger. This is evidence of overdispersion and is indicated as an underestimation of standard errors by the Poisson model. In contrast, estimated coefficients from negative binomial model were different than that of Poisson. However, coefficient standard errors from negative binomial were similar in value to that of quasi-Poisson. This showed the standard errors of the negative binomial model were also adjusted for overdispersion.

Table 4.8 Comparison of Poisson, quassi-Poisson, and Negative Binomial model outputs

Year	Variable	Poisson Coeff	quasi-Poisson Coeff	N. Binomial Coeff	SE Poisson Coeff	SE quasi-Poisson Coeff	SE N. Binomial Coeff
2011	(Intercept)	2.608	2.608	3.878	0.152	1.462	1.660
	Rural	0.000	0.000	-0.009	0.001	0.014	0.019
	HealthH	-0.212	-0.212	-0.176	0.017	0.161	0.143
	HospBeds	0.533	0.533	0.372	0.015	0.147	0.150
	Income	0.000	0.000	0.000	0.000	0.000	0.000
	Emply	-0.009	-0.009	-0.007	0.003	0.028	0.028
2012	(Intercept)	0.418	0.418	1.290	0.179	1.816	2.072
	Rural	-0.004	-0.004	-0.022	0.001	0.012	0.016
	HealthH	-0.042	-0.042	0.112	0.023	0.233	0.200
	HospBeds	0.535	0.535	0.280	0.012	0.124	0.162
	Income	0.000	0.000	0.000	0.000	0.000	0.000
	Emply	0.004	0.004	0.012	0.002	0.018	0.017
2013	(Intercept)	-0.020	-0.020	3.237	0.192	1.829	1.908
	Rural	-0.033	-0.033	-0.035	0.003	0.031	0.026
	HealthH	0.102	0.102	0.144	0.015	0.145	0.130
	HospBeds	0.371	0.371	0.033	0.012	0.113	0.131
	Income	0.000	0.000	0.000	0.000	0.000	0.000
	Emply	0.034	0.034	0.004	0.005	0.045	0.036
2014	(Intercept)	2.236	2.236	4.581	0.192	1.829	1.908
	Rural	-0.091	-0.091	-0.069	0.003	0.031	0.026
	HealthH	0.116	0.116	0.004	0.015	0.145	0.130
	HospBeds	-0.009	-0.009	-0.228	0.012	0.113	0.131
	Income	0.000	0.000	0.000	0.000	0.000	0.000
	Emply	0.095	0.095	0.055	0.005	0.045	0.036

The goodness of fit using the chi-squared statistic was reassessed for both the quasi-Poisson model and negative binomial model (Table 4.9) to determine which model provided a better fit. Results from the quasi-Poisson model showed that the chi-squared statistics remained the same. However, the chi-squared values for negative binomial were different and often below the critical value. The p-values for the negative binomial model were also above zero, proving the hypothesis that the data followed a negative binomial distribution should not be rejected.

Table 4.9 Goodness of Fit Test for quasi-Poisson and Negative Binomial

Year	Test	DF (n-p)	χ^2 q-Poisson	χ^2 neg.bin	χ^2 critical	p-value q-Poisson	p-value neg.bin
2011	Deviance	52	4370	67	70	0	0.081
	Pearson	52	4910	69			
2012	Deviance	51	5064	66	69	0	0.071
	Pearson	51	6519	70			
2013	Deviance	50	3875	65	68	0	0.074
	Pearson	50	5980	71			
2014	Deviance	49	4657	65	66	0	0.062
	Pearson	49	9132	96			

Since the quasi-Poisson and negative binomial models have different variance functions, mean-variance plots were generated to assess which model was able to better capture the mean-variance relationship so overdispersion can be minimized. Figure 4.8 illustrates the mean-variance relationship for quasi-Poisson and negative binomial responses. Firstly, values based on the linear predictor were computed using the negative binomial model. Then, the linear predictor values were grouped based on the percentile groups they belonged to. The percentile groups were organized into 20 groups from 0 to 100, with five percentile increments. The quasi-Poisson model was unable to fit the data, as the rate at which the variance increased was much higher than the mean. As shown in the plot, the negative binomial model was able to effectively capture the relationship of the data, particularly for data within the 25^{th} and 50^{th} percentile. Since the negative binomial model was able to better portray the data, it was selected as the model to analyze the significance of socioeconomic variables in predicting shigellosis incidence.

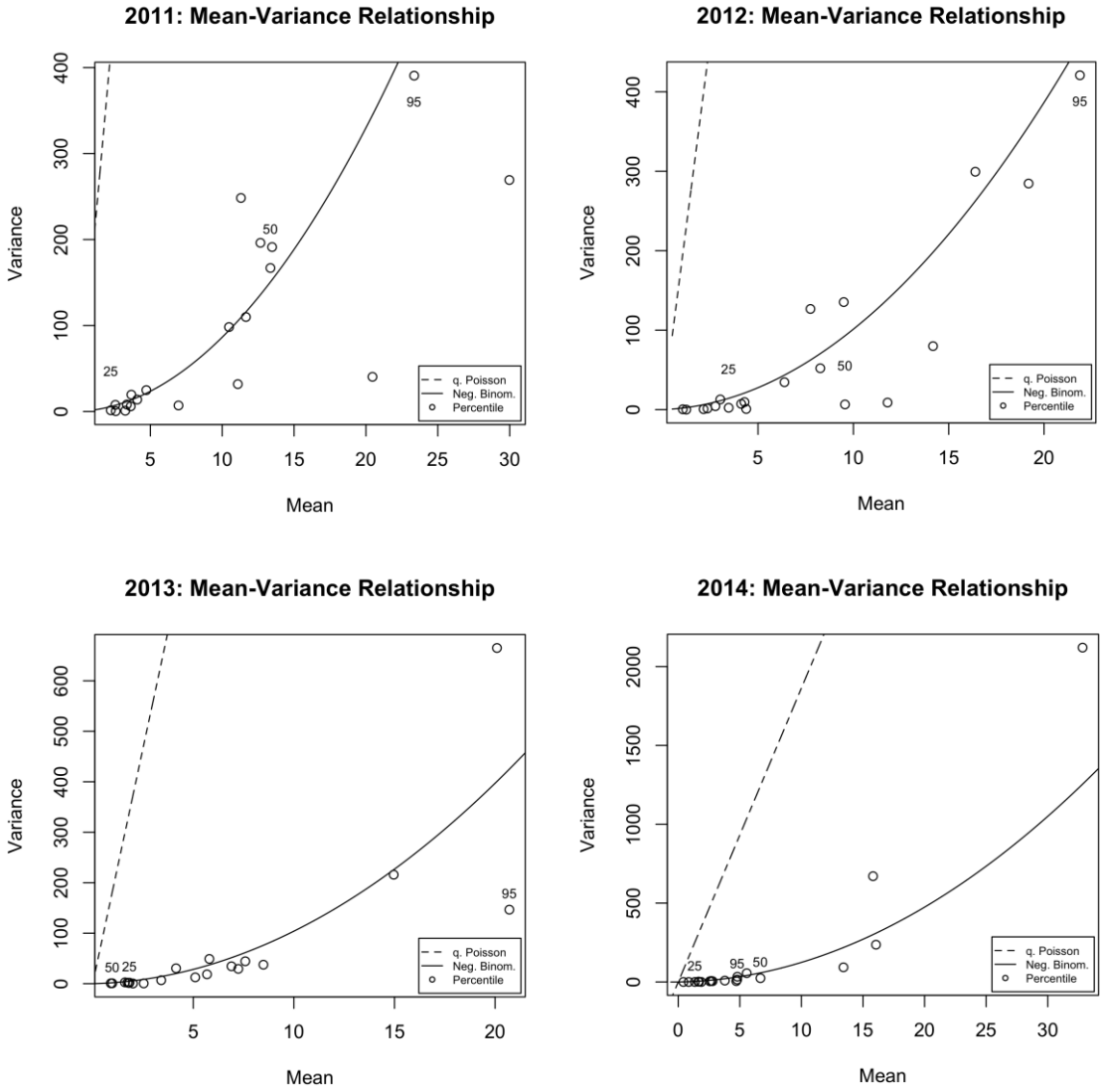

Figure 4.8 Mean-variance plots

To assess for model fit, residual and Q-Q plots were also generated (Appendix C – Figure 1 and Figure 2). It can be seen that the negative binomial model was able to better account for the mean and variance relationship of the data. The residual plots illustrated unbiased and homoscedastic behaviour. Normal QQ plots showed that the data do not fully follow normal behaviour. This was expected from a GLM, as residuals from a negative

binomial model are not assumed to follow a nearly normal distribution. Thus, the negative binomial model was chosen as the preferred model for shigellosis incidence.

4.3.5 Negative Binomial Regression Output

The negative binomial regression was performed to determine significant socioeconomic predictors. Only two variables were deemed significant using the negative binomial model. Univariate negative binomial analysis was performed on these predictors to determine its significance. The association is represented as an incidence rate ratio (IRR) as shown in Table 4.10.

Table 4.10 Association between shigellosis incidence and significant socioeconomic determinants

	Variables	Multivariate analysis		Univariate analysis	
		IRR (95% CI)	P-Value	IRR (95% CI)	P-Value
2011	#Hospital Beds per 1000 persons	1.45 (1.08,1.98)	≤ 0.01	1.56 (1.29, 1.90)	<0
2012	#Hospital Beds per 1000 persons	1.32 (0.99,1.80)	≤ 0.05	1.45 (1.20, 1.79)	<0.002
2014	#Hospital Beds per 1000 persons	0.784 (0.622,1.04)	≤ 0.05		Not Sig.
2014	% Rural Households	0.933 (0.888,0.98)	≤ 0.001		Not Sig.

For the 2011 dataset modeled with negative binomial regression, only the number of hospital beds per 1000 persons was determined to be significant ($p_{\text{neg.bin}} \leq 0.01$). The model was computed again with only the significant predictor as the sole predictor (Appendix A – Table 3). Results from this model showed that every increase in the number of hospital beds per 1000 persons was associated with an increase in incidence rate of shigellosis by a factor of 1.56 (95% CI: 1.05,1.95) ($p_{\text{neg.bin}} \ll 0.01$), given all other predictor variables were held

constant. This can be further converted into a percentage change, which is interpreted as a 56% increase in rate. For 2012, a 45% increase in rate was found ($p_{\text{neg.bin}} \ll 0.01$).

For the 2014 dataset modeled with negative binomial regression, both the number of hospital bed per 1000 persons and the percentage of rural households were determined to be significant ($p_{\text{neg.bin}} \leq 0.05$). The model was computed again with only the significant predictor as the sole predictor (Appendix A – Table 3). Results from this model showed that the variables were no longer significant.

4.3.6 Bayesian Spatial Regression Output

Since the negative binomial model availed a better fit, it was chosen for the Bayesian spatial regression model. Using the Bayesian negative binomial spatial model, the association between socioeconomic determinants and shigellosis incidence was also determined while taking in account the spatial relationships between counties (Table 4.11). All coefficient parameters were determined in the form of relative risk (RR), which is synonymous with IRR. The Bayesian negative binomial spatial model confirms the finding for percentage of rural households in 2011 using the negative binomial model, but was also able to determine additional associations between shigellosis incidence and other socioeconomic determinants.

Table 4.11 Coefficient Estimates of socioeconomic determinants using Bayesian Spatial Model

Variable	B	e(B)	SD	95% Credible Interval	
				2.5%	97.5%
Intercept	8.4021	4456.4154	2.1316	4.0026	12.4011
% Rural Household	-0.0357	0.9649	0.0319	-0.0982	0.0274
# Health Institutions per 10,000 people	0.3004	1.3504	0.2285	-0.1350	0.7649
# Hospital Beds per 1000 people	-0.0313	0.9692	0.1721	-0.3673	0.3098
Rural Income per capita	-0.0001	0.9999	0.0001	-0.0002	0.0001
% Rural Employment	-0.0479	0.9532	0.0452	-0.1384	0.0397

For example, a one percent increase in the percentage of rural households is associated with a decrease in the incidence rate of shigellosis by a factor of 0.96 (3.51% decrease in incidence rate). In addition, a one unit increase in the number of health institutions per 10,000 people is associated with a increase in incidence rate by a factor of 1.35 (35% increase in rate), while a one unit increase in number of beds per 1000 people is associated with a decrease in incidence rate by 0.969 (3.08% decrease in rate). A RR of 0.9999 or 1 for rural income implies that the factor has no effect. A one percent increase in rural employment is associated with a decrease in incidence rate by a factor of 0.953 (4.68% decrease in rate). In summary, while an increase in the percentage of rural households, number of hospital beds, and percentage of rural employment were associated with a decrease in shigellosis incidence rate, an increase in the number of health institutions was found to be associated with an increase in shigellosis incidence.

4.4 Chapter Summary

This chapter presented the results from the quantitative analyses. The spatial analysis on shigellosis incidence showed that in recent years, high shigellosis incidence was concentrated mainly in the northwestern regions and a few counties in the southern regions of Jiangsu province. The Global Moran's I indicator indicates significant positive spatial autocorrelation, indicating that high shigellosis incidence tend to cluster spatially. This is further confirmed by analyses using Local Moran's I and Getis Ord Gi*, which illustrates that statistically significant hot spots tend to cluster in the southwestern and southern regions.

Exploratory analyses using choropleth maps, scatterplots, and a Spearman rank correlation matrix were performed to understand the distribution and correlation of

socioeconomic indicators prior to the regression analysis that assessed the association between socioeconomic determinants and shigellosis incidence. Results from the Spearman rank correlation matrix showed that there was significant positive dependence between the percentage of rural population and the percentage of rural employees. In addition, counties with high percentage of rural households tended to have lower income and less number of hospital beds.

Two generalized linear regression models, negative binomial and quasi-Poisson, were employed to assess the association between socioeconomic indicators and shigellosis incidence. Results from the quasi-Poisson and negative binomial showed that the negative binomial was a more appropriate model, as it was able to better account for overdispersion in shigellosis incidence. Thus, the negative binomial model was chosen for the Bayesian Spatial Model, which was adopted to account for spatial dependencies between areas.

Using the non-spatial negative binomial model, the number of hospital beds was identified as a statistically significant indicator for 2011 and 2012 respectively. The model showed that increasing the number of hospital beds was associated with higher shigellosis incidence. The number of hospital beds and the percentage of rural households were found to be significant parameters in the multivariate model. The spatial negative binomial Bayesian model confirmed the finding for the percentage of rural households, but also found that increasing the percentage of rural employment and the number of hospital beds, respectively, was associated with decreased shigellosis incidence. In contrast, increasing the number of hospitals was correlated with increased shigellosis incidence.

CHAPTER 5: DISCUSSION AND CONCLUSIONS

5.1 Introduction

This concluding chapter summarizes key findings from this research, which aimed to address the following objectives:

1) To examine spatiotemporal variation of shigellosis incidence across Jiangsu province;
2) To explore the facilitators and barriers to safely managed water and sanitation;
3) To investigate the association between socioeconomic determinants and shigellosis incidence in rural areas of Jiangsu province

Findings on the spatial patterns of shigellosis and the association between socioeconomic conditions and shigellosis incidence are examined in the context of rural China. To add validity, findings are also compared with existing literature. Furthermore, contributions and policy implications are discussed. Lastly, limitations and recommendations for future research are reviewed.

5.2 Summary of Key Findings

Although past studies have explored the spatio-temporal variation of shigellosis incidence in China, no studies to date have explored the distribution of shigellosis incidence in recent years after 2011. Examining the incidence rate of shigellosis in recent years is critical to gauge its prevalence in rural China. Thus, a part of this thesis intends to fulfill this knowledge gap. The spatial analysis approaches used in this thesis identified the spatial clustering of shigellosis in Jiangsu from 2011 to 2015. Both the Local Moran's I and Getis-Ord Gi* were adopted to examine the size, intensity, and extent of shigellosis clustering across areas. Results show that clusters of high shigellosis incidence tend to concentrate in counties of the northwest

and south. This result confirms the findings from Tang et al. (2014) that high shigellosis incidence clusters tend to persist in Jiangsu's northwestern regions. While high incidence clusters were also evident in the south, particularly in the southwest, the intensity of those clusters has diminished greatly from 2011 to 2015. Cold spot clusters continue to dominate in northeastern, eastern, and central regions. The most notable difference between the results from the hot spot analysis and the incidence maps is that counties with the highest disease incidence do not always have the highest Gi* value. In Figure 4.3, this was evident amongst counties in the south, where some were not identified as a hot spot despite having a relatively high incidence in the incidence map. Instead, neighbouring counties with high incidence were highlighted as hot spots.

A previous study by Tang et al. (2014) identified one of the key reasons for the persistence of shigellosis in counties of the northwestern region is the relative lack of safe water supply and sanitation. County level survey data from the northwestern county of Suining illustrates that overall tap water access has been increasing (Suining Statistical Yearbook (2014), but observations from the field visit show that access in Suining still appears to vary by geography. This has been confirmed by Hongxing Li et al. (2015), as they have found an imbalance in the spatial distribution of rural water supply infrastructure in China, but noted that the imbalance has improved in recent years. While local differences in water access is noticeable and may contribute unevenly to shigellosis incidence, it is difficult to draw conclusions given the lack of empirical data.

Based on observations from a field visit in Suining and information shared during the field visit, none of the households had flushing toilets, as they are expensive to install and upkeep. Pour flush pit latrines are still the most common "improved" sanitation facility

adopted by most households in rural areas. Despite being considered an improved sanitation facility, pit latrines may encourage the process of cultivating untreated feces as night soil. This may have potential complications for drinking water quality as pathogens and contaminants from untreated night soil can leach and travel as runoff during storm events (Ling, 1993; Tong et al., 2017). The leaching of latrine contaminants into water has been confirmed by the findings of Ding et al. (2017). They found that dry latrines, septic tanks without covers, and fecal sewage drained into ponds and rivers were common in rural areas, especially near primary and secondary schools. Their study revealed that sanitation facilities close to water sources, as well as the lack of sterilization facilities contributed to *Shigella* infections. Thus, while the type of sanitation facility may contribute to shigellosis, it is difficult to draw strong conclusions given the lack of findings and field observations alone.

Access to safely managed water and sanitation are dependent on facilitators and barriers, which includes the socioeconomic drivers that influence the health of populations. This has been previously discussed in the context of a conceptual framework. To name a few, this include income, employment, school, and the existence of a social support network, in particular amongst children. These socioeconomic determinants can facilitate or hinder access to safely managed water and sanitation. This conceptual framework contributed to objective two of this thesis and set the foundation for quantitative analysis used to meet objective four.

The association between socioeconomic factors and shigellosis incidence has been explored both nationally across 31 provinces in China and at the county level in the southern province of Guangxi (Nie et al., 2014) and southwestern province of Sichuan (Ma et al., 2015), however no studies to date have investigated this relationship in the eastern province of Jiangsu, which has high shigellosis rates in some regions. This thesis aimed to fulfill this knowledge

gap by exploring the linkages between rural income per capita, percentage of rural employment, access to health institutions, access to hospital beds, and the percentage of rural households to shigellosis incidence across time and space. This thesis also aimed to understand how some of these factors could act as facilitators or barriers to safely managed water and sanitation. The modeling approaches adopted in this thesis are applicable, novel, and have not been applied in previous studies.

The negative binomial regression is an applicable approach that has been applied in this thesis to study the association between socioeconomic factors and shigellosis per year from 2011 to 2014. The non-spatial negative binomial regression results in this thesis showed that an increase in hospital beds per capita was associated with increased shigellosis incidence in 2011 and 2012. Along with the percentage of rural households, it was found that an increase in both variables were associated with reduced shigellosis incidence in 2014. Results from the negative binomial Bayesian spatial model confirmed with the previous finding using the negative binomial regression model that increasing percentage of rural households is associated with less reported cases of shigellosis. In addition, the Bayesian spatial model also found that higher number of health institutions was associated with higher shigellosis incidence. In contrast, increasing percentage of rural employment and more hospital beds per capita were associated with lower shigellosis incidence, which could point to areas with fewer, but larger hospitals.

The association between rural areas and lower shigellosis incidence found in this thesis may be explained by the fact that urban areas have higher shigellosis incidence as a result of constant monitoring and reporting while rural areas are susceptible to underreporting (Cheng et al., 2017). This supports the findings by Xia et al. (2011) and Wang et al. (2005) as they have also found underreporting present in rural areas in China. In rural areas, there may also be a

significant number of unreported cases of diarrhea, illustrating that passive disease surveillance methods used by the CDC are not effective. Data on shigellosis morbidity collected by the Jiangsu CDC may be susceptible to underreporting due to delayed or inconsistent diagnosis in rural settings (Lin Yin-Jun et al., 2013). In some cases, doctors might not order a laboratory stool test, and those that are tested may also return as false negative (Xia et al., 2011). In rural areas, many rural residents may also attempt to treat shigellosis at home due to time, distance, and economic constraints associated with seeking assistance from a doctor. In addition, rural health care services may lack appropriate medications for treatment.

Cases of shigellosis may be underreported due to several reasons. Correlogram results in this thesis showed that counties with a high percentage of rural households had generally lower income and lower number of hospital beds per capita, which affect access to health care. Rural areas in China have a lower rate on self-reported health status as "many rural areas are inhabited by children, the elderly, the chronically sick, and the less healthy" (Hesketh et al., 2008).

Limited access to health care services in rural areas of Jiangsu may have contributed to underreporting (Chao et al., 2017). In rural areas, village clinics, townships health centres, and hospitals at the county level constitute the basic rural health care system. In the case example of Suining, patients with shigellosis can only receive diagnosis through their local clinics. Although there are plenty of primary health care institutions, they only offer the most basic health care equipment. Furthermore, doctors working in village clinics have the most basic training and certification, as they are not required to take qualifying exams for advanced work. Health needs that cannot be met by these institutions are further referred to hospitals and

specialized health institutions at the county level; this process is known as the village-town-county three-tier referral system (Feng et al., 2017).

The persistence of underreporting in rural areas may also be influenced by behaviours and decisions of the individual. Despite the existence of the village-town-county referral system, Chinese patients do not have to follow this referral process and can choose their health institution based on their financial condition, location, convenience, and preference (Sun et al., 2017). In areas where diarrhea cases are common, families may become less inclined to seek treatment for two reasons. Firstly, they may become accustomed to diarrhea as they think that they are able to treat it at home. Secondly, due to the unavailability of effective antibiotics to treat shigellosis, individuals may find it futile to seek help from a doctor at a clinic or hospital. Thus, both decisions can lead to underreporting.

In rural areas, the decision to seek treatment amongst the rural population can also depend on health insurance coverage, which may influence underreporting. To offset treatment and health care costs, there exists a rural health insurance known as the New Rural Cooperative Medical Scheme (NRCMS). About 98.9% of the rural population is enrolled in the NRCMS, but enrollment is voluntary (Liu et al., 2016). Up until 2009, most rural residents had no access to basic pension and health care insurance until the social insurance reform in recent years. Chinese patients do not have a regular family physician that they can seek for health check-ups, and as mentioned previously, they can choose where they seek health care based on their financial situation. While richer individuals in rural areas can afford to dominate inpatient care, some individuals in poorer rural areas cannot even afford health care expenses even after reimbursement from their insurance. Even those that can afford health care can fail to receive a follow up due to the lack of coordination amongst the various levels of health care

services (Sun et al., 2017). Thus, those that are wealthy and have access to insurance care have been found to be more likely to seek treatment compared to those that are uninsured.

Similar to the relationship between rural households and shigellosis incidence, the negative binomial model in this thesis also found that an increase in percentage of rural employment was associated with lower shigellosis risk. While underreporting is a conjecture of this case, it can also be assumed the rural workers are less likely to consult a doctor due to time spent at work and the lack of accessibility to healthcare. It has been found that in rural areas of China, farmers are particularly susceptible to shigellosis (Xu et al., 2014; Nie et al., 2014), perhaps due to the lack of access to safely managed water and sanitation.

The model used to study the spatial relationships between socioeconomic factors and shigellosis is the Bayesian spatial model estimated using the integrated nested Laplace approximations (INLA). This novel approach was applied by Ma et al. (2015) in their study, but they used the Poisson distribution to model the data. In this thesis, the negative binomial Bayesian model was adopted since it was found to better account for overdispersed incidence counts and avail a better model fit.

Results from the Bayesian spatial model, which is based on the average shigellosis rate of all four years (2011-2014), found that increasing the number of hospital beds per 1,000 people were associated with lower shigellosis risk. This result is valid as hospital bed distribution is geographically clustered in China (Pan & Shallcross, 2016). Moreover, this finding is similar to what Ma et al. (2015) found in Sichuan province and what H. Zhang et al. (2017) found in the Southwestern provinces (Sichuan, Tibet, and Yunnan). In general, there are more hospital beds in the urban areas of southern Jiangsu. Increasing the number of hospital

beds would allow more room for inpatient care. This would give patients the option to stay over for additional treatment and to avoid the lack of follow-ups that come with one-day visits.

The association between health institutions and increasing shigellosis incidence found in this thesis using the Bayesian model confirms a similar finding by Nie et al. (2014). This may be due to the fact that areas with a higher number of health institutions per 10,000 people typically had smaller institutions such as clinics that are relatively less equipped. These areas may be rural as areas with more health institutions correspond to areas with a higher percentage of rural households, as shown in Figure 4.4.

The relationship between rural income and shigellosis incidence was rarely explored in previous studies. In this thesis, it was found that the relationship between these factors are insignificant using both the negative binomial generalized linear model and the Bayesian negative binomial model in this thesis. This finding provides a contrast with the finding from Tang et al. (2014), which found that household income was negatively associated with shigellosis. Their study was conducted using a matched a case control (n=1200 cases of shigellosis and n=1270 control) representing all cities in Jiangsu province. The association between household income and shigellosis was calculated using an odds ratio analysis.

The discrepancy in results between that of this thesis and the study by Tang et al. (2014) may be due to 1) choice of methodology, 2) sample size, and 3) choice of income indicator. Compared to a risk ratio, an odds ratio estimate is extremely sensitive when 1) the disease of interest is not rare (incidence rate is greater than 1%) and 2) its equivalent risk ratio is not close to 1. When the risk ratio is greater or less than 1, the odd ratio would be even greater or smaller, which may overestimate or underestimate the degree of risk. Secondly, the same size used in

this study is much smaller than what was used in the case control study. Lastly, Tang et al. (2014) assessed household income while this study assessed rural household income per capita, which may differ spatially. Therefore, more data and analysis are required to confirm the relationship between rural income and shigellosis using the negative binomial model and Bayesian negative binomial model.

The non-spatial negative binomial regression model in this thesis has only identified the number of hospital beds per 1,000 people and the percentage of rural households to be the only statistically significant indicators in predicting shigellosis incidence. By taking into account the spatial relationships between areas, the Bayesian spatial model was able to provide additional information on how socioeconomic determinants were correlated with shigellosis incidence. However, more analysis is required to understand how these factors can facilitate or hinder access to safely managed water and sanitation. Future work should aim to confirm the findings of this thesis, and to determine whether the associated factors will significantly contribute to shigellosis incidence.

5.3 Limitations

This study was unable to cover all the factors listed under the themes within the conceptual framework. In the future, a follow-up study could be done to address these factors. Nevertheless, this thesis was still able to explore how major socioeconomic determinants were linked to shigellosis through empirical models, and provided information on the linkages between rural water and sanitation access and shigellosis.

As previously discussed, the underreporting of shigellosis is a limitation of this study and a previous study by Ma et al. (2015). Future studies should 1) aim to understand and account

for the institutional and socioeconomic mechanisms behind diarrhea underreporting and 2) aim to utilize primary data that can precisely represent the number of shigellosis cases amongst the rural population. Nevertheless, data collected by the Chinese CDC still represents the most complete secondary source of incidence dataset to date, and thus was adopted in this thesis.

This thesis is susceptible to several limitations in the data due to 1) lack of information on survey data collection procedure, 2) weak data quality assurance and control, and 3) availability and accessibility of shigellosis data. Firstly, there is a lack of information on the data collection procedure as none of the statistical books contained any information on how certain indicators were sampled. For instance, data collected on rural water sources and sanitation facilities were sampled from a small sample of 40 households in 2014, however no information was provided on the sampling technique adopted to collect this data. However, it can be assumed that the local bureau of statistics has adopted a systematic data collection process throughout the years.

Secondly, city statistics bureaus were not consistent in selecting and reporting their indicators. For example, some cities included the number of first aid centers in the total count for the number of health institutions, while others did not. This could impact the regression results. In addition, when socioeconomic information was reported at the city and county level, data reported at the city level included information on the city and also on the nearby counties. Since this coinciding information was not clarified in the statistical yearbooks, more preprocessing steps were required. Despite this, none of the data was missing and the spatial distribution in socioeconomic conditions across counties can still be identified and assessed.

Thirdly, shigellosis data obtained in aggregated form limits the application of statistical analyses that can be performed to explore the distribution of shigellosis cases within each county. The aggregated form of shigellosis data (cases per county) also limits the usefulness of other effective regression approaches (e.g. geographically weighted regression) that may be more effective at studying the spatial association between facilitators and barriers and shigellosis incidence. However, a Bayesian spatial regression model was adopted to account for the spatial effects between areas.

It is important to note that survey data have inherent errors and are subject to reliability concerns when adopting them as a data source. A survey samples only a small portion of the population and is subject to sampling and non-sampling errors. Usually an estimated error threshold is determined; if an estimated sampling error is greater than the threshold, the data is considered too unreliable to be published. However, estimates that barely qualified can still be published. On the other hand, it is difficult to identify and evaluate the scope of many non-sampling errors. Non-sampling errors include coverage and response. Coverage errors arise when the sampled population is not representative of the target population, while response errors arise when there is a non-response from the individual and household level. These errors can greatly affect the accuracy of survey results however many of these errors are inherent in research reliant on survey data.

The modifiable areal unit problem (MAUP) exists and creates a statistical bias when conducting any spatial research using choropleth maps and small area data. The spatial aggregation of point-based measures into administrative divisions such as districts, counties, and provinces can significantly affect the results of hot spot and regression analyses (King, 1969). For instance, a hot spot analysis conducted on choropleth map created using

neighbourhood boundaries would yield significantly different results compared to a choropleth map created using county boundaries.

Regression results of socioeconomic variables can be extremely sensitive to MAUP as it is highly dependent on scale and zoning system. Fotheringham and Wong (1991) found in their study when data was aggregated to 800 zones, every 0.1 increase in the proportion of elderly yielded a $308 decrease in the predicted mean family income. When the same data was aggregated to 25 zones, the same increase resulted in a decrease of $2654 in predicted mean family income. In addition, MAUP can be unpredictable in multivariate regression analysis as it may interact with the covariances between the independent variables. However, the MAUP problem is common and present in all research adopting small area data.

Another issue associated with analyzing data using choropleth maps is the edge effect. This occurs when there is interdependence between counties near the study area and outside of the bounded region (Diggle, 2003). It is difficult to account for this effect because regions outside of the bounded area are not analyzed. Thus it is important not to overinterpret counties with high z-scores located near the edges of the study area as they are only based on neighbours within the bounded region (O'Sullivan & Unwin, 2010).

5.4 Challenges with Secondary Data in Data Deprived Areas

Collecting secondary data in rural areas can be particularly challenging, especially in China. These challenges occur during the following stages: 1) finding what data is available for use, 2) requesting data from relevant authorities, and 3) identifying the limitations in the data, reliability of the data, and the extent of its representation. In this section, the challenges faced

while collecting secondary data in Jiangsu, China will be discussed. In addition, best practices taken to verify the reliability of the data used in this study will also be discussed.

Determining what data is available to access for research can be particularly challenging. In China, secondary data on environmental and socioeconomic conditions in rural areas are not available online. Researching what data is available online through databases and past research articles can help to confirm what information is available. In this case, my research told me that infectious disease incidence data could be obtained from the Chinese CDC. Rural socioeconomic and water and sanitation data are available through county statistical yearbooks, but they are only available in hard copies. Thus, collecting this data may require a trip to the rural county and relevant agency. Having strong research connections and collaborators on the ground can significantly help to facilitate this process.

Accessing the data used in this thesis required contacting relevant authorities and sending in a request through a local institution. The data collected for this thesis was requested through a local academic and research institution in Jiangsu, Nanjing University. This process can be long as it can take weeks to gain approval from both the university and data agency. When accessing health and disease data, it is important to note that the authorities will choose what data can be shared. Health and disease data, for the most part, is confidential in China. According to Sorenson et al. (1996), “it is well-known that general practitioners and hospitals do not always respond or do not accept the use of their records for research” in epidemiology and health. In the case that certain data cannot be obtained, an alternate plan should be devised.

After obtaining access to the data, reliability of the data should be acknowledged. When evaluating the reliability of the data, it is important to understand how the information in the

dataset was generated, how it was coded, and how consistent coding was across sites and at different times (Kimberlein & Winterstein, 2008). Validity and rigour are also supported when the same data set have been used by other peer-reviewed studies. For example, shigellosis incidence data was collected from the Jiangsu CDC, which sources its data from all admission reports of shigellosis from clinics and hospitals in Jiangsu province. This data was presented as counts and rates and was consistent for all five years from 2011 to 2015. Furthermore, shigellosis data from the CDC in China have been used by several studies (Tang et al., 2014; Ma et al., 2015; Nie et al., 2014; Xu et al., 2014).

In addition, socioeconomic data adopted in this thesis was collected in count form. This applies to factors used - the number of health institutions, number of hospital beds, number of rural households, and number of rural employees. In the statistical yearbook, these counts were reported in exact values to ensure they accurately represent the population. In addition, the counts for these indicators reported in the county statistical yearbooks also match those reported in the provincial and national level statistical yearbooks. Furthermore, these indicators have also been adopted in other peer-reviewed publications (Nie et al., 2014; Ma et al., 2015). In this thesis, some indicators were converted to a proportion using the total count for each indicator so that comparisons between different areas can be made.

When socioeconomic and shigellosis incidence were geocoded into choropleth maps, they were also checked to ensure that they fulfilled the elements of spatial data quality (Guptill and Morrison et al., 2013). In the context of choropleth maps, this included lineage, positional accuracy, attribute accuracy, logical consistency, and completeness (Statistics Canada, 2015). The key limitation faced while working with the socioeconomic dataset was overcoming the challenge in completeness. During the regression analysis, socioeconomic data was not

available for each individual district located within a major city. In this case, socioeconomic values was aggregated and linked to all the districts within that city. This created a geographic discrepancy, as spatial analysis could no longer be fully conducted at the county and district level. However, this was fully accounted for in the regression through an offset term, which is the exposed population based on all districts within the city and was used to adjust the rate of shigellosis.

The challenge associated with the water and sanitation data in this thesis was verifying how this secondary data was collected as the sampling method was not disclosed in the statistical yearbook. There is no way of knowing if the sample is representative of the population. Additionally, previous studies have not adopted this data before for analysis. While these limitations can greatly affect the reliability of the data, it is also important to acknowledge that this is the current "best" available data from a secondary source that can be readily obtained by the public for research purposes. When adopting such data for research, it is important to interpret the results with caution and refrain from drawing strong conclusions. Thus, only an exploratory analysis was conducted using this data. Identifying this limitation also helps to aid researchers continuing this work to find ways to improve this data to establish rigour.

5.5 Contributions

This thesis makes several contributions to existing literature on shigellosis in China. Firstly, this thesis contributes information on the spatial and temporal distribution of shigellosis incidence at the county level across Jiangsu, China from 2011 to 2015. The findings

of this analysis are up to date and can be translated for key stakeholders in health and social policy to target areas of high shigellosis incidence.

Secondly, this thesis identified the facilitators and barriers to safely managed water and sanitation, which is directly linked to shigellosis incidence, through a conceptual framework. The socioeconomic determinants of health and environmental policy can act as facilitators and barriers to achieving access to safely managed water and sanitation, which affects the incidence of shigellosis. This conceptual framework was used to set the direction of the quantitative analyses, which 1) explored the association between socioeconomic determinants of health and shigellosis risk and 2) the relationships between water and sanitation access and shigellosis incidence. However more analysis is required to understand how characteristics of these themes can act as facilitators and barriers as not all the factors identified in the conceptual framework were explored in this thesis.

Furthermore, this thesis presents the only study that has investigated the spatio-temporal association between socioeconomic determinants and shigellosis incidence in the eastern province of Jiangsu. Results from the negative binomial model showed that in 2011, 2012, and 2014, the number of hospital beds was a reoccurring significant indicator of shigellosis risk. The spatial Bayesian negative binomial model in this thesis found that areas with a higher percentage of rural households were associated with lower shigellosis risk. Furthermore, increasing the percentage of rural employment and the number of hospital beds, respectively, was linked with decreased shigellosis risk, while increasing the number of hospitals was associated with increased shigellosis risk. The findings of this thesis confirm that of previous studies conducted in China. It is recommended that these relationships be further explored using various methods to confirm the validity of the findings and its implications.

This study has also illustrated the application of a negative binomial regression model and a negative binomial Bayesian spatial model in studying the spatio-temporal association between socioeconomic factors and shigellosis risk. Few studies have adopted a negative binomial regression model or a Bayesian negative binomial regression spatial model to study infectious disease or diarrhea incidence. This study used both models, which allowed results to be compared and contrasted. While these models have proven to be suitable for fitting overdispersed infectious disease counts, a bigger sample size of disease incidence data should be used in the future to avail more robust results.

5.6 Directions for Future Research

Since this thesis was predominantly an exploratory study, a few recommendations are made to assist future investigations. Firstly, the varying strength of the association between various predictors and disease based on location should be studied further. For instance, certain socioeconomic determinants may be considered significant predictors of a disease in one county, but not in other counties. To address this limitation, it would be ideal to incorporate a geographically weighted regression model with the current model to ensure that predictors are properly weighted based on their importance and relevance in each county.

Secondly, a more active surveillance method may be required to gauge the occurrence of shigellosis. Current passive disease surveillance methods have been identified to undermine the occurrence of shigellosis amongst rural populations. Future studies should aim to understand and account for the institutional and socioeconomic mechanisms behind diarrhea underreporting.

Thirdly, while the exploratory analysis in this thesis aimed to understand rural water and sanitation access, findings were inconclusive. Thus, future studies should aim to understand how water and sanitation access in rural townships contributes to shigellosis. In addition, the effectiveness of the water source and sanitation facility adopted should also be explored over time. For instance, in Suining county, it was observed that the existence of "improved" sanitation facilities such as pour flush pit latrines has not shown to alleviate shigellosis incidence. Considering the ubiquitous of the pour flush pit latrine in rural China, this should be further explored.

Furthermore, it is important for future work to incorporate data on hygiene. In a review conducted by Bartram and Cairncross (2010), hand washing and better hygiene was found to have the greatest effect on reducing exposure to diarrhea. Perception of hygiene, as well as advocacy efforts of the government for better hygiene and hand washing should be further explored in the rural setting. It is also important to identify the interaction and relationships between hygiene and sanitation.

In addition, because this study used data that was predominantly from the period of the MDGs, future research that aims to follow up on this study should look at changes in the SDG era. Progress made in socioeconomic determinants and water and sanitation access will continue to change and improve with time. Thus, future studies should continue to explore the changes in shigellosis over time and space and how its incidence is affected by socioeconomic and environmental factors.

Lastly, it would be more feasible to combine other theoretical approaches (e.g. social interactionist) for a follow-up study. Integrating another approach can help to gain a better

understanding of how people cope with current sanitation practices and how increased awareness of adequate sanitation can help to reduce the prevalence of waterborne diseases such as shigellosis.

REFERENCES

Abubakar, I. I., Tillmann, T., & Banerjee, A. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet, 385*(9963), 117-171.

Adams, E. A., Boateng, G. O., & Amoyaw, J. A. (2016). Socioeconomic and demographic predictors of potable water and sanitation access in Ghana. *Social Indicators Research, 126*(2), 673-687.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93-115.

Arku, R. E., Bennett, J. E., Castro, M. C., Agyeman-Duah, K., Mintah, S. E., Ware, J. H., . . . Ezzati, M. (2016). Geographical inequalities and social and environmental risk factors for under-five mortality in Ghana in 2000 and 2010: bayesian spatial analysis of census data. *PLoS Med, 13*(6), e1002038.

Baer, J. T., Vugia, D. J., Reingold, A. L., Aragon, T., Angulo, F. J., & Bradford, W. Z. (1999). HIV infection as a risk factor for shigellosis. *Emerging Infectious Diseases, 5*(6), 820-823.

Bardhan, P., Faruque, A., Naheed, A., & Sack, D. A. (2010). Decreasing shigellosis-related deaths without Shigella spp.-specific interventions, Asia. *Emerg Infect Dis, 16*(11), 1718-1723.

Bartram, J., & Cairncross, S. (2010). Hygiene, Sanitation, and Water: Forgotten Foundations of Health. *PLoS Med, 7*(11), e1000367.

Bayles, B., & Allan, B. (2014). Social-ecological factors determine spatial variation in human incidence of tick-borne ehrlichiosis. *Epidemiology and Infection, 142*(09), 1911-1924.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics, 43*(1), 1-20.

Bessell, P. R., Matthews, L., Smith-Palmer, A., Rotariu, O., Strachan, N. J., Forbes, K. J., ... & Innocent, G. T. (2010). Geographic determinants of reported human Campylobacter infections in Scotland. *BMC public health, 10*(1), 423.

Biao, X. (2007). How far are the left- behind left behind? A preliminary study in rural China. *Population, Space and Place, 13*(3), 179-191.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology, 7*, 39-55.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* Cambridge university press.

Carlton, E. J., Liang, S., McDowell, J. Z., Li, H., Luo, W., & Remais, J. V. (2012). Regional disparities in the burden of disease attributable to unsafe water and poor sanitation in China. *Bulletin of the World Health Organization, 90*(8), 578-587.

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T., Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of Biomedical Informatics, 51*, 287-298.

CDC (Centers for Disease Control and Prevention). (March 31, 2017). *Shigella* - Shigellosis. Retrieved from https://www.cdc.gov/shigella/general-information.html#definitions-symptoms

Chao, J., Lu, B., Zhang, H., Zhu, L., Jin, H., & Liu, P. (2017). Healthcare system responsiveness in Jiangsu Province, China. *BMC health services research*, *17*(1), 31.

Chen, T., Leung, R. K., Zhou, Z., Liu, R., Zhang, X., & Zhang, L. (2014). Investigation of key interventions for shigellosis outbreak control in China. *PloS One, 9*(4), e95006.

Chen, Y., Huang, X., Xiao, Y., Jiang, Y., Shan, X., Zhang, J., . . . Liu, J. (2015). Spatial analysis of schistosomiasis in Hubei Province, China: a GIS-based analysis of schistosomiasis from 2009 to 2013. *PloS One, 10*(4), e0118362.

Cheng, J., Xie, M. Y., Zhao, K. F., Wu, J. J., Xu, Z. W., Song, J., ... & Wen, L. Y. (2017). Impacts of ambient temperature on the burden of bacillary dysentery in urban and rural Hefei, China. *Epidemiology & Infection*, *145*(8), 1567-1576.

Chitunhu, S., & Musenge, E. (2016). Spatial and socio-economic effects on malaria morbidity in children under 5years in Malawi in 2012. *Spatial and spatio-temporal epidemiology, 16*, 21-33.

Clasen, T., Haller, L., Walker, D., Bartram, J., & Cairncross, S. (2007). Cost-effectiveness of water quality interventions for preventing diarrhoeal disease in developing countries. *Journal of Water and Health, 5*(4), 599-608.

Colombara, D. V., Cowgill, K. D., & Faruque, A. S. (2013). Risk factors for severe cholera among children under five in rural and urban Bangladesh, 2000–2008: a hospital-based surveillance study. *PloS One, 8*(1), e54395.

Delmelle, E., Hagenlocher, M., Kienberger, S., & Casas, I. (2016). A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta Tropica, 164*, 169-176.

Diggle, P. J. (2003). Introduction: Edge effects. *Statistical Analysis of Spatial Point Patterns.* (2nd ed). London, United Kingdom: Arnold.

Ding, Z., Zhai, Y., Wu, C., Wu, H., Lu, Q., Lin, J., & He, F. (2017). Infectious diarrheal disease caused by contaminated well water in Chinese schools: A systematic review and meta-analysis. *Journal of Epidemiology*.

Dreibelbis, R., Greene, L. E., Freeman, M. C., Saboori, S., Chase, R. P., & Rheingans, R. (2013). Water, sanitation, and primary school attendance: A multi-level assessment of determinants of household-reported absence in Kenya. *International Journal of Educational Development, 33*(5), 457-465.

ESRI ArcGIS. (2017). Hot Spot Analys (Getis Ord Gi*). Retrieved from http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/hot-spot-analysis.htm

Evans, G. W., & Kantrowitz, E. (2002). Socioeconomic status and health: the potential role of environmental risk exposure. *Annual Review of Public Health, 23*(1), 303-331.

Farrington, C. P., Andrews, N. J., Beale, A. D., & Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 547-563.

Feng, X. L., Martinez-Alvarez, M., Zhong, J., Xu, J., Yuan, B., Meng, Q., & Balabanova, D. (2017). Extending access to essential services against constraints: the three-tier health service delivery system in rural China (1949–1980). *International journal for equity in health, 16*(1), 49.

Fornace, K. M., Abidin, T. R., Alexander, N., Brock, P., Grigg, M. J., Murphy, A., ... & Cox, J. (2016). Association between landscape factors and spatial patterns of Plasmodium knowlesi infections in Sabah, Malaysia. *Emerging infectious diseases, 22*(2), 201.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025-1044.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2007). *Quantitative geography: perspectives on spatial data analysis*. Sage.

Fotso, J., & Kuate-Defo, B. (2005). Socioeconomic inequalities in early childhood malnutrition and morbidity: modification of the household-level effects by the community SES. *Health & Place, 11*(3), 205-225.

Fradelos, E. C., Papathanasiou, I. V., Mitsi, D., Tsaras, K., Kleisiaris, C. F., & Kourkouta, L. (2014). Health Based Geographic Information Systems (GIS) and their Applications. *Acta Informatica Medica, 22*(6), 402-405.

Fuhrmeister, E. R., Schwab, K. J., & Julian, T. R. (2015). Estimates of nitrogen, phosphorus, biochemical oxygen demand, and fecal coliforms entering the environment due to inadequate sanitation treatment technologies in 108 low and middle income countries. *Environmental science & technology*, *49*(19), 11604-11611.

Gatrell, A. C., & Elliott, S. J. (2014). *Geographies of health: An introduction*. West Sussex, United Kingdom: John Wiley & Sons.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*(3), 189-206.

Gondhalekar, D., Nussbaum, S., Akhtar, A., Kebschull, J., Keilmann, P., Dawa, S., . . . Dorje, S. (2013). Water-related health risks in rapidly developing towns: the potential of integrated GIS-based urban planning. *Water International*, *38*(7), 902-920.

Green, C. G., Krause, D. O., & Wylie, J. L. (2006). Spatial analysis of campylobacter infection in the Canadian province of Manitoba. *International Journal of Health Geographics*, *5*(1), 2.

Gu, H., Fan, W., Liu, K., Qin, S., Li, X., Jiang, J., ... & Jiang, Q. (2017). Spatio-temporal variations of typhoid and paratyphoid fevers in Zhejiang Province, China from 2005 to 2015. *Scientific Reports*, *7*.

Guo, C., DU, Y., Shen, S. Q., Lao, X. Q., Qian, J., & Ou, C. Q. (2017). Spatiotemporal analysis of tuberculosis incidence and its associated factors in mainland China. *Epidemiology & Infection*, 1-10.

Guptill, S. C., & Morrison, J. L. (Eds.). (2013). *Elements of spatial data quality*. Elsevier.

Haan, M. (2013). *Introduction to Statistics for Canadian Social Scientists 2ED*. Oxford University Press.

Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., . . . Gething, P. W. (2013). Global mapping of infectious disease. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1614).

Held, L., Höhle, M., & Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical modelling*, *5*(3), 187-199.

Hemson, D. (2007). 'The Toughest of Chores': Policy and Practice in Children Collecting Water in South Africa. *Policy Futures in Education*, *5*(3), 315-326.

Herrador, B. R. G., De Blasio, B. F., MacDonald, E., Nichols, G., Sudre, B., Vold, L., ... & Nygård, K. (2015). Analytical studies assessing the association between extreme precipitation or temperature and drinking water-related waterborne infections: a review. *Environmental Health*, *14*(1), 29.

Hesketh, T., Jun, Y. X., Lu, L., & Mei, W. H. (2008). Health status and access to health care of migrant workers in China. *Public health reports*, *123*(2), 189-197.

Hilbe, J. M. (1994). Generalized linear models. *The American Statistician*, *48*(3), 255-265.

Huang, D., Guan, P., Guo, J., Wang, P., & Zhou, B. (2008). Investigating the effects of climate variations on bacillary dysentery incidence in northeast China using ridge regression and hierarchical cluster analysis. *BMC Infectious Diseases, 8*(1), 1.

Hughes, G., & Gorton, R. (2015). Inequalities in the incidence of infectious disease in the North East of England: a population-based study. *Epidemiology and Infection, 143*(01), 189-201.

Hunter, P. R., Risebro, H., Yen, M., Lefebvre, H., Lo, C., Hartemann, P., ... & Jaquenoud, F. (2014). Impact of the provision of safe drinking water on school absence rates in Cambodia: a quasi-experimental study. *PloS one, 9*(3).

Hutton, G., & Haller, L. (2004). *Evaluation of the costs and benefits of water and sanitation improvements at the global level.* Water, Sanitation, and Health, Protection of the Human Environment, World Health Organization.

Jahiel, A. R. (1998). The organization of environmental protection in China. *The China Quarterly*, *156*, 757-787.

Jiang, Y. (2015). China's water security: Current status, emerging challenges and future prospects. *Environmental Science & Policy*, *54*, 106-125.

Karagiannis-Voules, D. A., Scholte, R. G., Guimarães, L. H., Utzinger, J., & Vounatsou, P. (2013). Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLoS neglected tropical diseases, 7*(5), e2213.

Kimberlin, C. L., & Winetrstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23).

King, L. J. (1969). *Statistical Analysis in Geography*. Pearson Education, Limited.

Kotloff, K. L., Winickoff, J. P., Ivanoff, B., Clemens, J. D., Swerdlow, D. L., Sansonetti, P. J., . . . Levine, M. M. (1999). Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bulletin of the World Health Organization*, *77*(8), 651-666.

Kleinman, K., Lazarus, R., & Platt, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology*, *159*(3), 217-224.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL.: University of Chicago Press.

Land, K. C., McCall, P. L., & Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models: With empirical applications to criminal careers data. *Sociological Methods & Research, 24*(4), 387-442.

Lee, J. H., Han, G., Fulp, W. J., & Giuliano, A. R. (2012). Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study. *Epidemiology & Infection, 140*(6), 1087-1094.

Li, Y., Li, X., Liang, S., Fang, L., & Cao, W. (2013). Epidemiological features and risk factors associated with the spatial and temporal distribution of human brucellosis in China. *BMC Infectious Diseases, 13*(1), 547.

Li, H., Yao, W., Dong, G., Wang, L., Luo, Q., Wang, S., ... & Zhang, Q. (2016). Water and sanitation interventions to control diarrheal disease in rural China. *Journal of Water Sanitation and Hygiene for Development, 6*(4), 640-649.

Li, H., Wei, Y., Dong, G., Li, W., Luo, Q., Wang, S., Xiong, C., Zhang, Q. (2016). Water and sanitation interventions to control diarrheal disease in rural China. *Journal of Water Sanitation and Hygiene for Development*, *6*(4), 640-649.

Li, Z., Zhang, X., Hou, X., Xu, S., Zhang, J., Song, H., & Lin, H. (2015). Nonlinear and threshold of the association between meteorological factors and bacillary dysentery in Beijing, China. *Epidemiology and Infection, 143*(16), 3510-3519.

Ling, B., Den, T., Lu, Z., Min, L., Wang, Z., & Yuan, A. (1993). Use of night soil in agriculture and fish farming.

Liu, J., Wu, X., Li, C., Xu, B., Hu, L., Chen, J., & Dai, S. (2017). Identification of weather variables sensitive to dysentery in disease-affected county of China. *Science of the Total Environment*, *575*, 956-962.

Liu, X., Wong, H., & Liu, K. (2016). Outcome-based health equity across different social health insurance schemes for the elderly in China. *BMC health services research*, *16*(1), 9.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., ... & Black, R. E. (2017). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, *388*(10063), 3027-3035.

Liu, X., Liu, Z., Zhang, Y., & Jiang, B. (2017). The Effects of Floods on the Incidence of Bacillary Dysentery in Baise (Guangxi Province, China) from 2004 to 2012. *International journal of environmental research and public health*, *14*(2), 179.

Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS one, 2*(2).

Lu S., Lin, Y., Vikse, J. H., & Huang, C. (2016). Well- being of migrant and left- behind children in China: Education, health, parenting, and personal values. *International Journal of Social Welfare, 25*(1), 58-68.

Luoto, J., Levine, D., Albert, J., & Luby, S. (2014). Nudging to use: Achieving safe water behaviors in Kenya and Bangladesh. *Journal of Development Economics, 110*, 13-21.

Ma, Y., Zhang, T., Liu, L., Lv, Q., & Yin, F. (2015). Spatio-Temporal Pattern and Socio-Economic Factors of Bacillary Dysentery at County Level in Sichuan Province, China. *Scientific Reports, 5*, 15264.

Maantay, J. (2007). Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research. *Health & Place, 13*(1), 32-56.

Mani, S., Wierzba, T., & Walker, R. I. (2016). Status of vaccine research and development for Shigella. *Vaccine, 34*(26), 2887-2894.

Marks, S. J., & Schwab, K. J. (2015). Water supply in rural communities. *Routledge Handbook of Water and Health* (pp. 336-344) Routledge.

Marshall, C. & Rossman, G.B. (1995). *Designing Qualitative Research, Second Edition*. Sage, London.

Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 421-441.

Mayer, H. B., & Wanke, C. A. (1994). Diagnostic strategies in HIV-infected patients with diarrhea. *Aids, 8*(12), 1639-1648.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models, Second Edition* Taylor & Francis.

Mock, N. B., Sellers, T. A., Abdoh, A. A., & Franklin, R. R. (1993). Socioeconomic, environmental, demographic and behavioral factors associated with occurrence of diarrhea in young children in the Republic of Congo. *Social Science & Medicine, 36*(6), 807-816.

Monmonier, M., and H. de Blij. "How to lie with maps . Chicago: U." (1996).

Montgomery, M. A., & Elimelech, M. (2007). Water and sanitation in developing countries: including health in the equation. *Environmental Science & Technology, 41*(1), 17-24.

Murphy, R., Zhou, M., & Tao, R. (2016). Parents' Migration and Children's Subjective Well-being and Health: Evidence from Rural China. *Population, Space and Place, 22*(8), 766-780.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2010). (2nd ed). *Generalized linear models: with applications in engineering and the sciences*. New Jersey, USA: John Wiley & Sons.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine, 24*(17), 2695-2717.

Nauges, C., & Strand, J.Water hauling and girls' school attendance: Some new evidence from Ghana. *Environmental and Resource Economics,* 1-24.

Nelson, K., & Williams, C. (2007). *Infectious disease epidemiology: theory and practice.* (2nd ed). Boston, USA: Jones and Bartlett Publishers.

Nie, C., Li, H., Yang, L., Zhong, G., & Zhang, L. (2014). Socio-economic factors of bacillary dysentery based on spatial correlation analysis in Guangxi Province, China. *PloS One, 9*(7), e102020.

Niyogi, S. K. (2005). Shigellosis. *Journal of Microbiology (Seoul, Korea), 43*(2), 133-143.

Odone, A., Crampin, A. C., Mwinuka, V., Malema, S., Mwaungulu, J. N., Munthali, L., & Glynn, J. R. (2013). Association between Socioeconomic Position and Tuberculosis in a Large Population-Based Study in Rural Malawi. *Plos One, 8*(10), e77740.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A, 16*(1), 17-31.

Ortiz-Correa, J. S., Resende Filho, M., & Dinar, A. (2016). Impact of access to water and sanitation services on educational attainment. *Water Resources and Economics, 14,* 31-43.

O'Sullivan, D., & Unwin, D. (2003). *Geographic Information Analysis.* Wiley.

Pan, J., & Shallcross, D. (2016). Geographic distribution of hospital beds throughout China: a county-level econometric analysis. *International journal for equity in health, 15*(1), 179.

Pazhani, G., Ramamurthy, T., Mitra, U., Bhattacharya, S., & Niyogi, S. (2005). Species diversity and antimicrobial resistance of Shigella spp. isolated between 2001 and 2004 from hospitalized children with diarrhoea in Kolkata (Calcutta), India. *Epidemiology and Infection, 133*(06), 1089-1095.

Pearson, J., & McPhedran, K. (2008). A literature review of the non-health impacts of sanitation. *Waterlines, 27*(1), 48-61.

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., & Clements, A. C. A. (2008). *Spatial Analysis in Epidemiology*. Oxford, United Kingdom: Oxford University Press.

Phung, D., Huang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., . . . Nguyen, T. (2015). Temporal and spatial patterns of diarrhoea in the Mekong Delta area, Vietnam. *Epidemiology and Infection, 143*(16), 3488-3497.

Potts, J. M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling, 199*(2), 153-163.

Public Health Agency of Canada. (2011a). What determines health? Retrieved from http://www.phac-aspc.gc.ca/ph-sp/determinants/index-eng.php

Public Health Agency of Canada. (2011b). Shigella SPP. Pathogen Safety Data Sheet - Infectious Substances. Retrieved from http://www.phac-aspc.gc.ca/lab-bio/res/psds-ftss/shigella-eng.php

Pruss-Ustun, A. (2008). *Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health.* Geneva, Switzerland: World Health Organization.

Qian, S. S. (2016). *Environmental and Ecological Statistics with R, Second Edition*. Boca Raton, USA: CRC Press.

Qiu, J. (2011). Environmental science. China to spend billions cleaning up groundwater. *Science (New York, N.Y.), 334*(6057), 745.

Qu, M., Zhang, X., Liu, G., Huang, Y., Jia, L., Liang, W., ... & Kan, B. (2014). An eight-year study of Shigella species in Beijing, China: serodiversity, virulence genes, and antimicrobial resistance. *The Journal of Infection in Developing Countries, 8*(07), 904-908.

Rego, R. F., Moraes, L. R., & Dourado, I. (2005). Diarrhoea and garbage disposal in Salvador, Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 99*(1), 48-54.

Rheinländer, T., Konradsen, F., Keraita, B., Apoya, P., & Gyapong, M. (2015). Redefining shared sanitation. *Bulletin of the World Health Organization, 93*(7), 509-510.

Richardson, D. B., Volkow, N. D., Kwan, M. P., Kaplan, R. M., Goodchild, M. F., & Croyle, R. T. (2013). Medicine. Spatial turn in health research. *Science (New York, N.Y.), 339*(6126), 1390-1392.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145-1165.

Roka, M., Goñi, P., Rubio, E., & Clavel, A. (2012). Prevalence of intestinal parasites in HIV-positive patients on the island of Bioko, Equatorial Guinea: Its relation to sanitary conditions and socioeconomic factors. *Science of the Total Environment, 432*, 404-411.

Rong, Z., HongXing, L., & XianFeng, W. (2009). Current situation analysis on China rural drinking water quality. *Journal of Environment and Health, 26*(1), 3-5.

Sack, R. B., Rahman, M., Yunus, M., & Khan, E. H. (1997). Antimicrobial resistance in organisms causing diarrheal disease. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 24 Suppl 1*, S102-5.

Saha, K., & Paul, S. (2005). Bias- corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics, 61*(1), 179-185.

Sanchez, T. H., Brooks, J. T., Sullivan, P. S., Juhasz, M., Mintz, E., Dworkin, M. S., . . . Adult/Adolescent Spectrum of HIV Disease Study Group. (2005). Bacterial diarrhea in persons with HIV infection, United States, 1992-2002. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 41*(11), 1621-1627.

Sarkar, R., Prabhakar, A. T., Manickam, S., Selvapandian, D., Raghava, M. V., Kang, G., & Balraj, V. (2007). Epidemiological investigation of an outbreak of acute diarrhoeal disease using geographic information systems. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 101*(6), 587-593.

SaTScan. (2005). Data Types and Methods. Retrieved from https://www.satscan.org/

Spatial data quality elements. (2017). *Statcan.gc.ca*. Retrieved 8 October 2017, from http://www.statcan.gc.ca/pub/92-195-x/2011001/other-autre/qua-eng.htm

Schmidt, W. P. (2014). The elusive effect of water and sanitation on the global burden of disease. *Tropical medicine & international health, 19*(5), 522-527.

Shaheed, A., Orgill, J., Montgomery, M. A., Jeuland, M. A., & Brown, J. (2014). Why? improved? water sources are not always safe. *Bulletin of the World Health Organization, 92*(4), 283-289.

Simonsen, J., Frisch, M., & Ethelberg, S. (2008). Socioeconomic risk factors for bacterial gastrointestinal infections. *Epidemiology, 19*(2), 282-290.

Shekhar, S., Yoo, E., Ahmed, S., Haining, R., & Kadannolly, S. (2017). Analysing malaria incidence at the small area level for developing a spatial decision support system: A case study in Kalaburagi, Karnataka, India. *Spatial and Spatio-Temporal Epidemiology, 20*, 9-25.

Sobsey, M. D., Stauber, C. E., Casanova, L. M., Brown, J. M., & Elliott, M. A. (2008). Point of use household drinking water filtration: a practical, effective solution for providing sustained access to safe drinking water in the developing world. *Environmental Science & Technology, 42*(12), 4261-4267.

Sorenson, S. B., Morssink, C., & Campos, P. A. (2011). Safe access to safe water in low income countries: water fetching in current times. *Social Science & Medicine, 72*(9), 1522-1526.

Sorenson, H. T., Sabroe, S., & OLSEN, J. (1996). A framework for evaluation of secondary data sources for epidemiological research. *International journal of epidemiology, 25*(2), 435-442.

Stauber, C., & Casanova, L. (2015). Drinking water contamination. *Routledge Handbook of Water and Health* (pp. 144-150). New York, USA: Routledge.

Suining Census Bureau. (2016). *Suining Statistical Yearbook*. China Statistics Press.

Sun, Y., Gregersen, H., & Yuan, W. (2017). Chinese health care system and clinical epidemiology. *Clinical Epidemiology, 9*, 167.

Tang, F., Cheng, Y., Bao, C., Hu, J., Liu, W., Liang, Q., . . . Chen, F. (2014). Spatio-Temporal Trends and Risk Factors for *Shigella* from 2001 to 2011 in Jiangsu Province, People's Republic of China. *Plos One, 9*(1), e83487.

Tao, T., & Xin, K. (2014). Public health: A sustainable plan for China's drinking water. *Nature, 511*(7511), 527-528.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(sup1), 234-240.

Tong, Y., Bu, X., Chen, C., Yang, X., Lu, Y., Liang, H., ... & Zhou, F. (2017). Impacts of sanitation improvement on reduction of nitrogen discharges entering the environment from human excreta in China. *Science of the Total Environment, 593*, 439-448.

Unkel, S., Farrington, C., Garthwaite, P. H., Robertson, C., & Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 175*(1), 49-82.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson Vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?. *Ecology, 88*(11), 2766-2772.

Von Seidlein, L., Kim, D. R., Ali, M., Lee, H., Wang, X., Thiem, V. D., . . . Bhutta, Z. A. (2006). A multicentre study of Shigella diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS Med*, *3*(9), e353.

Waddington, H., Snilstveit, B., White, H., & Fewtrell, L. (2009). Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries. *New Delhi: International Initiative for Impact Evaluation*.

Wagner, E. G., & Lanoix, J. N. (1958). Excreta disposal for rural areas and small communities. *Excreta Disposal for Rural Areas and Small Communities.*

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, *92*(438), 607-617.

Wang, Q., & Yang, Z. (2016). Industrial water pollution, water environment treatment, and health risks in China. *Environmental Pollution*, *218*, 358-365.

Wang, M., Webber, M., Finlayson, B., & Barnett, J. (2008). Rural industries and water pollution in China. *Journal of Environmental Management*, *86*(4), 648-659.

Wang, X. Y., Tao, F., Xiao, D., Lee, H., Deen, J., Gong, J., ... & Song, Y. (2006). Trend and disease burden of bacillary dysentery in China (1991-2000). *Bulletin of the World Health Organization*, *84*(7), 561-568.

Wang, X. Y., Du, L., Von Seidlein, L., Xu, Z. Y., Zhang, Y. L., Hao, Z. Y., ... & Han, C. Q. (2005). Occurrence of shigellosis in the young and elderly in rural China: results of a 12-month population-based surveillance study. *The American journal of tropical medicine and hygiene*, *73*(2), 416-422.

Weisent, J., Rohrbach, B., & Dunn, J. R. (2012). Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. *International Journal of Health Geographics*, *11*(1), 45.

Wen, M., & Lin, D. (2012). Child development in rural China: Children left behind by their migrant parents and children of nonmigrant families. *Child Development*, *83*(1), 120-136.

Wijayanti, S. P., Sunaryo, S., Suprihatin, S., McFarlane, M., Rainey, S. M., Dietrich, I., ... & Kohl, A. (2016). Dengue in Java, Indonesia: Relevance of Mosquito Indices as Risk Predictors. *PLoS neglected tropical diseases*, *10*(3), e0004500.

Wilking, H., Höhle, M., Velasco, E., Suckau, M., & Eckmanns, T. (2012). Ecological analysis of social risk factors for Rotavirus infections in Berlin, Germany, 2007–2009. *International Journal of Health Geographics*, *11*(1), 37.

Wolcott, H. (1995). *The Art of Fieldwork*. AltaMira Press, Walnut Creek. CA.

Woldemicael, G. (2001). Diarrhoeal Morbidity among Young Children in Eritrea: Environmental and Socioeconomic Determinants. *Journal of Health, Population and Nutrition, 19*(2), 83-90.

Wolf, J., Prüss- Ustün, A., Cumming, O., Bartram, J., Bonjour, S., Cairncross, S., . . . France, J. (2014). Systematic review: assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle- income settings: systematic review and meta-regression. *Tropical Medicine & International Health, 19*(8), 928-942.

WHO.(2017a). *Water-related diseases*. Retrieved from http://www.who.int/water_sanitation_health/diseases-risks/diseases/diarrhoea/en/

WHO. (2017b). *Diarrhoeal disease*. Retrieved from http://www.who.int/mediacentre/factsheets/fs330/en/

WHO (2015). *Millennium Development Goals (MDGs)*. Retrieved from http://www.who.int/mediacentre/factsheets/fs290/en/ [Accessed 4 Aug. 2017].

WHO/UNICEF. (2017). *Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines.* Geneva, Switzerland: World Health Organization.

WHO/UNICEF Joint Water Supply, Sanitation Monitoring Programme, World Health Organization, & UNICEF. (2008). *Progress on drinking water and sanitation: Special focus on sanitation.* Geneva, Switzerland: World Health Organization.

World Health Organization. (2005). *Guidelines for the control of shigellosis, including epidemics due to Shigella dysenteriae type 1*. Geneva, Switzerland: World Health Organization.

Xia, S., Xu, B., Huang, L., Zhao, J. Y., Ran, L., Zhang, J., ... & Hendriksen, R. S. (2011). Prevalence and characterization of human Shigella infections in Henan Province, China, in 2006. *Journal of clinical microbiology*, *49*(1), 232-242.

Xiao, G., Xu, C., Wang, J., Yang, D., & Wang, L. (2014). Spatial–temporal pattern and risk factor analysis of bacillary dysentery in the Beijing–Tianjin–Tangshan urban region of China. *BMC public health*, *14*(1), 998.

Xie, J. (2009). *Addressing China's water scarcity: recommendations for selected water resource management issues*, World Bank Publications.

Xu, Z., Hu, W., Zhang, Y., Wang, X., Tong, S., & Zhou, M. (2014). Spatiotemporal pattern of bacillary dysentery in China from 1990 to 2009: what is the driver behind? *PloS One*, *9*(8), e104329.

Yang, C., Sangthong, R., Chongsuvivatwong, V., McNeil, E., & Lu, L. (2009). Effect of village income and household income on sanitation facilities, hygiene behaviours and

child undernutrition during rapid economic growth in a rural cross-border area, Yunnan, China. *Journal of Epidemiology and Community Health, 63*(5), 403-407.

Yang, H., Wright, J. A., & Gundry, S. W. (2012). Water: Improve access to sanitation in China. *Nature, 488*(7409), 32-32.

Yi-Xin, H., & Manderson, L. (2005). The social and economic context and determinants of schistosomiasis japonica. *Acta Tropica, 96*(2–3), 223-231.

Yu, X., Geng, Y., Heck, P., & Xue, B. (2015). A review of China's rural water management. *Sustainability, 7*(5), 5773-5792.

Zhang, H., Si, Y., Wang, X., & Gong, P. (2016). Patterns of bacillary dysentery in China, 2005–2010. *International journal of environmental research and public health, 13*(2), 164.

Zhang, J., Jin, H., Hu, J., Yuan, Z., Shi, W., Yang, X., ... & Meng, J. (2014). Antimicrobial resistance of Shigella spp. from humans in Shanghai, China, 2004–2011. *Diagnostic microbiology and infectious disease, 78*(3), 282-286.

Zhang, J. (2012). The impact of water quality on health: Evidence from the drinking water infrastructure program in rural China. *Journal of Health Economics, 31*(1), 122-134.

Zhang, J., Mauzerall, D. L., Zhu, T., Liang, S., Ezzati, M., & Remais, J. V. (2010). Environmental health in China: progress towards clean air and safe water. *The Lancet, 375*(9720), 1110-1119.

Zhang, Y., Bi, P., Hiller, J. E., Sun, Y., & Ryan, P. (2007). Climate variations and bacillary dysentery in northern and southern cities of China. *Journal of Infection, 55*(2), 194-200.

Zhang, J., & Xu, L. C. (2016). The long-run effects of treated water on education: The rural drinking water program in China. *Journal of Development Economics, 122*, 1-15.

Zhao, X., Chen, J., Chen, M., Lv, X., Jiang, Y., & Sun, Y. (2014). Left-behind children in rural China experience higher levels of anxiety and poorer living conditions. *Acta Paediatrica, 103*(6), 665-670.

Zhao, J., Liao, J., Huang, X., Zhao, J., Wang, Y., Ren, J., . . . Ding, F. (2016). Mapping risk of leptospirosis in China using environmental and socioeconomic data. *BMC Infectious Diseases, 16.*

Zhao, L., Xiong, Y., Meng, D., Guo, J., Li, Y., Liang, L., ... & Zhang, L. (2017). An 11-year study of shigellosis and Shigella species in Taiyuan, China: Active surveillance, epidemic characteristics, and molecular serotyping. *Journal of Infection and Public Health.*

Zhang, H., Si, Y., Wang, X., & Gong, P. (2017). Environmental Drivers and Predicted Risk of Bacillary Dysentery in Southwest China. International Journal of Environmental Research and Public Health, 14(7), 782.

APPENDIX A: REGRESSION RESULTS

Table 1 Poisson regression model results

Year	Variable	Coefficient Estimate	SE	P – value (>|z|)	Significance
2011	(Intercept)	2.61E+00	1.52E-01	2.00E-16	***
	Rural	-1.33E-04	1.48E-03	2.00E-16	***
	HealthH	-2.12E-01	1.68E-02	2.00E-16	***
	HospBeds	5.33E-01	1.54E-02	2.00E-16	***
	Income	-9.29E-05	6.74E-06	3.12E-03	**
	Emply	-8.74E-03	2.96E-03	<2.00E-16	***
2012	(Intercept)	4.18E-01	1.79E-01	1.94E-02	*
	Rural	-4.14E-03	1.14E-03	2.63E-04	***
	HealthH	-4.20E-02	2.29E-02	6.68E-02	.
	HospBeds	5.35E-01	1.22E-02	<2.00E-16	***
	Income	-2.13E-05	7.01E-06	2.42E-03	**
	Emply	4.22E-03	1.77E-03	1.70E-02	*
2013	(Intercept)	-2.05E-02	1.95E-01	9.16E-01	
	Rural	-3.31E-02	2.71E-03	<2.00E-16	***
	HealthH	1.02E-01	2.12E-02	1.52E-06	***
	HospBeds	3.71E-01	1.59E-02	<2.00E-16	***
	Income	3.25E-05	6.15E-06	1.27E-07	***
	Emply	3.40E-02	4.27E-03	1.56E-15	***
2014	(Intercept)	2.24E+00	1.83E+00	2.27E-01	
	Rural	-9.08E-02	3.05E-02	4.46E-03	**
	HealthH	1.16E-01	1.45E-01	4.26E-01	
	HospBeds	-8.60E-03	1.13E-01	9.40E-01	
	Income	3.70E-05	5.83E-05	5.29E-01	
	Emply	9.51E-02	4.53E-02	4.07E-02	*

Signif. codes: 0‘***’, 0.001‘**’, 0.01‘*’, 0.05‘.’, 0.1‘ ’ ,1

Table 2 Negative binomial regression model results

Year	Variable	Coefficient Estimate	SE	P – value (>|z|)	Significance
2011	(Intercept)	3.88E+00	1.66E+00	0.0195	*
	Rural	-9.31E-03	1.89E-02	0.6222	
	HealthH	-1.76E-01	1.43E-01	0.2183	
	HospBeds	3.72E-01	1.50E-01	0.0127	*
	Income	-1.18E-04	7.72E-05	0.127	
	Emply	-7.14E-03	2.78E-02	0.7971	
2012	(Intercept)	1.29E+00	2.07E+00	0.5334	
	Rural	-2.21E-02	1.61E-02	0.1691	
	HealthH	1.12E-01	2.00E-01	0.5743	
	HospBeds	2.80E-01	1.62E-01	0.0834	.
	Income	3.32E-06	7.98E-05	0.9668	
	Emply	1.22E-02	1.73E-02	0.4806	
2013	(Intercept)	3.24E+00	2.25E+00	0.15	
	Rural	-3.46E-02	2.62E-02	0.186	
	HealthH	1.44E-01	1.91E-01	0.45	
	HospBeds	3.33E-02	1.75E-01	0.849	
	Income	-1.72E-05	7.00E-05	0.806	
	Emply	3.70E-03	3.55E-02	0.917	
2014	(Intercept)	4.58E+00	1.91E+00	0.01632	*
	Rural	-6.89E-02	2.55E-02	0.00693	**
	HealthH	3.95E-03	1.30E-01	0.97576	
	HospBeds	-2.28E-01	1.31E-01	0.08319	.
	Income	-3.95E-06	5.90E-05	0.94655	
	Emply	5.52E-02	3.65E-02	0.1301	

Signif. codes: 0‘***’, 0.001‘**’, 0.01‘*’, 0.05‘.’, 0.1‘ ’ ,1

Table 3 Negative binomial model with significant variable(s) only

Year	Variable	Model	Coefficient Estimate	SE	P – value (>|z|)	Significance
2011	(Intercept)	1	6.63E-01	3.92E-01	9.07E-02	.
	HospBeds		4.43E-01	1.12E-01	7.97E-05	***
2012	(Intercept)	1	5.21E-01	4.85E-01	2.83E-01	
	HospBeds		3.73E-01	1.23E-01	2.46E-03	**
2014	(Intercept)	1	4.24211	1.19757	0.000397	***
	HospBeds		-0.2685	0.13262	0.042907	*
	Rural		-0.02673	0.01159	0.021167	*

Signif. codes: 0‘***’, 0.001‘**’, 0.01‘*’, 0.05‘.’, 0.1‘ ’

APPENDIX B: SCATTER PLOTS

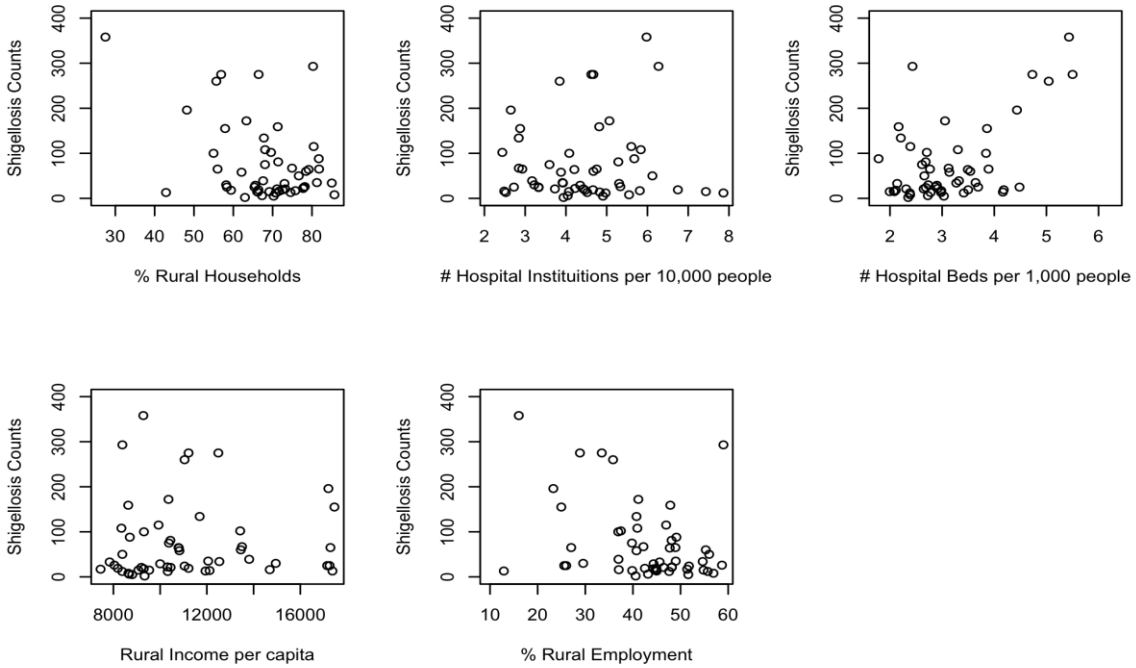

Figure 1 2011 scatterplots

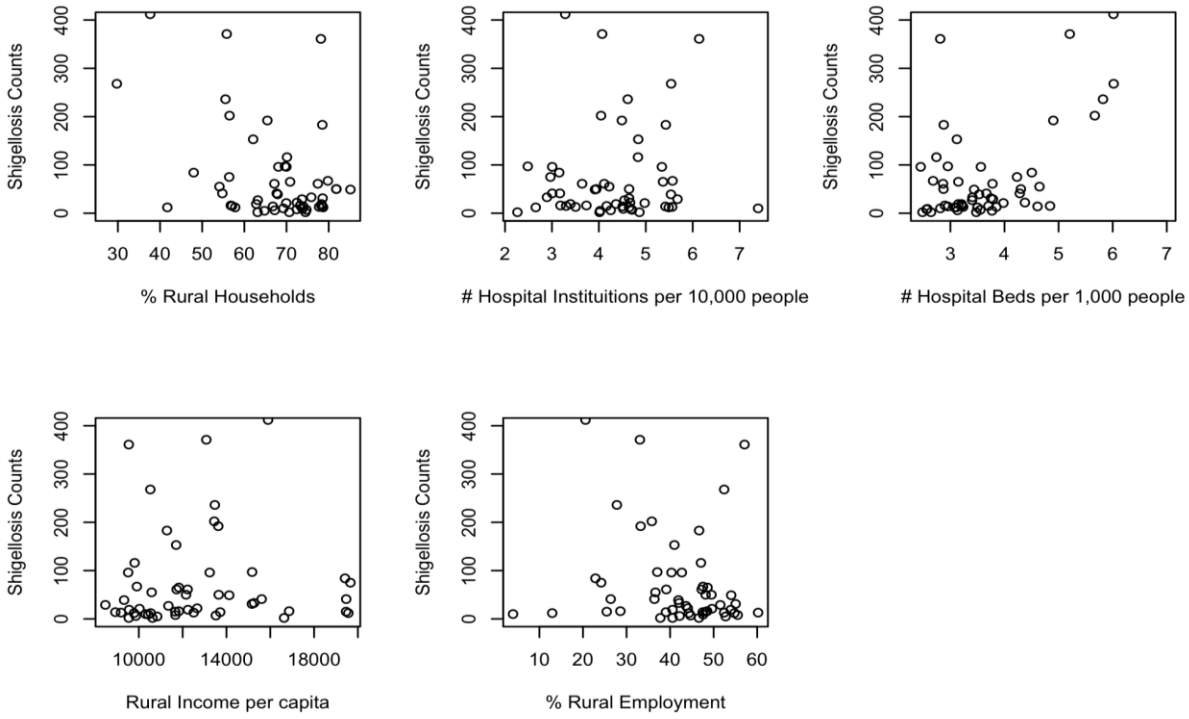

Figure 2 2012 scatterplots

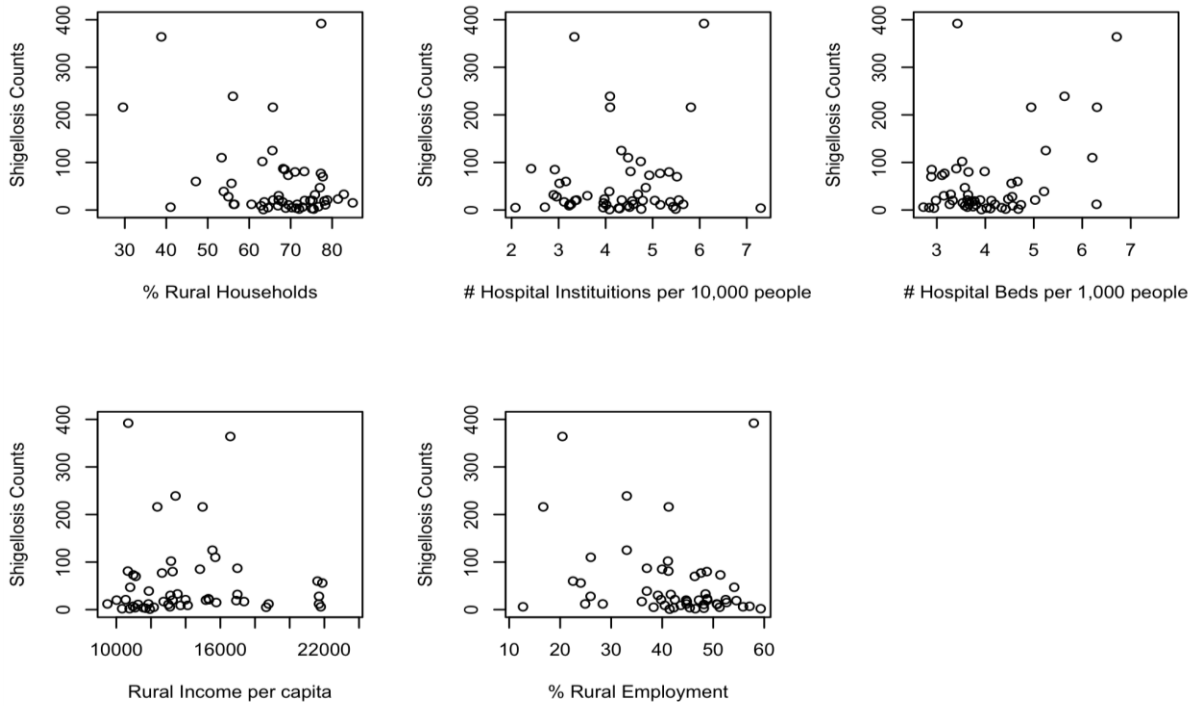

Figure 3 2013 scatterplots

APPENDIX C: EXPLORATORY PLOTS

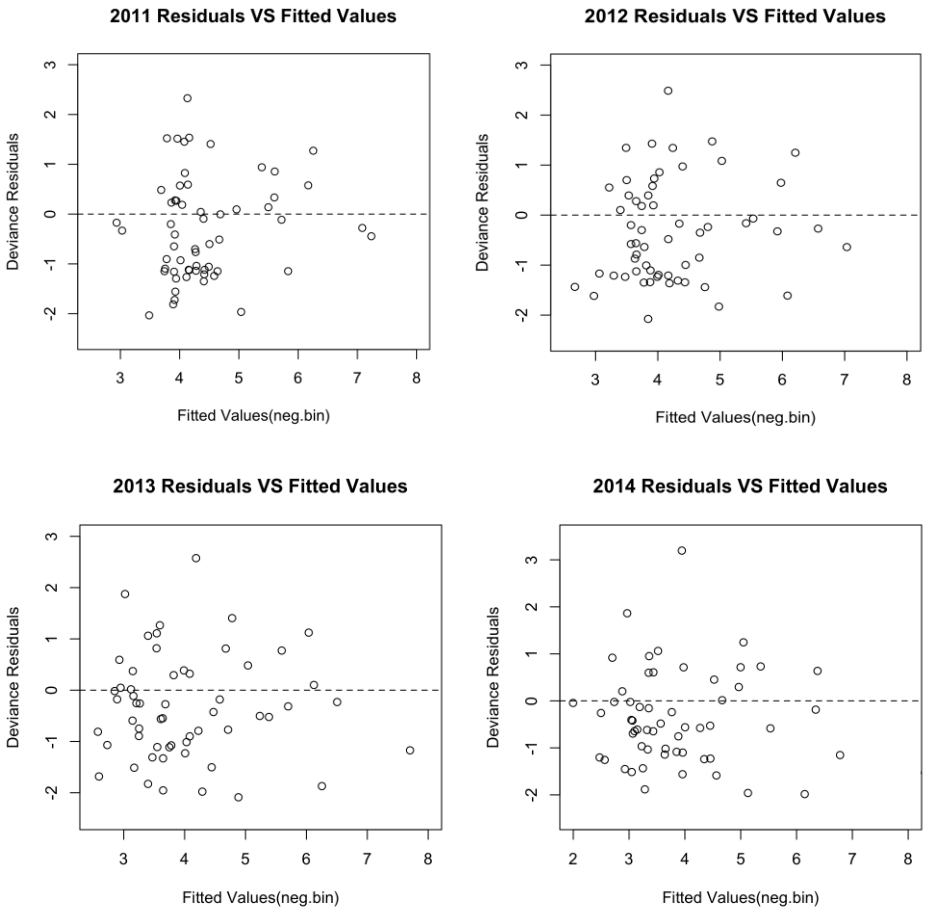

Figure 1 Deviance Residuals VS Fitted Values Plots

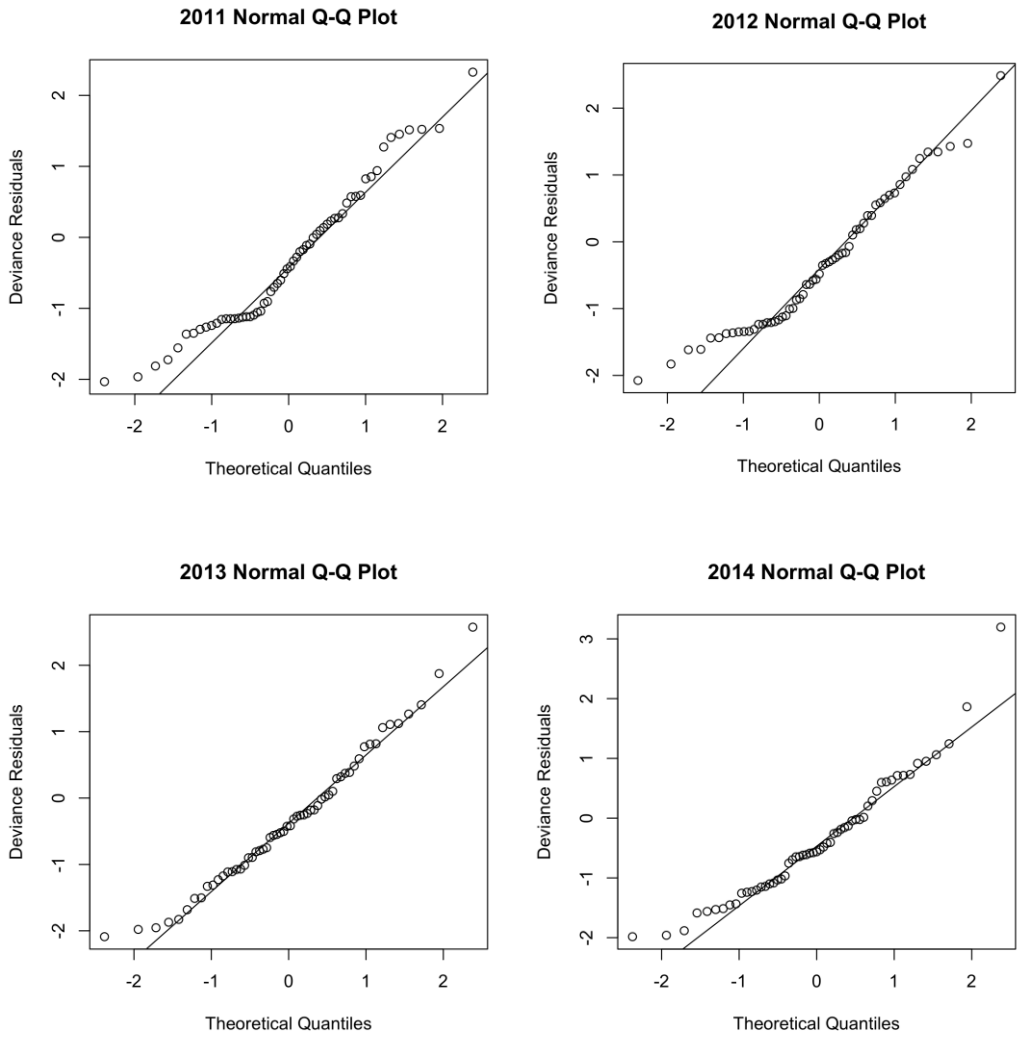

Figure 2 Normal Q-Q Plots

APPENDIX D: SOCIOECONOMIC DATA

2011

Obs		Rural Households%	Health Institutions/capita	Hospital beds/capita	Rural Income/capita	Rural Emply
1	Nanjing City	30.42	2.80	4.25	13108	15.05
2	Wuxi City (AD)	28.14	3.00	5.36	16438	12.13
3	Jiangyin City	57.91	2.88	3.86	17460	24.96
4	Yixing City	58.18	3.22	2.73	14949	29.56
5	Xuzhou City (AD)	71.55	4.83	6.27	10934	28.53
6	Feng County	85.71	5.55	2.39	8642	56.92
7	Pei County	65.60	4.35	2.90	10001	44.33
8	Suining County	80.27	6.27	2.43	8384	59.01
9	Xinyi City	71.30	4.82	2.17	8634	47.85
10	Pizhou City	80.44	5.60	2.39	9931	46.98
11	Changzhou City (AD)	54.51	2.13	3.81	15185	24.08
12	Liyang City	74.93	2.84	3.12	13505	42.21
13	Jintan City	67.59	3.17	3.32	13812	36.97
14	Suzhou City (AD)	26.25	2.54	4.63	17138	9.90
15	Changshu City	56.00	2.93	3.89	17289	27.04
16	Zhangjiagang City	58.36	3.33	4.48	17252	25.99
17	Kunshan City	42.89	2.53	2.78	17374	12.90
18	Wujiang	65.38	2.72	3.69	17150	25.57
19	Taicang City	48.20	2.65	4.43	17201	23.29
20	Nantong City (AD)	56.91	4.67	5.50	12491	28.88
21	Haian County	72.58	4.67	4.18	11216	44.94
22	Rudong County	81.81	4.75	2.77	10786	48.93
23	Qidong City	85.04	3.93	3.27	12535	54.65
24	Rugao City	77.75	4.22	2.92	10312	48.18
25	Haimen City	78.49	4.67	3.54	13453	55.30
26	Liangyungang City (AD)	27.48	5.98	5.43	9281	16.04
27	Ganyu County	69.17	7.43	2.08	9068	44.68
28	Donghai City	81.73	5.68	1.78	8701	49.12
29	Guanyun County	73.01	5.31	2.14	7839	45.61
30	Guannan County	75.77	5.81	2.98	7451	51.40
31	Huaian City (AD)	54.98	4.08	3.84	9307	36.92
32	Lianshui County	78.17	5.34	2.87	8043	58.70
33	Hongze County	66.05	4.07	1.99	9532	54.83
34	Xuyi County	70.25	4.91	3.03	8807	51.63
35	Jinhu County	62.96	3.94	2.35	9336	40.57
36	Yancheng City (AD)	56.37	4.06	4.51	11606	33.38
37	Xiangshui County	67.31	4.04	2.73	8673	43.17
38	Binhai County	73.19	4.43	2.31	9197	46.45
39	Funing County	59.42	4.47	2.11	9299	44.31
40	Sheyang County	68.05	3.60	2.61	10377	39.79
41	Jianhu County	63.33	5.07	3.06	10358	41.14
42	Dongtai City	81.26	3.92	3.64	12056	48.99
43	Dafeng City	74.56	4.52	2.98	11941	44.97
44	Yangzhou City (AD)	66.43	4.62	4.73	11217	33.46
45	Baoying County	70.74	4.97	2.38	10327	55.72
46	Yizheng City	62.10	3.88	3.14	10826	40.75
47	Gaoyou City	71.11	3.73	2.65	10449	48.16
48	Zhenjiang City (AD)	38.15	3.07	5.17	12165	21.24
49	Danyang City	69.59	2.44	2.71	13426	37.47
50	Yangzhong City	71.65	2.49	2.07	14692	37.03
51	Jurong City	67.78	2.84	2.21	11692	40.75
52	Taizhou City (AD)	55.68	3.85	5.04	11046	35.83
53	Xinghua City	71.40	5.29	2.68	10439	48.12
54	Jingjiang City	66.23	4.83	4.17	12116	39.85
55	Taixing City	77.87	3.33	2.69	11047	51.76
56	Jiangyan City	79.16	4.21	3.49	10802	47.74
57	Suqian City (AD)	68.08	5.84	3.30	8344	40.93
58	Shuyang County	76.65	6.12	2.66	8383	56.01
59	Siyang County	70.84	7.86	3.42	8379	47.62
60	Sihong County	66.38	6.75	3.50	8189	42.51

Legend

AD Represents all districts within the city

2012

Obs	County	Rural Household %	Health Institutions/capita	Hospital Beds/capita	Rural Income/capita	Rural Emply
1	Nanjing City	29.95	2.82	4.63	14786	14.88
2	Wuxi City (AD)	27.03	2.98	5.45	18830	11.77
3	Jiangyin City	56.41	2.97	4.23	19660	24.18
4	Yixing City	56.79	3.18	3.21	16862	28.59
5	Xuzhou City (AD)	51.77	4.64	6.98	12421	28.05
6	Feng County	78.67	5.49	3.50	9783	54.76
7	Pei County	63.18	4.55	3.40	11351	43.64
8	Suining County	78.15	6.14	2.81	9541	57.09
9	Xinyi City	70.11	4.83	2.74	9808	47.09
10	Pizhou City	78.49	5.43	2.88	11282	46.67
11	Changzhou City (AD)	51.53	2.17	4.71	17582	23.04
12	Liyang City	75.88	2.89	3.41	15261	42.08
13	Jintan City	67.67	3.17	3.67	15608	36.41
14	Suzhou City (AD)	34.02	2.70	4.65	19276	12.98
15	Changshu City	54.78	3.01	4.28	19467	26.37
16	Zhangjiagang City	57.07	3.30	4.84	19460	25.45
17	Kunshan City	41.73	2.65	3.08	19563	12.94
18	Taicang City	47.98	3.16	4.50	19411	22.86
19	Nantong City (AD)	55.52	4.61	5.82	13469	27.78
20	Haian County	72.47	4.67	4.38	12663	44.06
21	Rudong County	81.82	4.65	2.87	12156	48.15
22	Qidong City	85.16	3.92	3.44	14127	54.03
23	Rugao City	78.28	4.17	3.70	11663	48.17
24	Haimen City	78.56	4.64	3.75	15162	55.05
25	Liangyungang City (AD)	29.76	5.54	6.02	10525	52.41
26	Ganyu County	69.26	7.39	2.81	10310	3.94
27	Donghai City	79.81	5.57	2.68	9910	47.56
28	Guanyun County	73.97	5.42	2.95	8929	47.50
29	Guannan County	73.66	5.68	3.79	8472	51.54
30	Huaian City (AD)	54.09	4.22	4.65	10585	36.64
31	Lianshui County	77.73	5.56	3.85	9185	60.21
32	Hongze County	64.77	4.03	3.77	10838	52.75
33	Xuyi County	69.93	4.98	3.98	10031	49.60
34	Jinhu County	63.14	4.01	2.64	10624	40.57
35	Yancheng City (AD)	55.82	4.07	5.21	13081	33.07
36	Xiangshui County	67.20	4.25	3.13	9861	42.18
37	Binhai County	73.78	4.52	2.57	10429	47.59
38	Funing County	57.82	4.65	3.11	10545	44.31
39	Sheyang County	67.11	3.64	2.86	11726	39.13
40	Jianhu County	62.08	4.84	3.12	11705	40.99
41	Dongtai City	81.80	3.95	4.30	13647	49.51
42	Dafeng City	74.73	4.71	3.55	13517	44.69
43	Yangzhou City (AD)	65.46	4.49	4.90	13627	33.24
44	Baoying County	72.44	4.69	2.56	11670	55.52
45	Yizheng City	62.82	3.39	3.21	12244	40.55
46	Gaoyou City	73.26	3.73	2.90	11828	48.51
47	Zhenjiang City (AD)	37.66	3.28	6.01	15900	20.58
48	Danyang City	69.63	2.48	2.95	15171	37.06
49	Yangzhong City	70.66	2.26	2.48	16631	37.75
50	Jurong City	68.06	3.01	2.45	13235	40.32
51	Taizhou City (AD)	56.49	4.05	5.67	13444	35.80
52	Xinghua City	70.86	5.37	3.15	11827	48.60
53	Jingjiang City	66.77	4.51	4.61	13715	39.05
54	Taixing City	78.61	3.50	3.24	12505	52.41
55	Jiangyan City	77.43	4.11	3.78	12228	47.26
56	Suqian City (AD)	69.97	5.34	3.56	9514	42.70
57	Shuyang County	78.30	4.37	3.15	9557	53.94
58	Siyang County	74.50	4.86	3.48	9541	46.62
59	Sihong County	67.89	5.54	3.53	9327	41.92

Legend
AD Represents all districts within the city

2013

Obs	County	Rural Households	% Health Instituitions/capita	Hospital Beds/capita	Rural Income/capita	Rural Emply
1	Nanjing City	29.43	2.83	5.10	16531	14.69
2	Wuxi City (AD)	25.81	3.13	5.51	23637	11.23
3	Jiangyin City	55.73	3.02	4.54	21882	24.03
4	Yixing City	56.26	3.26	3.66	18783	28.35
5	Xuzhou City (AD)	31.16	4.84	7.78	13924	17.49
6	Feng County	76.70	5.44	3.75	10957	57.14
7	Pei County	63.64	5.37	3.67	12725	44.84
8	Suining County	77.36	6.09	3.42	10686	57.96
9	Xinyi City	69.40	4.93	3.11	10979	51.41
10	Pizhou City	77.29	5.16	3.16	12635	47.60
11	Changzhou City (AD)	50.47	2.16	4.87	19750	22.69
12	Liyang City	75.85	2.89	3.65	16985	41.69
13	Jintan City	68.06	3.12	3.73	17371	36.00
14	Suzhou City (AD)	32.75	2.72	5.06	21389	12.75
15	Changshu City	54.98	2.96	4.55	21691	25.96
16	Zhangjiagang City	56.44	3.24	6.30	21689	24.90
17	Kunshan City	41.03	2.71	3.63	21793	12.72
18	Taicang City	47.14	3.15	4.66	21605	22.52
19	Nantong City (AD)	53.34	4.48	6.21	15710	26.00
20	Haian County	66.94	4.47	4.57	14119	43.63
21	Rudong County	82.82	4.69	3.29	13529	48.50
22	Qidong City	84.96	3.96	3.53	15766	52.63
23	Rugao City	78.35	4.01	3.81	13004	48.17
24	Haimen City	78.13	4.57	3.78	16920	54.54
25	Liangyungang City (AD)	29.56	5.81	6.30	12366	16.68
26	Ganyu County	71.33	7.30	2.94	11564	45.40
27	Donghai City	77.74	5.52	2.89	11118	46.38
28	Guanyun County	73.39	5.05	2.98	10016	47.20
29	Guannan County	71.58	5.65	4.20	9488	50.70
30	Huaian City (AD)	53.86	4.08	5.21	11875	36.98
31	Lianshui County	75.56	5.49	4.42	10333	59.31
32	Hongze County	64.52	3.95	4.34	12160	51.26
33	Xuyi County	69.48	5.16	4.74	11255	50.80
34	Jinhu County	63.42	4.08	3.92	11931	41.46
35	Yancheng City (AD)	56.07	4.09	5.63	13416	33.06
36	Xiangshui County	68.82	4.28	4.03	11084	42.27
37	Binhai County	75.10	4.29	4.10	11702	48.18
38	Funing County	60.55	4.59	3.27	11853	44.77
39	Sheyang County	67.10	3.61	3.15	13121	39.17
40	Jianhu County	63.18	4.75	3.52	13156	41.12
41	Dongtai City	81.42	3.97	4.47	15312	48.80
42	Dafeng City	75.27	4.79	4.13	15166	44.69
43	Yangzhou City (AD)	65.59	4.33	5.25	15544	33.05
44	Baoying County	72.83	4.51	2.72	13093	55.85
45	Yizheng City	62.86	3.22	3.58	13701	40.51
46	Gaoyou City	74.72	3.35	3.33	13248	48.86
47	Zhenjiang City (AD)	38.81	3.33	6.71	16577	20.45
48	Danyang City	68.19	2.42	3.40	16983	36.99
49	Yangzhong City	70.29	2.08	2.84	18644	38.32
50	Jurong City	68.56	2.92	2.89	14824	39.98
51	Taizhou City (AD)	65.70	4.09	4.95	14976	41.27
52	Xinghua City	71.05	5.35	3.65	13247	48.75
53	Jingjiang City	67.12	4.34	5.03	15347	39.78
54	Taixing City	78.81	3.38	3.63	13993	52.45
55	Suqian City (AD)	73.29	4.53	3.99	10665	41.22
56	Shuyang County	77.04	4.85	3.58	10799	54.12
57	Siyang County	72.02	4.76	4.68	10765	46.49
58	Sihong County	65.79	5.56	3.89	10540	42.49

Legend

AD Represents all districts within the city

Imputed for missing data

2014

Obs	County	Rural Household	% Health Institutions/capita	Hospital Beds/capita	Rural Income/capita	Rural Emply
1	Nanjing City	29.18	2.90	5.32	17661	14.51
2	Wuxi City (AD)	24.68	2.54	4.86	26367	8.60
3	Jiangyin City	55.20	8.70	4.70	23965	23.57
4	Yixing City	56.00	3.48	3.88	20178	27.73
5	Xuzhou City (AD)	28.93	4.95	8.17	15100	16.10
6	Feng County	73.10	5.91	4.20	11757	56.22
7	Pei County	67.94	5.51	4.02	13249	43.98
8	Suining County	75.74	6.07	3.49	11600	58.71
9	Xinyi City	68.41	5.16	3.44	12140	50.10
10	Pizhou City	75.16	5.39	3.40	12846	46.76
11	Changzhou City (AD)	49.21	2.25	5.42	21332	23.34
12	Liyang City	75.81	3.05	5.01	18222	41.66
13	Jintan City	68.54	3.41	4.92	18733	35.82
14	Suzhou City (AD)	31.85	2.72	5.42	23296	12.53
15	Changshu City	54.60	3.01	4.67	23767	25.69
16	Zhangjiagang City	55.87	3.34	6.86	23722	24.51
17	Kunshan City	40.29	2.87	3.82	23921	12.63
18	Taicang City	46.11	3.20	5.05	23590	22.20
19	Nantong City (AD)	50.21	4.53	9.70	17051	24.73
20	Haian County	72.65	4.63	4.89	15155	43.23
21	Rudong County	82.62	4.69	3.35	14494	48.74
22	Qidong City	85.25	4.46	3.99	16762	52.46
23	Rugao City	78.11	4.07	4.13	14210	47.85
24	Haimen City	77.60	4.49	3.86	17419	54.29
25	Lianyungang City (AD)	51.64	6.63	4.85	12650	29.82
26	Donghai City	76.99	5.74	2.94	12171	45.97
27	Guanyun County	73.83	5.19	3.06	10864	47.64
28	Guannan County	70.92	5.85	4.41	10442	48.97
29	Huaian City (AD)	53.47	4.22	5.50	11922	36.57
30	Lianshui County	75.23	5.68	4.44	11206	58.45
31	Hongze County	63.92	4.22	4.46	13161	51.35
32	Xuyi County	70.65	5.44	4.78	12175	50.47
33	Jinhu County	63.54	4.45	4.47	13131	41.19
34	Yancheng City (AD)	55.96	4.24	6.09	17232	33.01
35	Xiangshui County	69.26	4.68	4.53	11964	42.53
36	Binhai County	74.18	4.42	4.66	12524	47.71
37	Funing County	60.13	4.67	4.39	12959	44.36
38	Sheyang County	66.76	3.78	3.65	13848	39.13
39	Jianhu County	64.10	4.93	4.53	14345	41.42
40	Dongtai City	81.39	4.45	5.32	16565	49.01
41	Dafeng City	75.86	4.93	4.58	16414	44.25
42	Yangzhou City (AD)	64.77	4.27	5.35	18141	33.70
43	Baoying County	73.04	4.49	2.92	14246	55.51
44	Yizheng City	62.94	2.89	3.69	14846	40.39
45	Gaoyou City	74.92	3.33	3.43	14335	49.05
46	Zhenjiang City (AD)	37.72	3.39	6.82	16725	20.58
47	Danyang City	69.99	2.52	3.39	18250	36.98
48	Yangzhong City	72.09	2.34	2.84	20078	38.29
49	Jurong City	68.48	3.11	2.94	15893	40.73
50	Taizhou City (AD)	66.58	3.98	5.55	15354	41.10
51	Xinghua City	71.43	5.32	3.41	14258	48.66
52	Jingjiang City	66.93	4.02	5.37	16570	39.68
53	Taixing City	78.73	3.62	3.68	15066	52.54
54	Suqian City (AD)	71.96	4.73	4.13	11678	41.62
55	Shuyang County	76.59	5.07	3.61	11828	53.04
56	Siyang County	74.04	5.01	5.13	11690	46.95
57	Sihong County	65.47	5.79	4.34	11405	42.68

Legend

AD Represents all districts within the city