Geographic and Socioeconomic Risk Factors for Sporadic Cryptosporidiosis and *E. coli* infection in East Tennessee

A Dissertation Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Ingrid Elizabeth Luffman
August 2013

Copyright © 2013 by Ingrid Elizabeth Luffman
All rights reserved

ACKNOWLEDGEMENTS

I have many people to thank for their assistance in completing this dissertation. First, thanks go to my advisor, Liem Tran, for his mentorship and for inspiring me to see not just the numbers, but also the beauty to be found in statistical modeling. I also thank my committee members Nicholas Nagle, Yingkui Li, and Agricola Odoi for their thoughtful comments and questions which helped to provide a clear focus for the dissertation.

The data used in this research was made available through a partnership with Judy Manners, L. Rand Carpenter, and John Dunn of the State of Tennessee Department of Health in Nashville, TN. My thanks especially go to Judy Manners for answering myriad questions along the way, and providing information on diagnostic methods for *E. coli* O157 and cryptosporidiosis.

I am very grateful to my colleagues in the Department of Geosciences, East Tennessee State University, for their unwavering support and encouragement.

Finally, and most importantly, I thank my family. I may not have pursued this route had it not been for the example set by my father, Roger Gough, who showed me what was possible. My husband, Robert Luffman, and my children, Ben and Sydney Luffman, have provided endless encouragement and understanding of the late nights, and missed scout meetings and band concerts. Thank you Robert, for not only keeping our house clean and our children fed, but for your patience, strength, and love. Thank

you, Ben and Sydney for cheerfully learning to take on household responsibilities, and for being so visibly proud of me. This truly is a group effort.

ABSTRACT

This research examines risk factors for sporadic cryptosporidiosis and *Escherichia coli* (*E. coli*) O157 infection in East Tennessee, using case-control and retrospective ecological approaches. Multiple models and approaches are used to identify risk factors for the two diseases, and to examine the effect of scale on risk for disease in the individual and in the population. Risk factors examined are animal density, land use, geology, surface water impairment, poverty rate and availability of private water supply. The research objectives are, first, to identify risk factors for *E. coli* O157 and cryptosporidiosis in East Tennessee by relating disease data to environmental data through statistical regression models and second, to examine the effect of scale by comparing risk factors for disease in the individual (case-control approach) and the population (ecological approach).

At the individual level Generalized Linear Models (GLMs), Generalized Additive Models (GAMs), and Spatial Logistic Regression Models are compared. At the population level, Spatial Lag, GLMs and GAMs are developed using Gamma, Tweedie and Poisson distributions.

Beef cow population density and proximity to karst geology are positively associated with both diseases at the individual scale. Land use variables representing developed land and pasture land are positively associated with both diseases at both scales. Poverty rate is positively associated with both diseases at the regional scale, and availability of private water supply is negatively associated with both diseases at both scales.

The results presented here show that the significance of environmental variables as risk factors for cryptosporidiosis and *E. coli* O157 depend on scale, and that an examination of risk factors for these diseases in the individual and the population can reveal the scale at which variables are important.

These results can be used to identify important environmental risk factors for the diseases and to identify the communities where background risk is highest. Limited public health resources can then be targeted to the risk factors and communities most at risk. These results can also be used as the framework upon which to develop a comprehensive epidemiological study that focuses on risk factors important at the individual level.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Introduction and overview of research purpose 1
1.2 Statement of the problem 2
1.3 Research questions and objectives 4
1.4 Organization 4

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW 6
2.1 Study area 6
2.2 Epidemiological background 7
2.2.1 Description of diseases and impact 11
2.2.2 Risk factors for cryptosporidiosis and *E. coli* O157 infection 12
2.3 Analytical methods for modeling disease 16
2.4 Special considerations of health datasets 26
2.4.1 Non-reporting and incomplete or inconsistent reporting 26
2.4.2 Incubation period 29

CHAPTER 3 METHODS 31
3.1 Assembling the case dataset 31
3.1.1 Case data 34
3.1.2 Explanatory variables for individual models 35
3.1.3 Control data 37
3.2 Assembling the aggregate dataset 40
3.2.1 Rate standardization 40
3.2.2 Explanatory variables for regional models 41
3.3 Exploratory mapping 47
3.3.1 Cryptosporidiosis 47
3.3.2 *E. coli* O157 55
3.4 Individual models 62
3.4.1 Selection of covariates 62
3.4.2 Model development 63

3.5 Regional models 66
3.5.1 Exploratory data analysis 66
3.5.2 Model development 67
3.5.3 Model development using non-zero rates partition 69

CHAPTER 4 CRYPTOSPORIDIOSIS RESULTS 71
4.1 Individual models 71
4.1.1 Exploratory analysis 71
4.1.2 Individual regression models for cryptosporidiosis 76
4.1.3 Comparison of cryptosporidiosis individual models 80
4.2 Regional models 82
4.2.1 Exploratory analysis 82
4.2.2 Regional regression model results for cryptosporidiosis 89
4.2.3 Comparison of cryptosporidiosis regional models and treatments 99
4.3 Chapter summary 101

CHAPTER 5 *E. COLI* O157 RESULTS 103
5.1 Individual models 103
5.1.1 Exploratory analysis 103
5.1.2 Individual regression models for *E. coli* O157 107
5.1.3 Comparison of *E. coli* O157 individual models 112
5.2 Regional models 114
5.2.1 Exploratory analysis 114
5.2.2 Regional regression model results for *E. coli* O157 122
5.2.3 Comparison of *E. coli* O157 regional models and treatments 136
5.3 Chapter summary 137

CHAPTER 6 DISCUSSION 140
6.1 Risk factors for Cryptosporidiosis 141
6.2 Risk factors for *E. coli* O157 146
6.3 Importance of scale in identification of risk factors 151

6.3.1 Methodological insights 151
6.3.2 Epidemiological insights 154
6.4 Other considerations 155

CHAPTER 7 CONCLUSION 158

LIST OF REFERENCES 161

VITA 175

LIST OF TABLES

Table 1. Candidate explanatory variables for cryptosporidiosis and *E. coli* O157.....38

Table 2: Description of explanatory variables used in aggregate models.....42

Table 3. Comparison of explanatory variables used in individual and regional models.....46

Table 4: Statistical point models and platforms.....64

Table 5. Regional (ecological) regression models and platforms.....70

Table 6. Results (p-values) of a comparison of distributions of cryptosporidiosis case-control data using the Mann-Whitney U Test and Kolmogorov-Smirnov Test. H_0 = the distributions are the same.....72

Table 7. Spearman correlation coefficients for cryptosporidiosis explanatory variables (n=496). Significant correlations are marked with a double asterisk, p-values are provided in parentheses, and correlations above 0.7 are shaded.....75

Table 8. Descriptive statistics for explanatory variables selected for inclusion in cryptosporidiosis models (statistics are given for case data only, n=246).....76

Table 9. Individual model coefficients (p-values follow in parentheses) and diagnostics for cryptosporidiosis models (not including spatial logistic regression models). Shaded cells represent excluded variables and diagnostic results that were unavailable.....77

Table 10. Spatial logistic regression model (SLRM) coefficients (p-values in parentheses) for cryptosporidiosis individual models. Shaded cells represent excluded variables. Model diagnostics (AIC) are shown on the bottom row.....79

Table 11. Spearman correlation of cryptosporidiosis standardized rates and explanatory variables. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01. P-values are in parentheses. Shaded cells indicate correlations above r=0.7.....83

Table 12. A comparison of coefficients (p-values) for cryptosporidiosis regional models and data partitions. Smoothed variables in GAMs have multiple coefficients and are not reported.....91

Table 13. Spearman correlation for non-zero partition of cryptosporidiosis standardized rates and explanatory variables. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01 (p-values are in parentheses). Shaded cells indicate correlations above r=0.7.....96

Table 14. Tally of explanatory variable significance in cryptosporidiosis aggregate models.....100

Table 15. Summary of statistically significant risk factors for cryptosporidiosis at the individual and regional scales. "+" indicates positive association with disease, "-" indicates negative association with disease.....102

Table 16. Results (p-values) of a comparison of distributions of *E. coli* O157 case and control data using the Mann-Whitney U Test and Kolmogorov-Smirnov Test. H_0 = the distributions are the same.....104

Table 17. Spearman correlation coefficients for explanatory variables (n=500). Significant correlations are marked with a single (p<0.05) or double (p<0.01) asterisk (2-tailed test), and p-values are in parentheses. Correlations greater than r=0.7 are shaded.....106

Table 18. Descriptive statistics for variables selected for inclusion in *E. coli* models (statistics are given for case data only, n=250).....107

Table 19. Individual model coefficients (p-values follow in parentheses) and diagnostics for *E. coli* O157 models (does not include spatial logistic regression models). Shaded cells represent excluded variables and diagnostic results that were unavailable.....108

Table 20. Spatial logistic regression model coefficients (p-values in parentheses) for *E. coli* O157 individual models. Shaded cells represent excluded variables. Model diagnostics (AIC) are shown in the bottom row. Model E3 included all possible interaction terms (not shown in table). 111

Table 21. Spearman correlation of *E. coli* O157 standardized rates and explanatory variables. Significant correlations are marked with a single (p<0.05) or double (p<0.01) asterisk (two-tailed, n=180). Correlations above 0.7 are shaded, p-values are in parentheses.....115

Table 22. Comparison of coefficients and p-values (in parentheses) for *E. coli* O157 regional models and data partitions. Smoothed variables in GAMs have multiple coefficients and are not reported. 124

Table 23. Spearman correlations for non-zero partition of *E. coli* O157 standardized rates. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01 (p-values are in parentheses). Correlations greater than r=0.7 are shaded. 132

Table 24. Tally of significance of explanatory variables in regional *E. coli* O157 models. Significant explanatory variables are shaded. 137

Table 25. Summary of significant environmental risk factors for E. coli O157 infection at individual and regional scales. “+” indicates positive association with disease, “-“ indicates negative association with disease.....139

LIST OF FIGURES

Figure 1. The project study area consists of the twenty-seven (27) eastern most counties in Tennessee. 6

Figure 2. A comparison of Tennessee cryptosporidiosis rate to study area and Greene County rate (1995–2011). 9

Figure 3. A comparison of Tennessee *E. coli* O157 rate to study area rate from 1995 through 2011. 11

Figure 4. Overlay of cryptosporidiosis rates and beef cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 48

Figure 5. Overlay of cryptosporidiosis rates and milk cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 48

Figure 6. Overlay of cryptosporidiosis rates and hog population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 49

Figure 7. Overlay of cryptosporidiosis rates and sheep population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 49

Figure 8. Overlay of cryptosporidiosis rates and developed land use. Data source: National Land Cover Dataset 2006. 50

Figure 9. Overlay of cryptosporidiosis rates and forest land. Data source: National Land Cover Dataset 2006. 51

Figure 10. Overlay of cryptosporidiosis rates and agricultural land. Data source: National Land Cover Dataset 2006. 51

Figure 11. Overlay of cryptosporidiosis rates and karst geologic areas. 52

Figure 12. Overlay of cryptosporidiosis rates and surface water quality. The water quality data source is the EPA approved 2008 303(d) list for Tennessee. 53

Figure 13. Overlay of cryptosporidiosis rates, karst areas, and *E. coli* impaired stream segments. Water quality data source is the EPA approved 2008 303(d) list for Tennessee. 53

Figure 14. Overlay of cryptosporidiosis rates and ZTCA poverty rate. Data source: 2010 US Census. 54

Figure 15. Overlay of cryptosporidiosis rates and county well permit density. 55

Figure 16. Overlay of *E. coli* O157 rates and beef cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 56

Figure 17. Overlay of *E. coli* O157 rates and milk cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 57

Figure 18. Overlay of *E. coli* O157 rates and hog population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 57

Figure 19. Overlay of *E. coli* O157 rates and sheep population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level. 58

Figure 20. Overlay of *E. coli* O157 rates and percent developed land. Data source: National Land Cover Dataset 2006). 58

Figure 21. Overlay of *E. coli* O157 rates and percent forested land. Data source: National Land Cover Dataset 2006. 59

Figure 22. Overlay of *E. coli* O157 rates and percent agricultural land. Data source: National Land Cover Dataset 2006. 59

Figure 23. Overlay of *E. coli* O157 rates and karst geologic areas. 60

Figure 24. Overlay of *E. coli* O157 rates and surface water quality. The water quality data source is the EPA approved 2008 303(d) list for Tennessee. 60

Figure 25. Overlay of *E. coli* O157 rates and *E. coli* impaired stream segments (from EPA 2008 305(b) list) over karst geology. 61

Figure 26. Overlay of *E. coli* O157 rates and ZCTA poverty rate. Data Source: US 2010 Census. 61

Figure 27. Overlay of *E. coli* O157 rates and county well permit density (number of permits per 10,000 population). 62

Figure 28. Box plots of case (Group 1 - right) and control (Group 0 - left) data for sheep population density (SheepDen2). 73

Figure 29. Spatial logistic regression model (C1) risk map for cryptosporidiosis (see Table 10 for coefficients). 81

Figure 30. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and agricultural animal population density (beef cow, milk cow, calf, hog, and sheep). 84

Figure 31. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and agricultural animal population density (beef cow, milk cow, calf, hog and sheep) after removal of outliers. 84

Figure 32. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and the percent of land area classified as developed, forested, and pasture land (outliers removed). 85

Figure 33. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and geology, water quality, poverty and well permit variables (outliers removed). 86

Figure 34. Map of cryptosporidiosis rates showing Local Indicators of Spatial Autocorrelation (LISA) generated using GeoDa software. Outliers have been removed. 87

Figure 35. Getis Gi cluster map of cryptosporidiosis rate data. The map was generated in GeoDa using a 40 km distance spatial weights matrix. 88

Figure 36. Histogram of disease rates at the zip code level showing (a) standardized cryptosporidiosis rates per 10,000 people (CR_St10K) and (b) log-transformed rates per 10,000 people (log(CR_std10K)). 89

Figure 37. R output for regional GAM using cryptosporidiosis rate complete dataset (n=180). R code is printed in red and preceded by the ">" character. 93

Figure 38. Scattergram of non-zero cryptosporidiosis rates and agricultural variables (n=75). .. 97

Figure 39. Scattergram of non-zero cryptosporidiosis rates and land use variables (n=75) 97

Figure 40. Scattergram of non-zero cryptosporidiosis rates and geology, water quality, poverty and well permit variables (n=75) 98

Figure 41. Box plots of *E. coli* O157 case (Group 1 - right) and control (Group 0 - left) data for milk cow population density 104

Figure 42. Spatial logistic regression model (E4) risk map for *E. coli* O157 (see Table 20 for coefficients). 113

Figure 43. Scattergram of *E. coli* O157 rate with agricultural variables (n=180). 116

Figure 44. Scattergram of *E. coli* O157 rate with agricultural variables after removal of outliers and zip code areas with zero population (n= 167). 117

Figure 45. Scattergram of *E. coli* O157 rate with land use variables (n=180). 117

Figure 46. Scattergram of *E. coli* O157 rate with land use variables after removal of outliers and zip code areas with zero population (n= 167). 118

Figure 47. Scattergram of *E. coli* O157 rate with poverty, water quality, geology and well density variables (n=180). 118

Figure 48. Scattergram of *E. coli* O157 rate with poverty, water quality, geology and well density variables after removal of outliers and zip code areas with zero population (n= 167) 119

Figure 49. Cluster map showing LISA clusters. The map was generated in GeoDa using a distance based spatial weights matrix at 40km. 121

Figure 50. Getis Gi Cluster map of *E. coli* O157 rate data. The map was generated in GeoDa using a 40km distance spatial weights matrix. 121

Figure 51. Histograms showing the distribution of a) standardized *E. coli* O157 rates and b)log-transformed rates for outlier removed aggregate data (n=167). 122

Figure 52. R output for GAM model using complete *E. coli* O157 rate dataset (n=180). R code is printed in red and preceded by the ">" character. 126

Figure 53. R output for GAM model using dataset with outliers removed (n=167). R code is printed in red and preceded by the ">" character. 130

Figure 54. Scattergram of non-zero *E. coli* O157 rates and agricultural variables (n=87) 133

Figure 55. Scattergram of non-zero *E. coli* O157 rates and land use variables (n=87) 133

Figure 56. Scattergram of non-zero *E. coli* O157 rates and geology, water quality, poverty and well permit variables (n=87) 134

CHAPTER 1 INTRODUCTION

1.1 Introduction and overview of research purpose

This research examines environmental and socioeconomic risk factors for sporadic cryptosporidiosis and *Escherichia coli* (*E. coli*) O157 infection in East Tennessee, using case-control and retrospective ecological approaches. The aim for epidemiologic research in general is to identify associations between exposures and outcomes to maximize health or to prevent disease (Dohoo, Martin and Stryhn 2003). Exposure to cryptosporidiosis and *E. coli* O157 may occur through consumption of tainted food, through contact with an infected object, person or animal, or through incidental contact as a result of an individual's environment. The probability of human infection by these pathogens depends on three factors: 1) the nature of the pathogen and its ability to infect the host; 2) the susceptibility of the host to infection; and 3) the environment, which includes the survivability of the pathogen in the environment and opportunities for host-pathogen interaction (Robertson and Yasvinski 2012). A geographic approach to assessment of risk for cryptosporidiosis and *E. coli* that includes the use of GIS and spatial statistical modeling can therefore be a powerful method to infer associations between environment and health (Jarup 2000).

1.2 Statement of the problem

Since the first outbreak of cryptosporidiosis related to recreational water in the United States was reported in 1988, *Cryptosporidium* has emerged as the most recognized cause of disease outbreak associated with recreational water (Hlavsa et al. 2011). An estimate of the overall annual number of *Cryptosporidium* spp. cases (adjusted for non-reporting) in 2011 suggests an annual average of 57,616 cases in the US, with a 90% credible interval of 12,060 to 166,771 cases, a hospitalization rate of 25% and a death rate of 0.3% (Scallan et al. 2011). From 2006 - 2009, Tennessee reported 315 cryptosporidiosis cases statewide. Forty-nine (15.7%) were from the northeast region, though this region represents only 6% of the state's population. Forty-seven of the forty-nine cases (96%) in the northeast region were in two counties, warranting further investigation. While some of these cases were attributed to a specific exposure (such as contaminated food or water), the sources of infections of most of the cases remained unexplained.

E. coli 0157 is a pathogen first identified in 1982 as the cause of two outbreaks of disease in Oregon and Michigan, USA (Sehgal, Kumar and Kumar 2008). Since that time the disease has become widely distributed throughout the United States and the rest of the world because of the high survival rate of the pathogen and low infectious dose (between 10 and 100 organisms) (Chalmers, Aird and Bolton 2000). *E. coli* 0157 infection causes an estimated 73,000 illnesses in the United States each year, seventeen percent of which can require hospitalization (Rangel et al. 2005). From 2000-2010, 903 cases of *E. coli* O157 were reported in Tennessee. In 21 cases, onset of symptoms was preceded by

international travel, and only 2 cases were associated with a known outbreak. Therefore, in the majority of Tennessee cases, the cause of illness is unknown.

Because risk factors for cryptosporidiosis and *E. coli* O157 can be associated with an individual's environment in addition to exposure through food and human contact, analysis of these datasets benefit from explicitly including space and spatial relationships between potential risk factors and these diseases in the selected analytical technique. Proximity to a known risk factor may increase the incidence of cryptosporidiosis or *E. coli* O157 in a population and therefore, epidemiologic research involving these diseases should take into account the spatial relationship between the individual, the environment, and other individuals, keeping in mind the relationship between and among cases of these diseases. In other words, the datasets are not sets of independent cases, but rather correlated sets of points in two-dimensional (or three-dimensional) space. The spatial nature of the known risk factors for these diseases lends itself well to analysis using geographic methods.

Spatial epidemiological research therefore creates a bridge between physical geography, human geography, and health because consideration of the geographic relationships between physical and social environments, and cryptosporidiosis and *E. coli* O157 is essential to understanding patterns of incidence for these diseases.

1.3 Research questions and objectives

The purpose of this research is to examine the occurrence of cryptosporidiosis and *E. coli* O157 in East Tennessee from 2000-2010 to identify statistically significant environmental and socioeconomic risk factors for the two diseases at the individual and the regional scale. This research will develop multiple regression models to investigate risk factors for 247 cryptosporidiosis cases and 250 *E. coli* O157 cases reported in East Tennessee between 2000 and 2010. Risk factors to be investigated for potential association with cryptosporidiosis and *E. coli* O157 infection risks are: land use, agricultural animal population densities, surface water quality, geology, poverty rate, and availability of private drinking water source.

The research questions to be investigated are:

1) Identify risk factors for *E. coli* O157 and cryptosporidiosis in East Tennessee by relating disease data to environmental data through statistical regression models; and
2) Examine the effect of scale on risk factors for *E. coli* O157 and cryptosporidiosis by comparing risk factors for disease in the individual (case-control approach) and the population (ecological approach).

1.4 Organization

The chapter which follows contains a literature review and background information on the epidemiology of cryptosporidiosis and *E. coli* O175. Chapter 2 also presents examples of the use of regression modeling in epidemiology. A description of

the methods used to construct the point and aggregate datasets and develop the point and aggregate models is presented in Chapter 3. Chapters 4 and 5 contain the model results for the cryptosporidiosis and *E. coli* O157 datasets, respectively. Discussion of the model results and risk factors for each disease is contained in Chapter 6, including a discussion of the effect of scale in spatial epidemiological modeling. The conclusion is presented in Chapter 7.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

2.1 Study area

The study area consists of the following Tennessee Department of Health (TDH) regions: East Tennessee, Knoxville/Knox County, Northeast Tennessee, and Sullivan County, plus Bradley, Polk and McMinn counties, a total of twenty-seven (27) of the ninety-five (95) counties in the state (Figure 1).

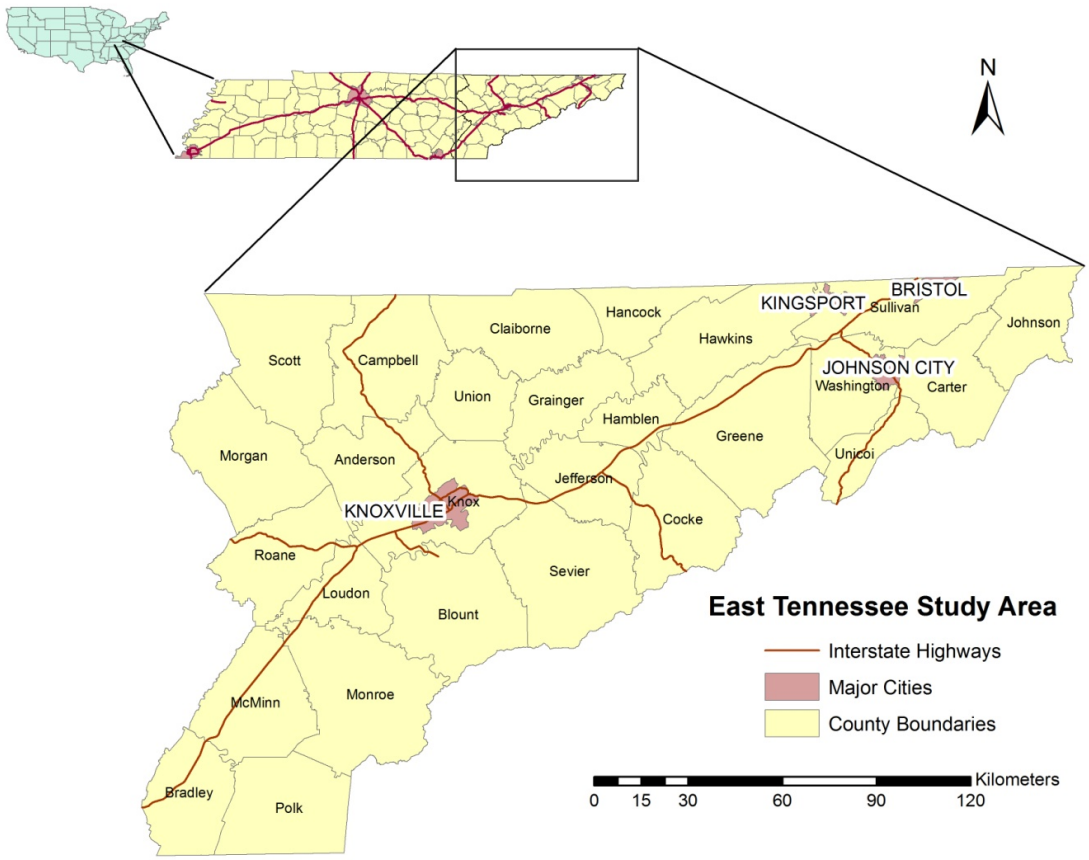


Figure 1. The project study area consists of the twenty-seven (27) eastern most counties in Tennessee.

2.2 Epidemiological background

Cryptosporidiosis is one of the most common waterborne diseases worldwide, infecting humans in 106 countries and found in over 150 mammal species (Fayer 2008). Worldwide, reporting is not universally required, diagnostic methods vary and the decision of whether or not to seek treatment may depend on patient access to health care (Suresh et al. 2012). It is therefore difficult to make a precise assessment of the global prevalence of this disease, which is estimated at 1 to 4.5% for developed countries, and at 3 to 20% for developing countries (Center for Food Security and Public Health 2005). A recent review of 199 global waterborne protozoan outbreaks reported between 2004 and 2010 attributed over 60 percent (120) of the outbreaks to *Cryptosporidium* spp. (Baldursson and Karanis 2011).

The Center for Food Security and Public Health estimates that in North America, approximately 80% of the population has been exposed to cryptosporidiosis and 2% of the population is infected (Center for Food Security and Public Health 2005). In the United States, cryptosporidiosis incidence increased from 2,972 cases in 1995 to 11,657 cases in 2007 (Yoder and Beach 2009), with 7, 656 and 8,951 cases reported in 2009 and 2010, respectively (Yoder et al. 2012). This increase has been attributed to: improved surveillance, improved awareness of the connection between disease and environmental exposure, an increase in the number of cases detected due to changing health care practices, and an actual increase in disease (Yoder and Beach 2009). The cost of hospitalization alone for cryptosporidiosis in the United States is estimated at $45.8

million annually (Yoder et al. 2012), not including lost productivity or other related expenses. There is therefore a definite economic benefit to be gained from an understanding of the risk factors and by implementing prevention and control measures to reduce infection. Cryptosporidiosis is geographically widespread throughout the country and incidence was generally higher in the Midwest and Northwest regions (4.3 and 3.4 cases per 100,000 population, respectively) and lower in the South and Southwest regions (1.9 and 1.5 cases per 100,000 population, respectively) (Yoder et al. 2012).

In Tennessee, standardized data and statistics from the Tennessee Department of Health Communicable Disease Interactive Data utility (accessed on 05/06/13 at health.state.tn.us/Ceds/WebAim/) show an increase in cases from zero in 1995 to 99 cases of cryptosporidiosis in 2011 (1.5 cases per 100,000 population), with a maximum of 136 cases occurring in 2007 (2.2 cases per 100,000 population) (Figure 2). East Tennessee bears a larger share of the state burden of cryptosporidiosis, having a higher disease rates within the study area from 1996 through 2011, except for 2001 through 2003. Greene County in particular continues to experience high incidence of disease given its low population, especially from 2004 through 2011. An epidemiologic study by the Tennessee Department of Health examined exposure data from 2006 through 2011 obtained through interviews of cryptosporidiosis patients residing in Greene County and the neighboring Washington County. Exposure to cattle was more likely (OR=2.2; 95% confidence interval of 1.1-4.5) in patients in these counties compared to patients statewide (Judy Manners, Tennessee Department of Health, personal communication).

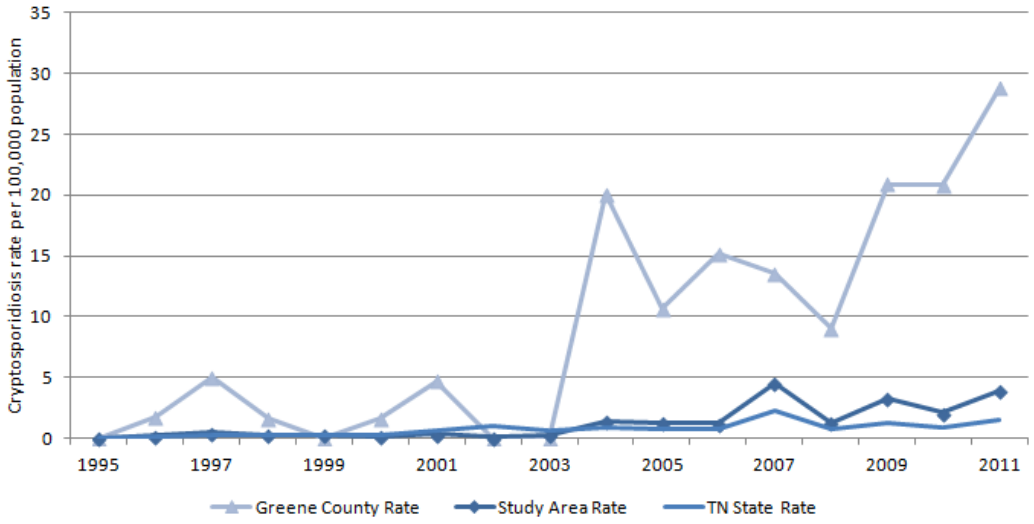


Figure 2. A comparison of Tennessee cryptosporidiosis rate to study area and Greene County rate (1995–2011).

Epidemiologic studies that involve patient interviews are costly in terms of manpower and time. This study examines the distribution of cases and disease rate in relation to the distribution of known environmental risk factors, using a Geographic Information System and statistical modeling, to provide an alternative method to examine risk factors for cryptosporidiosis in East Tennessee.

Globally, *E. coli* O157 infection varies from country to country and from region to region, from less than one case to greater than twenty cases per 100,000 population members. This variation has been related to regional differences in: surveillance and diagnostic capabilities, cattle density, beef consumption (including cultural practices that promote the consumption of uncooked or undercooked beef), occupational exposure, contact with animals, consumption of raw milk, and exposure through drinking water and recreational water (Suresh et al. 2012).

From 2000-2010, 5688 cases of *E. coli* O157 were reported to ten FoodNet sites in the United States, with 2446 (43%) hospitalizations and 33 (0.6%) deaths (Gould et al. 2013). The ten FoodNet surveillance states represent only 15% of the U. S. population, and the burden of *E. coli* O157 nationally is estimated at 63,153 cases, with a 90% credible interval of 17,587 to 149,631 cases and a death rate of 0.5% (Scallan et al. 2011). In the United States, as in the rest of the world, the estimated rate of disease is higher in children less than five years old (40.5 cases per 100,000 population) than in the general population (32.3 cases per 100,000 population), as is the estimated rate of hospitalization (3.6 per 100,000 versus 1.1 per 100,000 in the general population) and the estimated death rate (0.06 per 100,000 versus 0.01 per 100,000 in the general population) (Scallan et al. 2013).

In Tennessee, reported *E. coli* O157 incidence increased from 0 cases in 1995 to 111 cases in 2011 (a rate of 1.7 cases per 100,000 population), with a maximum of 143 cases (2.4 per 100,000 population) in 2006 (Figure 3). Within the study area, *E coli* O157 rate has historically exceeded the state rate, except in 1997 and 2006 when nationwide outbreaks associated with contaminated ground beef (1997) and spinach (2006) occurred. No research has been published to examine why parts of East Tennessee bear a larger burden of *E. coli* O157 compared to the state as a whole. Examination of known risk factors and their spatial distribution relevant to the distribution of this disease in East Tennessee can shed light on the environmental risk factors for *E. coli* O157 in the region.

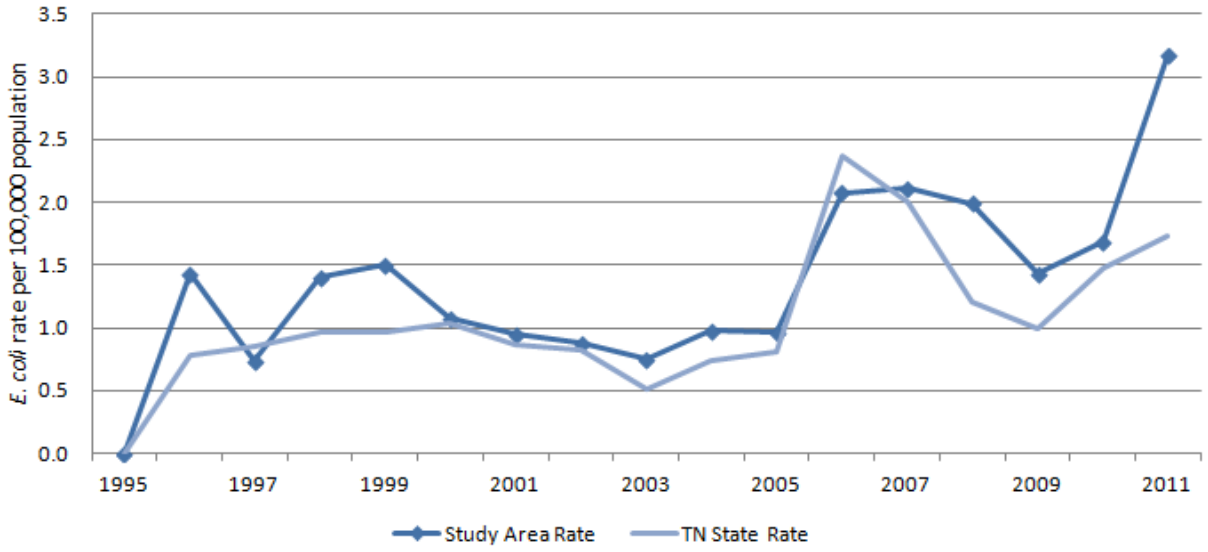


Figure 3. A comparison of Tennessee *E. coli* O157 rate to study area rate from 1995 through 2011.

2.2.1 Description of diseases and impact

Cryptosporidiosis is a gastrointestinal illness caused by the protozoa *Cryptosporidium*, a waterborne single-celled parasite that infects the host by fecal-oral transmission through person-to-person contact, consumption of contaminated food or water, contact with contaminated recreational water, and contact with infected farm animals (Dietz and Roberts 2000, Suresh et al. 2012). The two main species of *Cryptosporidium* that affect humans are *Cryptosporidium parvum*, associated with animal-to-human transmission, and *Cryptosporidium hominus*, associated with human-to-human transmission (Suresh et al. 2012, Hunter et al. 2007). *Cryptosporidium* oocysts are able to survive a wide range of environmental conditions, including freezing and treatment

with chlorine and therefore require micro-filtration for removal from drinking water (Suresh et al. 2012). Symptoms include explosive watery diarrhea, nausea, vomiting and stomach cramps, and are more severe in immunocompromised individuals (Yoder et al. 2012). Cryptosporidiosis is diagnosed as a probable case in the clinic by rapid immunochromatographic cartridge assay (rapid card assays) and confirmed in the laboratory by detection of *Cryptosporidium* oocysts or antigens in stool (Ova and Parasite Exam) or intestinal fluids by enzyme immunoassay (EIA), direct fluorescent antibody (DFA), polymerase chain reaction (PCR), or culture (Yoder et al. 2012, Bradley and Atkinson 2011).

E. coli is a bacterium present in the intestinal tract of animals, and while most strains of *E. coli* are harmless, one strain, *E. coli* O157, has been implicated in disease outbreaks in humans and produces bloody diarrhea and colon ulcers accompanied by severe abdominal cramping and pain. In extreme cases (5-10% of cases) *E. coli* O157 infection can result in Hemolytic Uremic Syndrome leading to renal failure with a mortality rate of up to 5%, disproportionately affecting young children and the elderly (Rangel et al. 2005). *E. coli* O157 is a shiga-toxin producing *E. coli* (STEC), and is detected through various lab protocols, including isolation by culture, shiga toxin enzyme immunoassay (Stx EIA), and PCR (Hanna et al. 2010).

2.2.2 Risk factors for cryptosporidiosis and *E. coli* O157 infection

Sporadic cryptosporidiosis infections and local and national outbreaks have been associated with food borne and waterborne sources (Said et al. 2003), however,

environmental factors such as surface water quality, agricultural activity, geology, and drinking water source have been identified in other parts of the United States and in the British Isles as important risk factors. Similarly, *E. coli* O157 infection is associated with consumption of contaminated water or food such as undercooked beef, dairy products, and salads, however, a connection between environmental exposure and *E. coli* O157 outbreaks has also been established (Chalmers et al. 2000, Strachan et al. 2006, Michel et al. 1999, Valcour et al. 2002, Kassenborg et al. 2004).

Environmental contamination of surface water supplies, resulting in the presence of fecal bacteria in streams and lakes is known to increase the risk of bacteria-related sickness and death in humans (Benham et al. 2006, Robertson and Yasvinski 2012). Currently pathogens are the most common cause of waterbody impairment for assessed waters in the United States; of 41,944 impaired stream segments in the United States in 2008, 10,249 were impaired due to pathogens (United States Environmental Protection Agency 2009). Pathogens enter surface water bodies through contaminated runoff associated with livestock grazing or manure application, through effluent from failing septic systems, and through direct deposition from livestock, domestic animals and wildlife (Parajuli, Mankin and Barnes 2009). While only specific strains of pathogens, for example, *E. coli* O157, pose a serious human health risk, it is not practical to test for a specific pathogen (Robertson and Yasvinski 2012). Instead, indicator organisms such as fecal coliform bacteria and more recently *E. coli* are used as markers for the presence of *E. coli* O157 and *Cryptosporidium* in surface water (Jamieson et al. 2004, Benham et al. 2006). Recreational contact with impaired surface water has been associated with

outbreaks of cryptosporidiosis in the United States (Hlavsa et al. 2011) and the United Kingdom (Beach 2008).

Contact with agricultural animals was identified as a risk factor for cryptosporidiosis in a study of *Cryptosporidium parvum* infections in England and Wales (Hunter et al. 2007). An outbreak of cryptosporidiosis at a summer camp in North Carolina, USA, was linked to contact with livestock (Centers for Disease Control and Prevention 2011a).

Socioeconomic risk factors have also been identified. In a case-control study of risk factors for cryptosporidiosis conducted in England and Wales, high socioeconomic status was positively associated with disease (Lake et al. 2007). In contrast, a negative association between the both poverty rate and unemployment rate and disease was found in a study of cryptosporidiosis in US counties (Chang et al. 2009). This finding may result from differing access to healthcare among different socioeconomic classes.

Cryptosporidiosis has been associated with both low population density (Jagai et al. 2010) and high population density (Lake et al. 2007), and well water consumption/private water supplies (Said et al. 2003).

Surficial geology has been associated with cryptosporidiosis in the United States through an index of geologic sensitivity (Klumb et al. 2010). Geologic sensitivity refers to the presence of karst topography, which forms in limestone regions, and is characterized by landforms such as sinkholes, caves and springs. Surface water and groundwater are highly connected in karst regions, and thus surface contaminants can readily be transported into the groundwater. Recent research in *cryptosporidium parvum*

infections in Minnesota identified an association between illness and geologic sensitivity (Klumb et al. 2010).

A study of 350 *E. coli* O157 outbreaks in the United States from 1982-2002 determined that transmission routes for disease were 52% foodborne, 14% waterborne (broken down as 9.5% recreational water exposure and 4.5% drinking water), 9% person-person, 3% cattle contact and 21% unknown (Rangel et al. 2005). This result indicates that environmental pathways for disease transmission (in this example, through exposure to recreational water or agricultural animals) are significant. Perhaps more significant is that over one fifth of the outbreaks had unknown transmission routes. It is important to identify transmission routes for disease so that risky behavior can be altered and disease risk reduced.

From 1994 through 2003, 35 outbreaks of *E. coli* O157 were reported in Scotland, of which 19 (54%) were attributed to environmental pathways of transmission such as farm visits, camping on pasture land, contaminated private water supplies from surface runoff, contact with manure and soil, and recreational water exposure (Strachan et al. 2006). Strachan modeled risk factors for *E. coli* O157 using linear regression, and found a positive association between disease and the use of private water supplies, rural areas, and both sheep and cattle density. Other studies have linked *E. coli* O157 infection in humans to cattle population density (Michel et al. 1999), and the ratio of cattle population to human population (Valcour et al. 2002).

Socioeconomic indicators such as income and standard of living were found to be negatively correlated with *E. coli* O157 infection (Sakuma, Mitsuyoshi and Nobuhiko

2006) and the presence of *E. coli* O157 antibodies (Hasin et al. 2007), indicating that those with a higher standard of living have a lower risk for disease.

2.3 Analytical methods for modeling disease

Strachan and others outline four methods that can be employed to assess the relative importance of different disease pathways: epidemiological methods such as in depth patient interviews to identify the cause of an outbreak, case-control studies to assess the importance of a given risk factor, spatial disease mapping to assess the importance of environmental risk factors, and quantitative microbial risk assessment to determine the risk of disease from a given pathway (Strachan et al. 2006). This research focuses on the second and third of these methods by developing regression models to assess the importance of environmental and socioeconomic risk factors that are unevenly distributed in space.

Regression models are commonly employed to identify risk factors for disease using incidence and exposure data (Friesema et al. 2010, Valcour et al. 2002, Strachan et al. 2006, Kistemann et al. 2004, O'Brien, Goutam and Gilham 2001, Young et al. 2009, Hu, Mengersen and Tong 2009). The first step in regression analysis is to examine the distribution of the data, the relationship between explanatory variables, and the relationship between the input and output variables. Disease data can be problematic for a number of reasons. First, the data are rarely normally distributed and one is generally dealing with a binary outcome (disease or no disease) or count data (number of cases of disease). Binary data are typically modeled using a Binomial distribution,

while count data are Poisson distributed. Second, disease data are spatial data when they are linked to a location in space and to risk factors unevenly distributed throughout the environment. In other words, spatial proximity and spatial location relative to risk factors and to other cases of disease may have an impact on outcome. Third, the relationship between input (risk factors) and output (outcome) variables may not be linear.

The familiar Ordinary Least Squares (OLS) regression model has been used to model *E. coli* O157 incidence using rural risk factors related to farm animal densities and drinking water source in Scotland (Strachan et al. 2006), to cattle *E. coli* O157 prevalence and farm density in Sweden (Kistemann et al. 2004), and to sociodemographic factors in the United States (Chang et al. 2009). The OLS regression model

$$y = b_0 + BX + \varepsilon$$

expresses the expectation of the dependent variable (y) as a linear combination of independent variables $X= (x_1,x_2,\ldots,x_n)$ with coefficients $B=(b_1,b_2,\ldots,b_n)$ where b_0 is the y-intercept. OLS regression assumes 1) normal distribution of the errors and stationary variance (homoscedasticity), 2) independence of the errors, and 3) a linear relationship between inputs and outcomes. Epidemiological data do not tend to adhere to these assumptions: the relationship between risk factors and disease incidence may not be linear; disease data are binary or count data, which are appropriately modeled with Binomial, Poisson, or even Negative Binomial distributions if the data are overdispersed; and spatial data are rarely independent. Tobler's first law of geography

tells us that everything is related to everything else, but near things are more related than far things (Tobler 1970). Hence, Tobler's first law expresses the property of spatial autocorrelation common to many spatial datasets, which violates the assumption of independence. A spatial dataset containing case data for a given disease and known risk factors must therefore be analyzed with methods that allow for non-linear relationships to be modeled, without relying on assumptions of normality and independence.

An alternative to OLS regression that allows for non-normal distributions is the Generalized Linear Model (GLM)

$$y = g^{-1}[b_0 + BX] + \varepsilon$$

where *g()* is a link function, and y, B, X and ε are as in OLS regression. Valcour and others modeled the association between livestock density indicators and disease in Ontario, Canada, using generalized linear regression to capture the skewed (Poisson) distribution of the disease data. They found a positive association between *E. coli* O157 counts and two variables: the ratio of human to beef cow population, and the use of manure as solid or liquid fertilizer. They also found a negative association between disease and pig density (Valcour et al. 2002). While the GLM does not assume the data are normally distributed, independence in the residuals is assumed, and hence the model does not account for spatial autocorrelation in the dataset. One advantage of the GLM over OLS regression is that the modeled relationship between input and output variables can be nonlinear (because a nonlinear link function can be used in the model to transform the input variable).

When the purpose of a research study is to identify risk factors associated with a binary outcome, logistic regression is appropriate (O'Brien et al. 2001, Tikkinen et al. 2009, Olsen et al. 2002) because the logit function is superior to a linear function in predicting the probability of a binary outcome (Pohlmann and Leitner 2003). Studies that compare disease incidence between two groups (one having contracted the disease, and one group disease-free) have a binary outcome, and are termed case-control studies. By definition, case-control studies require a dual dataset containing disease occurrence data (case data) plus data about where disease is absent (control data). Comparison of explanatory variables for cases and controls provides information about the risk of infection given exposure to an explanatory variable (risk factor). The risk is expressed as an odds ratio, whereby persons exposed to a risk factor would be more or less likely to contract the disease. If control data are not readily available and point data are used to represent disease, the absence data can be randomly generated if one assumes that disease reporting is complete and spatially consistent. In this case, controls are generated as a randomly distributed set of points representing the absence of disease (Bivand, Pebesma and Gomez-Rubio 2008) (p. 173).

Logistic regression has been employed to model the probability of *E. coli* O157 infection and to identify behavioral risk factors for sporadic infection, using results from an epidemiological study of cases in England and Wales in 1996 and 1997 (O'Brien et al. 2001). Risk factors that emerged as significant for increasing the odds of disease included consumption of certain foods (rare chicken, watercress, peaches), contact with farm animals, and wading in surface water streams.

Logistic regression is a special case of the GLM, where the logit transform of the probability of disease (positive outcome), is modeled as a linear combination of risk factors so that the outcome Y_i (i = 1,…,n) has a value of 1 with probability p_i, and value 0 with probability *1-p*, and follows a Bernouilli distribution. p_i can be expressed as

$$p_i = \frac{1}{1 + e^{-z}}$$

where

$$z = b_0 + BX + \varepsilon$$

and b_0, X, B, and ε are as defined previously for OLS regression (King and Zeng 2001). The Bernouilli and binomial distributions are related, but different. The Bernoilli distribution models the result of one trial of a binary outcome, while the binomial distribution models the sum of the result of n Bernouilli trials. The logistic regression model assumes that the data are binary, that they are independent and that a linear relationship exists between the logit of disease and the explanatory variables (Bivand et al. 2008).

A second alternative to OLS regression is the Geographically Weighted Regression (GWR) model, based on the OLS regression model,

$$y = b_0 + B(s)X + \varepsilon\,.$$

This model differs from OLS in that the coefficients B are a function of a location vector s. GWR can therefore model a relationship that varies in space, but as an OLS-based regression model, it assumes 1) normality, 2) independence, and 3) linearity. GWR has been used to investigate associations between myocardial infarction (heart attack) and

ozone levels in Florida (Young et al. 2009). One characteristic of GWR is the need for an extensive dataset to generate meaningful values and estimates of standard errors for the spatially varying coefficients.

Third, spatial lag and spatial error models are linear models that incorporate spatial autocorrelation in the dataset into the model. The spatial lag model incorporates spatial autocorrelation in the dependent variable that can be explained by the explanatory variables

$$y = b_o + XB + \rho WY + \varepsilon$$

where WY is a vector of lagged observations (observations at nearby points), ρ is the spatial autoregressive parameter, XB is a vector of explanatory variables and associated coefficients, and ε is a random error term. The spatial error model is often used to resolve problems with a dataset, for example, when the process under study and the available data are at different spatial scales (Anselin 2002), because the model captures spatial autocorrelation in the error term resulting from spatial autocorrelation in the explanatory variables not reflected in the dependent variable. The spatial error model is represented by

$$y = b_o + XB + \lambda W\varepsilon + \zeta$$

where XB is a vector of explanatory variables and coefficients (typical of the OLS model), λ is a coefficient of lagged autoregressive errors, $W\varepsilon$ is a spatially lagged error term that regresses on errors associated with nearby points weighted by proximity or contiguity, and ζ is a random error term. The spatial lag and error models assume: 1) normality in the dependent variable, however this assumption can be relaxed for large

datasets; 2) spatial autocorrelation; and 3) a linear relationship between inputs and outputs. Spatial lag and spatial error models were used to study the spatial relationship between livestock density and *E. coli* 0157 cases in Ontario, Canada (Michel et al. 1999).

Spatial lag and error models are described as simultaneous autoregressive (SAR) models. In contrast, for conditional autoregressive (CAR) models the estimate of the probability of values at any location is conditional on neighboring values. SAR and CAR models applied to identical dataset have been found to provide no real difference in results (de Smith, Goodchild and Longley 2009). SAR and CAR models were designed for datasets with a regular lattice structure and infinite extent, which results in problems when using these models with area data that have irregularly shaped polygons with varying number of neighbors.

While spatial lag and error models are useful because of their ability to model spatial relationships between variables, they are linear models and so for the same reason logistic regression is preferred to OLS regression for binary datasets, spatially autocorrelated binary datasets are best modeled with spatial logistic regression rather than a spatial model based on a linear function. Spatial logistic regression was used by Agterberg to model the existence of mineral deposits in a unit area, using the location of known deposits and geologic variables as predictors (Agterberg 1974). Agterberg expressed the presence data as a raster of cells with value 1 if one or more mineral deposits occurred in a cell, and 0 otherwise. The explanatory data (geologic variables) took the form of a raster of cells containing the percent coverage of a geologic unit, with one raster per geologic unit of interest. The logistic model generated a probability of

occurrence for each cell from the value of the covariate rasters at that cell and the cell probabilities were summed and contoured to generate a prediction map expressing the number of deposits likely to be found. The spatial logistic regression model is appropriate to use with a binary response variable when the outcome is likely to be dependent on the spatial configuration of the occurrence data (mineral deposits or cases of disease) and the explanatory variables. This method can therefore be extended to epidemiological studies to generate a risk map from a set of image files representing the explanatory variables (risk factors).

Next, the Additive Model (AM) and its extensions (specifically, the Generalized Additive Model (GAM) and the geo-additive model) models non-linear processes as a linear combination of non-linear functions

$$y = b_0 + F(X) + \varepsilon$$

where *F(X)* is a matrix of are smooth functions and ε is the error term. The AM model assumes: 1) normality; 2) independence; and 3) a non-linear relationship between input and output. The GAM includes a link function to handle non-normal datasets

$$y = \mathrm{g}^{-1}(b_0 + F(X))$$

where g is a link function and *F(X)* is defined as above. The GAM is able to model, for example, a Poisson-distributed dependent variable as a nonlinear combination of smooth functions representing independent risk factors. The geo-additive model (Kammann and Wand 2003) adds spatial lag and spatial error terms to the AM to account for spatial autocorrelation

$$y = b_0 + F(X) + f_{spat}(S_i) + w_i\gamma + \varepsilon$$

where $f_{spat}(S_i)$ is a spatial lag term that captures the spatial relationship between areas S_i, and $w_i\gamma$ is a spatial error term. The geo-additive model therefore assumes: 1) normality; 2) spatial autocorrelation; and 3) non-linear relationship between inputs and outcomes.

A logical next step is to combine the various models into one that provides sufficient flexibility to model non-linear relationships in non-parametric spatially autocorrelated datasets. A recent study by Friesema and others combined the GLM with a spatial lag term to evaluate the geographical association between human *E. coli* O157 infection and livestock density in the Netherlands. This hierarchical study found that disease in humans was associated with increased cattle density by separating cases of disease into spatial units and further subdividing the spatial units into groups according to age and gender. The expected number of cases from area *i* in subgroup *j* was modeled using fixed effects (independent), structured effects (spatially correlated), and unstructured white noise (random error), with an offset term for population (Friesema et al. 2010). The study by Friesema illustrates one way to model nonlinear relationships in a spatially autocorrelated count dataset using the generalized linear model.

Last, the Classification and Regression Tree (CART) approach is a decision model that partitions the dataset at each level of a decision tree according to a set of rules. The tree is constructed using a learning database, then pruned so that it can be generally applied to an unknown dataset. CART models are useful because they do not make assumptions about the data distribution or independence, nor the type of function that relates inputs to outputs. In addition, CART models can work with nonstationary and

heteroscedastic datasets. Host sources for *E. coli* in manure slurry were identified through the development and use of a CART model (Lu et al. 2005).

Bel and others (Bel et al. 2009) developed a CART model that used Kriging to assess the spatial relationship between pollen data and tree species by assigning weights to each data record to deal with spatially clustered observations. In their model, clustered observations were given less weight because closer observations are expected to be similar, while remote observations were given more weight. The weights were incorporated into the CART at each level when the dataset was partitioned (weights were used in the decision rule) but no data from surrounding observations were used.

Hu and others (Hu et al. 2009) developed a Spatial CART model to study the risk factors associated with cryptosporidiosis infection in Brisbane, Australia at the statistical local area (SLA), and found a positive association between socioeconomic advantages, low education and disease. After the data were geo-coded and aggregated to the SLA, Spatial Empirical Bayes (SEB) smoothing was used to reduce variability in rates associated with small populations by shifting the rate closer to the mean for low population SLAs. Because SEB factors in the distance between centroids for each area as well as population at risk, the method is termed spatial, even though output at nearby SLAs is not incorporated into the model.

Chelgoum and others (Chelghoum, Zeitouni and Boulmakoul 2002) developed a Spatial CART model of traffic risk that incorporates nearby observations in the decision rule by storing the spatial relationship between observations in a set of relational tables similar to those used in a Geographic Information System.

Each of the methods outlined above assume that the population to be studied has certain characteristics in its distribution (normal or otherwise), spatial autocorrelation, and the relationship between the dependent variable and the various risk factors (linear or otherwise). These characteristics, along with the research question to be answered by the model determine which statistical methods are most appropriate.

2.4 Special considerations of health datasets

Health datasets present unique challenges given that the datasets may be incomplete due to under-reporting, diagnostic methods vary, and the dataset may reflect exposure at an earlier time or at another place, especially when the disease of interest has a long latency time or incubation period.

2.4.1 Non-reporting and incomplete or inconsistent reporting

Diseases for which regular and timely information is considered necessary for control of the disease are termed notifiable diseases. This means that data collection is mandated at the state and local levels, and data are reported weekly to the state health departments through various surveillance systems. Both *E. coli* O157 and cryptosporidiosis are notifiable diseases in the United States (Adams et al. 2010), therefore when a patient presents to the hospital or doctor's office with symptoms of disease, and lab tests confirm the diagnosis, information about the case is input into the surveillance database. For a case of disease to be reported to public health authorities, a number of sequential steps must therefore occur: the patient must seek medical care, a

specimen must be taken and analyzed by a laboratory, and the cause must be identified and then reported through the surveillance network. All of these steps must be complete for the report to be made, and because most surveillance data include only laboratory-confirmed cases of illness, underreporting and underdiagnosis are common (Scallan et al. 2011, Adams et al. 2010).

Variability exists in the diagnostic tests used for identification of confirmed cases of *E. coli* O157 and cryptosporidiosis, which can impact the number of reported cases (and disease incidence). Test sensitivity (the ability of a test to correctly identify those who have the disease) and test specificity (the ability of a test to correctly identify those who do not have the disease) are important in validating a diagnostic method (Gordis 2009). The predictive values of a test incorporate sensitivity, specificity and also the prevalence of the disease in the population to provide a quantitative measure of how well a test predicts disease given the prevalence of disease in a population.

Diagnostic tests for cryptosporidiosis have changed over the course of this study time period. The adoption of rapid card assays was followed by an increase in confirmed cases, later associated with an increased number of false positives associated with a low positive predictive value (PPV) of rapid card assays (56%) compared to the overall PPV of the nonrapid assays PCR, DFA and Ova and Parasite Exam (97%) (Robinson et al. 2010). Of 117 Tennessee labs that responded to a survey of diagnostic testing methods for cryptosporidiosis, twenty (17%) test on site using one of four methods: rapid card assay (8%), EIA (2%), DFA (3%), and by microscopic examination of a stool sample (4%). A shift from DFA to the rapid card assay in three labs was

followed by a doubling in the number of cryptosporidiosis cases reported (Judy Manners, Tennessee Department of Health, personal communication). Whether this finding suggests a reduction in underreporting (related to the new method) or an increase in false positives (associated with rapid card assays) (Yoder et al. 2012) is unclear. An evaluation of the positive predictive value (PPV) of rapid card assays found a PPV of 56% for rapid card assays versus 97% for nonrapid assays (Robinson et al. 2010), which suggests that the increase in cryptosporidiosis incidence may result in part from the switch to rapid card assay method of diagnosis.

The gold standard diagnostic test for identification of *E. coli* O157 is culture, typically accompanied by EIA and PCR to confirm the presence of shiga-toxin, which is harmful to humans. EIA has been associated with a high rate of false positives (Amy Woron, Molecular biologist, Tennessee Department of Health, personal communication). Of 132 Tennessee clinical laboratories responding to a 2007 survey of testing methods for *E. coli* 0157 and STEC , fifty-six (42%) tested onsite using the following methods: culture based methods only (47 labs, 36%), non-culture based methods only (1 lab, 0.7%), or both culture and non-culture based methods (eight labs, 6%) (Hanna et al. 2010).

In addition to variability in reporting from lab to lab, underreporting and underdiagnosis have been related to socioeconomic status, as individuals with limited access to healthcare are less likely to seek treatment (Lake et al. 2007). In addition, symptoms vary in severity between cases, and in the case of cryptosporidiosis, generally resolve themselves within two to three weeks (Yoder et al. 2012). If symptoms

are mild, a case may go unreported if the patient does not seek medical attention (Said et al. 2003) and therefore cases of cryptosporidiosis are more likely to go unreported than cases of *E. coli* O157because the symptoms are generally less severe.

Records of disease prevalence in humans can therefore be incomplete because of variability in diagnostic methods, access to healthcare, and reporting (Suresh et al. 2012, Fayer 2008). False positives associated with rapid card assays, for example, may increase the number of cryptosporidiosis cases reported from labs using this method, which may affect the spatial distribution of disease reporting and be a source of bias.

2.4.2 Incubation period

When working with health data, it is necessary to consider the time between exposure and development of disease. Disease latency can present problems when working with diseases such as cancers, where latency times can span decades (Jarup 2000). In cases such as these, it can be difficult to link environmental exposures to disease because of migration into or out of the study area over time. For cryptosporidiosis and *E. coli* O157, symptoms develop within days to weeks after exposure (Naumova, O'Neil and MacNeil 2005, Centers for Disease Control and Prevention 2011b), and therefore disease incubation periods are a concern only for patients who report recent travel (defined as having occurred during the incubation period for the disease), which suggests the disease could have been contracted outside of the study area and be unrelated to the explanatory variables under investigation. To

negate the effect of incubation periods for disease in this research, the case records for patients having reported recent travel were excluded.

CHAPTER 3 METHODS

The research was accomplished in three steps. First the spatial databases of disease data and explanatory variables were assembled, and exploratory mapping was done. In this step the cases were geocoded, rates were calculated and standardized for each zip code, and explanatory variables were extracted for each case and zip code, and overlayed with the disease data. In the second step, the regression models were developed. Two modeling approaches were used; a case-control approach (termed the individual model) examined the risk for disease in the individual and an ecological model (termed the regional model) examined the risk for disease in the population. In the third step, model results were compared within and between modeling methods, with the dual goal of identifying risk factors for each disease and assessing the effect of scale, respectively.

3.1 Assembling the case dataset

Both cryptosporidiosis and *E. coli* O157 are notifiable diseases, and since 1999, laboratory-confirmed cases of common food-borne pathogens such as STEC (one strain of which is *E. coli* 0157) and the protozoan cryptosporidium have been recorded in the Tennessee Department of Health's Foodborne Disease Active Surveillance Network (FoodNet). Cases of reportable disease are tracked in this national electronic surveillance system, and if the case is related to a known outbreak or attributed to a specific source, this is recorded.

For this study, a dataset of *E. coli* O157 and cryptosporidiosis cases occurring in Tennessee from 2000-2010 was extracted from the FoodNet database by the Tennessee Department of Health. The cryptosporidiosis and *E. coli* O157 case databases contained 555 and 903 patient records, respectively, each of which included information on patient age, gender, date of specimen, address, recent travel history, and whether the infection was associated with an outbreak.

Appropriate safeguards were taken to protect patient privacy and confidentiality, which are outlined in agreements with the Institutional Review Boards of all institutions involved in the study (East Tennessee State University, University of Tennessee, and the Tennessee Department of Health). These safeguards included use of secure file transfer methods, password-protection of files and databases to control access to patient data, introducing random spatial error in maps of case locations, and aggregation of cases to the zip code level.

Each record was entered into a spreadsheet, and manually examined for completeness and for characteristics that would render it unacceptable for inclusion in the dataset to be modeled: missing data, recent travel, duplicate records, or association with an identified outbreak. These will be discussed in the following paragraphs.

Some record attributes (such as patient address or age) were essential to the study, while other attributes were non-essential. For eighteen cryptosporidiosis cases (7%) and four *E. coli* O157 cases (2%), the date of specimen was missing, and therefore the date of hospitalization or date of onset of symptoms was used as a proxy.

International or out-of-state travel within the twenty one (21) days (cryptosporidiosis) or fifteen (15) days (*E. coli* O157) prior to the specimen date was cause for discarding a case, because it was uncertain whether infection occurred at home or out of state. The incubation period for cryptosporidiosis is one to twelve (12) days, with a range of one to nine (9) days before the patient seeks treatment (Hunter et al. 2007). Seventeen (17) cryptosporidiosis case records were discarded due to international travel, and two records of disease reported from other states were discarded. The incubation period for *E. coli* O157 ranges from one to ten (10) days, with typically another one to five days before the patient seeks treatment (Centers for Disease Control and Prevention 2011b), therefore fifteen (15) days was selected as the *E. coli* O157 threshold for travel. Twenty one (21) *E. coli* O157 cases exceeded the threshold and were discarded.

Some patients provided multiple stool specimens during the course of their infection (i.e. within one week), which resulted in duplicate records. In this case, the record associated with the earliest specimen date was retained and the remaining duplicate records were discarded. Two duplicate cryptosporidiosis records and twenty five duplicate *E. coli* O157 records were discarded.

Two *E. coli* O157 cases attributed to foodborne outbreaks were discarded because the source of the infection was known and not attributed to environmental risk factors. All Tennessee cases located outside the study area counties were discarded leaving 251 cryptosporidiosis records and 270 *E. coli* O157 records.

The environmental risk factors selected as explanatory variables include geology, surface water impairment, agricultural animal population density by zip code, land use (30 meter resolution raster data), groundwater well permit density by county, and poverty rate at the block group level. Each explanatory variable was extracted at the best resolution available. These data are publicly available as spatially referenced spreadsheet data (related to a geographic unit such as a county or census unit) or as shapefiles that were directly imported into ArcGIS 10.0 (Environmental Systems Research Institute 2011).

3.1.1 Case data

Latitude and longitude coordinates for the disease cases were geocoded from the case record address data using the University of South California WebGIS Services batch geocoder (Goldberg and Wilson 2011). The geocoder employed 2008 Tiger/Line files (United States Census Bureau 2012) for streets, places, cities, Zip Code Tabulation Area (ZCTA), county subdivisions, counties, and states. Records with a match score less than 98% and records matched to state, county, city or ZCTA centroid (regardless of match score) were examined individually and correct latitude and longitude coordinates were determined by correcting common misspellings of street names and/or manually mapping each record using Google Earth 6.0.

Krieger and others suggest that zip codes in health datasets obtained by patient self-reporting may not directly correspond to the ZCTAs (Krieger et al. 2002). To combat this potential source of bias, cases in which patient self-reported zip code did

not match geocoded zip codes were examined individually and corrected. Four cryptosporidiosis records and twenty *E. coli* O157 records remained unresolved and were discarded, and the resulting datasets contained 247 and 250 records respectively.

3.1.2 Explanatory variables for individual models

The geocoded case records were mapped in ArcGIS 10.0 and county, zip code and census block group 2010 Tiger/Line files were added.

USDA Agricultural Census Data from 2007 (United States Department of Agriculture 2009) were joined to the ZCTA, and animal population densities were calculated by dividing population by ZCTA area which was measured using the calculate geometry tool in ArcGIS 10.0.

Karst regions were delineated by selecting polygons classified as limestone or dolomite in the 1:250,000 Geology of Tennessee shapefile (Greene and Wolfe 2000) and creating a new layer of the karst-prone regions of Tennessee. The near tool in ArcGIS was used to measure the distance from each case to a karst region. Cases falling within a karst region had a distance of zero.

Surface water shapefiles and attribute data for the Year 2008 305(b) Report (Denton et al. 2008) and Year 2008 303(d) Report (Tennessee Department of Environment and Conservation 2008) were downloaded from the Environmental Protection Agency Reach Address Database (EPA RADims accessed at http://epamap32.epa.gov/radims/). Stream segments that were identified as impaired for the presence of *E. coli*, fecal coliform, or pathogens, were extracted from the

impaired waters database. The near tool in ArcGIS was used to determine the distance from each case to the nearest impaired stream segment.

The 2006 National Land Cover Dataset raster (Fry et al. 2011) was downloaded from the ArcGIS add-on ArcGIS online (downloaded 03/26/12) and the extract values to points tool was used to associate each case record with land cover. Three dummy variables were generated to identify presence within developed land (NLCD classifications 21, 22, 23, and 24), forest land (NLCD classifications 41, 42 and 43) and agricultural land (NLCD classifications 81 and 82), and each case fell within at most one of the three classes. Because the spatial logistic regression model could not accept binary explanatory variables, the categorical NLCD raster was reclassified and smoothed to produce three rasters representing percent cover of each of three land use classes (PASTURE, FOREST, and DEVELOPED) in a 500 m X 500 m raster cell.

A database of well permits was obtained from Tennessee Department of Environment and Conservation, and because detailed geographic identifiers were incomplete for most records in the database (but county location was complete for all records), a count of well permits in each county was made and joined to the 2010 Tiger/Line county layer.

Poverty rates for each Census block group were extracted from Table B17017 (2010 US Census) using the Summary File Retrieval Tool available from http://www.census.gov/acs/www/data_documentation/summary_file/ (retrieved 04/10/12) and were joined to the block group layer.

The case records databases were joined to the geographic data using a spatial join, the study area was outlined by selecting all counties lying within the study area, and all layers were clipped to the study area outline. Candidate explanatory variables are outlined in Table 1.

3.1.3 Control data

The objective of the case-control modeling approach is to develop a series of models to identify risk factors for cryptosporidiosis and *E. coli* O157 in the individual from a candidate list of likely environmental and socioeconomic variables. While it is appropriate to consider risk factors that have emerged as significantly associated with disease in other studies, it is useful to assess whether these risk factors are likely to be significant in the study area.

Likely explanatory variables can be identified by comparing the distributions of these variables at case (presence) locations with the distributions of these variables at control locations where disease is not present, because this is an indication of a potential statistical relationship between the explanatory variable and disease. One limitation of doing a case-control study using an existing database, is that control data may not be available, in which case random absence data can be generated and used as the control dataset (Bivand et al. 2008). Four options for generation of the control data for use in the first group of individual models were considered, and are outlined below.

Table 1. Candidate explanatory variables for cryptosporidiosis and *E. coli* O157.

Variable	Description	Units
BEEF[1]	Beef cow population density (zip code level data)	Animals/km^2
MILK[1]	Milk cow population density (zip code level data)	Animals/km^2
CALF[1]	Calf population density (zip code level data)	Animals/km^2
COW[1]	Total cow population density includes beef cows, milk cows and calves (zip code level data)	Animals/km^2
HOG[1]	Hog population density (zip code level data)	Animals/km^2
SHEEP[1]	Sheep population density (zip code level data)	Animals/km^2
AGDEN[1]	Total agricultural animal population density includes all cows, hogs and sheep (zip code level data)	Animals/km^2
DEVELOPED[6]	Binary variable for presence within a developed land use classification or percent cover in a 500 m X 500 m raster cell (for spatial logistic regression models only)	Dimensionless
FOREST[6]	Binary variable for presence within a forest land use classification or percent cover in a 500 m X 500 m raster cell (for spatial logistic regression models only)	Dimensionless
PASTURE[6]	Binary variable for presence within agricultural land use classification or percent cover in a 500 m X 500 m raster cell (for spatial logistic regression models only)	Dimensionless
KARST[3]	Distance to nearest karst geology	Kilometers (Euclidean distance)
STREAM[4]	Distance to nearest *E. coli* contaminated stream segment	Kilometers (Euclidean distance)
POVRATE[2]	Poverty rate (block group level data)	Percent, expressed as decimal
WELLS[5]	Number of well permits by population (county level data)	Permits/10,000 population

Data Sources:
1 USDA 2007 Agricultural Census
2 US 2010 Census
3 USGS 1:250,000 Geology of Tennessee
4 EPA Year 2008 303(d) and 305(b) reports
5 TDEC Well permit database
6 NLCD 2006 Dataset

First, the Create Random Points tool was used in ArcGIS to generate 250 points throughout the study area. The disadvantage of this option is that points representing absence data may be located where disease is unlikely to be reported because population is low or nil.

Second, to incorporate population distribution, a large number of random points were generated and weighted by population density and sorted by weight. The top 250 points (by weight) were selected as the absence data. This resulted in a clustered set of absence data, with the degree of clustering dependent upon the initial number of random points, n. Three different values for n were used (n=1000, n=2000 and n=3000), and clustering increased with increasing n. The disadvantage of this method is the dependence of the control data on the selected value for n.

Third, to better preserve the randomness of the control data, each generated point was assigned a pseudo-random weight, calculated by multiplying the population density for each point by a random number between zero and one. The weights were then sorted, and the top 250 weights were selected as the absence data. The disadvantage of this method is that the absence data were not well correlated with population density and the objective of generating control data in areas where people live was not achieved.

Fourth, amid concerns of too much data manipulation, a new approach was sought, which was to identify a population threshold below which disease is absent, and then generate random absence data from within census block groups with populations above the identified threshold. Block group population data (2010 US

Census) were normalized by land area to generate population density, and were spatially joined to the disease datasets in ArcGIS. The case record data were sorted by population density and the minimum block group population density where disease occurred was selected as the threshold below which disease was assumed to be absent. Block groups below this threshold were masked out of the study area, and the generate random points tool in ArcGIS was used to generate 250 random points from within the candidate area. These random points served as control locations for the first group of individual models.

For the final individual model, a spatial logistic regression model was used, which eliminated the problems of generating control data because this method did not require that controls be specified.

3.2 Assembling the aggregate dataset

3.2.1 Rate standardization

Because disease incidence generally increases with the number of people exposed, disease is often reported as a rate instead of a count. The Zip Code Tabulation Area (ZCTA) was selected as the base spatial unit for data aggregation because the agricultural variables emerged during earlier analyses as important risk factors for disease, and they are reported at the zip code level. Morbidity rates were therefore generated from the geocoded case data for the 172 ZCTAs in east Tennessee, by counting the number of cases in each ZCTA and dividing by the ZCTA population (2010 US Census). Rates are expressed as cases per 10,000 population.

E. coli O157 and cryptosporidiosis are diseases that disproportionately affect the very young or the very old, and it is therefore important to take population distribution by age into account when reporting morbidity rates. For example, a geographic area with a high percentage of preschool-age children may have a higher disease rate simply because young people make up a larger proportion of the population. To eliminate population age distribution as source of bias, morbidity rates for cryptosporidiosis and *E. coli* O157 were standardized by age and gender using Indirect Standardization (Dohoo, Martin and Stryhn 2003). The reference datasets were disease incidence in Tennessee from 2000-2010 (FoodNet database) and population breakdown by age for Tennessee (US 2010 Census).

3.2.2 Explanatory variables for regional models

Data sources for explanatory variables have been outlined in a previous section and in Table 1, therefore this section will explain how the explanatory variables were aggregated from the individual level to zip-code level areal data. These variables are summarized in Table 2.

Agricultural variables The agricultural risk factors were animal population density for beef cows (BEEF), milk cows (MILK), calves (CALF), hogs (HOG), and sheep (SHEEP). These variables were reported at the zip code level from the agricultural census and incorporated in the aggregate (regional) models without transformation. Agricultural census data were joined to the standardized disease rates by matching the ZCTA field

Table 2: Description of explanatory variables used in aggregate models.

Variable	Description
Lat	Latitude of zip centroid
Long	Longitude of zip centroid
CR_std10K	Standardized cryptosporidiosis rate per 10,000 people
ER_std10K	Standardized *E. coli* O157 rate per 10,000 people
BEEF	Beef cow population density (by zip code area)
MILK	Milk cow population density (by zip code area)
CALF	Calf population density (by zip code area)
COW	All cows population density (by zip code area)
HOG	Hog population density (by zip code area)
SHEEP	Sheep population density (by zip code area)
AGDEN	All agricultural animals population density (by zip code area)
DEVELOPED	Percent developed land (medium and high intensity)
FOREST	Percent forested land
PASTURE	Percent pasture land
KARST_PCT	Percent karst geology
STREAM_PCT	Percent of assessed stream miles impaired by ecoli
POVRATE	Poverty rate (weighted sum of block group poverty rates)
WELLS	Well permit density for zip code (weighted sum of county well permit density by county population).

in the Tiger/Line files with the zip code field in the agricultural census data. Populations of beef cattle, milk cattle, calves, total cattle, hogs, sheep and total agricultural population were normalized by the ZCTA area, which was calculated using the calculate geometry tool in ArcGIS to generate seven agricultural variables (BEEF, MILK, CALF, COWS, HOG, SHEEP, and AGDEN, respectively). It is important to mention that there can be a spatial disconnect between ZCTA and zip codes, especially when comparing data between decadal censuses (for example, between the 1990 Census and 2000 Census by ZCTA). This disconnect occurs because of changes in coverage areas over time on both the part of the US Postal Service and the US Census Bureau (Krieger et al. 2002). Because the US Census of Agriculture reports by USPS ZIP code and the US Census shapefiles are based on ZCTAs, these discrepancies can be an unavoidable source of bias.

Geology At the regional level, geology was included as the percent of karst land surface area for each zip code (KARST_PCT). The proportion of karst geology within each zip code area was calculated in ArcGIS 10.0 by breaking up the karst polygons at the ZCTA boundaries using the intersect tool, and calculating the area of the new polygons using the calculate area tool. The karst polygon area was summed for each ZCTA and divided by the total area to generate the variable (KARST_PCT). Note that this variable differs from the distance to karst (KARST) variable used in the individual models, and the interpretation of the coefficient is different. For example, a positive coefficient for KARST indicates increased disease with increasing distance from karst geology,

however a positive coefficient for KARST_PCT indicates increasing disease with increased karst surface area.

Surface water quality Surface water quality was included at the regional level as the percent of assessed stream miles found to be impaired with *E. coli* (STREAM_PCT). To generate this variable, *E. coli* impaired stream segments were extracted from the EPA 305b shapefiles of assessed surface water stream segments using ArcGIS 10.0. Stream polylines were broken at ZCTA boundaries using the intersect tool, and the length of the new reaches was calculated using the calculate geometry tool, and summed for each ZCTA. The water quality variable (STREAM_PCT) was calculated as the ratio of *E. coli* impaired to assessed stream miles. Careful interpretation of the water quality parameters in the individual and regional models must be made because higher values of STREAM (individual models) indicate increased distance from the risk factor, yet higher values of STREAM_PCT (regional models) indicate increasing coverage of the risk factor. Therefore, all things being equal, one might expect a positive coefficient for one variable and a negative coefficient for the other due to the way the variables are defined in the individual and regional models.

Land use variables In the regional models, the NLCD raster was used to generate statistics on the percent land cover for each land use class in each zip code area. The tabulate area tool was used to measure the amount of area devoted to a given land use in each zip code, which was subsequently normalized by the total land area in the zip

code. Three variables: percent forest (FOREST), percent agricultural (PASTURE), and percent developed (DEVELOPED) land were generated.

Poverty rate US Census poverty data for ZCTA in Tennessee for the 2010 census were not yet released at the time the areal dataset was assembled (Erran Persley, US Census office Nashville TN, personal communication) and was not expected to be available for several months. Poverty rates in each ZCTA were estimated using an areal-weighted sum of the poverty rates in the block groups intersecting with each ZCTA.

Well permit density The number of well permits issued in a county (using all permits recorded in the Tennessee Department of Environment and Conservation well database) was normalized by the 2010 county population. An area-weighted average of the well permit density was calculated for zip codes that straddled multiple counties.

Slight differences exist between some of the explanatory variables used in the individual and regional models because of aggregation. These differences are summarized in Table 3.

Table 3. Comparison of explanatory variables used in individual and regional models.

Explanatory Variable	**Individual (case-control) model**	**Regional (ecological) model**
Agricultural variables (BEEF, MILK, CALF, HOG, SHEEP)	Animal population density of zip code	Animal population density of zip code
POVRATE	Poverty rate of block group	Areal-weighted average of poverty rate of block groups falling within zip code area
KARST/KARST_PCT	Distance to nearest karst geology	Percent karst geology in zip code area
STREAM/STREAM_PCT	Distance to nearest *E. coli* contaminated stream segment	Percent of assessed stream miles contaminated with *E. coli*
WELLS	Number of well permits by population (county level data)	Areal-weighted average of well permits by county population for zip codes falling within one or more counties.
Land use variables (DEVELOPED, FOREST, PASTURE)	Binary variable for presence within a the relevant land use classification or percent cover in a 500 m X 500 m raster cell (SLRM only)	Percentage of land area within zip code.

3.3 Exploratory mapping

Mapping disease rates over candidate explanatory variables is one way to explore the spatial relationship between disease locations and risk factors by visually examining trends or patterns visible on the maps. This is not a substitute for statistical analysis, but a preliminary step necessary to understand the data. The maps presented in Section 3.3.1 and 3.3.2 are raw rates, not smoothed rates, however the use of a smoothing technique such as Spatial Empirical Bayes can have advantages for exploratory mapping of rates for rare diseases. Variability in the population at risk, combined with a low number of cases, introduces variability in disease rates such that a small change in the number of cases of disease can have a significant effect on the rate. SEB smoothing adjusts rates with high variances towards the overall mean (Anselin, Kim and Syabri 2004).

3.3.1 Cryptosporidiosis

Overlays of covariates and cryptosporidiosis raw rates (Figure 4 through Figure 15) provide a visual analysis of the potential spatial relationship between disease rates and environmental and socioeconomic factors. Beef cow (Figure 4) and milk cow (Figure 5) population densities tend to be higher in the north-eastern portion of the study area, coinciding with high rate ZCTAs, and lower in the north-western portion of the study area, coinciding with low rate ZCTAs. There does not appear to be a clear relationship between sheep and hog population densities and cryptosporidiosis (Figure 6 and Figure 7).

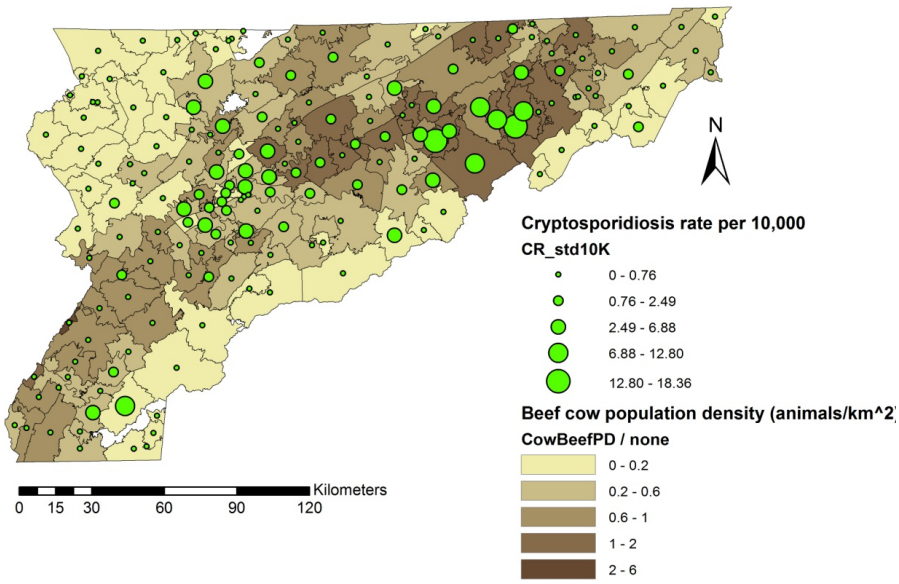


Figure 4. Overlay of cryptosporidiosis rates and beef cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

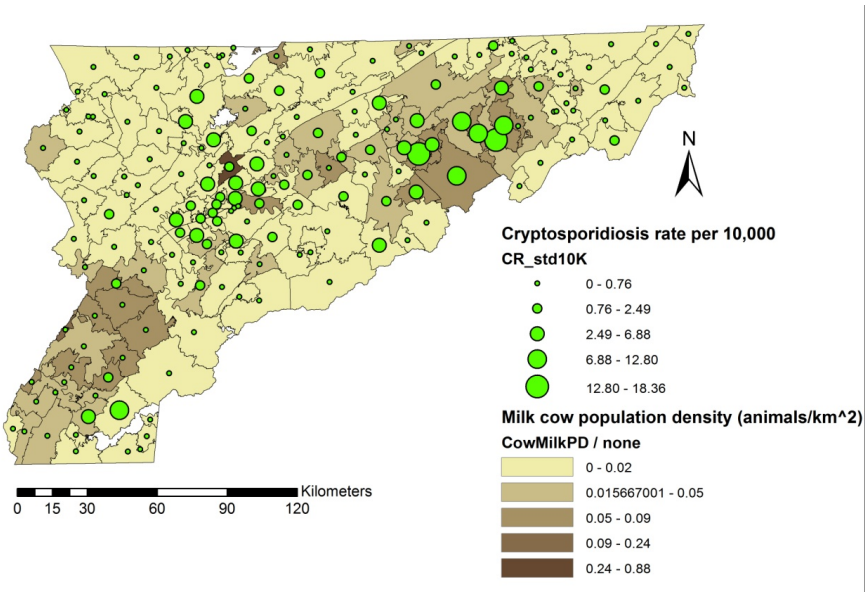


Figure 5. Overlay of cryptosporidiosis rates and milk cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

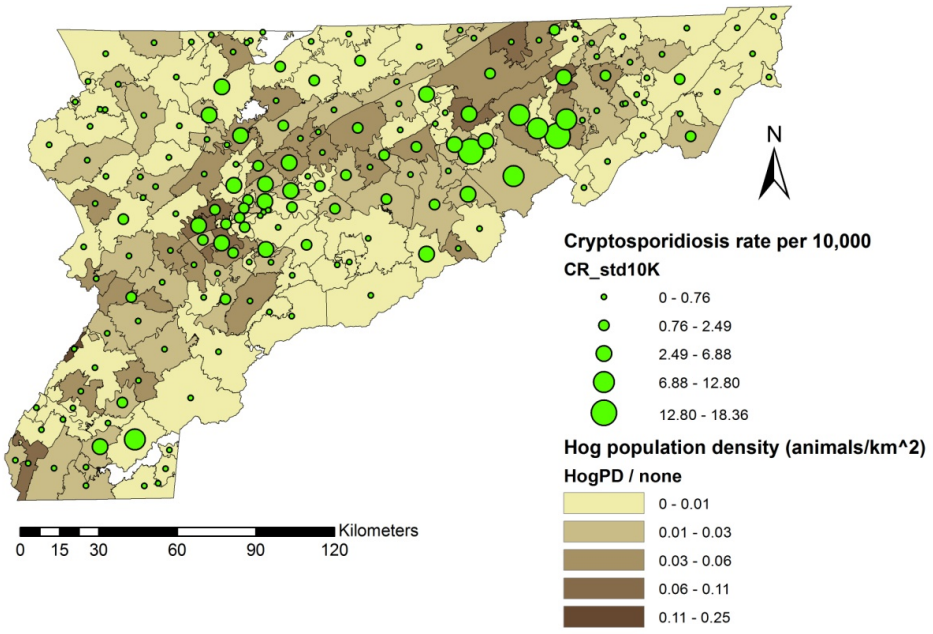


Figure 6. Overlay of cryptosporidiosis rates and hog population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

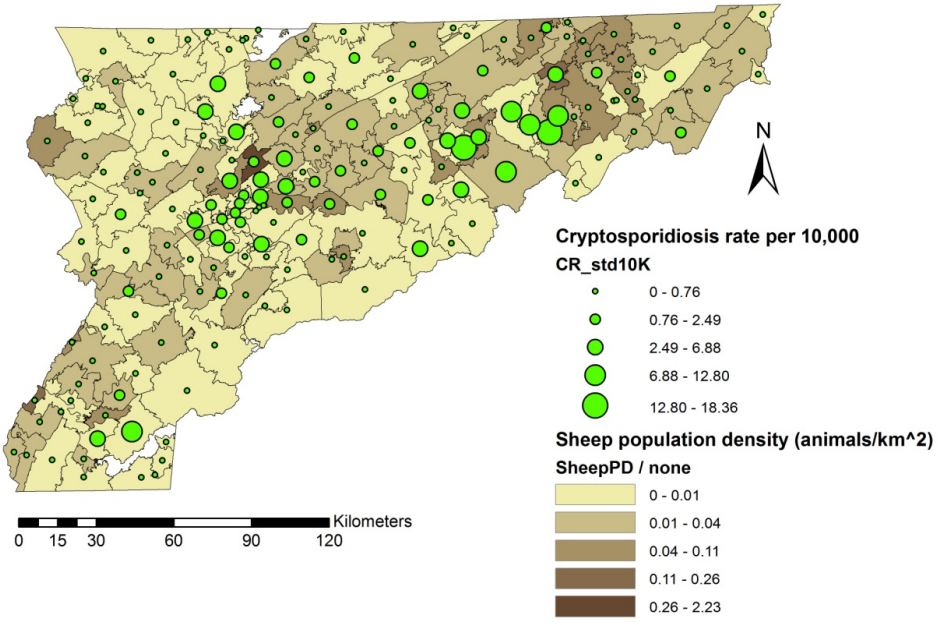


Figure 7. Overlay of cryptosporidiosis rates and sheep population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

An examination of overlays of cryptosporidiosis rate on the land use parameters shows that cryptosporidiosis may be positively associated with developed land (Figure 8), negatively associated with forested land (Figure 9), and positively associated with agricultural land (Figure 10).

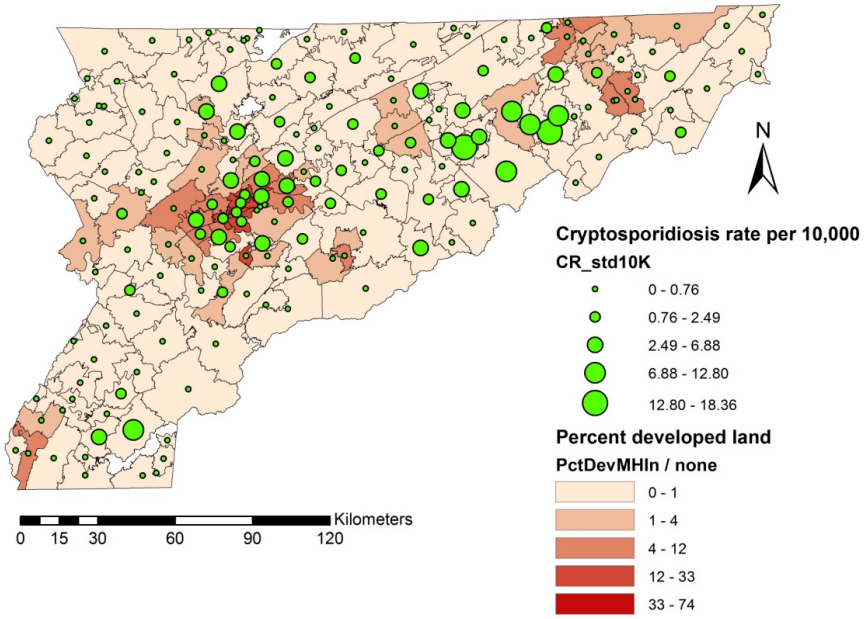


Figure 8. Overlay of cryptosporidiosis rates and developed land use. Data source: National Land Cover Dataset 2006.

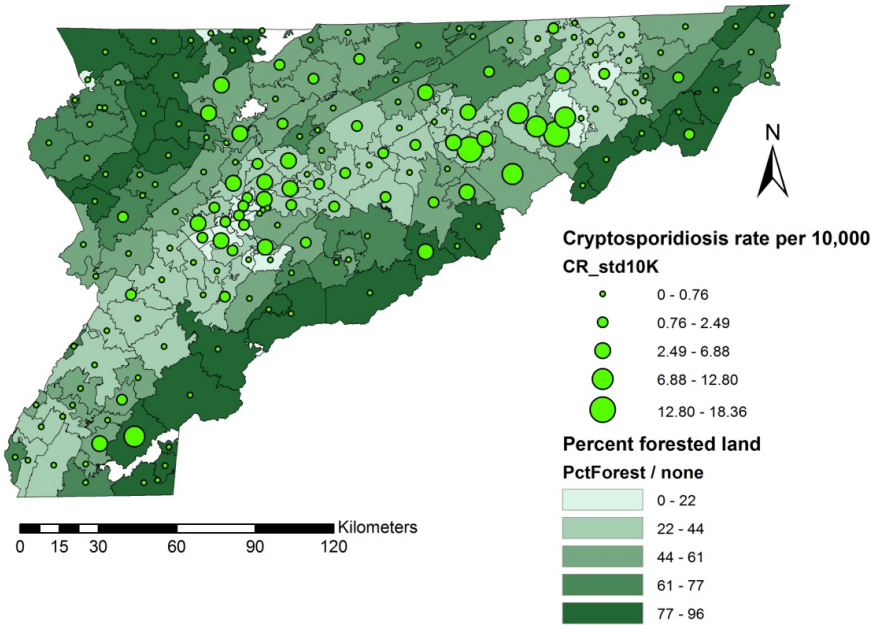


Figure 9. Overlay of cryptosporidiosis rates and forest land. Data source: National Land Cover Dataset 2006.

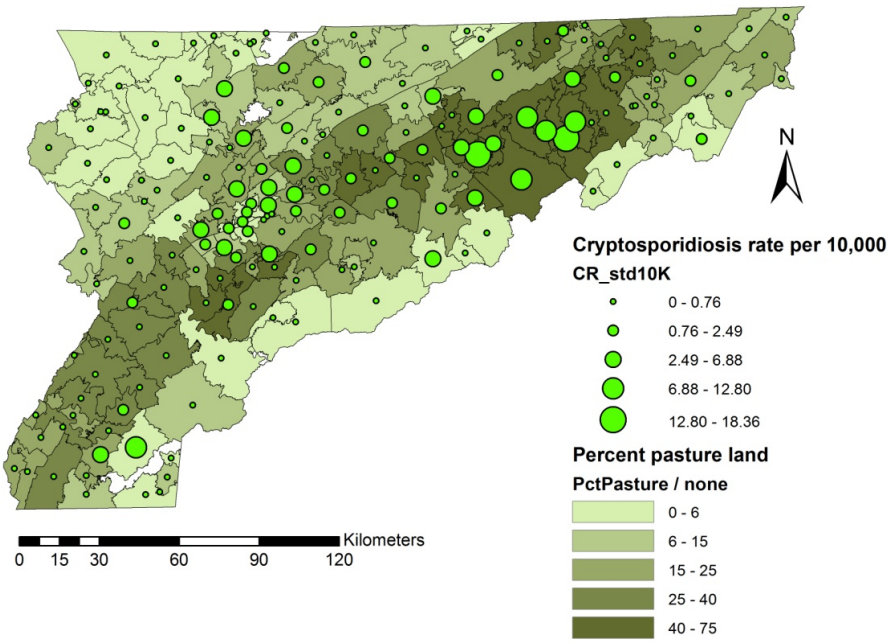


Figure 10. Overlay of cryptosporidiosis rates and agricultural land. Data source: National Land Cover Dataset 2006.

Exploratory mapping of cryptosporidiosis rates with geology and surface water shows that most high rate ZCTAs coincide with areas of abundant karst (Figure 11), but the relationship with regional water quality (Figure 12) is less clear. Figure 13 provides an alternate view of geology and surface water quality by displaying the boundaries of the karst polygons and the line segments representing *E. coli* impaired surface water streams used in the individual models.

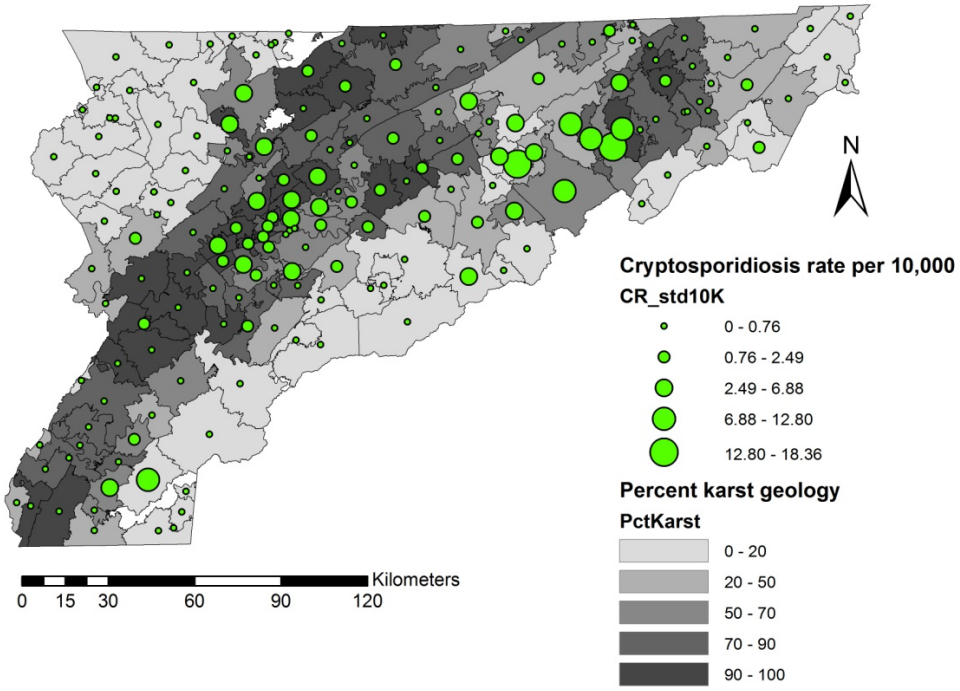


Figure 11. Overlay of cryptosporidiosis rates and karst geologic areas.

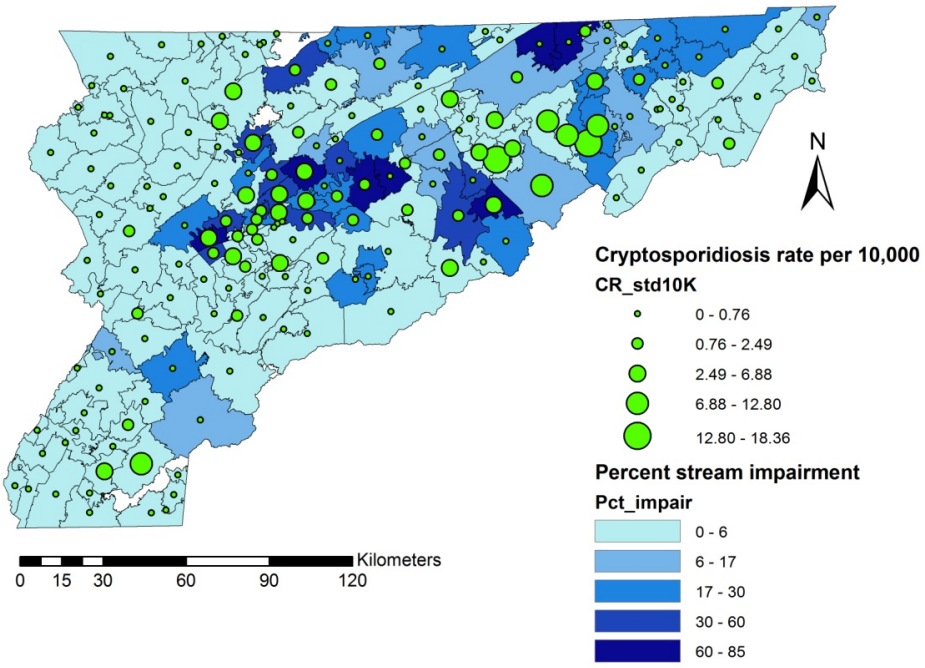


Figure 12. Overlay of cryptosporidiosis rates and surface water quality. The water quality data source is the EPA approved 2008 303(d) list for Tennessee.

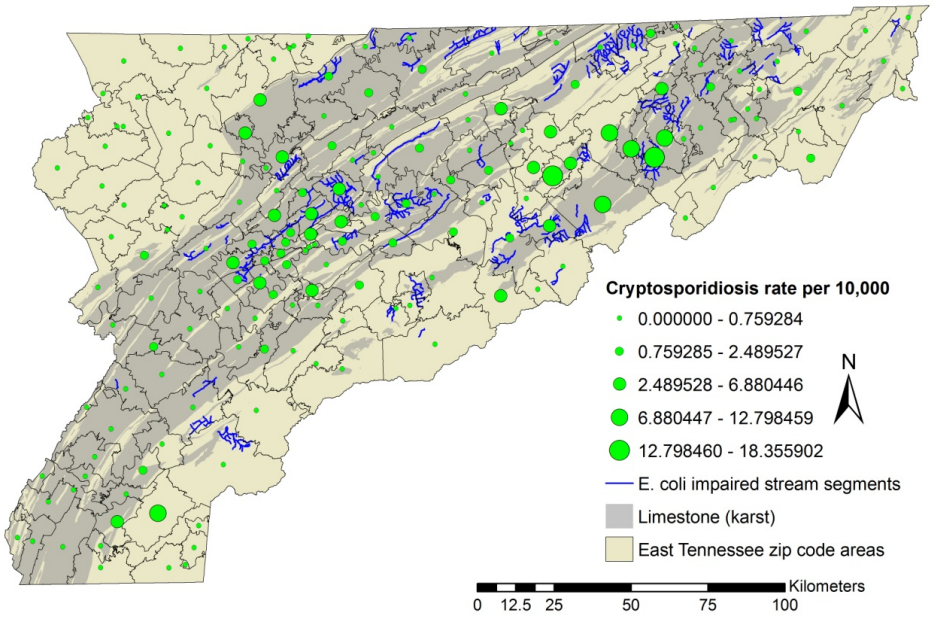


Figure 13. Overlay of cryptosporidiosis rates, karst areas, and *E. coli* impaired stream segments. Water quality data source is the EPA approved 2008 303(d) list for Tennessee.

The overlay of poverty rate and cryptosporidiosis rate does not reveal any obvious relationships (Figure 14). Well permit density is lower in counties with a higher number of cryptosporidiosis cases (Figure 15).

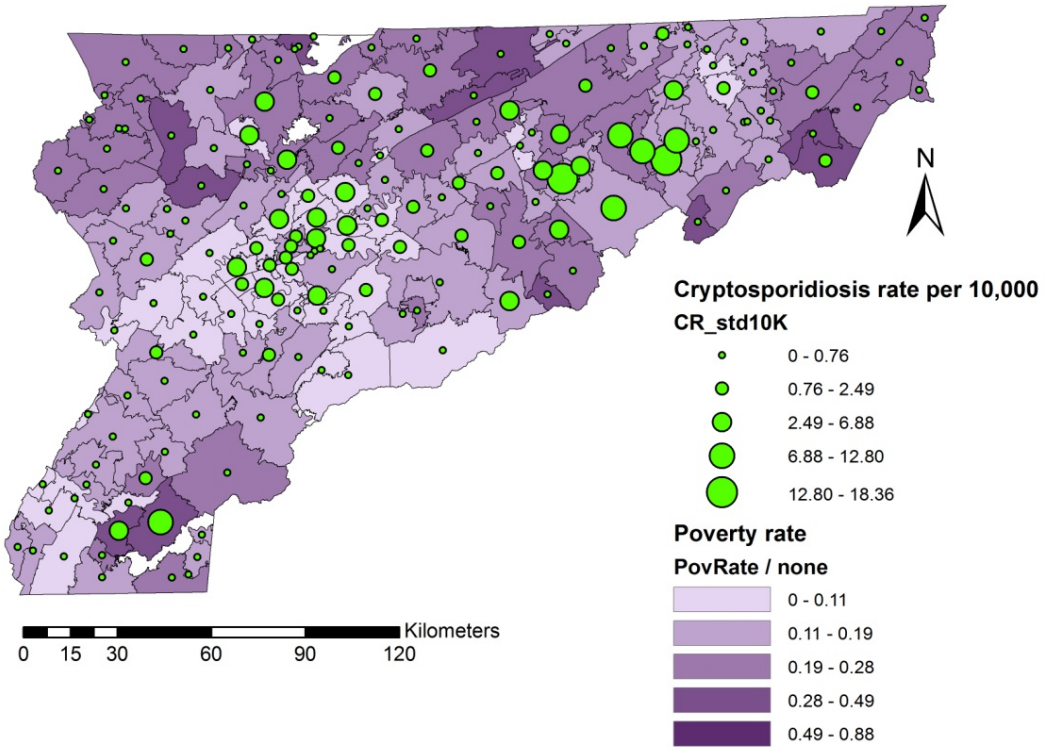


Figure 14. Overlay of cryptosporidiosis rates and ZTCA poverty rate. Data source: 2010 US Census.

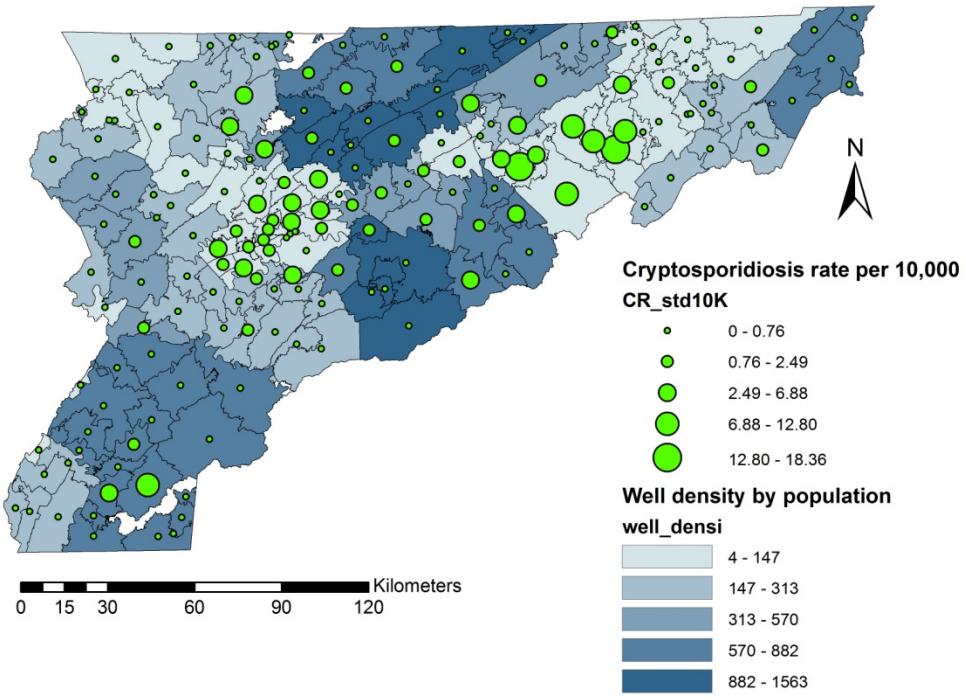


Figure 15. Overlay of cryptosporidiosis rates and county well permit density.

3.3.2 *E. coli* O157

Using a similar overlay of *E. coli* O157 raw rates and explanatory variables (Figure 16 through Figure 27), *E. coli* O157 appears to be positively related to beef cow (Figure 16) and hog (Figure 18) population densities, and percent developed (Figure 20) or pasture (Figure 22) land, while negatively related to percent forest land cover (Figure 21).

Areas dominated by karst geology coincide with areas having high *E. coli* O157 rates (Figure 23). The relationship between impaired surface water and *E. coli* O157 rate is unclear from an examination of two map overlays showing the proportion of

impaired stream miles (Figure 24) and impaired stream segments (Figure 25) .

Accessibility of private water supplies (well permit density) appears to be negatively associated with *E. coli* O157 rate (Figure 27).

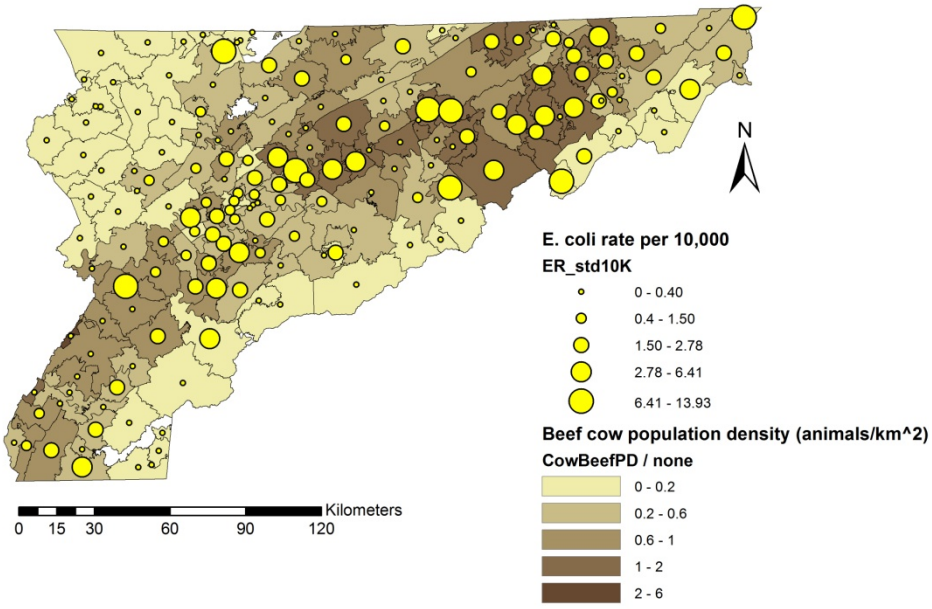


Figure 16. Overlay of *E. coli* O157 rates and beef cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

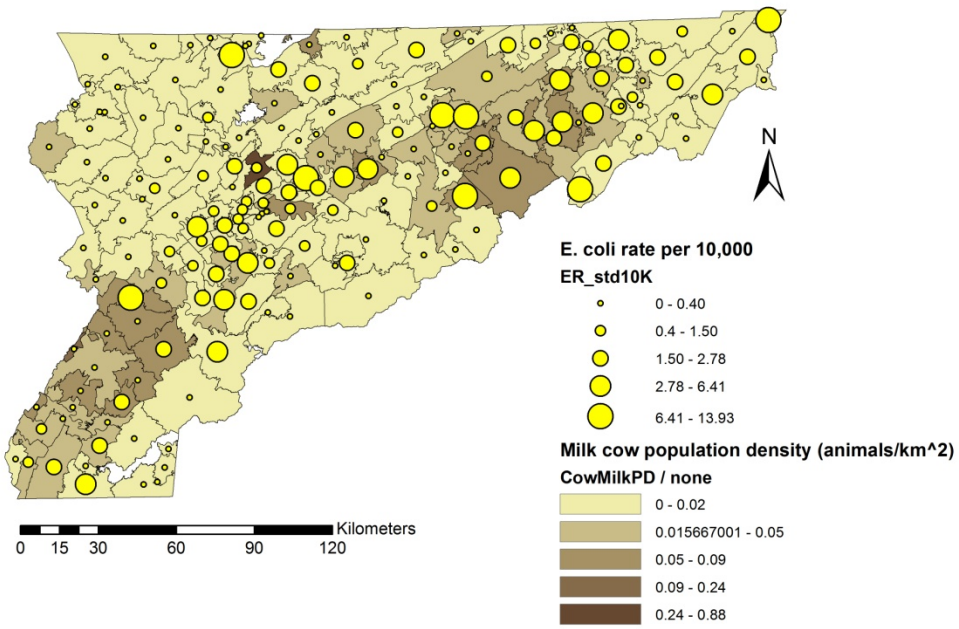


Figure 17. Overlay of *E. coli* O157 rates and milk cow population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

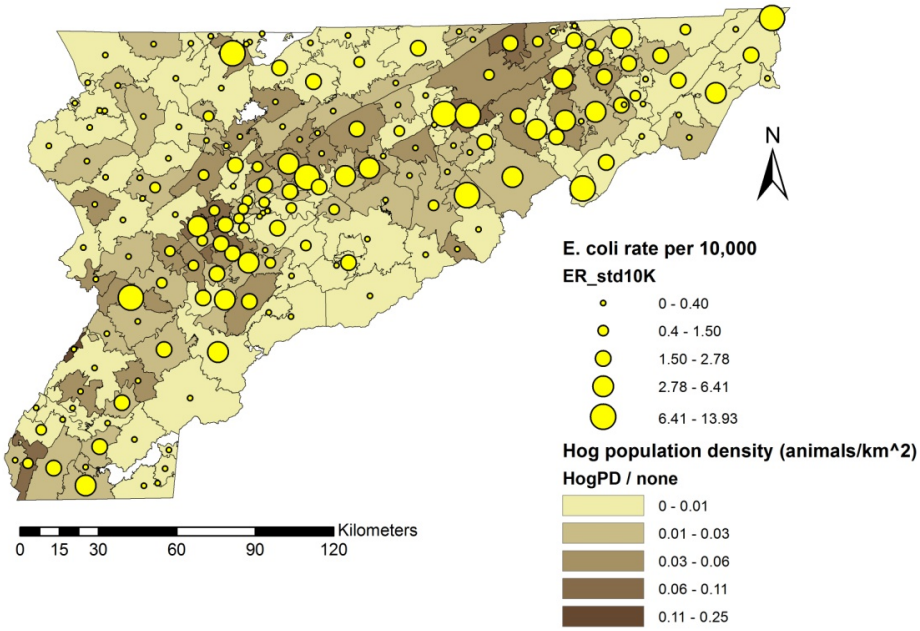


Figure 18. Overlay of *E. coli* O157 rates and hog population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

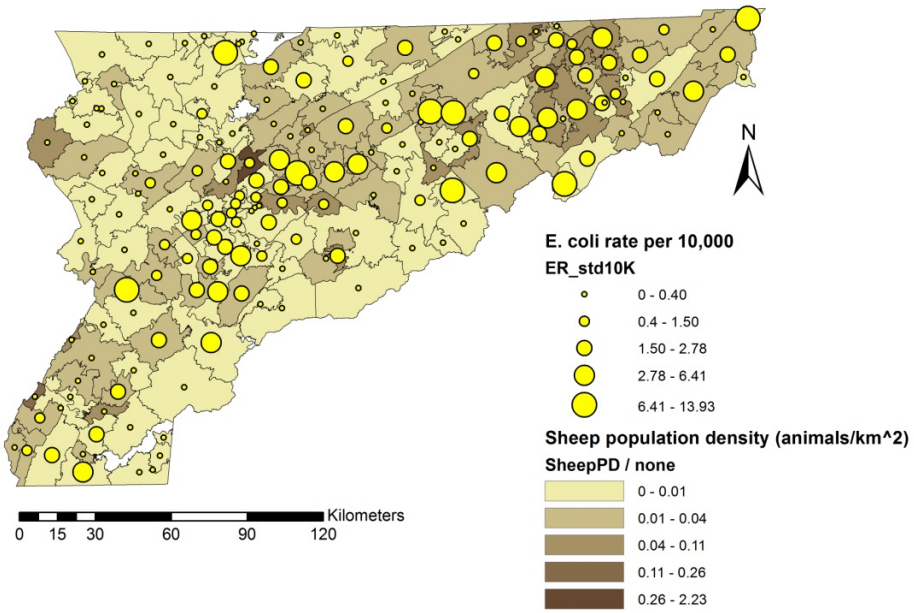


Figure 19. Overlay of *E. coli* O157 rates and sheep population density (animals/km^2). Data source: USDA 2007 Census of Agriculture at zip code level.

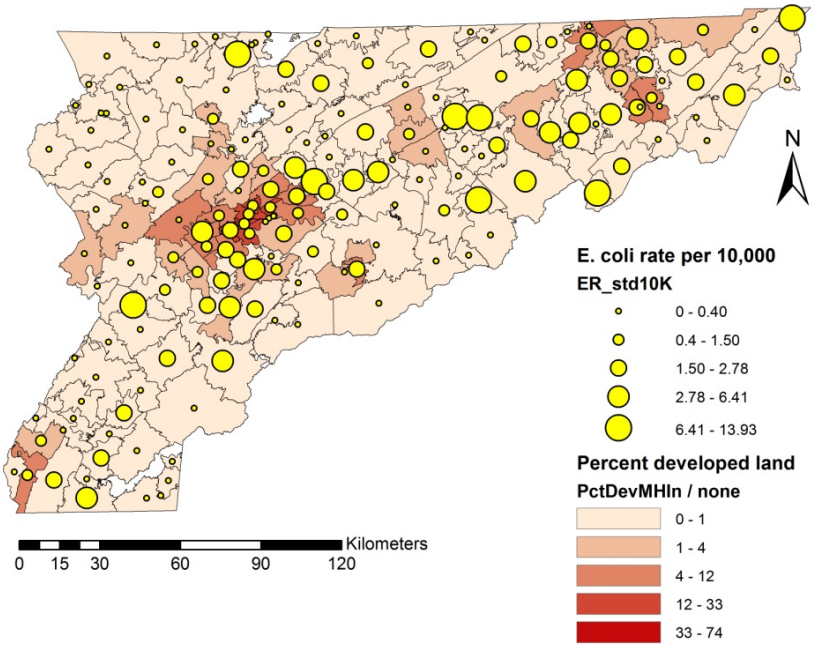


Figure 20. Overlay of *E. coli* O157 rates and percent developed land. Data source: National Land Cover Dataset 2006).

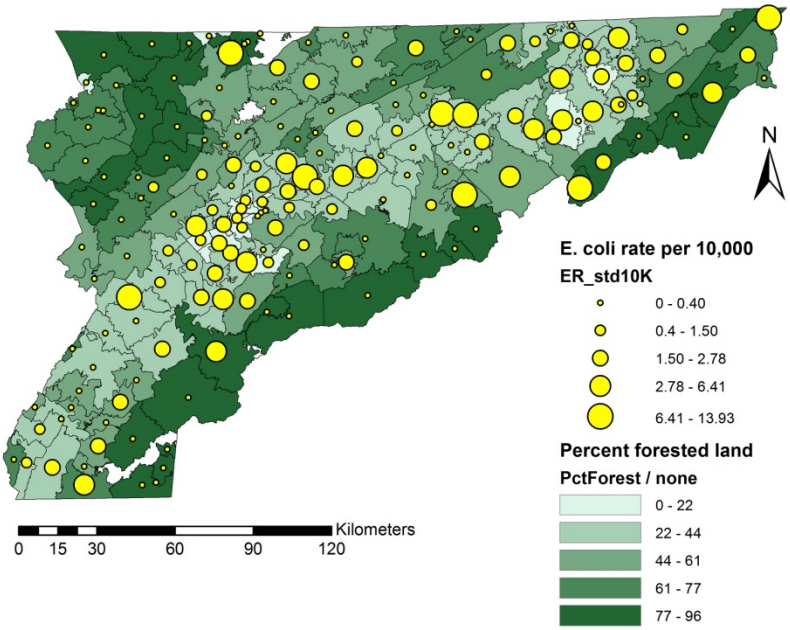


Figure 21. Overlay of *E. coli* O157 rates and percent forested land. Data source: National Land Cover Dataset 2006.

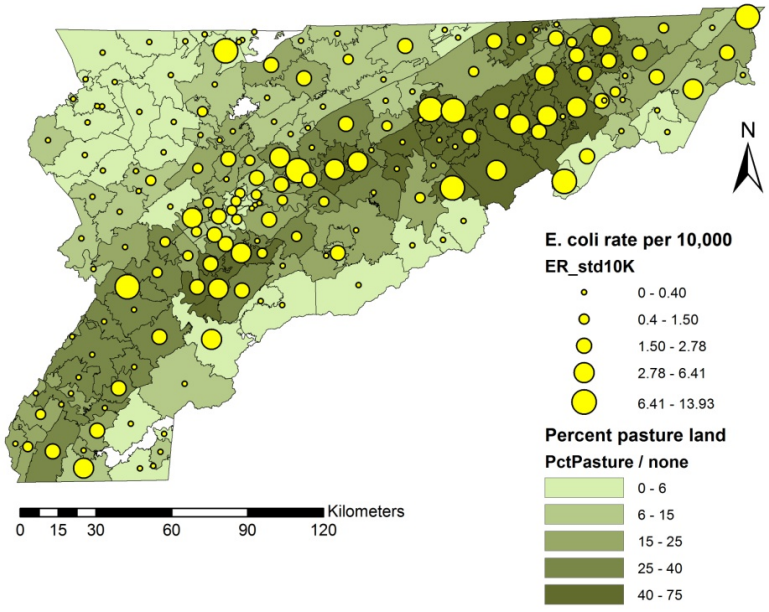


Figure 22. Overlay of *E. coli* O157 rates and percent agricultural land. Data source: National Land Cover Dataset 2006.

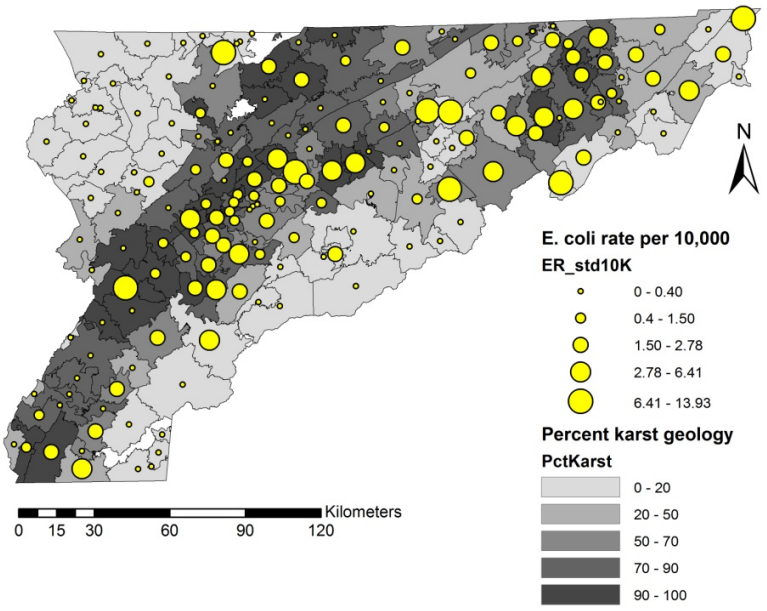


Figure 23. Overlay of *E. coli* O157 rates and karst geologic areas.

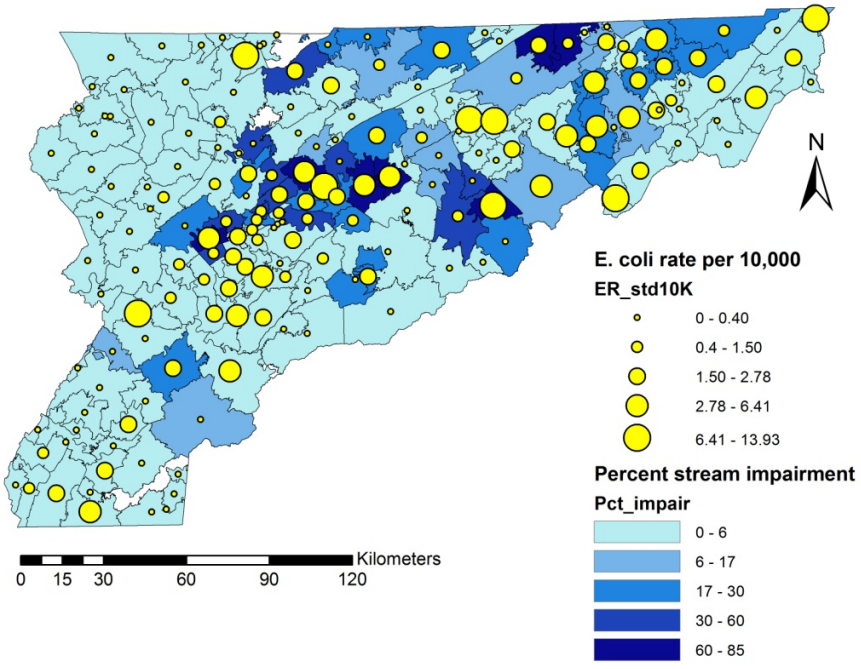


Figure 24. Overlay of *E. coli* O157 rates and surface water quality. The water quality data source is the EPA approved 2008 303(d) list for Tennessee.

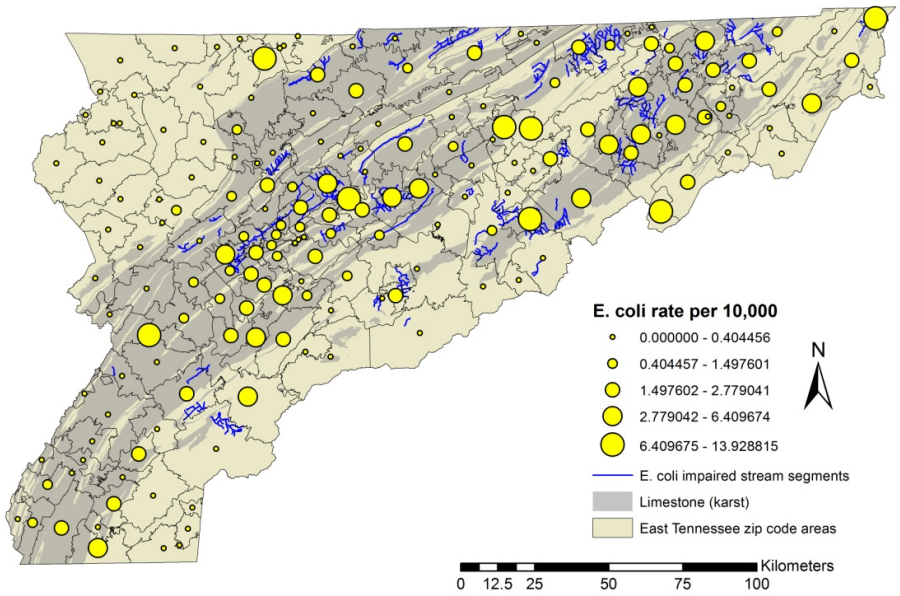


Figure 25. Overlay of *E. coli* O157 rates and *E. coli* impaired stream segments (from EPA 2008 305(b) list) over karst geology.

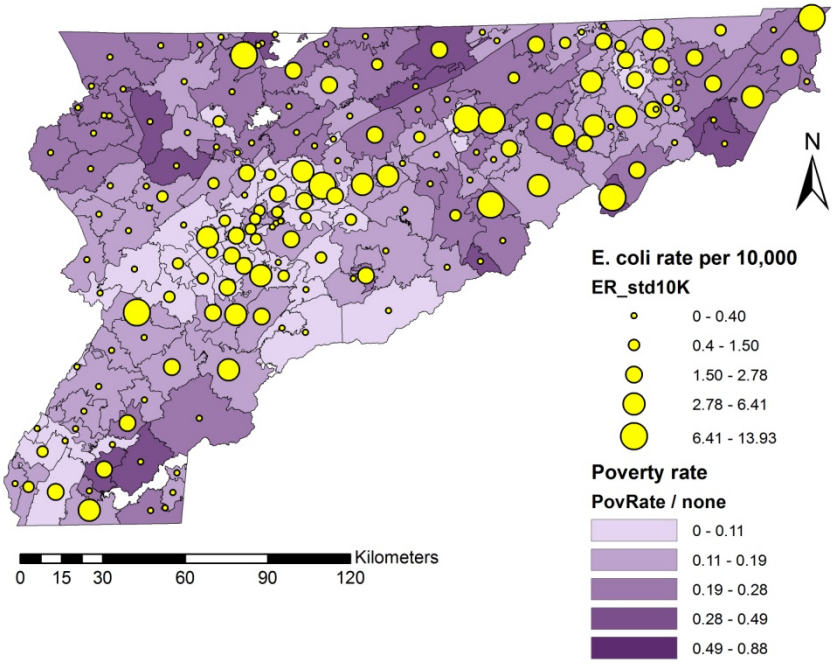


Figure 26. Overlay of *E. coli* O157 rates and ZCTA poverty rate. Data Source: US 2010 Census.

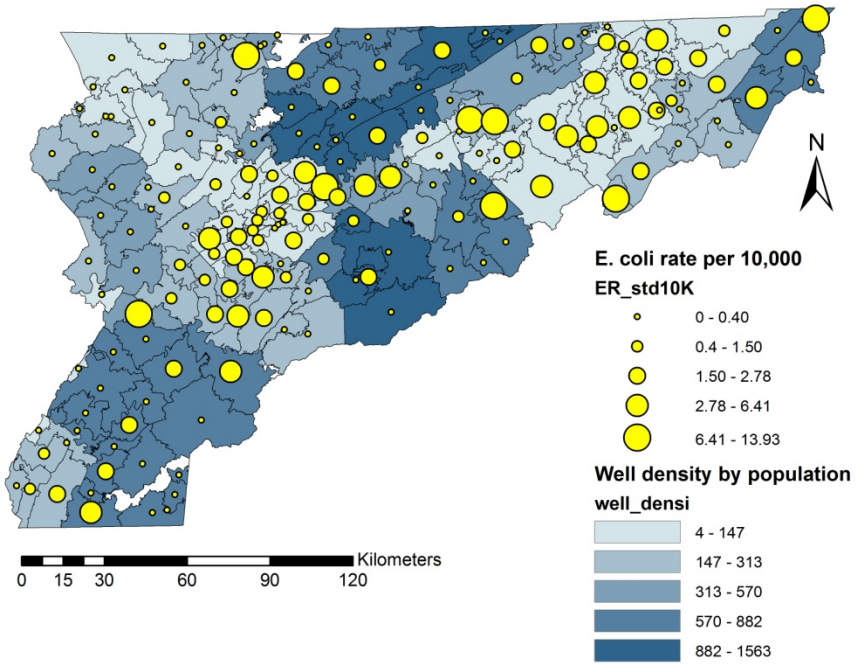


Figure 27. Overlay of *E. coli* O157 rates and county well permit density (number of permits per 10,000 population).

3.4 Individual models

3.4.1 Selection of covariates

Using the controls generated with the fourth method described in Section 3.1.3, distributions of risk factors for the presence (case) and absence(control) datasets were compared in SPSS 20 (IBM SPSS for Windows Release 20 2011) using the Mann-Whitney U-test and Kolmogorov-Smirnov test for continuous data (using Analyze>Nonparametric Tests>Independent Samples), and the Chi-squared test for the land use (categorical) data (using Analyze>Descriptive Statistics>Crosstabs).The risk factors for which case distributions were identified as statistically different from the

control distributions were selected as potential explanatory variables for the presence of disease.

Spearman correlation coefficients were calculated in SPSS 20 for the selected explanatory variables to assess the potential for multicollinearity. Variable pairs with high correlation coefficients were examined, and only one of the pair was included in the model.

Likely explanatory variables were therefore identified with the following attributes: the variable's distribution for the case locations (presence data) statistically differed from a distribution of control points (random absence data), and the correlation between the variable and other explanatory variables was small and no explanatory variable could be expressed as a linear combination of the other explanatory variables.

3.4.2 Model development

Multiple statistical models (Table 4) were generated using SPSS 20 and *R* (R Core Team 2012) to assess consistency of the model output. Multiple models were generated to assess whether a given risk factor was retained in the model because it truly is significant, or whether it is an artifact of the model. By generating multiple models, consistency between models increases confidence in the model results.

Table 4: Statistical point models and platforms.

Model structure	Platform
Logistic Regression	SPSS
Generalized Additive Model	R
Spatial Logistic Regression	R

Each of the models listed in Table 4 was developed for a binary outcome (disease or no disease) using both the cryptosporidiosis and *E. coli* O157 datasets. These models were selected because each provides key pieces of information in understanding the relationship between risk factors and disease incidence.

A logistic regression model was developed in SPSS 20 using the 'forward stepwise' variable selection method. Next a GLM and GAM were developed in R using the `glm` and `gam` functions, respectively, specifying a binomial distribution. For the GAM, a logit link function was used and the residual spatial variation was modeled using a smooth spatial function (Bivand et al. 2008).

A limitation of logistic regression is the assumption of independence in the residuals. To incorporate spatial explanatory variables, a spatial logistic regression model was created in R, using the `slrm` function (R package `spatstat`). R's `slrm` function requires the point data to be entered as a point process, and the covariate data to be input as image files that span the area of interest. The continuous explanatory variables (BEEF, MILK, SHEEP, WELLS, KARST, STREAM, and POVRATE) were converted to tiff files in ArcGIS using the feature to raster tool and imported into the R workspace. To incorporate the categorical land cover data into the model, land use

variables (PASTURE, DEVELOPED, and FOREST) were converted to a percent coverage index by reclassifying the 30X30m NLCD raster into three binary rasters -- one for each variable of interest. Next, each raster was resampled using the block statistics tool to a 500X500m cell and scaled arithmetically to return a percentage cover of the variable of interest in the new, larger cell. A risk map was generated in ArcGIS 10.0 from the model coefficients using the raster calculator tool and the equation (Baddeley et al. 2010)

$$P = \lfloor 1 + \exp(-\log a - BX) \rfloor^{-1}$$

where *P* is the risk for disease at any given raster cell, *log a* is an offset term representing the raster cell size, B is the vector of model coefficients and X is the vector of covariate values. The choice of raster cell size as the offset assumes homogeneous distribution of population and an alternate choice for offset could be the population density of the raster cell, similar to the use of the log transformed population as an offset in Poisson regression.

Each model in Table 4 has differing strengths and limitations, and each is appropriate under specific assumptions. The advantage of using multiple models to study a dataset is that when models are consistent in teasing out the relevant explanatory variables, and are consistent in the sign and magnitude of the coefficients, more confidence can be placed in the model results because one can be more confident that model selection is not masking the true relationship.

The models were compared for consistency, and the best model was selected by comparing model diagnostics such as the Coefficient of Determination (R-squared) and

Akaike's An Information Criterion (AIC) (Akaike 1974), and also by weighing the characteristics of the data with the appropriateness of the approach.

3.5 Regional models

3.5.1 Exploratory data analysis

Spearman correlation coefficients were calculated to identify explanatory variables that were significantly correlated to disease rates, and also to examine the relationship between explanatory variables. Explanatory variables that were highly correlated with each other were noted, but were retained in the model variables for the first round of modeling.

Scatterplots of disease rates and explanatory variables were generated in SPSS and examined to identify nonlinear relationships not evident from examination of the correlation coefficients. In some cases it was necessary to remove outliers and reexamine the relationship between and among the variables. One outlier each in BEEF, MILK, CALF, HOG, DEVELOPED and POVRATE, and two outliers in SHEEP were discarded. Because some zip code areas were outliers in more than one explanatory variable, only five records were discarded to remove the eight outliers. An additional eight zip code areas with zero population (typically associated with university campuses or large commercial or industrial works) were also removed, leaving n=167 records.

Disease rates were plotted using GeoDa (Anselin, Syabri and Kho 2006) and the local Moran's *I* and Getis gi statistics were computed to examine spatial clustering of areas of high and low disease.

3.5.2 Model development

Working with each disease dataset separately, histograms of disease rates and log-transformed rates were examined. Both datasets contain a high proportion of zero rates associated with areas where no disease was reported. These excess zeros can be problematic because model bias can be introduced when the modeling distribution does not well fit the data. When this occurs, unrealistically small standard errors can be reported, which can result in selection of unimportant explanatory variables, and overconfidence in model fit (Potts and Elith 2006). While the focus of this research is on the identification of risk factors for disease, and less so on generating a prediction model, it is nonetheless important to develop the best possible models by addressing the characteristics of the data and the nature of the relationship between dependent and independent variables. Specifically, this involves 1) the use of appropriate methods to handle the distribution of the dependent variable, 2) handling spatial autocorrelation in the dataset, and 3) ensuring that the model captures the relationship between dependent and independent variable (linear or otherwise). These will be discussed in the following paragraphs.

To address the distribution of the dependent variable, multiple model distributions were used (Poisson on the count data with an offet of log(population), and

Gamma and Tweedie using the rate data as a dependent variable). These distributions were selected because the Poisson distribution is traditionally used to model counts and rate data, the Gamma distribution appeared to be a reasonable fit to the skewed rate data, and the Tweedie showed promise as a way to handle the excess zeroes in the dataset.

Spatial lag models were developed using GeoDa to handle spatial autocorrelation. While the spatial lag and spatial error models are linear models and assume normally distributed data, they can provide insight into the nature of the spatial relationship between risk factors and disease.

To allow both linear and nonlinear relationships to be modeled, linear regression models were developed in R and GeoDa as discussed above, and nonlinear GAMs were developed in R with Gamma, Tweedie, and Poisson distributions and a log link function. The GAM models nonlinear relationships as a linear combination of smooth functions of each explanatory variable, and can also incorporate spatial relationships though the use of a spatial smoothing term, a smoothed function of the Cartesian coordinates of the centroid of each zip code area. The smoothing function s(x,y) is an isotropic thin plate regression spline which adds a smooth spatial structure from the model residuals to the fit (Bivand et al. 2008) (p. 180 and 297).

The models discussed above, including their characteristics and the platforms under which they were developed, are summarized in Table 5. The GLM and GAM employed a logarithmic link function, and so the rates data were increased by adding 1

to each rate prior to modeling, to avoid problems associated with calculating the log of zero.

All models were run on both the complete dataset plus the dataset with outliers removed to enable comparison of the effect of outlier removal. In addition, all models were run using a non-zero partition of the dataset to examine the effect of overdispersion.

3.5.3 Model development using non-zero rates partition

In an effort to isolate the effect of the zero rate areas, the outlier-removed dataset was partitioned into zero and non-zero rate areas, and additional models were developed on the non-zero rates only.

First, exploratory analyses were completed in SPSS: Spearman correlation coefficients were calculated, and scatterplots were examined. Second, all models outlined in Table 5 were developed as described in the previous section.

Table 5. Regional (ecological) regression models and platforms.

Model	Distribution	Spatial	Linear/non-linear	Platform
GLM	Gamma with log link	No	Linear	SPSS
GLM	Tweedie with log link	No	Linear	SPSS
Spatial lag	Normal	Yes	Linear	GeoDa
GAM	Gamma with log link	Yes	Non-linear	R
GAM	Tweedie with log link	Yes	Non-linear	R
GAM	Poisson with log link, offset=log(population)	Yes	Non-linear	R

CHAPTER 4 CRYPTOSPORIDIOSIS RESULTS

4.1 Individual models

4.1.1 Exploratory analysis

Kolmogorov-Smirnov (KS) test results for cryptosporidiosis case and control data indicate that the distributions of the all of the agricultural population variables except hog population density differ between the case and control data. Additionally, the distributions of well permit density, distance to karst geology, distance to impaired streams, and poverty rate differ between the case and control data. Mann-Whitney (MW) tests indicate no difference in the median rank of the presence and absence data for sheep and hog population density (Table 6).

The KS and the MW tests produce conflicting results for sheep population density because they use different test statistics; the KS test computes the discrepancy between the cumulative distributions of the two datasets, in this sense the KS test is sensitive to differences in the shape of the distribution (i.e. skewness and kurtosis) as well as the median. The MW test is a mean rank test and does not examine the shape of the distribution, but the value of the mean ranked entry. It is possible to have two distributions with a similar mean rank, but different shapes, which is the case for sheep population density (Figure 28).

Table 6. Results (p-values) of a comparison of distributions of cryptosporidiosis case-control data using the Mann-Whitney U Test and Kolmogorov-Smirnov Test. H_0 = the distributions are the same.

Explanatory variable	**Mann-Whitney U Test**	**Kolmogorov-Smirnov Test**
Beef cow population density (BEEF)	.000	.000
Milk cow population density (MILK)	.022	.001
Calf population density (CALF)	.000	.000
Cow population density (COW)	.000	.000
Hog population density (HOG)	.067	.109
Sheep population density (SHEEP)	.198	.014
Agricultural animal population density (AGDEN)	.000	.000
Well permit density (WELLS)	.000	.000
Distance to karst geology (KARST)	.000	.000
Distance to impaired stream (STREAM)	.000	.000
Poverty rate (POVRATE)	.000	.000

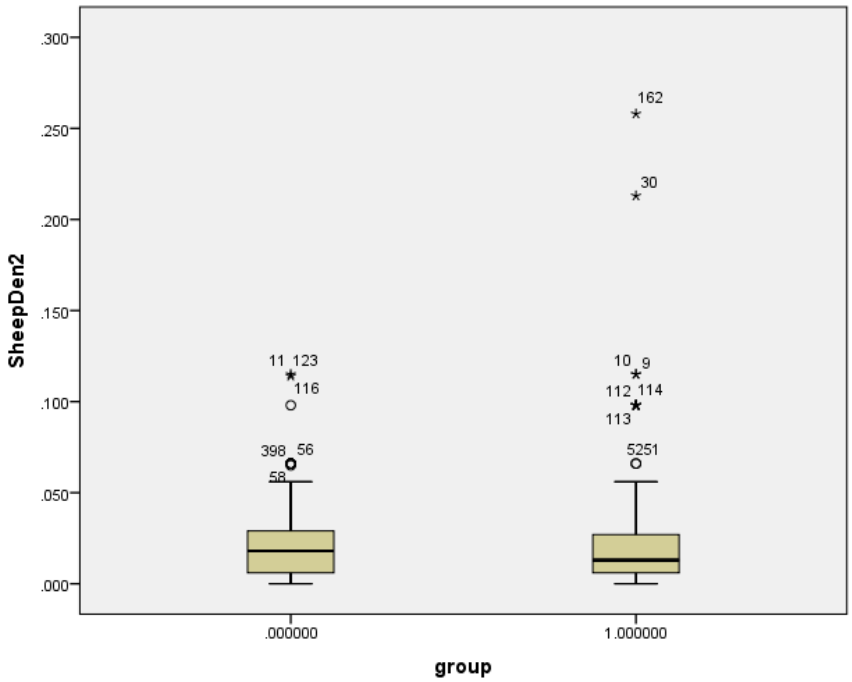


Figure 28. Box plots of case (Group 1 - right) and control (Group 0 - left) data for sheep population density (SheepDen2).

Chi-squared test results to compare distributions of the binary land use classifications of agricultural land (PASTURE = 1), developed land (DEVELOPED=1) or forested land (FOREST=1) showed that all land use variables for the presence data were different from the absence data ($p<0.001$).

Spearman correlation (Table 7) is high between all cattle variables (beef and milk cow density, calf density, total cow density), and because the total cow density and total agricultural animal density can be expressed as a linear combination of the other agricultural covariates, they were discarded to avoid multicollinearity. Similarly calf population density was discarded because of its near perfect correlation with beef cow population density. Hog population density was discarded because the case and control

distributions were not significantly different from each other (Table 6). Therefore of the agricultural census variables, only beef cow, dairy cow, and sheep population density (BEEF, MILK, and SHEEP) were selected for inclusion in the point models for cryptosporidiosis.

The explanatory variables selected for inclusion in preliminary cryptosporidiosis models were: beef and milk cow population density; sheep population density; the land use variables for developed land, pasture land and forest; distance to karst geology; distance to impaired stream; well permit density; and poverty rate (Table **8**).

Table 7. Spearman correlation coefficients for cryptosporidiosis explanatory variables (n=496). Significant correlations are marked with a double asterisk, p-values are provided in parentheses, and correlations above 0.7 are shaded.

	BEEF	MILK	CALF	COW	SHEEP	AGDEN	POVRATE	KARST	STREAM	WELLS
BEEF	1 .	.675** (.000)	.996** (.000)	.998** (.000)	.428** (.000)	.997** (.000)	-0.083 (.064)	-.355** (.000)	-.283** (.000)	-0.058 (.198)
MILK	.675** (.000)	1 .	.683** (.000)	.692** (.000)	.349** (.000)	.687** (.000)	.100* (.026)	-.178** (.000)	-0.069 (.124)	0.026 (.570)
CALF	.996** (.000)	.683** (.000)	1 .	.999** (.000)	.428** (.000)	.998** (.000)	-0.081 (.071)	-.359** (.000)	-.289** (.000)	-0.055 (.218)
COW	.998** (.000)	.692** (.000)	.999** (.000)	1 .	.423** (.000)	.999** (.000)	-0.081 (.073)	-.358** (.000)	-.286** (.000)	-0.059 (.189)
SHEEP	.428** (.000)	.349** (.000)	.428** (.000)	.423** (.000)	1 .	.440** (.000)	-0.074 (.097)	-0.085 (.057)	-.154** (.001)	-.154** (.001)
AGDEN	.997** (.000)	.687** (.000)	.998** (.000)	.999** (.000)	.440** (.000)	1 .	-0.087 (.052)	-.359** (.000)	-.289** (.000)	-0.06 (.183)
POVRATE	-0.083 (.064)	.100* (.026)	-0.081 (.071)	-0.081 (.073)	-0.074 (.097)	-0.087 (.052)	1 .	.205** (.000)	.246** (.000)	.155** (.000)
KARST	-.355** (.000)	-.178** (.000)	-.359** (.000)	-.358** (.000)	-0.085 (.057)	-.359** (.000)	.205** (.000)	1 .	.360** (.000)	.171** (.000)
STREAM	-0.05 (.285)	-0.069 (.124)	-0.055 (.22)	-.286** (.000)	-.154** (.001)	-.289** (.000)	.246** (.000)	.360** (.000)	1 .	.224** (.000)
WELLS	-0.058 (.198)	0.026 (.570)	.126** (.005)	-0.059 (.189)	-.154** (.001)	-0.06 (.183)	.155** (.000)	.171** (.000)	.224** (.000)	1 .

Table 8. Descriptive statistics for explanatory variables selected for inclusion in cryptosporidiosis models (statistics are given for case data only, n=246).

Variable	Units	Minimum	Maximum	Median
BEEF	Animals/km²	0.034	1.579	0.651
MILK	Animals/km²	0	0.093	0.017
SHEEP	Animals/km²	0	0.258	0.013
POVRATE	Percent	0	0.595	0.141
KARST	km	0	8.901	0.000
STREAM	km	0.03	49.34	3.859
WELLS	Permits/10,000 population	40.18	1215	115.44
DEVELOPED	0/1	0	1	1
FOREST	0/1	0	1	0
PASTURE	0/1	0	1	0

4.1.2 Individual regression models for cryptosporidiosis

Forward stepwise logistic regression retained the land use variables DEVELOPED and PASTURE, plus beef cow population density (BEEF), well permit density (WELLS), and distance to karst geology (KARST) ($R^2 = 0.437$) (Table 9). The logistic regression model predicted 82.5% correct classification for presence and 80.4% correct classification for absence, with an overall model score of 81.5% correct classification and a Nagelkerke R-Square of 0.552.

Similar to the explanatory variables retained in the logistic regression model, the GLM retained the land use variables for developed land and agricultural land

Table 9. Individual model coefficients (p-values follow in parentheses) and diagnostics for cryptosporidiosis models (not including spatial logistic regression models). Shaded cells represent excluded variables and diagnostic results that were unavailable.

Model Variable	**Logistic forward stepwise (SPSS)**	**GLM binomial (Logistic) (R)**	**GAM binomial (R) spatial smoothing**
Constant	3.291 (.000)	-0.592 (.309)	-0.312 (.638)
BEEF	0.872 (.010)	1.008 (.024	0.530 (.327)
MILK	(.298)	-5.164 (.466)	-3.417 (.683)
SHEEP	(.862)	-2.098 (.664)	2.933 (.588)
DEVELOPED	3.139 (.000)	2.676 (.000)	2.309 (.000)
PASTURE	1.474 (.000)	1.083 (.026)	0.747 (.136)
FOREST	(.216)	-0.580 (.245)	-0.910 (.085)
WELLS	-0.002 (.000)	-0.0022 (.000)	-0.0023 (.000)
KARST	-0.154 (.022)	-0.137 (.042)	-0.167 (.041)
STREAM	(.247)	-0.012 (.298)	0.0056 (.813)
POVRATE	(.464)	-0.843 (.484)	-0.373 (.782)
spatial term			(.029)
Model Diagnostics			
R^2	0.552		
Adj. R^2			0.496
AIC	442.575	440.56	425.97

(DEVELOPED and PASTURE, respectively), plus beef cow population density (BEEF), well permit density (WELLS), and distance to karst geology (KARST).

The GAM (adjusted $R^2 = 0.496$) retained the land use variable DEVELOPED, plus the well density (WELLS) and distance to karst geology (KARST) variables. A comparison of Akaike's AIC for the models generated to this point (Table 9), shows that the GAM emerges as the best model from the first group (which does not include the spatial logistic regression models).

Multiple spatial logistic regression models were generated in R using the `slrm` function, and different combinations of covariates and interaction terms (Table 10). Model C1 includes all candidate covariates (as listed previously in Table 8), C2 includes only those variables significant in model C1, and C3 includes all variables significant in C1 plus all possible interaction terms. Interactions were examined but were not statistically significant ($p>0.05$). Model results were generally consistent between models, and model C1 was the best model, using Akaike's AIC as a diagnostic tool. The risk factors for cryptosporidiosis identified in model C1 are beef cow population density (BEEF), the land use variables DEVELOPED and PASTURE, well density (WELLS), distance to karst geology (KARST) and distance to *E. coli* impaired stream (STREAM).

Table 10. Spatial logistic regression model (SLRM) coefficients (p-values in parentheses) for cryptosporidiosis individual models. Shaded cells represent excluded variables. Model diagnostics (AIC) are shown on the bottom row.

Covariates	**SLRM coefficients (p-values)**		
	C1	**C2**	**C3**
(Intercept)	-18.58868	-19.28357	-19.54150
BEEF	0.1899209 (.000)	0.415818 (.000)	0.384648 (.000)
MILK	9.700272 (.599)		
SHEEP	-6.439952 (.973)		
DEVELOPED	0.02983753 (.000)	0.034451 (.000)	0.041139 (.000)
PASTURE	0.01080318 (.000)	0.016177 (.000)	0.012557 (.000)
FOREST	-0.0055259 (.761)		
WELLS	-0.001641 (.000)	-0.001450 (.000)	-0.000690 (.000)
KARST	-0.0001505 (.000)	-0.0001545 (.000)	-0.000518 (.000)
STREAM	-0.0000237 (.004)	-0.0000198 (.014)	-0.0000029 (.014)
POVRATE	-0.3994621 (.538)		
Model Diagnostics			
AIC	8957.298	8961.791	8968.142

4.1.3 Comparison of cryptosporidiosis individual models

When selecting a modeling approach, it is important to consider the structure of the data (which will inform the modeling method based on model assumptions), simplicity, and diagnostic parameters which may include the Coefficient of Determination (R-squared), Akaike's AIC or others.

Compared with other approaches outlined in Table 9, the SLRM is a more suitable approach because it takes into account the characteristics of the dataset (binary outcome, nonlinear relationships, and spatial autocorrelation). The simplest model (model C1), which included all explanatory variables, was the best SLRM; successive refinements to C1 by exclusion of non-significant variables (C2) and inclusion of interaction factors (C3) reduced model performance using AIC as the diagnostic tool. Given the order of the AIC, the range of 8957 to 8968 between the three models is not large, indicating minimal additional loss of information with increased model complexity.

The risk factors that were most significant in predicting cryptosporidiosis presence were beef cow population density, residence within developed or agricultural land, proximity to karst areas, proximity to *E. coli* impaired streams and a lower number of well permits by population.

A map showing background environmental risk generated using model C1 (Figure 29) with an offset of log(raster cell size) shows elevated risk (as probability of disease) along developed corridors, in urban centers, and in one predominantly rural

area (Greene County area). Because cryptosporidiosis is a rare disease the probabilities are low, but differences in high versus low risk are apparent. A map such as this can be used to develop the framework for a more focused epidemiological study, allowing limited resources to be used more effectively.

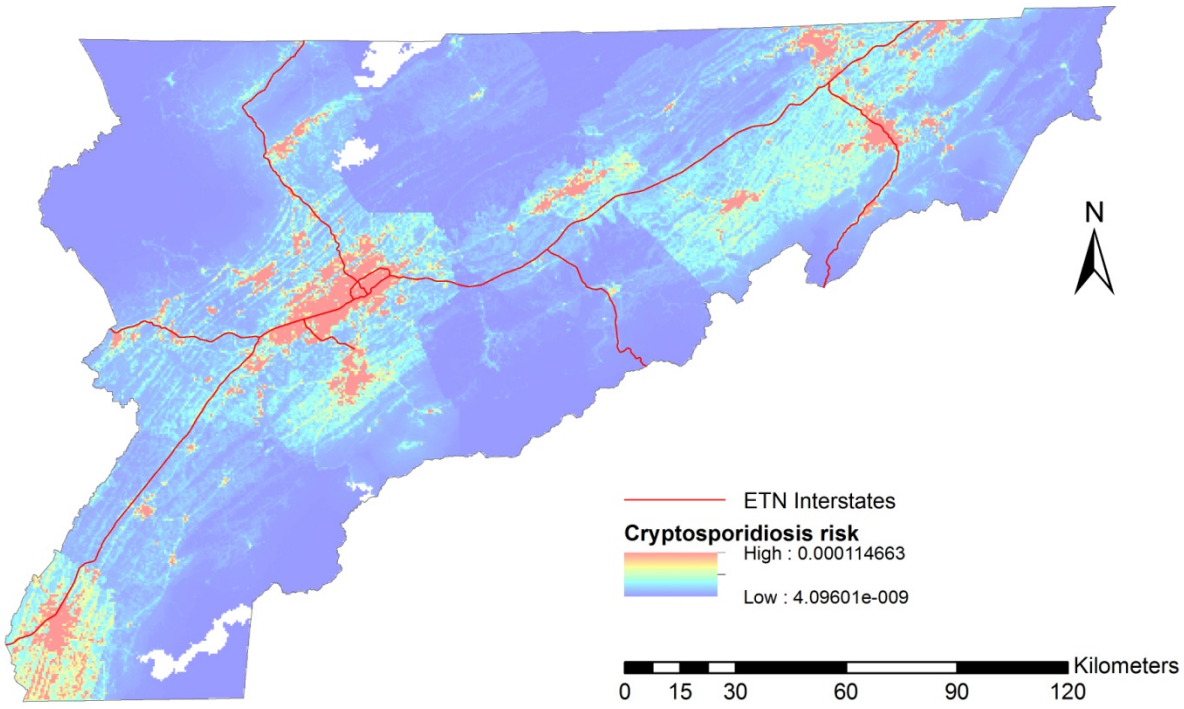


Figure 29. Spatial logistic regression model (C1) risk map for cryptosporidiosis (see Table 10 for coefficients).

4.2 Regional models

4.2.1 Exploratory analysis

Spearman correlation coefficients (Table 11) quantify the relationship between the individual candidate explanatory variables and 1) disease rate, and 2) each other. Because cow population and agricultural animal population can be expressed as linear combinations of the populations of beef cows, milk cows, calves, sheep and hogs, the explanatory variables for cow population density (COW) and agricultural animal population density (AGDEN) were discarded to avoid problems with multicollinearity. Although beef cow and calf population density (BEEF and CALF, respectively) were correlated with each other and with agricultural land use (PASTURE), they were retained in the models because it was preliminary to preferentially retain one variable over the other without understanding the portion of the variability that could be explained by each.

Scatterplots of the agricultural explanatory variables with (Figure 30) and without (Figure 31) outliers show that after removal of outliers, the points are more scattered. A positive relationship between cryptosporidiosis rate and the beef cow (BEEF) and milk cow (MILK) population density looks likely, and there is no real indication of nonlinearity. Removal of outliers had minimal impact on Spearman correlation coefficients.

Table 11. Spearman correlation of cryptosporidiosis standardized rates and explanatory variables. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01. P-values are in parentheses. Shaded cells indicate correlations above r=0.7.

	CRYPTO_RT	BEEF	MILK	CALF	HOG	SHEEP	DEVELOPED	FOREST	PASTURE	KARST_PCT	STREAM_PCT	POVRATE	WELLS
CRYPTO_RT	1.000	.341** (.000)	.240** (.001)	.364** (.000)	.206** (.006)	.152* (.042)	.296** (.000)	-.328** (.000)	.339** (.000)	.348** (.000)	.323** (.000)	-.176* (.018)	-.172* (.021)
BEEF	.341** (.000)	1.000	.545** (.000)	.967** (.000)	.398** (.000)	.490** (.000)	.181* (.015)	-.407** (.000)	.811** (.000)	.443** (.000)	.342** (.000)	-.315** (.000)	-.063 (.400)
MILK	.240** (.001)	.545** (.000)	1.000	.585** (.000)	.298** (.000)	.399** (.000)	-.091 (.223)	-.154* (.039)	.594** (.000)	.232** (.002)	.225** (.002)	-.077 (.302)	.055 (.462)
CALF	.364** (.000)	.967** (.000)	.585** (.000)	1.000	.408** (.000)	.521** (.000)	.207** (.005)	-.424** (.000)	.815** (.000)	.475** (.000)	.372** (.000)	-.328** (.000)	-.092 (.217)
HOG	.206** (.006)	.398** (.000)	.298** (.000)	.408** (.000)	1.000	.344** (.000)	.081 (.279)	-.098 (.192)	.398** (.000)	.263** (.000)	.193** (.009)	-.145 (.052)	.002 (.983)
SHEEP	.152* (.042)	.490** (.000)	.399** (.000)	.521** (.000)	.344** (.000)	1.000	.084 (.260)	-.145 (.053)	.566** (.000)	.298** (.000)	.311** (.000)	-.214** (.004)	.027 (.715)
DEVELOPED	.296** (.000)	.181* (.015)	-.091 (.223)	.207** (.005)	.081 (.279)	.084 (.260)	1.000	-.394** (.000)	.157* (.035)	.432** (.000)	.223** (.003)	-.292** (.000)	-.388** (.000)
FOREST	-.328** (.000)	-.407** (.000)	-.154* (.039)	-.424** (.000)	-.098 (.192)	-.145 (.053)	-.394** (.000)	1.000	-.394** (.000)	-.641** (.000)	-.220** (.003)	.326** (.000)	.407** (.000)
PASTURE	.339** (.000)	.811** (.000)	.594** (.000)	.815** (.000)	.398** (.000)	.566** (.000)	.157* (.035)	-.394** (.000)	1.000	.412** (.000)	.344** (.000)	-.327** (.000)	.002 (.978)
KARST_PCT	.348** (.000)	.443** (.000)	.232** (.002)	.475** (.000)	.263** (.000)	.298** (.000)	.432** (.000)	-.641** (.000)	.412** (.000)	1.000	.296** (.000)	-.466** (.000)	-.147* (.048)
STREAM_PCT	.323** (.000)	.342** (.000)	.225** (.002)	.372** (.000)	.193** (.009)	.311** (.000)	.223** (.003)	-.220** (.003)	.344** (.000)	.296** (.000)	1.000	-.145 (.052)	-.044 (.556)
POVRATE	-.176* (.018)	-.315** (.000)	-.077 (.302)	-.328** (.000)	-.145 (.052)	-.214** (.004)	-.292** (.000)	.326** (.000)	-.327** (.000)	-.466** (.000)	-.145 (.052)	1.000	.241** (.001)
WELLS	-.172* (.021)	-.063 (.400)	.055 (.462)	-.092 (.217)	.002 (.983)	.027 (.715)	-.388** (.000)	.407** (.000)	.002 (.978)	-.147* (.048)	-.044 (.556)	.241** (.001)	1.000

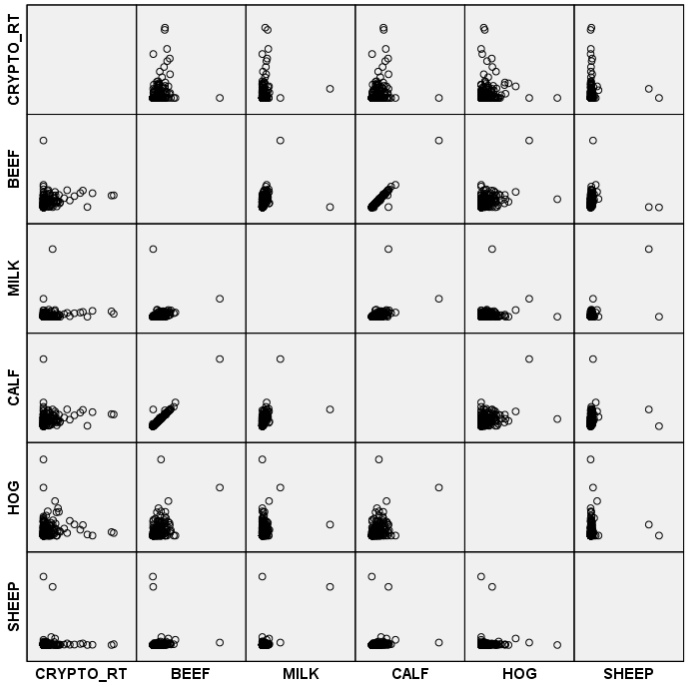


Figure 30. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and agricultural animal population density (beef cow, milk cow, calf, hog, and sheep).

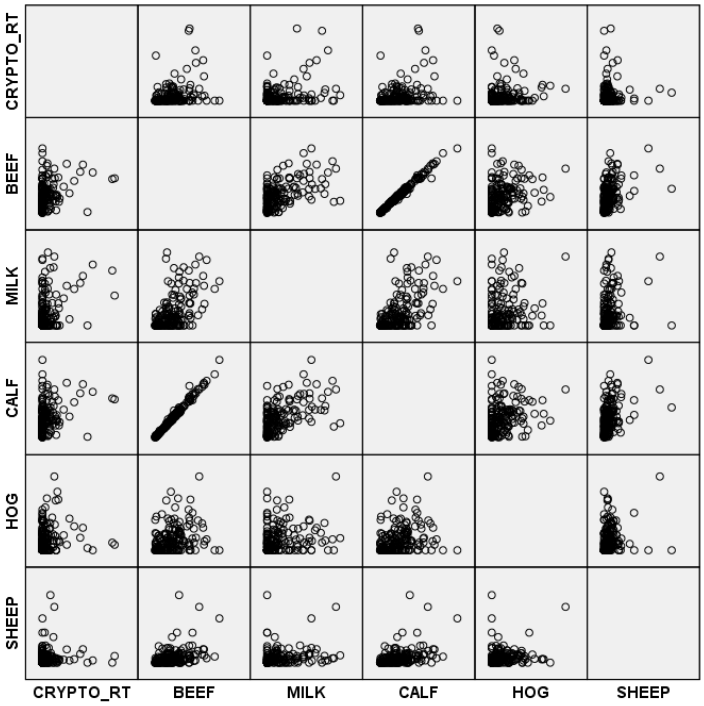


Figure 31. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and agricultural animal population density (beef cow, milk cow, calf, hog and sheep) after removal of outliers.

Similarly, outlier-removed scatterplots of cryptosporidiosis rate with land use variables for developed land (DEVELOPED), agricultural land (PASTURE), and forested land (FOREST) (Figure 32), and cryptosporidiosis rate with percent karst geology (KARST_PCT), percent impaired streams (STREAM_PCT), poverty rate (POVRATE), and well permit density (WELLS) (Figure 33) shed little additional insight into the relationship between disease and the explanatory variables.

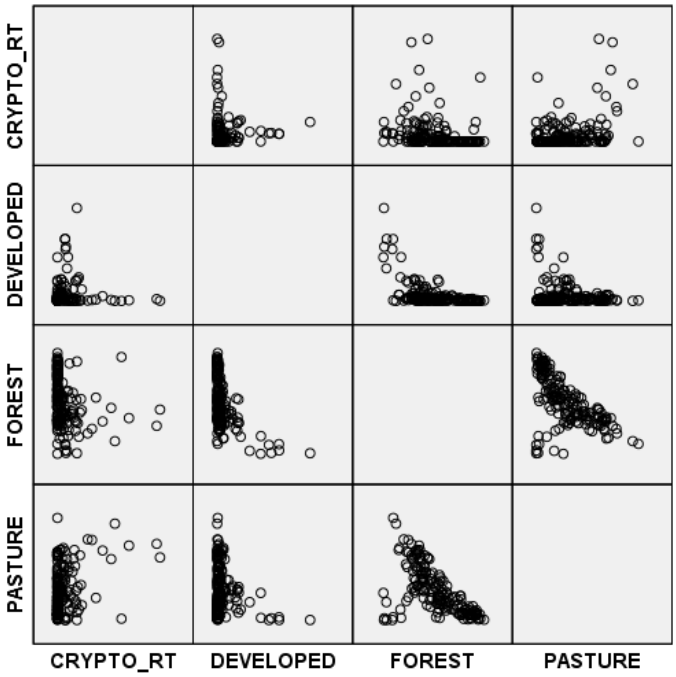


Figure 32. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and the percent of land area classified as developed, forested, and pasture land (outliers removed).

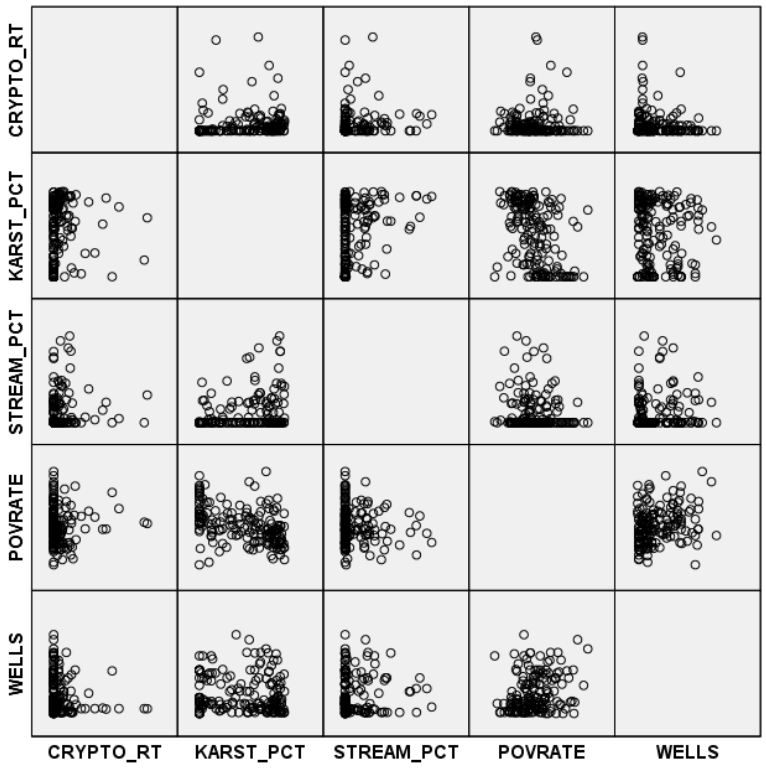


Figure 33. Scattergram of cryptosporidiosis rate (CRYPTO_RT) and geology, water quality, poverty and well permit variables (outliers removed).

Candidate variables included, therefore: population densities of beef cows (BEEF), milk cows (MILK), calves (CALF), hogs (HOG), and sheep (SHEEP); land use as a percentage of developed land (DEVELOPED), forested land (FOREST), and agricultural land (PASTURE); percent of bedrock classified as karst geology (KARST_PCT); percent of streams classified as impaired for the presence of *E. coli* (STREAM_PCT); poverty rate (POVRATE); and well permit density by county population (WELLS).

Using a distance based spatial weights matrix (distance = 40 km, which allows each zip code area to have at least one neighbor), LISA (local Moran's $I = 0.1363$, estimated p=0.004) and Getis Gi cluster maps suggest three clusters. In the northeast

corner and western edge of the study area there are low-low clusters (areas of low disease rate are spatially proximate). A high-high cluster exists in the east-central part of the study area, where zip code areas with high disease rate are spatially near each other (Figure 34 and Figure 35). The distance-based spatial weight was used in lieu of rook or queen contiguity because of the differing sizes of the zip code areas between urban and rural environments.

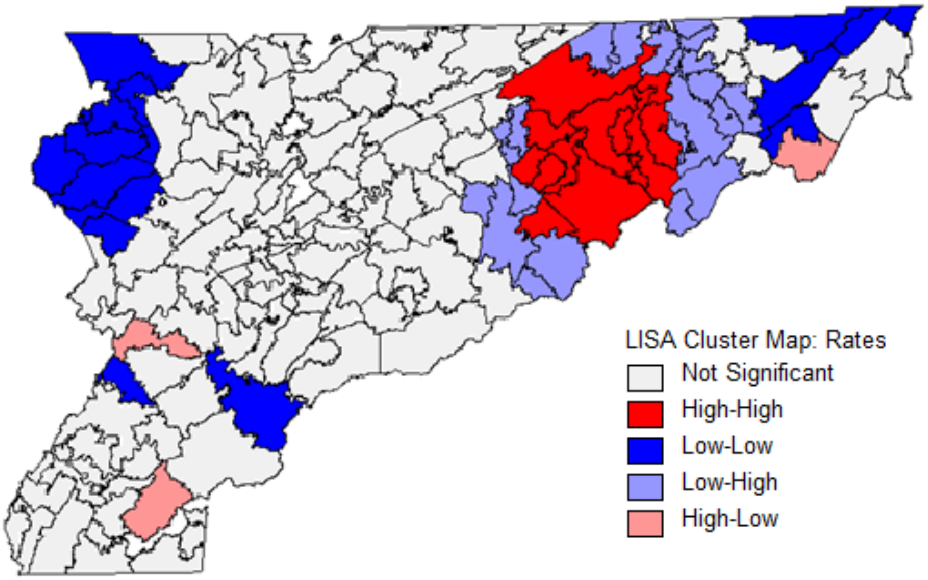


Figure 34. Map of cryptosporidiosis rates showing Local Indicators of Spatial Autocorrelation (LISA) generated using GeoDa software. Outliers have been removed.

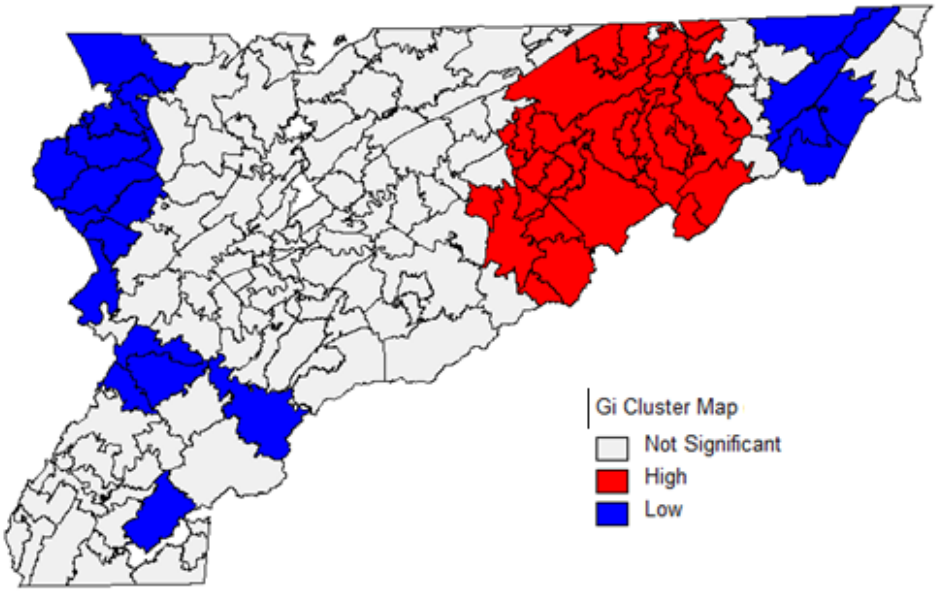


Figure 35. Getis Gi cluster map of cryptosporidiosis rate data. The map was generated in GeoDa using a 40 km distance spatial weights matrix.

Histograms of the distribution of standardized rates and log-transformed rates show marked skewness and zero-inflation (Figure 36). Log transformation of the rate plus 1 does not remove the skewness because of the large number of zeros in the dataset.

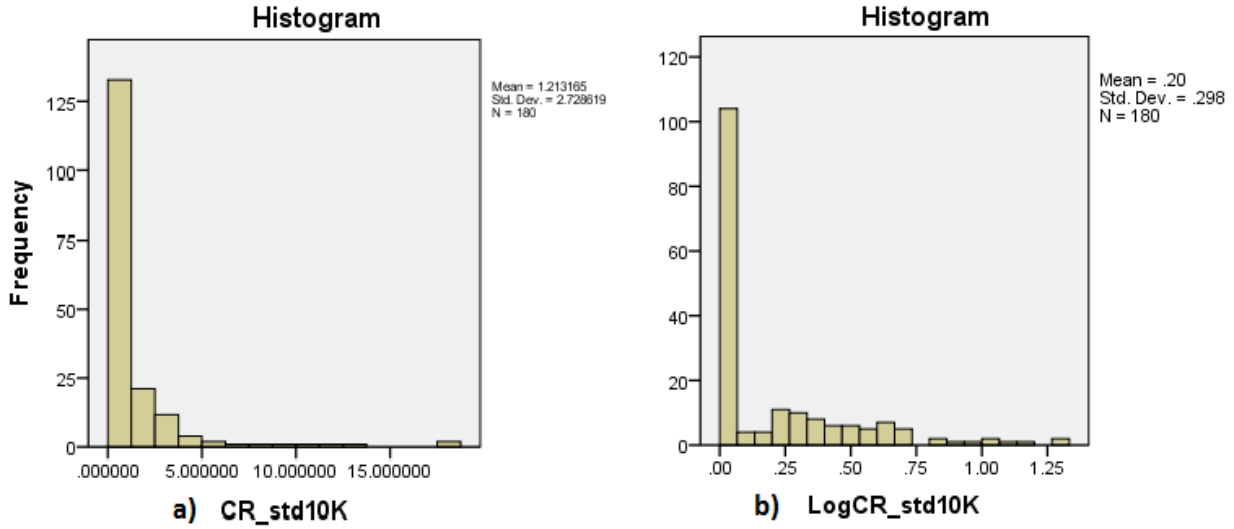


Figure 36. Histogram of disease rates at the zip code level showing (a) standardized cryptosporidiosis rates per 10,000 people (CR_St10K) and (b) log-transformed rates per 10,000 people (log(CR_std10K)).

4.2.2 Regional regression model results for cryptosporidiosis

The dataset was subjected to three treatments, and models were developed for each treatment. First, the models listed in Table 5 were developed for the full dataset (all zip codes, n=180). Second, the models were developed for the full dataset with outliers and zip codes with zero population removed (n=167). Third, the zip code areas where disease was absent (rate equal to zero) were discarded, and the models were developed on the zip codes with non-zero rates areas only (n=75). Only the Poisson-distributed GAMs included an offset term for the log of population. The following paragraphs and summary table (Table 12) describe these modeling results.

Model results using complete dataset A GLM with a Gamma distribution and log link function (AIC = 584) retained the agricultural land use variable PASTURE as a significant explanatory variable (p=0.05), and milk cow population density (MILK), percent developed land (DEVELOPED) and well permit density (WELLS) as borderline significant (using p=0.1). PASTURE, MILK, and DEVELOPED had positive coefficients, while the coefficient for WELLS was negative. The GLM with a Tweedie distribution (AIC=640) retained milk cow density (MILK), percent agricultural land (PASTURE), and well permit density (WELLS) as significant variables with positive coefficients at p=0.05 (the coefficient for WELLS was near zero). Calf population density (CALF) was marginally significant (p=0.1) with a negative coefficient.

Table 12. A comparison of coefficients (p-values) for cryptosporidiosis regional models and data partitions. Smoothed variables in GAMs have multiple coefficients and are not reported.

	GLM Gamma	GLM Tweedie	Spatial Lag	GAM Gamma	GAM Tweedie	GAM Poisson
All data (n=180)						
BEEF	1.251 (.213)	1.667 (.111)	3.212 (.364)	-0.596 (.515)	-0.596 (.515)	-0.145 (.914)
MILK	3.566 (.057)	4.522 (.020)	10.387 (.116)	-0.927 (.587)	-0.927 (.587)	(.797)
CALF	-1.19 (.188)	-1.569 (.099)	-3.055 (.341)	0.567 (.490)	0.567 (.490)	0.159 (.896)
HOG	-0.825 (.696)	-1.152 (.593)	-3.426 (.643)	0.950 (.624)	0.950 (.624)	(.781)
SHEEP	-0.371 (.187)	-0.446 (.217)	-1.070 (.325)	-0.033 (.901)	-0.033 (.901)	-0.360 (.463)
DEVELOPED	0.021 (.090)	0.013 (.205)	0.019 (.579)	0.023 (.108)	0.023 (.108)	(.000)
FOREST	0.004 (0.262)	0.002 (.500)	0.001 (.876)	(.103)	(.103)	-0.034 (.000)
PASTURE	0.020 (.000)	0.020 (.000)	0.046 (.004)	0.006 (.353)	0.006 (.353)	(.000)
KARST_PCT	0.001 (.505)	0.000 (.883)	-0.002 (.736)	0.002 (.300)	0.002 (.300)	-0.001 (.831)
STREAM_PCT	0.004 (.229)	0.004 (.224)	0.002 (.831)	-0.0002 (.945)	-0.0002 (.945)	0.001 (.721)
POVRATE	1.092 (.123)	0.928 (.173)	1.955 (.394)	(.136)	(.136)	(.801)
WELLS	0.000 (.053)	0.000 (.024)	-0.001 (.112)	-0.0003 (.204)	-0.0003 (.204)	(.468)
Outliers removed (n=167)						
BEEF	0.143 (.913)	-0.034 (.979)	-1.221 (.777)	-1.055 (.328)	-1.055 (.328)	-0.754 (.605)
MILK	3.755 (.286)	5.030 (.114)	21.972 (.054)	0.759 (.824)	0.759 (.824)	(.149)
CALF	0.019 (.987)	0.187 (.866)	1.451 (.700)	0.887 (.334)	0.887 (.334)	(.515)
HOG	2.348 (.394)	2.198 (.425)	3.238 (.757)	2.197 (.411)	2.197 (.411)	4.568 (.107)
SHEEP	-1.300 (.386)	-1.893 (.249)	-7.668 (.232)	(.331)	(.331)	0.200 (.925)
DEVELOPED	0.060 (.023)	0.052 (.034)	0.140 (.130)	(.552)	(.552)	-0.018 (.494)
FOREST	0.008 (.260)	0.007 (.383)	0.020 (.466)	(.091)	(.091)	-0.011 (.234)
PASTURE	0.022 (.013)	0.021 (.015)	0.055 (.077)	0.005 (.671)	0.005 (.671)	-0.018 (.107)
KARST_PCT	0.002 (.406)	0.001 (.651)	-0.001 (.910)	(.200)	(.200)	0.002 (.600)
STREAM_PCT	0.002 (.536)	0.002 (.556)	-0.002 (.879)	-0.001 (.729)	-0.001 (.729)	-0.001 (.750)
POVRATE	2.493 (.003)	2.640 (.005)	4.719 (.141)	2.471 (.018)	2.471 (.018)	(.396)
WELLS	0.000 (.082)	0.000 (.040)	-0.001 (.197)	-0.0004 (.149)	-0.0004 (.149)	-0.0003 (.404)
Non-zero partition (n=75)						
BEEF	-0.187 (.917)	-0.312 (.861)	-4.796 (.597)	-0.562 (.649)	-0.562 (.649)	0.123 (.946)
MILK	3.005 (.478)	3.549 (.373)	16.014 (.421)	3.117 (.402)	3.117 (.402)	(.306)
CALF	0.578 (.728)	0.706 (.671)	5.305 (.531)	0.686 (.554)	0.686 (.554)	0.168 (.920)
HOG	0.156 (.959)	-0.943 (.753)	-8.932 (.572)	-1.502 (.570)	-1.502 (.570)	2.408 (.472)
SHEEP	-0.290 (.857)	-1.371 (.421)	-12.069 (.183)	(.570)	(.570)	-1.462 (.570)
DEVELOPED	-0.001 (.986)	0.006 (.838)	0.073 (.631)	(.903)	(.903)	-0.016 (.607)
FOREST	0.001 (.894)	0.004 (.686)	0.031 (.555)	(.086)	(.086)	-0.005 (.669)
PASTURE	-0.001 (.897)	0.003 (.820)	0.040 (.505)	0.005 (.632)	0.005 (.632)	-0.014 (.316)
KARST_PCT	-0.006 (.061)	-0.005 (.104)	-0.016 (.388)	-0.006 (.146)	-0.006 (.146)	-0.003 (.560)
STREAM_PCT	0.005 (.164)	0.004 (.214)	0.010 (.564)	0.001 (.771)	0.001 (.771)	0.0005 (.907)
POVRATE	3.876 (.005)	4.040 (.005)	12.388 (.101)	(.021)	(.021)	(.335)
WELLS	-0.001 (.005)	-0.001 (.005)	-0.003 (.053)	-0.0001 (.808)	-0.0001 (.808)	(.316)

The spatial lag model retained PASTURE only, with R^2=0.21, and the spatial autoregressive parameter (distance based spatial weights at 40 km) was significant, suggesting that the spatial lag model can capture some of the spatial variability in the dependent variable that can be explained by the independent variables.

The R output for the GAM displays the estimated degrees of freedom (edf) for each smoothed variable, which can serve as a diagnostic to identify whether the smooth function of a variable is linear (edf = 1) or nonlinear (edf > 1). In this sense, the GAM can provide more information about the relationship between the covariates and the response variable. Explanatory variables with edf>1 were included as smoothed variables, and multiple coefficients were generated for these variables. Only the single coefficients for the un-smoothed variables are reported in Table 12.

For the Gamma and Tweedie models, only the spatial smoothing term s(long, lat) was significant at p=0.05 although the smoothed land use variables DEVELOPED (p=0.108) and FOREST (p=0.103) were borderline significant at p=0.1 (Figure 37). The Coefficient of Determination for the Gamma and Tweedie models (R^2 = 0.431) is suspiciously high, given the lack of significant explanatory variables, indicating that the model may be misspecified.

For the Poisson model, all three land use parameters (DEVELOPED, FOREST and PASTURE) were retained (p<0.05), with forested land associated with reduced risk for disease (negative coefficient). Differences between the Gamma, Tweedie, and Poisson models stem from differences in how the distributions are defined.

```
> gamcr1 <-gam(CR_std10Kplus1 ~ 1 + s(long,lat)+ s(BEEF) + s(MILK) + s(CALF) + s(HOG)
+ s(SHEEP) + s(DEVELOPED) + s(FOREST) + s(PASTURE) + s(KARST_PCT) + s(STREAM_PCT) +
s(POVRATE) + s(WELLS), data=rates, family=Gamma(link=log))
> summary(gamcr1)

Family: Gamma Link function: log

Formula: CR_std10Kplus1 ~ 1 + s(long, lat) + s(BEEF) + s(MILK) + s(CALF) +
    s(HOG) + s(SHEEP) + s(DEVELOPED) + s(FOREST) + s(PASTURE) +
    s(KARST_PCT) + s(STREAM_PCT) + s(POVRATE) + s(WELLS)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57834    0.04019   14.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df     F  p-value
s(long,lat)  21.450 25.785 3.002 1.37e-05 ***
s(BEEF)       1.000  1.000 0.425    0.515
s(MILK)       1.000  1.000 0.296    0.587
s(CALF)       1.000  1.000 0.480    0.490
s(HOG)        1.000  1.000 0.241    0.624
s(SHEEP)      1.000  1.000 0.016    0.901
s(DEVELOPED)  1.000  1.000 2.619    0.108
s(FOREST)     4.763  5.796 1.813    0.103
s(PASTURE)    1.000  1.000 0.867    0.353
s(KARST_PCT)  1.000  1.000 1.082    0.300
s(STREAM_PCT) 1.000  1.000 0.005    0.945
s(POVRATE)    3.159  3.967 1.784    0.136
s(WELLS)      1.000  1.000 1.629    0.204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.431   Deviance explained = 65.7%
GCV score = 0.37483  Scale est. = 0.29076   n = 180

> AIC(gamcr1)
[1] 499.8599
```

Figure 37. R output for regional GAM using cryptosporidiosis rate complete dataset (n=180). R code is printed in red and preceded by the “>” character.

Model results using dataset with outliers removed To investigate whether the presence of outliers interfered with model fit, the models were run again after removal of outliers and zip code areas with zero population (n=167). Spearman correlations were minimally affected, and scatterplots became more clear as the cluster of points near the

origin were more dispersed after removal of the outliers (see Figure 31, Figure 32, and Figure 33).

The GLM with Gamma distribution and log link function (AIC = 549) retained two land use variables, DEVELOPED and PASTURE, and poverty rate (POVRATE), all with positive coefficients and significant at p=0.05. Well permit density (WELLS) was marginally significant at p=0.1, ant its coefficient was reported as 0, due to truncation, however because the values of this variable are large, it can contribute significantly to explaining the variability in the data, even though the coefficient is small. Substitution of the Tweedie distribution for the Gamma distribution in the GLM did not change the general results, although WELLS became significant at p=0.05 and the AIC increased to 599.

Milk cow population density (MILK) and the agricultural land use variable PASTURE were significant at p=0.1 in the spatial lag model, and both were positively associated with disease.

The GAM with Gamma distribution and log link function retained poverty rate (POVRATE) with a positive coefficient at p=0.05 and the smoothed land use variable s(FOREST) at p=0.1. Substitution of the Gamma distribution with a Tweedie distribution in the GAM produced no change in output, and there were no differences in model diagnostics between the two GAMs (R^2=0.424, AIC=477). The GAM with Poisson distribution and log link function retained only the land use variable PASTURE at p=0.1.

Model results for non-zero partition of rates dataset Spearman correlation coefficients for the non-zero partition (n=75) indicate that after removal of zip code areas with cryptosporidiosis rate equal to zero, only milk cow population density (MILK, r=0.253), percent developed land use (DEVELOPED, r=-0.270), and poverty rate (POVRATE, r=0.253) are significantly correlated to disease, and these candidate variables are not strongly correlated with each other (Table 13). All variables were retained as candidates in the model because although the Spearman correlation indicates no significant linear relationship, a non-linear relationship (even a weak one) cannot be ruled out.

An examination of scatterplots (Figure 38, Figure 39, and Figure 40) shows the potential for a linear or nonlinear relationship between explanatory variables and the dependent variable, and so both linear (GLM and spatial lag) and nonlinear (GAM) models were developed.

Table 13. Spearman correlation for non-zero partition of cryptosporidiosis standardized rates and explanatory variables. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01 (p-values are in parentheses). Shaded cells indicate correlations above r=0.7.

	CRYPTO_RT	BEEF	MILK	CALF	HOG	SHEEP	DEVELOPED	FOREST	PASTURE	KARST_PCT	STREAM_PCT	POVRATE	WELLS
CRYPTO_RT	1	.178 (.126)	.247* (.032)	.180 (.122)	.099 (.400)	-.047 (.692)	-0.275* (.017)	-.109 (.353)	.080 (.493)	-.165 (.156)	.196 (.092)	.267* (.020)	-.107 (.361)
BEEF	.178 (.126)	1	.544** (.000)	.991** (.000)	.266* (.021)	.355** (.002)	-.359** (.002)	-.365** (.001)	.781** (.000)	.130 (.266)	.158 (.175)	.009 (.936)	.031 (.794)
MILK	.247* (.032)	.544** (.000)	1	.567** (.000)	.289* (.012)	.335** (.003)	-.444** (.000)	-.092 (.430)	.651** (.000)	-.153 (.191)	.196 (.092)	.346** (.002)	.112 (.339)
CALF	.180 (.122)	.991** (.000)	.567** (.000)	1	.264* (.022)	.382** (.001)	-.368** (.001)	-.372** (.001)	.791** (.000)	.135 (.248)	.167 (.152)	.016 (.890)	.027 (.815)
HOG	.099 (.400)	.266* (.021)	.289* (.012)	.264* (.022)	1	.218 (.060)	-.121 (.302)	-.121 (.300)	.361** (.001)	.133 (.254)	.060 (.610)	-.033 (.781)	-.052 (.660)
SHEEP	-.047 (.692)	.355** (.002)	.335** (.003)	.382** (.001)	.218 (.060)	1	-0.246* (.033)	.004 (.972)	.437** (.000)	.071 (.547)	.186 (.110)	-.035 (.766)	.064 (.585)
DEVELOPED	-0.275* (.017)	-.359** (.002)	-.444** (.000)	-.368** (.001)	-.121 (.302)	-0.246* (.033)	1	-.459** (.000)	-.487** (.000)	.433** (.000)	.002 (.998)	-.504** (.000)	-.598** (.000)
FOREST	-.109 (.353)	-.365** (.001)	-.092 (.430)	-.372** (.001)	-.121 (.300)	.004 (.972)	-.459** (.000)	1	-.212 (.067)	-.505** (.000)	-.177 (.128)	.387** (.001)	.595** (.000)
PASTURE	.080 (.493)	.781** (.000)	.651** (.000)	.791** (.000)	.361** (.001)	.437** (.000)	-.487** (.000)	-.212 (.067)	1	-.095 (.418)	.069 (.557)	.037 (.756)	.132 (.258)
KARST_PCT	-.165 (.156)	.130 (.266)	-.153 (.191)	.135 (.248)	.133 (.254)	.071 (.547)	.433** (.000)	-.505** (.000)	-.095 (.418)	1	.171 (.418)	-.474** (.000)	-0.243* (.036)
STREAM_PCT	.196 (.092)	.158 (.175)	.196 (.092)	.167 (.152)	.060 (.610)	.186 (.110)	.002 (.998)	-.177 (.128)	.069 (.557)	.171 (.418)	1	-.139 (.236)	-.180 (.123)
POVRATE	.267* (.020)	.009 (.936)	.346** (.002)	.016 (.890)	-.033 (.781)	-.035 (.766)	-.504** (.000)	.387** (.001)	.037 (.756)	-.474** (.000)	-.139 (.236)	1	.448** (.000)
WELLS	-.107 (.361)	.031 (.794)	.112 (.339)	.027 (.815)	-.052 (.660)	.064 (.585)	-.598** (.000)	.595** (.000)	.132 (.258)	-0.243* (.036)	-.180 (.123)	.448** (.000)	1

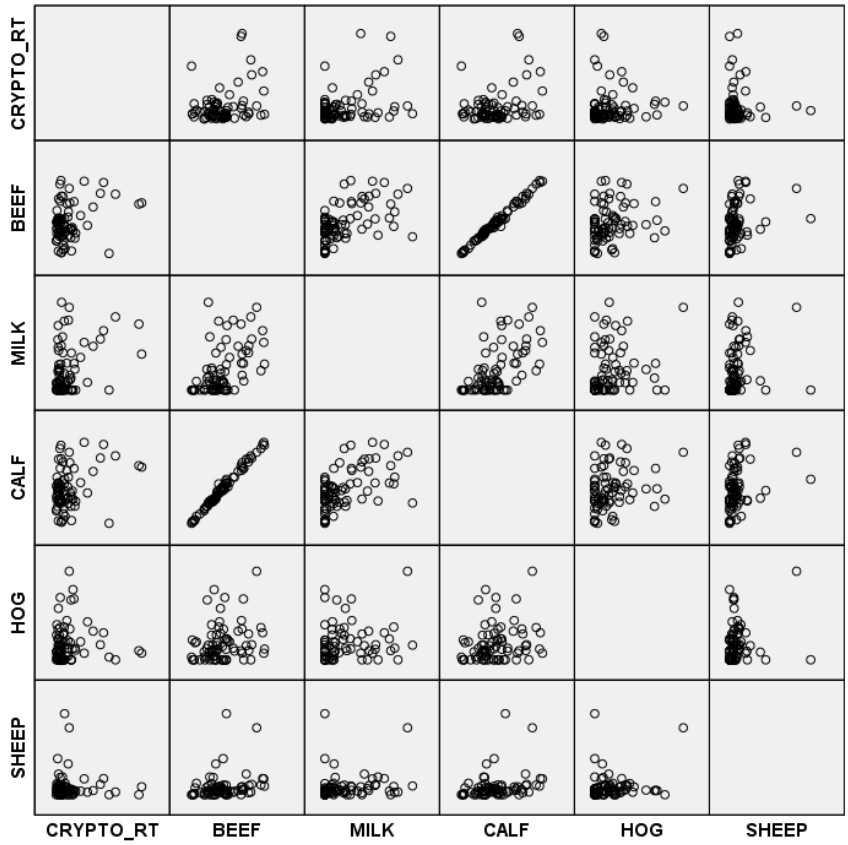


Figure 38. Scattergram of non-zero cryptosporidiosis rates and agricultural variables (n=75).

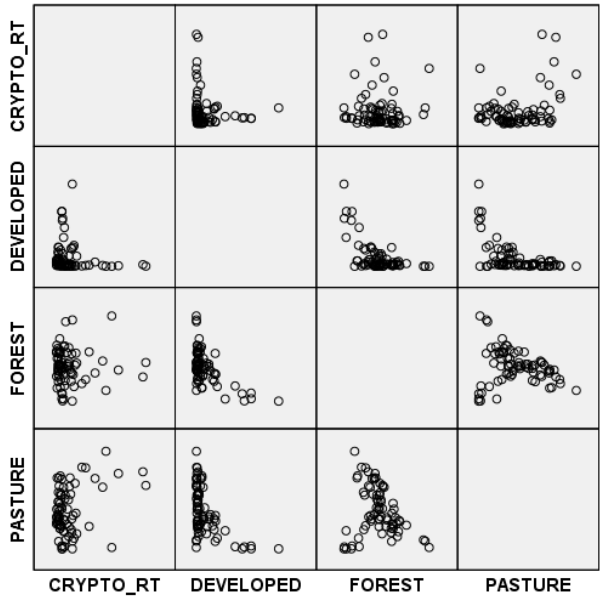


Figure 39. Scattergram of non-zero cryptosporidiosis rates and land use variables (n=75)

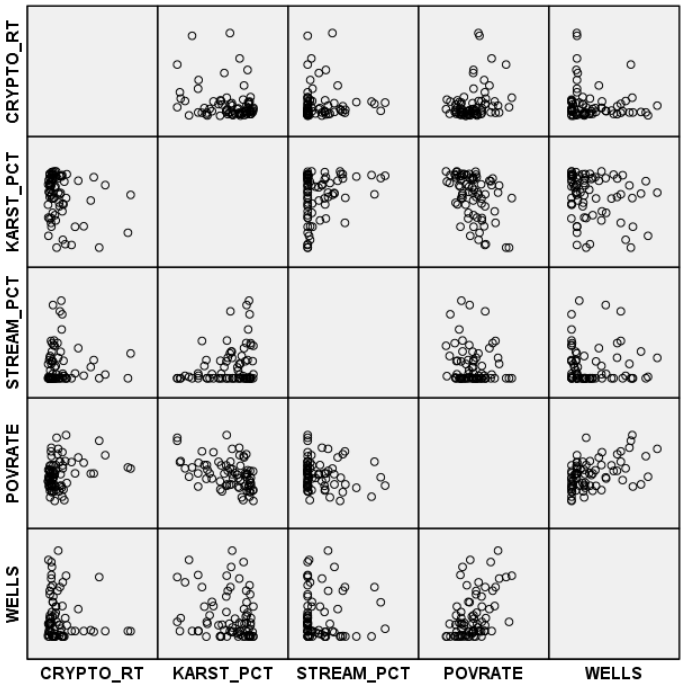


Figure 40. Scattergram of non-zero cryptosporidiosis rates and geology, water quality, poverty and well permit variables (n=75)

The GLM with Gamma distribution and log link retained poverty rate (POVRATE) (positive coefficient) and well permit density (WELLS) (negative coefficient) at p=0.05, and percent karst geology (KARST_PCT) (negative coefficient) at p=0.1. Repeating the GLM with a Tweedie distribution returned the same results, except percent karst geology (KARST_PCT) was marginally significant (p=0.104) at the more permissive p=0.1 threshold.

The spatial lag model positively associated poverty rate (POVRATE) and negatively associated well permit density (WELLS) with cryptosporidiosis in the

population, using a significance level of p=0.1. No explanatory variables were significant at p=0.05.

The GAM with Gamma distribution and log link retained the smoothed covariates poverty rate s(POVRATE) (at p=0.05) and percent forest land cover s(FOREST) (at p=0.1). As with the previous data treatments, the GAM with a Tweedie distribution retained the same variables that were retained with the Gamma distribution. No variables were retained in the Poisson model.

4.2.3 Comparison of cryptosporidiosis regional models and treatments

None of the aggregate models were good predictive models, suggesting that aggregating disease data to the zip code level did not capture the local scale variability that may be important in a predictive model of background environmental risk. Even though a good predictive model could not be developed, identification of important risk factors for cryptosporidiosis can be made by examining the explanatory variables that were consistently retained in the models. Table 14 displays a tally of the number of instances in which a variable was retained in the models. Important risk factors were chosen as those retained in at least six of the eighteen models because when a variable is consistently significant between models and treatments, we have more confidence that the variable is truly significant and not an artifact of the model or selected treatment. The important risk factors for cryptosporidiosis in the population are percent developed land (DEVELOPED), forested land (FOREST), and agricultural land (PASTURE), and the ZCTA poverty rate (POVRATE) and county well permit density (WELLS).

The land use variables DEVELOPED and PASTURE were positively associated with disease, indicating that an increased percentage of developed land and agricultural land use in a zip code area was a risk factor for cryptosporidiosis. Similarly, poverty rate (POVRATE) was positively associated with disease in the population.

The coefficients for both the percentage of forested land (FOREST) and well permit density (WELLS) were negative, indicating a protective effect of forest land use and availability of private water sources is a protective factor for cryptosporidiosis.

Table 14. Tally of explanatory variable significance in cryptosporidiosis aggregate models.

	All data (n=180)	**Outliers removed (n=167)**	**Non-zero partition (n=75)**	**Total (out of a possible 18)**
BEEF	0	0	0	0
MILK	2	1	0	3
CALF	1	0	0	1
HOG	0	0	0	0
SHEEP	0	0	0	0
DEVELOPED	4	2	0	6
FOREST	3	2	2	7
PASTURE	4	4	0	8
KARST_PCT	0	0	2	2
STREAM_PCT	0	0	0	0
POVRATE	0	4	5	9
WELLS	2	2	3	8

4.3 Chapter summary

In this chapter, two groups of models were generated for the cryptosporidiosis case data and rate data. The rate data were partitioned using three treatments (all data, outliers removed, and zero-rate areas removed), and multiple models were developed for each partition. The purpose of model development was to identify significant risk factors for cryptosporidiosis in the individual (using the case data) and in the population (using the rate data).

The spatial logistic regression model positively associated beef cow population density (BEEF), percent developed (DEVELOPED) and percent agricultural land (PASTURE) with cryptosporidiosis in the individual. The distance from karst geology (KARST) and impaired surface waters (STREAM) were negatively associated with cryptosporidiosis, indicating that distance from these features had a protective effect, reducing risk for disease in the individual. The number of well permits normalized by county population (WELLS) was negatively associated with cryptosporidiosis, suggesting that availability of private water supplies is also associated with a protective effect at the individual level.

Important risk factors for cryptosporidiosis in the population were identified by examining which explanatory variables were consistently statistically significant in the different modeling approaches. The risk factors positively associated with cryptosporidiosis in the population are percent developed land (DEVELOPED), percent agricultural land (PASTURE), and poverty rate (POVRATE). The risk factors associated

with decreased risk for cryptosporidiosis in the population are percent forested land (FOREST) and well permit density (WELLS).

These results are summarized in Table 15 and their implications will be discussed in Chapter 6.

Table 15. Summary of statistically significant risk factors for cryptosporidiosis at the individual and regional scales. "+" indicates positive association with disease, "-" indicates negative association with disease.

Risk factor	Individual model	Regional model
Agricultural animals	+Beef cow population density	
Land use	+Percent developed land in cell +Percent pasture land in cell	+Percent developed land in zip +Percent pasture land in zip +Percent forested land in zip
Geology	+Proximity to limestone bedrock	
Surface water	+Proximity to *E. coli* impaired stream	
Poverty rate		+Poverty rate in zip
Well permits	-Well permits by county population	-Well permits by county population

CHAPTER 5 *E. COLI* O157 RESULTS

5.1 Individual models

5.1.1 Exploratory analysis

Kolmogorov-Smirnov (KS) tests to compare the distributions of case and control data indicate that the distributions of the following agricultural variables differ: population densities of beef cows (BEEF), milk cows (MILK), calves (CALF), total cows (COW), hogs (HOG), and total agricultural animals (AGDEN)). In addition, the distributions of distance to karst geology (KARST), distance to impaired stream (STREAM), poverty rate (POVRATE) and well permit density (WELLS) differ between the case and control data. For milk cow population density (MILK), the Mann-Whitney and Kolmogorov-Smirnov test results did not agree (Table 16). These results are not unreasonable because the two tests compare different characteristics of the distributions; the Mann-Whitney U-test compares the mean rank for each group, while the Kolmogorov-Smirnov test compares the shape of the distribution (i.e. the higher order moments that describe skewness and kurtosis).

A comparison of box plots for MILK for the case (Group 1) and control (Group 0) data highlights these differences in the shape of the distributions. In general, the case variables have wider and more skewed distributions than the control variables, even though the mean ranks are close (Figure 41).

Table 16. Results (p-values) of a comparison of distributions of *E. coli* O157 case and control data using the Mann-Whitney U Test and Kolmogorov-Smirnov Test. H_0 = the distributions are the same.

Explanatory variable	Mann-Whitney U Test	Kolmogorov-Smirnov Test
BEEF	.000	.000
MILK	.152	.002
CALF	.000	.000
COW	.000	.000
HOG	.001	.002
SHEEP	.173	.200
AGDEN	.000	.000
WELLS	.000	.000
KARST	.000	.000
STREAM	.000	.000
POVRATE	.000	.000

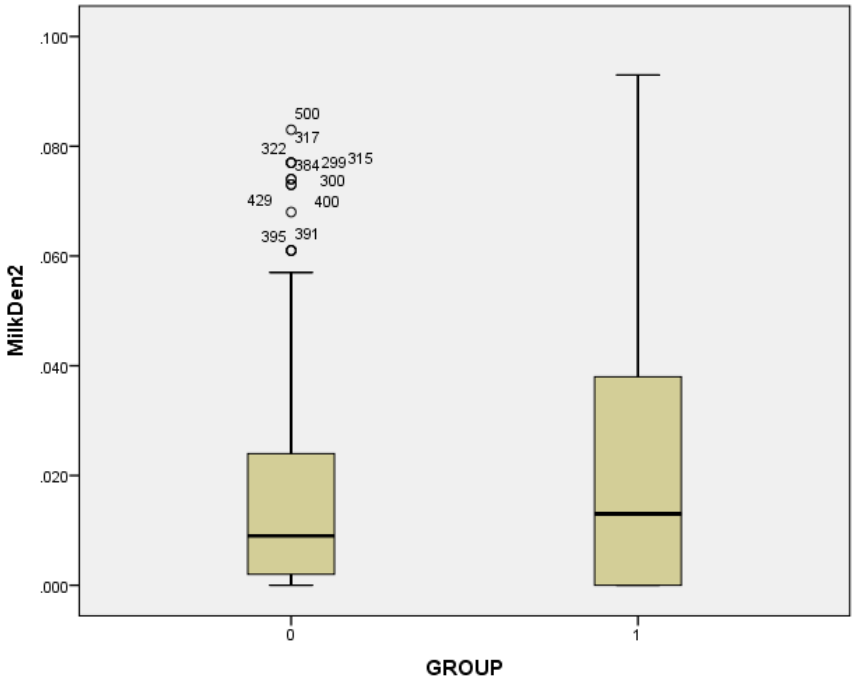


Figure 41. Box plots of *E. coli* O157 case (Group 1 - right) and control (Group 0 - left) data for milk cow population density.

Chi-squared test results to compare distributions of the binary land use variables PASTURE, DEVELOPED, and FOREST showed that all land use variables for the case data were different from the control data ($p<0.001$).

Spearman correlation coefficients were calculated for all variable pairs found to differ between case and control data (Table 17). Correlation between agricultural variable pairs was high, and because the total cow population density (COW) and agricultural population density (AGDEN) could be expressed as linear combinations of the other agricultural variables, they were discarded. Calf population density (CALF) was also discarded because it had a near perfect correlation with beef cow population density (BEEF).

After removal of highly correlated variables, the explanatory variables selected for inclusion in preliminary models were: agricultural variables beef cow, milk cow, and hog population density; land use variables percent developed, forested, and agricultural land; distance to karst geology; distance to impaired surface water; poverty rate; and well permit density (Table 18).

Table 17. Spearman correlation coefficients for explanatory variables (n=500). Significant correlations are marked with a single (p<0.05) or double (p<0.01) asterisk (2-tailed test), and p-values are in parentheses. Correlations greater than r=0.7 are shaded.

	BEEF	MILK	CALF	COW	SHEEP	AGDEN	POVRATE	KARST	STREAM	WELLS
BEEF	1	.631** (.000)	.993** (.000)	.997** (.000)	.365** (.000)	.517** (.000)	.997** (.000)	-.140** (.002)	-.377** (.000)	-.330** (.000)
MILK	.631** (.000)	1	.643** (.000)	.650** (.000)	.333** (.000)	.385** (.000)	.650** (.000)	0.059 (.185)	-.178** (.000)	-.146** (.001)
CALF	.993** (.000)	.643** (.000)	1	.998** (.000)	.355** (.000)	.522** (.000)	.997** (.000)	-.143** (.001)	-.374** (.000)	-.345** (.000)
COW	.997** (.000)	.650** (.000)	.998** (.000)	1	.361** (.000)	.522** (.000)	.999** (.000)	-.138** (.002)	-.377** (.000)	-.339** (.000)
SHEEP	.365** (.000)	.333** (.000)	.355** (.000)	.361** (.000)	1	.297** (.000)	.382** (.000)	-0.064 (.151)	-.118** (.008)	-.170** (.000)
AGDEN	.517** (.000)	.385** (.000)	.522** (.000)	.522** (.000)	.297** (.000)	1	.535** (.000)	-0.031 (.486)	-.152** (.001)	-.232** (.000)
POVRATE	.997** (.000)	.650** (.000)	.997** (.000)	.999** (.000)	.382** (.000)	.535** (.000)	1	-.141** (.002)	-.375** (.000)	-.341** (.000)
KARST	-.140** (.002)	0.059 (.185)	-.143** (.001)	-.138** (.002)	-0.064 (.151)	-0.031 (.486)	-.141** (.002)	1	.202** (.000)	.246** (.000)
STREAM	-.377** (.000)	-.178** (.000)	-.374** (.000)	-.377** (.000)	-.118** (.008)	-.152** (.001)	-.375** (.000)	.202** (.000)	1	.239** (.000)
WELLS	-.330** (.000)	-.146** (.001)	-.345** (.000)	-.339** (.000)	-.170** (.000)	-.232** (.000)	-.341** (.000)	.246** (.000)	.239** (.000)	1

Table 18. Descriptive statistics for variables selected for inclusion in *E. coli* models (statistics are given for case data only, n=250).

Variable	Units	Minimum	Maximum	Median
BEEF	Animals/km^2	0	1.904	0.700
MILK	Animals/km^2	0	0.093	0.013
HOG	Animals/km^2	0	0.119	0.017
DEVELOPED	0/1	0	1	1
FOREST	0/1	0	1	0
PASTURE	0/1	0	1	0
WELLS	Permits/10,000 population	0	1563	101.78
KARST	km	0	9.49	0
STREAM	km	0.024	61.40	4.534
POVRATE	Percent	0	0.853	0.132

5.1.2 Individual regression models for *E. coli* O157

Regression model results for the risk of *E. coli* O157 in the individual are summarized in Table 19. Forward stepwise logistic regression retained two categorical land use variables (DEVELOPED and PASTURE), plus beef cow population density (BEEF), distance to karst geology (KARST), and well permit density (WELLS). The logistic regression model predicted 82.2% correct classification (cases 81.2% and controls 83.2%), with a Nagelkerke R-Square of 0.540.

Table 19. Individual model coefficients (p-values follow in parentheses) and diagnostics for *E. coli* O157 models (does not include spatial logistic regression models). Shaded cells represent excluded variables and diagnostic results that were unavailable.

Model Variable	Logistic forward stepwise (SPSS)*	GLM binomial (Logistic) (R)	GAM binomial (R) spatial smoothing
Constant	3.080 (.000)	-1.0288 (.085)	-0.7616 (.313)
BEEF	1.078 (.001)	1.339 (.002)	0.7564 (.206)
MILK	(.158)	-9.337 (.173)	7.189 (.499)
HOG	(.134)	7.940 (.169)	6.210 (.379)
DEVELOPED	3.139 (.000)	2.508 (.000)	2.344 (.000)
PASTURE	1.505 (.000)	0.870 (.079)	0.9481 (.074)
FOREST	(.062)	-0.866 (.081)	-0.9631 (.065)
WELLS	-0.001 (.000)	-0.0013 (.000)	-0.00092 (.034)
KARST	-0.310 (.001)	-0.304 (.001)	-0.227 (.033)
STREAM	(.924)	0.00415 (.721)	-0.0345 (.247)
POVRATE	(.357)	-0.6136 (.616)	-0.219 (.887)
spatial term			(.0575)
Model Diagnostics			
R square	0.540		
Adj. R Square			0.513
AIC	455.434	447.8	432.18

*coefficients not reported for variables not retained in logistic regression model

The GLM performed similarly to the logistic model, retaining beef cow population density (BEEF), developed land use (DEVELOPED), distance to karst geology (KARST), and well permit density (WELLS).

The GAM (adjusted R^2=0.513) did not retain any agricultural variables, and developed land use (DEVELOPED) was positively associated with disease. Distance to karst geology (KARST) and well permit density were negatively associated with *E. coli* O157. There were minor differences in model coefficients when a quasibinomial distribution was substituted for the binomial in both the GLM and GAM (these results are not reported).

A comparison of model diagnostics (Table 19), which include R-squared values and Akaike's AIC values for the models generated to this point, shows that the GAM is the best model of the first group of models (which does not include the spatial logistic regression models (SLRM)).

SLRMs were generated in R for various combinations of covariates and interaction terms (Table 20). Results were generally consistent between models. The first SLRM (E1) retained all variables in Table 20, with no interaction terms. Model E2 included only the significant variables from E1 and a comparison of AIC shows E2 is a slight improvement over E1. Model E3 was constructed using only the variables in E2 plus all possible interaction terms. Only two interaction terms (DEVELOPED*PASTURE and HOG*KARST) emerged as significant, and the AIC increased due to increased complexity in the model.

The final model (E4) was the best SLRM (using the AIC as a diagnostic tool), and included only significant covariates from previous models. The risk factors for *E. coli* O157 identified in model E4 are: beef cow (BEEF), milk cow (MILK), and hog (HOG) population density; percent developed land (DEVELOPED); percent agricultural land (PASTURE); distance to karst geology (KARST); and well permit density (WELLS). The interaction terms HOG*KARST and DEVELOPED*PASTURE were also significant ($p<0.05$). Disease is positively associated with increasing values of BEEF, HOG, DEVELOPED and PASTURE, and negatively associated with increasing values of MILK, WELLS and KARST.

Table 20. Spatial logistic regression model coefficients (p-values in parentheses) for *E. coli* O157 individual models. Shaded cells represent excluded variables. Model diagnostics (AIC) are shown in the bottom row. Model E3 included all possible interaction terms (not shown in table).

Covariates	**SLRM coefficient (p-values)**			
	E1	**E2**	**E3**	**E4**
Constant	-19.43082	-19.27842	-19.56200	-19.14675
BEEF	0.3765216 (.000)	0.3863660 (.000)	1.633936 (.000)	0.3046392 (.000)
MILK	-2.96153 (.001)	-0.299347 (.001)	-20.07985 (.001)	-1.319051 (.001)
HOG	3.016985 (.023)	2.954141 (.023)	4.032882 (.023)	1.794563 (.023)
DEVELOPED	0.0322240 (.000)	0.0307588 (.000)	0.0320371 (.000)	0.0282152 (.000)
PASTURE	0.0166063 (.000)	0.0150557 (.000)	0.0093708 (.000)	0.0103302 (.000)
FOREST	0.0187083 (.737)			
WELLS	-0.00122612 (.000)	-0.00122976 (.000)	-0.000922645 (.000)	-0.00128114 (.000)
KARST	-0.000219501 (.000)	-0.000216085 (.000)	-0.000380838 (.000)	-0.000336388 (.000)
STREAM	0.0000023 (.721)			
POVRATE	0.184116 (.774)			
HOG*KARST			0.00698929 (.036)	0.00518194 (.064)
DEVELOPED* PASTURE			0.000405056 (.005)	0.000354069 (.003)
Model Diagnostics				
AIC	9190.348	9184.642	9194.744	9176.378

5.1.3 Comparison of *E. coli* O157 individual models

Similar to the cryptosporidiosis point models, the SLRM approach was the most suitable to use given the characteristics of the *E. coli* O157 point dataset (binary dependent variable, non-linear relationships, and spatial autocorrelation). Refinements to the base model E1 (Table 20) that removed non-significant explanatory values and included two interaction terms generated the best SLRM (E4), determined by a comparison of the AIC. The AIC ranged from 9194 in the most complex model that included all interaction terms to 9176 in the best model, a small range of only 18 units, indicating that there is little change in information lost from model to model.

A map of background environmental risk generated using model E4 (Figure 42) and an offset of log(raster cell size) shows elevated risk along developed corridors and in urban centers, in areas of agricultural land use, and in areas of karst. Like cryptosporidiosis, *E. coli* O157 is a rare disease and so the values that identify risk are very small, but differences between high and low risk areas are apparent. A map such as this one can be used to target a more detailed epidemiological study.

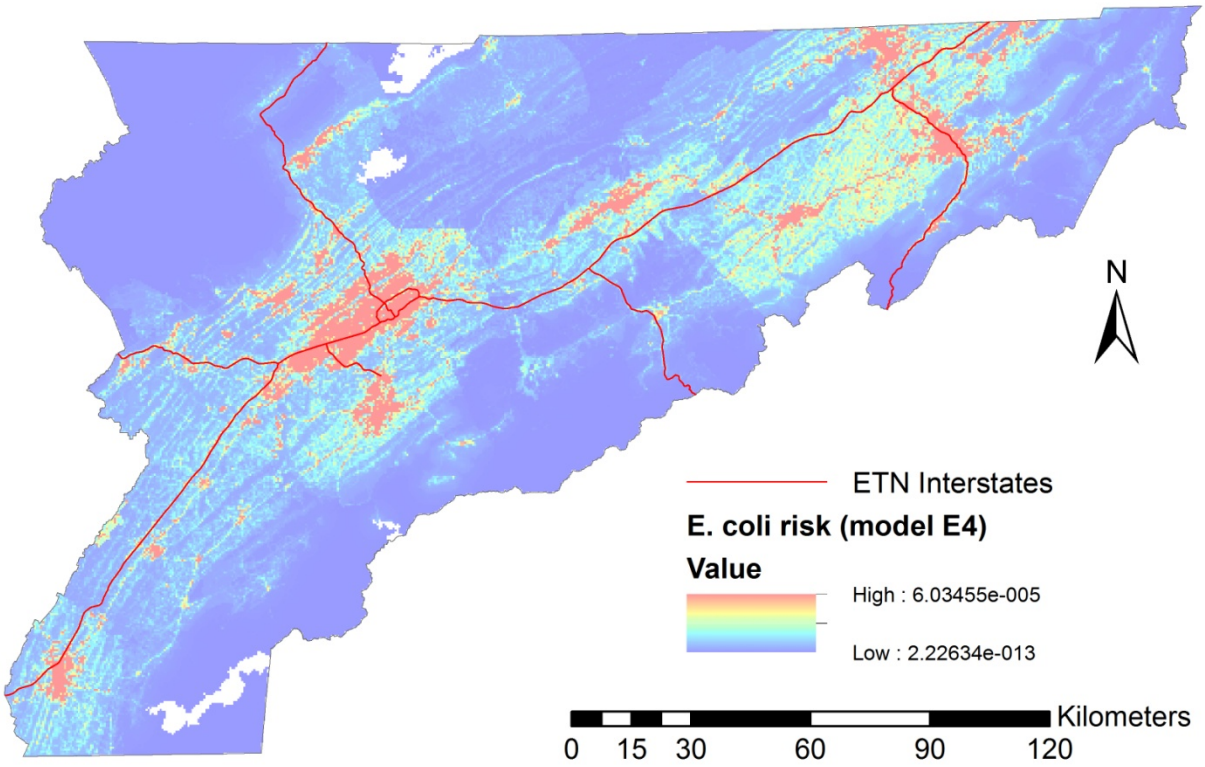


Figure 42. Spatial logistic regression model (E4) risk map for *E. coli* O157 (see Table 20 for coefficients).

5.2 Regional models

5.2.1 Exploratory analysis

Spearman correlation coefficients (Table 21) indicate that all explanatory variables are correlated with the *E. coli* O157 disease rate data (variable ECOLI_RT). Total cow population density (COW) and total agricultural population density (AGDEN) were removed because they can be expressed as a linear combination of the other agricultural variables, and would therefore introduce multicollinearity effects into the model, which is not desirable. After removal of these two variables, the explanatory variables selected for inclusion in preliminary models were: beef cow (BEEF), milk cow (MILK), calf (CALF), hog (HOG), and sheep (SHEEP) population densities; land use variables representing percentage of developed (DEVELOPED), forest (FOREST) and agricultural (PASTURE) land; percent karst geology (KARST_PCT); percent impaired surface streams (STREAM_PCT), poverty rate (POVRATE), and well permit density (WELLS). While the agricultural variables BEEF, CALF, and PASTURE, are highly correlated, they were retained in the models because it was preliminary to exclude one over the other.

Scatterplots of disease rates and explanatory variables are shown in the following pages (Figure 43, Figure 45, and Figure 47). These scatterplots include all variables, and outliers are visible for BEEF, MILK, CALF, HOG (2 outliers), SHEEP (2 outliers), DEVELOPED, and POVRATE.

Table 21. Spearman correlation of *E. coli* O157 standardized rates and explanatory variables. Significant correlations are marked with a single (p<0.05) or double (p<0.01) asterisk (two-tailed, n=180). Correlations above 0.7 are shaded, p-values are in parentheses.

	ECOLI_RT	BEEF	MILK	CALF	COW	HOG	SHEEP	AGDEN	DEVELOPED	FOREST	PASTURE	KARST_PCT	STREAM_PCT	POVRATE	WELLS
ECOLI_RT	1	.366** (.000)	.276** (.000)	.385** (.000)	.380** (.000)	.203** (.006)	.261** (.000)	.366** (.000)	.268** (.000)	-.281** (.000)	.423** (.000)	.325** (.000)	.433** (.000)	-.157* (.035)	-.160* (.032)
BEEF	.366** (.000)	1	.545** (.000)	.967** (.000)	.973** (.000)	.398** (.000)	.490** (.000)	.947** (.000)	.181* (.015)	-.407** (.000)	.811** (.000)	.443** (.000)	.342** (.000)	-.315** (.000)	-0.063 (.400)
MILK	.276** (.000)	.545** (.000)	1	.585** (.000)	.592** (.000)	.298** (.000)	.399** (.000)	.577** (.000)	-0.091 (.223)	-.154* (.039)	.594** (.000)	.232** (.002)	.225** (.002)	-0.077 (.302)	0.055 (.462)
CALF	.385** (.000)	.967** (.000)	.585** (.000)	1	.999** (.000)	.408** (.000)	.521** (.000)	.976** (.000)	.207** (.005)	-.424** (.000)	.815** (.000)	.475** (.000)	.372** (.000)	-.328** (.000)	-0.092 (.217)
COW	.380** (.000)	.973** (.000)	.592** (.000)	.999** (.000)	1	.412** (.000)	.519** (.000)	.977** (.000)	.202** (.007)	-.421** (.000)	.818** (.000)	.473** (.000)	.371** (.000)	-.326** (.000)	-0.086 (.251)
HOG	.203** (.006)	.398** (.000)	.298** (.000)	.408** (.000)	.412** (.000)	1	.344** (.000)	.412** (.000)	0.081 (.279)	-0.098 (.192)	.398** (.000)	.263** (.000)	.193** (.009)	-0.145 (.052)	0.002 (.983)
SHEEP	.261** (.000)	.490** (.000)	.399** (.000)	.521** (.000)	.519** (.000)	.344** (.000)	1	.565** (.000)	0.084 (.26)	-0.145 (.053)	.566** (.000)	.298** (.000)	.311** (.000)	-.214** (.004)	0.027 (.715)
AGDEN	.366** (.000)	.947** (.000)	.577** (.000)	.976** (.000)	.977** (.000)	.412** (.000)	.565** (.000)	1	.224** (.003)	-.448** (.000)	.828** (.000)	.504** (.000)	.362** (.000)	-.330** (.000)	-0.094 (.212)
DEVELOPED	.268** (.000)	.181* (.015)	-0.091 (.223)	.207** (.005)	.202** (.007)	0.081 (.279)	0.084 (.26)	.224** (.003)	1	-.394** (.000)	.157* (.035)	.432** (.000)	.223** (.003)	-.292** (.000)	-.388** (.000)
FOREST	-.281** (.000)	-.407** (.000)	-.154* (.039)	-.424** (.000)	-.421** (.000)	-0.098 (.192)	-0.145 (.053)	-.448** (.000)	-.394** (.000)	1	-.394** (.000)	-.641** (.000)	-.220** (.003)	.326** (.000)	.407** (.000)
PASTURE	.423** (.000)	.811** (.000)	.594** (.000)	.815** (.000)	.818** (.000)	.398** (.000)	.566** (.000)	.828** (.000)	.157* (.035)	-.394** (.000)	1	.412** (.000)	.344** (.000)	-.327** (.000)	0.002 (.978)
KARST_PCT	.325** (.000)	.443** (.000)	.232** (.002)	.475** (.000)	.473** (.000)	.263** (.000)	.298** (.000)	.504** (.000)	.432** (.000)	-.641** (.000)	.412** (.000)	1	.296** (.000)	-.466** (.000)	-.147* (.048)
STREAM_PCT	.433** (.000)	.342** (.000)	.225** (.002)	.372** (.000)	.371** (.000)	.193** (.009)	.311** (.000)	.362** (.000)	.223** (.003)	-.220** (.003)	.344** (.000)	.296** (.000)	1	-0.145 (.052)	-0.044 (.556)
POVRATE	-.157* (.035)	-.315** (.000)	-0.077 (.302)	-.328** (.000)	-.326** (.000)	-0.145 (.052)	-.214** (.004)	-.330** (.000)	-.292** (.000)	.326** (.000)	-.327** (.000)	-.466** (.000)	-0.145 (.052)	1	.241** (.001)
WELLS	-.160* (.032)	-0.063 (.400)	0.055 (.462)	-0.092 (.217)	-0.086 (.251)	0.002 (.983)	0.027 (.715)	-0.094 (.212)	-.388** (.000)	.407** (.000)	0.002 (.978)	-.147* (.048)	-0.044 (.556)	.241** (.001)	1

Removal of outliers provides a greater spread of points in the scatterplot, and consequently makes a trend easier to identify (if present). Note that some zip code areas were outliers for multiple variables, and so removal of this one area resulted in the removal of multiple outliers. In total, nine outliers were removed, corresponding to five zip code areas. Scatterplots of the outlier-removed data are shown in Figure 44, Figure 46, and Figure 48.

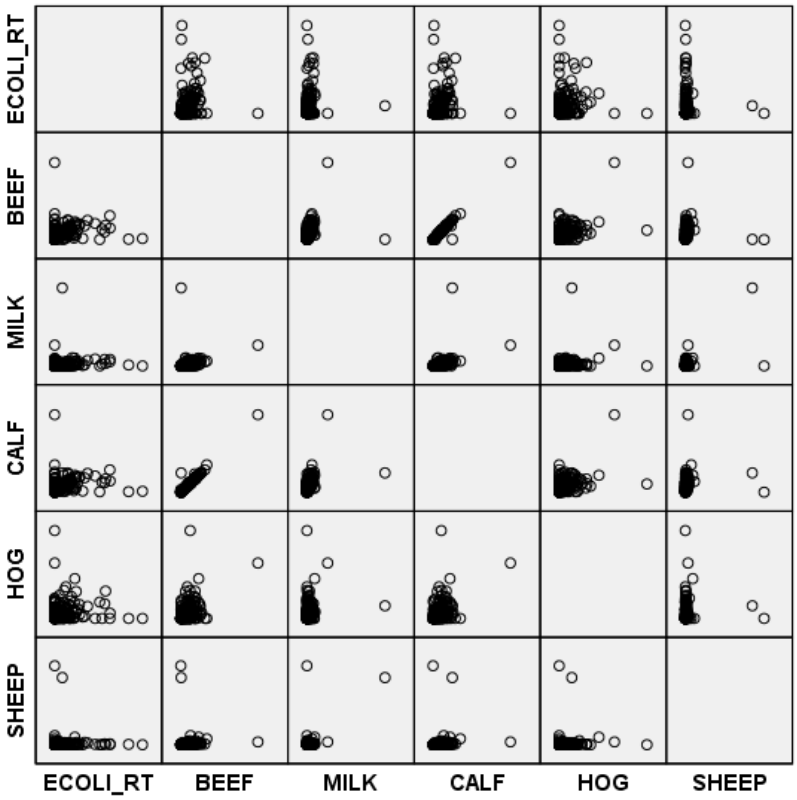


Figure 43. Scattergram of *E. coli* O157 rate with agricultural variables (n=180).

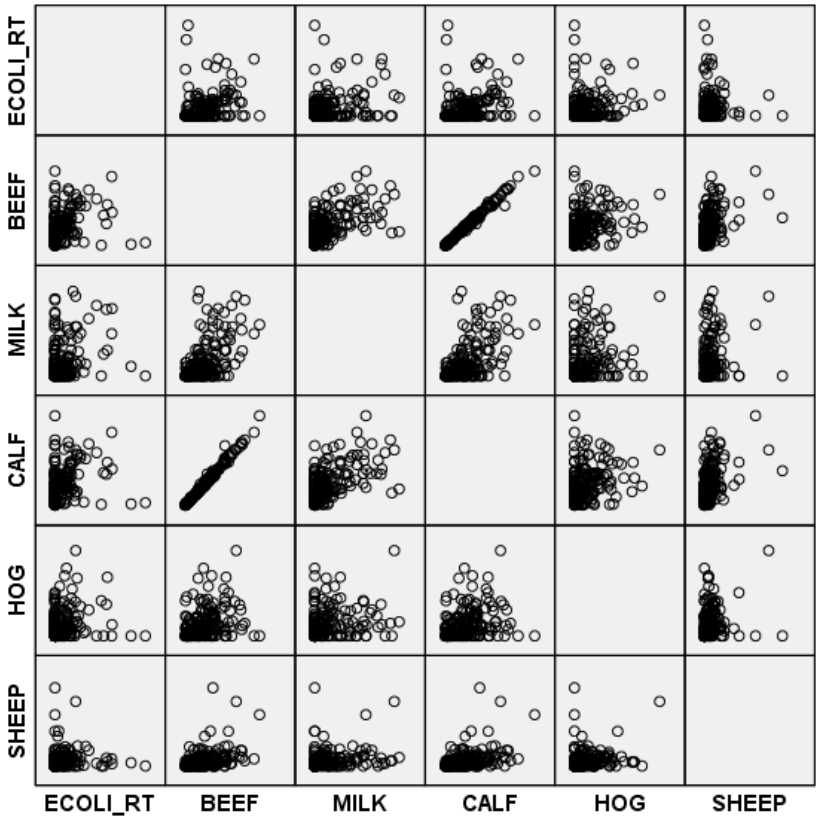


Figure 44. Scattergram of *E. coli* O157 rate with agricultural variables after removal of outliers and zip code areas with zero population (n= 167).

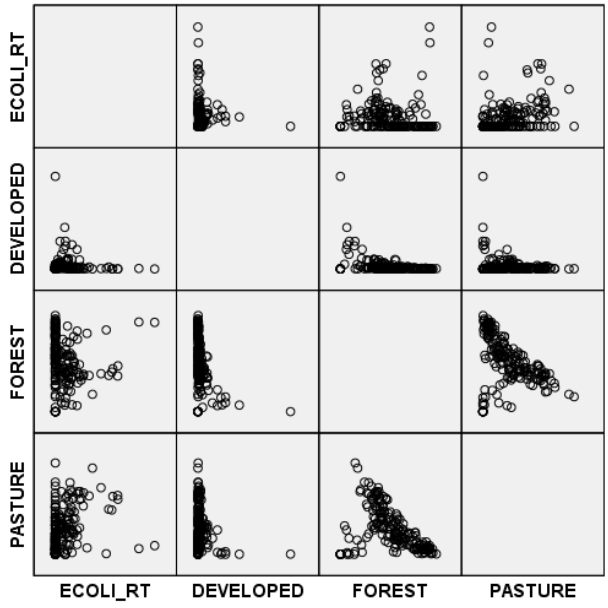


Figure 45. Scattergram of *E. coli* O157 rate with land use variables (n=180).

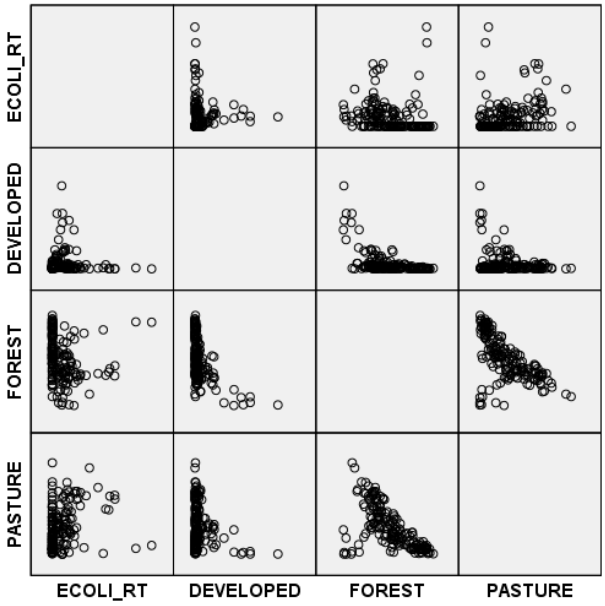


Figure 46. Scattergram of *E. coli* O157 rate with land use variables after removal of outliers and zip code areas with zero population (n= 167).

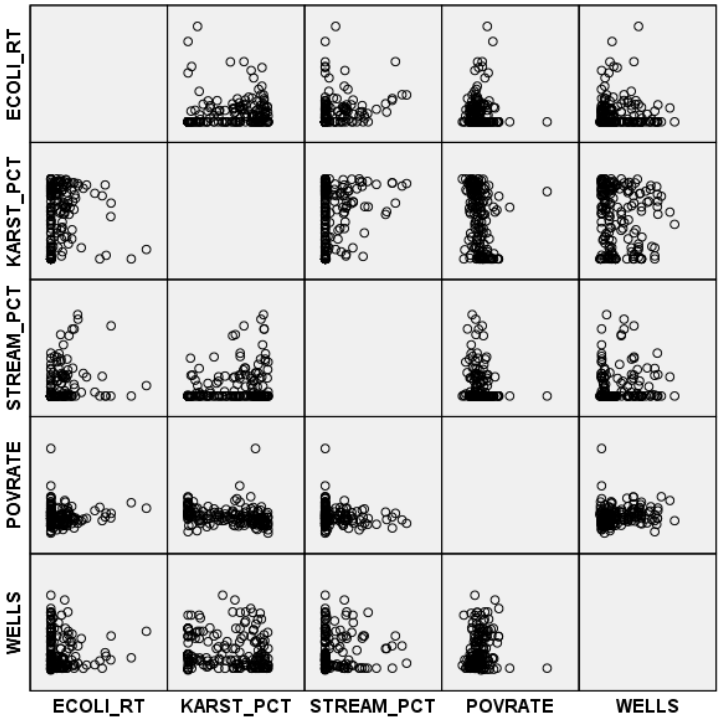


Figure 47. Scattergram of *E. coli* O157 rate with poverty, water quality, geology and well density variables (n=180).

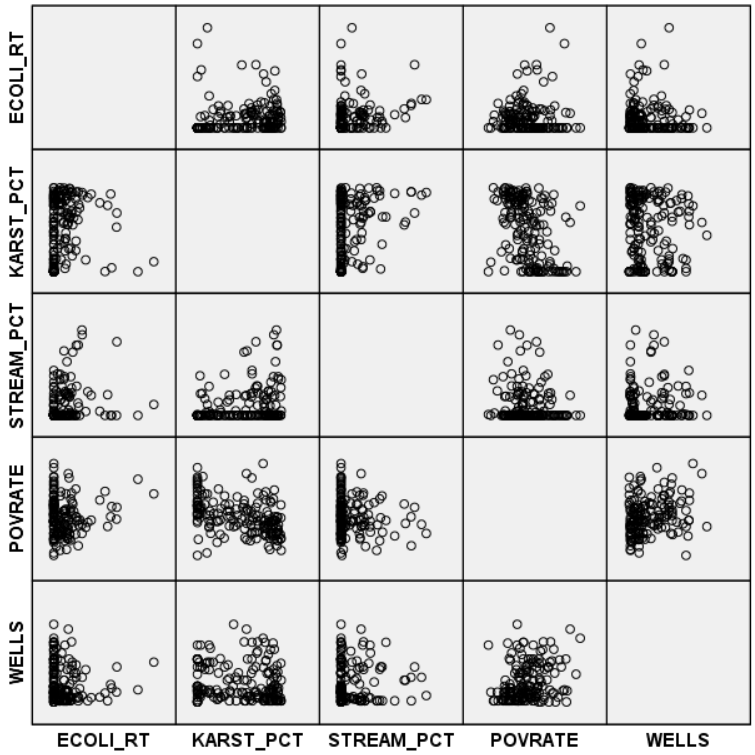


Figure 48. Scattergram of *E. coli* O157 rate with poverty, water quality, geology and well density variables after removal of outliers and zip code areas with zero population (n= 167).

Candidate variables for modeling the risk for *E. coli* O157 in the population included: population densities of beef cows (BEEF), milk cows (MILK), calves (CALF), hogs (HOG), and sheep (SHEEP); land use as a percentage of developed land (DEVELOPED), forested land (FOREST), and agricultural land (PASTURE); percent of bedrock classified as karst geology (KARST_PCT); percent of streams classified as impaired for the presence of *E. coli* (STREAM_PCT); poverty rate (POVRATE); and well permit density by county population (WELLS).

Using a distance based spatial weights matrix (distance = 40 km, which allows each zip code area to have at least one neighbor), the local Moran's I was calculated at 0.072 (estimated p=0.004) using GeoDa. Cluster maps of Local Indicators of Spatial Autocorrelation (LISA) (Figure 49) and Getis Gi (Figure 50) suggest two disease clusters, a small one in the north, and a larger one just south of the first, in the east central portion of the study area. To the west and south, low-low clusters are present, indicating lower disease incidence (after normalization by population) in these areas.

A histogram of the distribution of standardized rates shows marked skewness in the *E. coli* O157 rate data (Figure 51). Log transformation does not remove the skewness because of the large number of zeros in the rates dataset.

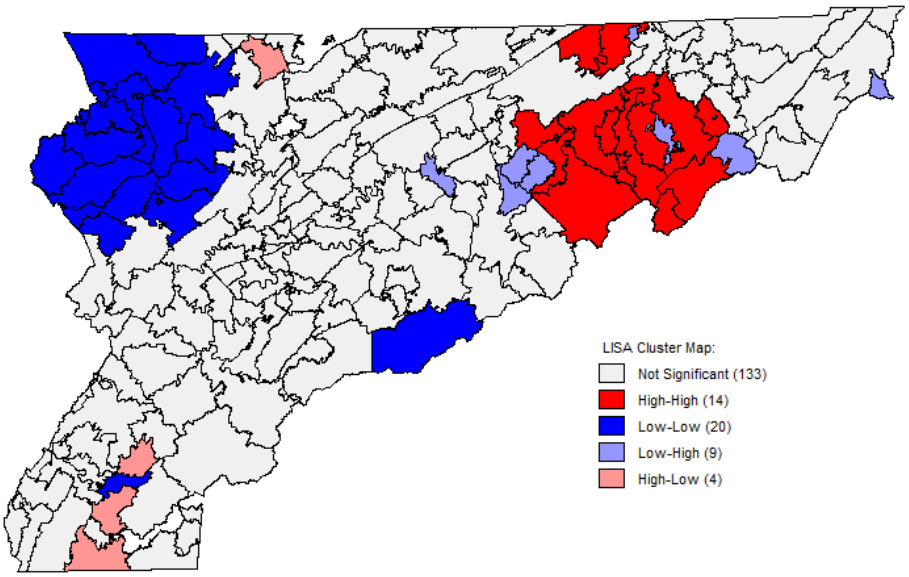


Figure 49. Cluster map showing LISA clusters. The map was generated in GeoDa using a distance based spatial weights matrix at 40km.

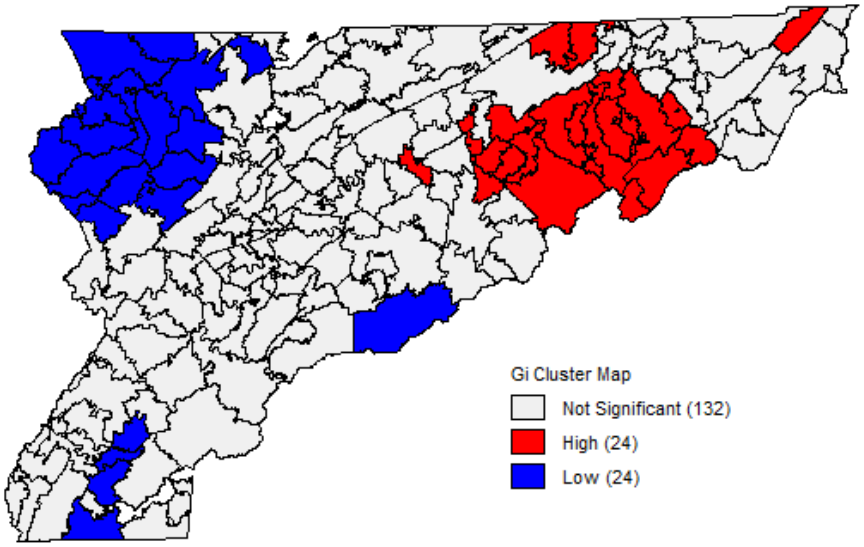


Figure 50. Getis Gi Cluster map of *E. coli* O157 rate data. The map was generated in GeoDa using a 40km distance spatial weights matrix.

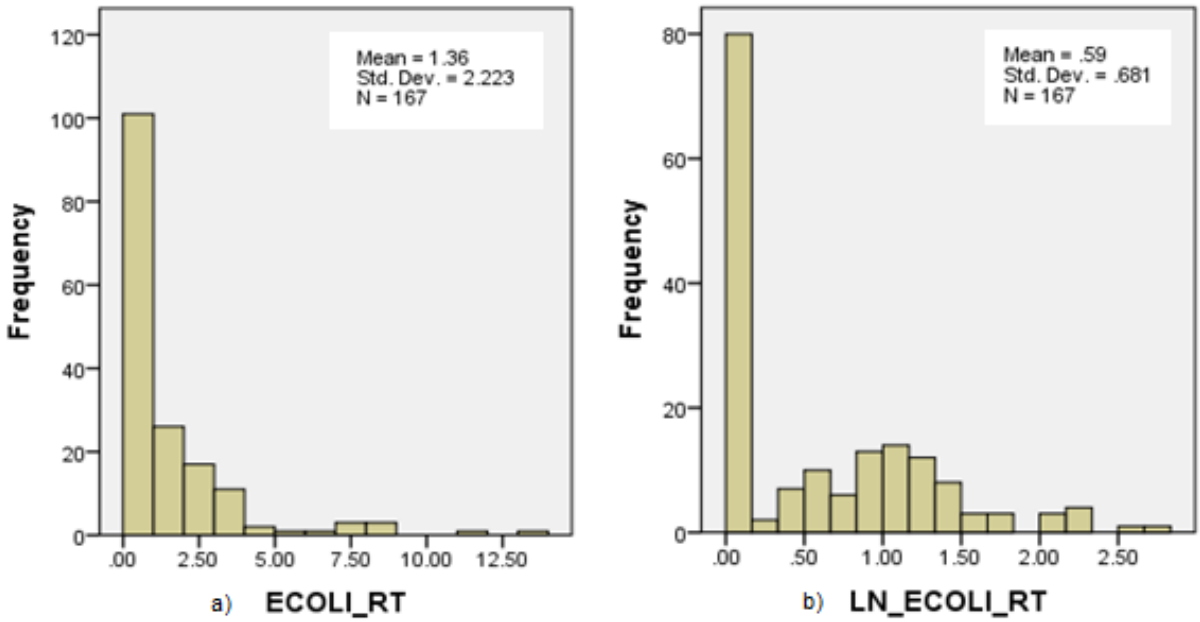


Figure 51. Histograms showing the distribution of a) standardized *E. coli* O157 rates and b)log-transformed rates for outlier removed aggregate data (n=167).

5.2.2 Regional regression model results for *E. coli* O157

The set of models listed in Table 5 were developed for three partitions of the dataset: the complete dataset, the dataset with outliers and non-zero population areas removed, and the non-zero rate partition. The results of the three data partitions are presented in the following pages, and the coefficients and p-values for all models generated with the three partitions are summarized in Table 22. Only the Poisson-distributed GAMs included an offset term for the log of population.

E. coli O157 model results using complete dataset The spatial lag model retained the land use variable percent agricultural land (PASTURE) and percent impaired stream miles (STREAM_PCT) with $R^2 = 0.16$ and AIC=785. Interestingly, the spatial weights variable was not significant in the model, and given that the spatial parameter was significant in the GAM, this suggests that the linear model is not a good fit for the data and the spatial lag model is misspecified.

The GLM with Gamma distribution and log link function (AIC=590) retained the land use variables percent forested land (FOREST) and agricultural land use (PASTURE), percent impaired streams (STREAM), and well permit density (WELLS). Poverty rate (POVRATE) was marginally significant at p=0.01. The Tweedie GLM retained the same variables, with AIC=630, indicating a greater loss of information in the Tweedie model.

Table 22. Comparison of coefficients and p-values (in parentheses) for *E. coli* O157 regional models and data partitions. Smoothed variables in GAMs have multiple coefficients and are not reported.

	Spatial Lag	GLM Gamma	GLM Tweedie	GAM Gamma	GAM Tweedie	GAM Poisson
All data (n=180)						
BEEF	1.539 (.595)	0.270 (.771)	0.467 (.621)	(.357)	(.357)	-1.967 (.069)
MILK	3.983 (.464)	1.085 (.542)	1.489 (.406)	(.693)	(.693)	-1.722 (.407)
CALF	-1.49 (.568)	-0.341 (.688)	-0.506 (.557)	0.634 (.455)	0.634 (.455)	1.760 (.071)
HOG	-1.09 (.856)	0.051 (.979)	-0.120 (.951)	(.319)	(.319)	(.124)
SHEEP	-0.80 (.370)	-0.338 (.188)	-0.384 (.254)	(.274)	(.274)	(.229)
DEVELOPED	0.005 (.835)	0.018 (.124)	0.009 (.361)	(.061)	(.061)	-0.066 (.000)
FOREST	0.013 (.139)	0.008 (.004)	0.008 (.011)	0.005 (.080)	0.005 (.080)	(.000)
PASTURE	0.036 (.005)	0.020 (.000)	0.018 (.000)	(.102)	(.102)	(.122)
KARST_PCT	0.003 (.537)	0.002 (.314)	0.002 (.280)	0.002 (.365)	0.002 (.365)	-0.001 (.745)
STREAM_PCT	0.023 (.010)	0.008 (.007)	0.008 (.003)	0.006 (.031)	0.006 (.031)	0.004 (.077)
POVRATE	2.695 (.154)	1.242 (.070)	1.105 (.084)	(.162)	(.162)	1.337 (.125)
WELLS	-0.000 (.113)	0.000 (.040)	0.000 (.033)	-0.0004 (.082)	-0.0004 (.082)	-0.0004 (.026)
Outliers removed (n=167)						
BEEF	-2.540 (.467)	-0.873 (.405)	-1.014 (.340)	(.296)	(.296)	-2.934 (.016)
MILK	6.096 (.509)	0.25 (.932)	0.402 (.887)	5.161 (.110)	5.161 (.110)	-0.094 (.976)
CALF	2.625 (.389)	0.926 (.333)	1.048 (.272)	0.846 (.347)	0.846 (.347)	2.890 (.007)
HOG	7.245 (.393)	3.898 (.099)	3.603 (.142)	(.223)	(.223)	5.185 (.044)
SHEEP	-6.625 (.203)	-3.159 (.040)	-3.093 (.069)	-3.161 (.061)	-3.161 (.061)	-4.119 (.039)
DEVELOPED	0.172 (.022)	0.096 (.000)	0.086 (.000)	(.036)	(.036)	0.015 (.539)
FOREST	0.055 (.017)	0.026 (.000)	0.025 (.001)	-0.001 (.900)	-0.001 (.900)	0.002 (.731)
PASTURE	0.065 (.009)	0.033 (.000)	0.032 (.000)	(.036)	(.036)	0.003 (.675)
KARST_PCT	0.010 (.158)	0.005 (.028)	0.005 (.028)	0.001 (.608)	0.001 (.608)	0.001 (.470)
STREAM_PCT	0.018 (.062)	0.006 (.063)	0.006 (.041)	0.005 (.057)	0.005 (.057)	0.002 (.294)
POVRATE	5.326 (.041)	2.444 (.003)	2.484 (.004)	1.928 (.046)	1.928 (.046)	0.951 (.395)
WELLS	-0.001 (.176)	0.000 (.091)	0.000 (.080)	-0.0003 (.125)	-0.0003 (.125)	-0.0003 (.080)
Non-zero partition (n=87)						
BEEF	-1.494 (.803)	-1.096 (.329)	-1.010 (.394)	-1.854 (.082)	-1.854 (.082)	-2.748 (.055)
MILK	23.904 (.064)	4.779 (.085)	5.379 (.049)	8.425 (.005)	8.425 (.005)	2.760 (.467)
CALF	2.423 (.659)	1.263 (.226)	1.200 (.278)	2.009 (.037)	2.009 (.037)	2.892 (.027)
HOG	4.307 (.684)	2.182 (.315)	2.209 (.325)	(.094)	(.124)	4.828 (.083)
SHEEP	-21.021 (.011)	-5.537 (.001)	-5.848 (.001)	-6.663(.000)	-6.663 (.000)	-6.227 (.022)
DEVELOPED	0.044 (.695)	0.017 (.515)	0.019 (.448)	-0.007 (.761)	-0.007 (.761)	-0.0004 (.987)
FOREST	0.045 (.265)	0.011 (.211)	0.014 (.119)	(.060)	(.060)	0.002 (.840)
PASTURE	0.022 (.560)	0.007 (.374)	0.008 (.358)	0.005 (.469)	0.005 (.469)	-0.001 (.890)
KARST_PCT	-0.009 (.481)	-0.002 (.464)	-0.002 (.568)	(.065)	(.065)	0.001 (.610)
STREAM_PCT	0.009 (.399)	0.002 (.299)	0.002 (.330)	0.000 (.801)	0.001 (.801)	0.000 (.899)
POVRATE	11.679 (.021)	2.606 (.016)	2.534 (.020)	1.363 (.221)	1.363 (.221)	1.212 (.454)
WELLS	-0.002 (.068)	0.000 (.068)	0.000 (.062)	(.359)	(.359)	-0.0004 (.175)

The Generalized Additive Model (GAM) with a Gamma distribution and log link function retained percent impaired streams (STREAM_PCT) (significant at p=0.05), and percent developed land (DEVELOPED), percent forested land (FOREST), percent agricultural land (PASTURE), and well permit density (WELLS) (significant at p=0.1). Figure 52 displays the R output for the GAM model using a Gamma distribution and log link. The first column of the table displays the edf, revealing that the significant variables FOREST, STREAM_PCT and WELLS have a linear relationship with *E. coli* O157 rate, and therefore the use of a smooth function of these variables as a term in the model is unnecessary, and the variables can be included as simple covariates.

Re-running the model without smoothes on CALF, FOREST, STREAM_PCT, KARST_PCT and WELLS, produces no change in significance of variables however, because single coefficients are reported for non-smoothed variables the nature of the relationship between risk factors and disease can be better understood. FOREST and STREAM_PCT have positive coefficients, indicating a positive relationship with disease (as the percentage of land devoted to forest increases, and as the percentage of stream miles impaired with *E. coli* increases, so does disease incidence). The coefficient for WELLS is negative indicating that as the number of well permits increases in a county, the risk for disease decreases. Because DEVELOPED is a smoothed model parameter, multiple coefficients are generated, some positive and some negative.

```
> gamer1a <- gam(ERATEplus1 ~ 1+ s(long,lat) + s(BEEF) + s(MILK) +
s(CALF) + s(HOG) + s(SHEEP) + s(DEVELOPED) + s(PASTURE) + s(FOREST) +
s(KARST_PCT) + s(STREAM_PCT) + s(POVRATE) + s(WELLS), data=rates,
family=Gamma(link=log))
> summary(gamer1a)

Family: Gamma     Link function: log

Formula:ERATEplus1 ~ 1 + s(long, lat) + s(BEEF) + s(MILK) + s(CALF) +
    s(HOG) + s(SHEEP) + s(DEVELOPED) + s(PASTURE) + s(FOREST) +
    s(KARST_PCT) + s(STREAM_PCT) + s(POVRATE) + s(WELLS)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.66553    0.04088   16.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(long,lat)  17.834 22.587 1.943  0.0103 *
s(BEEF)       2.914  3.477 1.085  0.3573
s(MILK)       1.468  1.705 0.318  0.6931
s(CALF)       1.000  1.000 0.561  0.4551
s(HOG)        2.344  2.809 1.165  0.3194
s(SHEEP)      1.536  1.787 1.265  0.2740
s(DEVELOPED)  1.692  1.995 2.847  0.0614 .
s(PASTURE)    3.061  3.797 1.991  0.1022
s(FOREST)     1.000  1.000 3.110  0.0799 .
s(KARST_PCT)  1.000  1.000 0.824  0.3655
s(STREAM_PCT) 1.000  1.000 4.726  0.0313 *
s(POVRATE)    1.955  2.412 1.773  0.1628
s(WELLS)      1.000  1.000 3.064  0.0822 .
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.278   Deviance explained = 56.6%
GCV score = 0.38343  Scale est. = 0.30077   n = 180
> AIC(gamer1a)
[1] 534.6554
```

Figure 52. R output for GAM model using complete *E. coli* O157 rate dataset (n=180). R code is printed in red and preceded by the “>” character.

Examination of the coefficient of determination (R^2) shows an increase with each tweak to the smoothed variables, however, the largest value of the adjusted $R^2 = 0.278$ for the GAM model leaves over 70% of the variability still to be explained.

A substitution of the Tweedie distribution in place of the Gamma distribution in the GAM had no effect on model output, and the effect of removing smoothes on some variables impacted significance only at the 4th decimal place. Model diagnostics were equal in the Tweedie and Gamma GAMs (adjusted R^2=0.278 and AIC=534), indicating that neither the Gamma nor the Tweedie distributions are good fits for the full *E. coli* O157 rates dataset.

The Poisson-distributed GAM (adjusted R^2=0.826 and AIC=585.67) returned the agricultural variables beef cow (BEEF) and milk cow (MILK) population density, and the water quality variable (STREAM_PCT) as marginally significant (at $p<0.1$). The land use variables DEVELOPED and FOREST plus well permit density (WELLS) were significant at $p<0.05$. Both well permit density (WELLS) and percent developed land (DEVELOPED) were associated with a protective effect (negatively associated with disease in the population).

Model results using dataset with outliers removed After removal of outliers and zip code areas with zero population (n=167), Spearman correlations and scatterplots were minimally affected. The spatial lag model (using a 40 km distance spatial weights matrix) included all land use variables (DEVELOPED, FOREST, and PASTURE), and

poverty rate (POVRATE) at p=0.05. All coefficients had positive coefficients. The regional water quality variable STREAM_PCT was significant at p=0.1.

The GLM with Gamma distribution and a log link function retained one agricultural variable (sheep population density (SHEEP)), all land use variables (DEVELOPED, FOREST, and PASTURE), percent karst geology (KARST_PCT), and poverty rate (POVRATE) at a significance level of p=0.05. Hog population density (HOG), surface water quality (STREAM_PCT), and well permit density (WELLS) were significant at p=0.1. All coefficients were positive except the coefficient for SHEEP.

The GLM with a Tweedie distribution and a log link function returned the same group of significant covariates as the GLM with Gamma distribution. Similarly, all coefficients were positive except for SHEEP. While the GLM does not return a Coefficient of Determination (R^2), the AIC indicated that the Gamma distribution produced a model with a better fit (AIC = 550) than the model with the Tweedie distribution (AIC = 587).

The GAM with Gamma distribution and log link function retained two land use parameters (DEVELOPED and PASTURE) and poverty rate (POVRATE) as significant at p=0.05. Sheep population density (SHEEP) and surface water quality (STREAM_PCT) were significant at p=0.1. An examination of the estimated degrees of freedom in the R output for the first GAM run (Figure 53) shows that the relationship between *E. coli* O157 rate and the significant variables SHEEP, STREAM_PCT, and POVRATE is linear. A second model which removes the smoothes on these variables provides single coefficients; STREAM_PCT and POVRATE have positive coefficients, while the

coefficient for SHEEP is negative, indicating that increased sheep population density is associated with a reduced incidence of disease. The smoothed variable s(DEVELOPED) had multiple coefficients with varying signs. As observed with models generated using the full dataset, the coefficient of variation ($R^2 = 0.286$) increased with tweaks to the dataset and model, but over 70% of the variability remained unexplained.

The GAM with Tweedie distribution and log link function produced identical output to the GAM with a Gamma distribution. Differences arose in p-values at the 4th decimal place when removing smoothes on some variables. There are no difference in model diagnostics between the two GAMs ($R^2 = 0.286$, AIC = 517).

The GAM with Poisson distribution and log link function returned the agricultural animal population density variables BEEF, CALF, HOG and SHEEP as significant (at p=0.05) and well permit density (WELLS) significant at p=0.1. The agricultural variables BEEF and SHEEP were negatively associated with disease in the population (a protective effect) as was well density (WELLS).

```
Family: Gamma Link function: log

Formula:
ERATEplus1 ~ 1 + s(long, lat) + s(BEEF) + s(MILK) + s(CALF) +
    s(HOG) + s(SHEEP) + s(DEVELOPED) + s(PASTURE) + s(FOREST) +
    s(KARST) + s(STREAM) + s(POVRATE) + s(WELLS)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.70104    0.04282   16.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(long,lat)  16.844 21.702 1.905  0.0140 *
s(BEEF)       2.000  2.508 1.219  0.2962
s(MILK)       1.000  1.000 2.581  0.1105
s(CALF)       1.000  1.000 0.890  0.3472
s(HOG)        2.473  2.973 1.476  0.2236
s(SHEEP)      1.000  1.000 3.547  0.0618 .
s(DEVELOPED)  1.878  2.340 3.209  0.0366 *
s(PASTURE)    3.128  3.884 1.799  0.1341
s(FOREST)     1.000  1.000 0.016  0.9006
s(KARST)      1.000  1.000 0.263  0.6088
s(STREAM)     1.000  1.000 3.663  0.0577 .
s(POVRATE)    1.000  1.000 4.025  0.0468 *
s(WELLS)      1.000  1.000 2.379  0.1253

R-sq.(adj) =  0.286   Deviance explained = 55.2%
GCV score = 0.39168  Scale est. = 0.30982   n = 169
> AIC(gamer1)
[1] 517.0182
```

Figure 53. R output for GAM model using dataset with outliers removed (n=167). R code is printed in red and preceded by the “>” character.

Model Results using non-zero partition Spearman correlation coefficients (Table 23) for the non-zero partition of the dataset show that after removal of zip code areas with *E. coli* O157 disease rate equal to zero, only the variables BEEF, MILK, CALF, DEVELOPED, PASTURE and POVRATE were significantly correlated to disease (p=0.05, n=87), but all variables were retained as candidate variables because the Spearman correlation coefficient measures the strength of the linear relationship, and a weak but significant non-linear relationship cannot be ruled out. The variables BEEF, CALF and PASTURE were highly correlated with each other, but were retained as candidate variables because it was preliminary to exclude one variable over another.

Scattergrams (Figure 54, Figure 55, and Figure 56) graphically show the relationship between *E. coli* O157 rate and explanatory variables. These can be more useful than the correlation coefficient in identifying non-linear relationships, although the individual scatterplots below show only the strong linear relationship between BEEF and CALF, which was identified in the correlation matrix (Table 23), and non-linear relationships between variables are difficult to discern.

Table 23. Spearman correlations for non-zero partition of *E. coli* O157 standardized rates. Significant correlations are marked with a single asterisk (*) for p=0.05 and a double asterisk (**) for p=0.01 (p-values are in parentheses). Correlations greater than r=0.7 are shaded.

	ECOLI_RT	BEEF	MILK	CALF	HOG	SHEEP	DEVELOPED	FOREST	PASTURE	KARST_PCT	STREAM_PCT	POVRATE	WELLS
ECOLI_RT	1	.289** (.006)	.356** (.000)	.308** (.003)	0.132 (.224)	-0.045 (.676)	-.527** (.526)	0.064 (.554)	.286** (.007)	-0.141 (.191)	0.109 (.313)	.340** (.001)	0.128 (.238)
BEEF	.289** (.006)	1	.527** (.590)	.993** (.886)	.254* (.017)	.449** (.287)	-.308** (.003)	-.375** (.000)	.819** (.347)	.250* (.019)	.257* (.016)	-0.073 (.498)	-0.069 (.524)
MILK	.356** (.000)	.527** (.590)	1	.541** (.458)	.218* (.042)	.321** (.002)	-.463** (.343)	-0.016 (.885)	.597** (.051)	-0.059 (.590)	0.183 (.090)	.279** (.008)	0.200 (.063)
CALF	.308** (.003)	.993** (.886)	.541** (.458)	1	.262* (.014)	.460** (.409)	-.314** (.003)	-.369** (.000)	.822** (.622)	.249* (.020)	.261* (.014)	-0.067 (.540)	-0.067 (.539)
HOG	0.132 (.224)	.254* (.017)	.218* (.042)	.262* (.014)	1	.218* (.042)	-0.108 (.320)	-0.090 (.405)	.336** (.001)	0.146 (.175)	0.086 (.429)	-0.064 (.557)	-0.014 (.900)
SHEEP	-0.045 (.676)	.449** (.287)	.321** (.002)	.460** (.409)	.218* (.042)	1	-0.198 (.065)	-0.036 (.742)	.529** (.409)	0.034 (.751)	.237* (.026)	0.029 (.791)	-0.022 (.840)
DEVELOPED	-.527** (.526)	-.308** (.003)	-.463** (.343)	-.314** (.003)	-0.108 (.320)	-0.198 (.065)	1	-.516** (.119)	-.424** (.310)	.405** (.000)	0.001 (.993)	-.474** (.587)	-.574** (.351)
FOREST	0.064 (.554)	-.375** (.000)	-0.016 (.885)	-.369** (.000)	-0.090 (.405)	-0.036 (.742)	-.516** (.119)	1	-.250* (.019)	-.657** (.677)	-0.137 (.205)	.506** (.616)	.565** (.200)
PASTURE	.286** (.007)	.819** (.347)	.597** (.051)	.822** (.622)	.336** (.001)	.529** (.409)	-.424** (.310)	-.250* (.019)	1	0.116 (.285)	0.151 (.163)	-0.046 (.672)	0.044 (.688)
KARST_PCT	-0.141 (.191)	.250* (.019)	-0.059 (.590)	.249* (.020)	0.146 (.175)	0.034 (.751)	.405** (.000)	-.657** (.677)	0.116 (.285)	1	0.111 (.307)	-.558** (.966)	-.333** (.001)
STREAM_PCT	0.109 (.313)	.257* (.016)	0.183 (.090)	.261* (.014)	0.086 (.429)	.237* (.026)	0.001 (.993)	-0.137 (.205)	0.151 (.163)	0.111 (.307)	1	-0.022 (.840)	-0.134 (.215)
POVRATE	.340** (.001)	-0.073 (.498)	.279** (.008)	-0.067 (.540)	-0.064 (.557)	0.029 (.791)	-.474** (.587)	.506** (.616)	-0.046 (.672)	-.558** (.966)	-0.022 (.840)	1	.438** (.205)
WELLS	0.128 (.238)	-0.069 (.524)	0.200 (.063)	-0.067 (.539)	-0.014 (.900)	-0.022 (.840)	-.574** (.351)	.565** (.200)	0.044 (.688)	-.333** (.001)	-0.134 (.215)	.438** (.205)	1

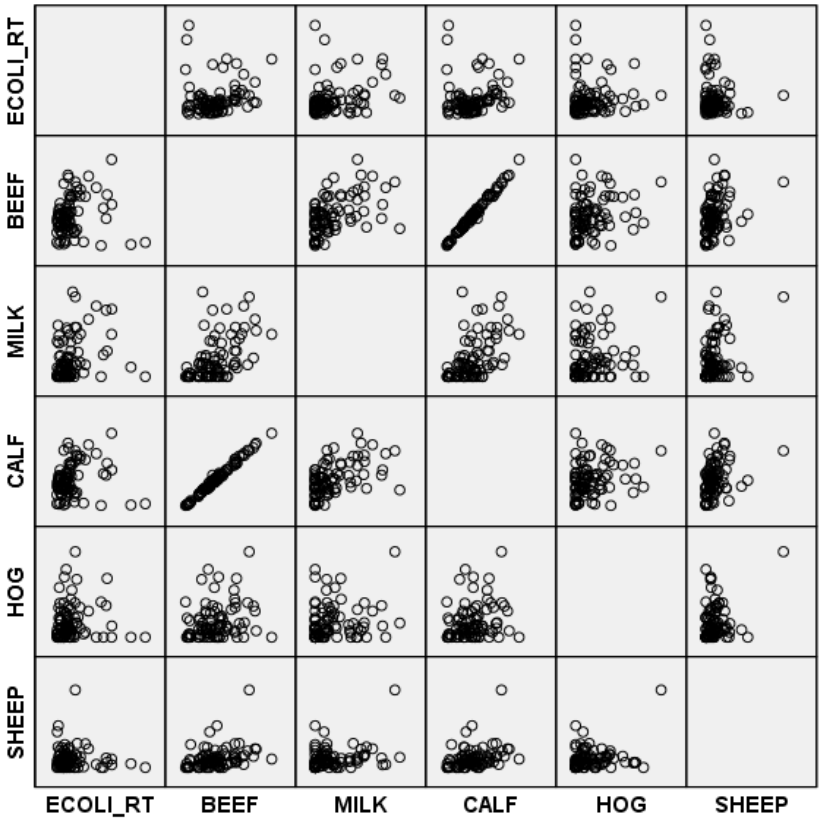


Figure 54. Scattergram of non-zero *E. coli* O157 rates and agricultural variables (n=87)

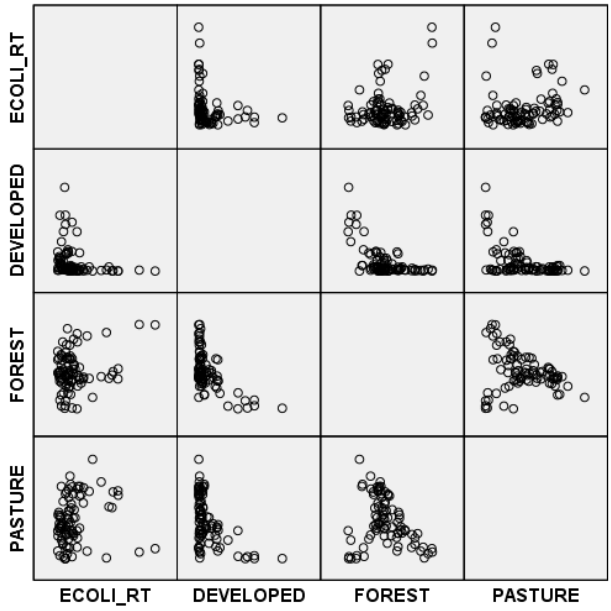


Figure 55. Scattergram of non-zero *E. coli* O157 rates and land use variables (n=87)

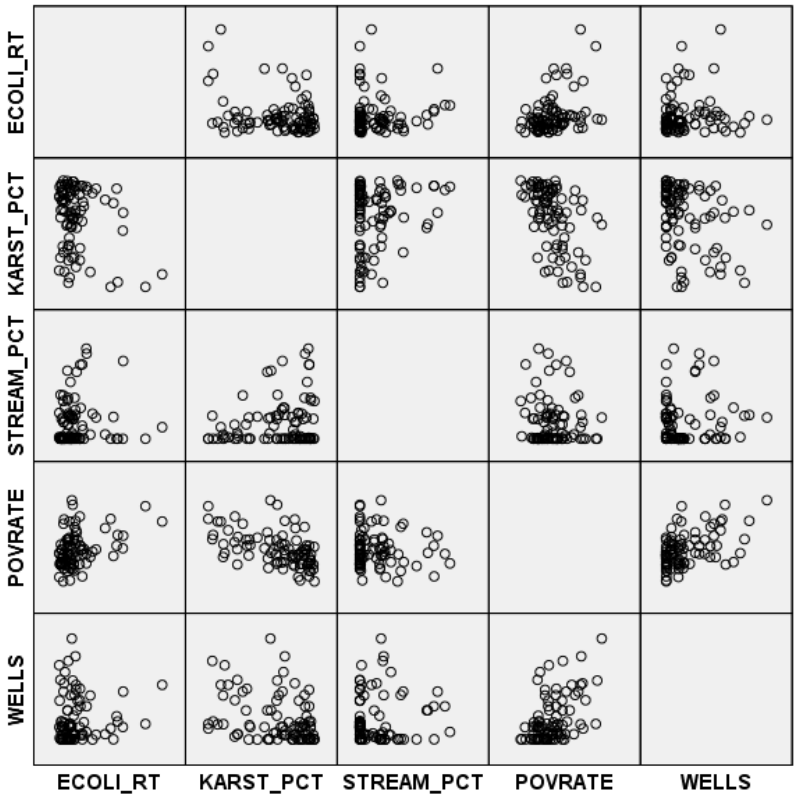


Figure 56. Scattergram of non-zero *E. coli* O157 rates and geology, water quality, poverty and well permit variables (n=87)

The spatial lag model retained sheep population density (SHEEP) and poverty rate (POVRATE) at p=0.05, and milk cow population density (MILK) and well permit density (WELLS) at p=0.1. Variables SHEEP and WELLS had negative coefficients.

Similar to the spatial lag model, the GLM with Gamma distribution and log link retained SHEEP and POVRATE at p=0.05, and MILK and WELLS at p=0.1. Only the coefficient for SHEEP was negative, although the coefficient for well permit density (WELLS) was reported as zero due to truncation. The GLM with Tweedie distribution and log link retained milk cow population density (MILK), sheep population density (SHEEP), and poverty rate (POVRATE) at a significance level of p=0.05, and retained well permit density (WELLS) at p=0.1. All coefficients were positive except the coefficient for the variable SHEEP. For comparison, the Gamma GLM had AIC=325, while the GLM with the Tweedie distribution had AIC=337. Because these models used the same input variables it is safe to compare AIC values, and we conclude that the Gamma GLM with log link function is the better of the two GLM models.

The GAM with Gamma distribution and log link retained the agricultural population density variables MILK (positive coefficient), CALF (positive coefficient) and SHEEP (negative coefficient) at p=0.05. Beef cow population density (BEEF) (negative coefficient), hog population density (HOG), forested land use (FOREST), and percent karst geology (KARST_PCT) were significant at p=0.1. The latter three (HOG, FOREST and KARST_PCT) were included in the GAM as smoothed parameters indicating a nonlinear relationship with *E. coli* O157. The GAM with Tweedie

distribution and log link reported the same results as the Gamma distributed model with the exception that HOG was not retained in the Tweedie model.

The GAM with Poisson distribution and log link retained the agricultural variables CALF and SHEEP ($p<0.05$), while BEEF and HOG were marginally significant ($p<0.1$). Both BEEF and SHEEP were associated with reduced risk of disease in the population, while CALF and HOG were positively associated with disease.

5.2.3 Comparison of *E. coli* O157 regional models and treatments

In contrast to the individual models for *E. coli* O157 infection, no viable prediction model was developed for the population. Instead useful information was extracted by examining the overall consistency of the regional model output. Table 24 displays a tally of the number of times each variable was retained in a model. Explanatory variables retained in at least six of the eighteen models are considered important risk factors for *E. coli*O157 at the regional scale. By identifying which variables were consistently significant, the important risk factors for disease were identified as sheep population density (SHEEP), the land use variables DEVELOPED, FOREST, and PASTURE, surface water quality (STREAM_PCT), poverty rate (POVRATE), and well permit density (WELLS).

Sheep population density (SHEEP) and well permit density (WELLS) were negatively associated with disease, indicating that these factors have a protective effect. All land use variables (DEVELOPED, FOREST, and PASTURE) were positively

associated with disease, as were percent surface water impairment (STREAM_PCT) and poverty rate (POVRATE).

5.3 Chapter summary

In this chapter, two groups of models were generated for the *E. coli* O157 case data and rate data. Three partitioning treatments were applied to the rate data (all data, outliers removed, and non-zero rate areas removed), and multiple models were generated. The purpose of the modeling was to identify the risk factors for *E. coli* O157 infection in the individual (using a case-control approach) and in the population (using an ecological approach).

Table 24. Tally of significance of explanatory variables in regional *E. coli* O157 models. Significant explanatory variables are shaded.

	All data (n=180)	**Outliers removed (n=167)**	**Non-zero partition (n=87)**	**Total (out of a possible 18)**
BEEF	1	1	3	4
MILK	0	0	5	5
CALF	1	1	3	5
HOG	0	1	2	3
SHEEP	0	5	6	11
DEVELOPED	3	5	0	8
FOREST	5	3	2	10
PASTURE	5	5	0	10
KARST_PCT	0	2	2	4
STREAM_PCT	6	5	0	11
POVRATE	2	5	3	10
WELLS	5	3	3	11

To identify risk factors for disease in the individual, the most appropriate approach was the spatial logistic regression model. This model positively associated beef cow population density, hog population density, percent developed land, percent agricultural land, and proximity to karst geology with increased risk for *E. coli* O157 in the individual. Milk cow population density and well permit density were associated with reduced risk for disease in the individual.

Risk factors for *E. coli* O157 in the population were those variables that were consistently statistically significant throughout the different modeling approaches and data treatments. The risk factors positively associated with disease in the population are percent developed land, percent agricultural land, percent forested land, percent surface water impairment, and poverty rate. The sheep population density and well permit density variables were negatively associated with *E. coli* O157 at the regional level.

Table 25 summarizes the *E. coli* O157 results for both individual and regional models, and their implications will be discussed in section 6.2.

Table 25. Summary of significant environmental risk factors for E. coli O157 infection at individual and regional scales. "+" indicates positive association with disease, "-" indicates negative association with disease.

Risk factor	Individual model	Regional model
Agricultural animals	+Beef cow population density - Milk cow population density +Hog population density	-Sheep population density
Land use	+Percent developed land in cell +Percent pasture land in cell	+Percent developed land in zip +Percent pasture land in zip +Percent forested land in zip
Geology	+Proximity to limestone bedrock	--
Surface water	--	+Proportion of *E. Coli* impaired stream miles in zip
Poverty rate	--	+Poverty rate in zip
Well permits	-Well permits by county population	-Well permits by county population

CHAPTER 6 DISCUSSION

This research focuses on two goals:

1) to identify risk factors for *E. coli* O157 and cryptosporidiosis in East Tennessee by relating disease data to environmental data through statistical regression models; and

2) to examine the effect of scale on risk factors for *E. coli* O157 and cryptosporidiosis by comparing risk factors for disease in the individual (case-control approach) and the population (ecological approach).

Risk factors for both diseases at the individual and the regional scales were identified using a multiple model-multiple approach method. This research extends the application of the spatial logistic regression model from mineral exploration to epidemiology, to generate a background risk map for disease at the individual level.

This chapter will discuss individual and regional risk factors for each disease separately, beginning with cryptosporidiosis (Section 6.1) and following with risk factors for *E. coli* O157 (Section 6.2). The chapter will end with a discussion of the implications of scale (Section 6.3) and the wider implications of this research (Section 6.4).

6.1 Risk factors for Cryptosporidiosis

This section contains a discussion of the risk factors and their significance in the individual (case-control) and regional (ecological) cryptosporidiosis models, which were presented in Chapter 4.

Agricultural variables In the individual models, only the beef cow population density (BEEF) were associated with increased risk for disease, in agreement with Hunter et al. (2007) and Said et al. (2008), who positively associated cattle contact with cryptosporidiosis. This study found that although beef cow population density was associated with cryptosporidiosis, milk cow population density was not significantly associated with disease at either the individual or the regional scales. It is therefore important to differentiate between beef and milk cattle. Other agricultural variables such as hog and sheep population density were not significant, likely because their populations are an order of magnitude smaller than beef cow populations, and so their impact would understandably be less than the impact of beef cattle; the greater numbers of beef cows is more likely to have an effect on those living nearby in terms of opportunity for contact and environmental reservoirs for cryptosporidiosis. Yoder and Beach (2010) suggest that risk for cryptosporidiosis might be reduced by reducing contact between humans and infected animals, and ensuring that hand-washing is completed after contact with animals.

In the regional models, beef cow population density (BEEF) was not a significant risk factor for cryptosporidiosis in the population, even though the scale of the variable

did not change between models. High correlation between beef cow population density and agricultural land use (PASTURE) at the zip code level suggests that PASTURE explains the same portion of the variance that is explained by BEEF, forcing this variable out of the aggregate model. This was confirmed by sequentially adding variables to a regional cryptosporidiosis model, beginning with beef cow population density. When the agricultural land use variable was added, beef cow population density was no longer retained in the model.

Land use variables At both the individual and regional scale, land use was important. Percent agricultural land (PASTURE) and percent developed land (DEVELOPED) were associated with increased risk for cryptosporidiosis in the individual and in the population, suggesting multiple pathways of transmission. A high percentage of developed land can be thought of as a proxy for high population density, and a higher number of cases of disease would be expected in these areas due to the number of residents and consequent opportunities for human-to-human transmission. Agricultural land was also associated with increased disease in the individual and in the population, likely due to the presence of agricultural animals and opportunities for animal-to-human transmission (Hunter et al. 2007, Centers for Disease Control and Prevention 2011a).

The data used in this study were not sufficiently detailed to speciate the cryptosporidiosis cases into *c. hominis* and *c. parvum,* which are the two species commonly associated with human-to-human contact, and animal-to-human contact,

respectively. Others have found a positive association between developed land and *c. hominis,* and between agricultural land and with *c. parvum* (Lake et al. 2007) , but there is no way to test this using the cryptosporidiosis case data available for this study.

Karst In most individual models KARST emerged as a small but consistently significant risk factor for disease, such that increased proximity to karst was associated with increased risk. This finding is in agreement with Klumb (2010), and is not surprising for a number of reasons. First, karst areas are underlain by limestone that is less resistant to weathering and erosion. Karst areas therefore coincide with valley bottoms in East Tennessee, which is where development is concentrated due to ease of construction and access to transportation corridors (roads and rivers). This may lead to increased opportunities for human-to-human transmission. Second, and perhaps more importantly, in karst areas surface water and groundwater are well connected through sinkholes and springs, and so surface contamination readily enters the groundwater system and can emerge elsewhere as a spring. Patients may unwittingly contract cryptosporidiosis or another waterborne disease through contact with an impaired spring. Proximity to karst geology is therefore an important risk factor and should be included in future studies of environmental risk factors for cryptosporidiosis.

In the aggregate models, the regional variable percent karst geology was not consistently significant. Because the variables used in the individual and regional models are different (in the individual model KARST is a distance measure, and in the regional model KARST_PCT represents percent coverage), Kolmogorov-Smirnov and

Mann-Whitney tests were done on the regional geology variable for zero rate and non-zero rate zip code areas. Test results confirmed that aggregating geology to a regional percent karst variable did not remove the differences between the high and low rate areas, and that the lack of significance of the regional karst geology variable is not due to aggregation. Therefore, proximity to karst geology is an important risk factor for cryptosporidiosis infection in the individual, but living in a karst region is not an important risk factor in the population.

Overlap exists in the explanatory power of the regional variables, which was identified by generating single variable GLMs using the cryptosporidiosis rate data, beginning with the beef cow population density variable, and sequentially adding a second variable to see whether one or both were retained. The beef cow population density, geology, water quality and agricultural land use variables explain similar portions of the variance, which is one reason for the exclusion of KARST_PCT in the regional model.

Surface water quality In the individual models, proximity to an impaired stream (STREAM) was a significant risk factor in the spatial logistic regression models, but not in the non-spatial models, suggesting that there is a significant spatial relationship between surface water quality and disease. Proximity to an *E. coli* impaired stream was associated with increased risk for disease, which is in agreement with risk factors reported in other studies (Hlavsa et al. 2011, Beach 2008, Hoek et al. 2008). Contact with contaminated water may be more likely the closer the patient resides to the stream,

although it is important to note that proximity to a stream does not necessarily mean that contact occurs between an individual and the contaminant. Surface water streams identified by the state as impaired with *E. coli* are posted with signs warning against bodily contact. Such signs may be less than effective given the positive association between impaired stream proximity and disease risk.

In the aggregate models, the proportion of surface water streams that were impaired with *E. coli* (STREAM_PCT) was not significant. As with the geology variables, Kolmogorov-Smirnov and Mann-Whitney Tests revealed that this difference did not result from the differences between the variables, but may result from the overlap of the portion of the variability explained by the beef cow population density, geology, and agricultural land use regional variables.

Poverty rate Included in the individual model as the aggregated block group average, poverty rate was not a significant risk factor for disease in the individual, possibly due to the use of an aggregate variable. If socioeconomic status at the individual level were available, this finding may change. The significance of ZCTA poverty rate at the regional level suggests that the regional socioeconomic condition is a better predictor for disease in the population. This result agrees with the trend for developing countries (with higher poverty rates), where cryptosporidiosis prevalence is three times that of developed countries (Center for Food Security and Public Health 2005). In contrast, others found that indicators of economic advantage were associated with increased

disease, likely related to access to health care and increased diagnosis (Hu et al. 2009, Lake et al. 2007).

Well permits The protective effect associated with the availability of private water supply (well permit density) at both the individual and regional scale does not contradict results of a Tennessee Department of Health clinical epidemiological study on cryptosporidiosis conducted in the Greene County, Tennessee region from 2009-2011. This study used a supplemental form to gather drinking water source, recreational water exposure, food, travel, and animal exposures at the individual level. No association with private water supplies was found.

In contrast, a positive association was identified between well water consumption and cryptosporidiosis at the individual scale in Minnesota (Klumb et al. 2010), and between the prevalence of well water use and cryptosporidiosis at the regional scale in England and Wales (Said et al. 2003).

6.2 Risk factors for *E. coli* O157

This section discusses the risk factors associated with *E. coli* O157 in the individual and the population (regional model), which were presented in Chapter 5.

Agricultural variables The positive association of beef cow population density and negative association of milk cow population density with *E. coli* O157 infection in the individual suggests that it is important to treat beef cattle and dairy cattle separately.

All cattle-related studies reviewed for this research treated cattle as one group, but this research suggests that cattle population should be partitioned into beef and dairy. The protective effect of milk cow population density may result from regulation of farming practices in dairy farms, specifically the controls required to capture and treat runoff, which reduce the pathogen load in the environment. In contrast, beef cattle are more common in East Tennessee, and no environmental regulations exist to control runoff from pasture land nor to control access to streams by cattle. Consequently, runoff from agricultural fields is a non-point source of surface water pollution which increases the environmental pollution load.

The positive association between beef cow population density and disease agrees with other studies that have positively associated *E. coli* O157 infection with cattle population (Strachan et al. 2006, Michel et al. 1999, Valcour et al. 2002). The positive association found between hog population and disease contrasts with two other studies that found a negative association (Friesema et al. 2010) and no association (Valcour et al. 2002) between hog population and *E. coli* O157.

Land use variables The positive association of developed and pasture land with *E. coli* O157 at both the individual and the regional scale suggests that land use is important for disease in the individual and in the population, and scale is less important for this variable. The association of developed land with *E. coli* O157 infection contrasts with the findings of Strachan et al. (2006) who positively associated disease with rural areas and Kistemann et al. (2004) who positively associated disease with farm density. Their

findings agree, however, with the positive association between disease and pasture land found in this study.

The significance of all land use variables as risk factors for disease at either the individual or regional level suggest multiple pathways for infection. In developed land, transmission may be by human-to-human contact, while in pasture and forested lands, transmission may be through environmental contamination or animal-to-human contact. The significance of the interaction term between developed land and agricultural land represents proximity of developed and agricultural lands, (i.e. locations where developed land and agricultural land are both dominant) and an opportunity for contact between residents and agricultural activity, which has been related to increased risk for *E. coli* O157 (Strachan et al. 2006, O'Brien et al. 2001).

Karst In karst areas, sinkholes and springs provide a conduit for surface runoff to enter the groundwater system and emerge elsewhere with little opportunity for natural filtering that occurs in non-karst regions. Karst geology can therefore be related to an increased contaminant load in the environment, and proximity to karst geology was a significant risk factor for *E. coli* O157 in the individual.

Percent karst geology (the regional geology variable) was not associated with disease in the population, and this difference cannot be explained by aggregation of geology to a regional variable (KS and MW tests show that differences exist in the distributions of the regional geology variable for zero-rate and non-zero rate areas). This suggests that geology should be included in environmental risk models for *E. coli*

O157 as a proximity measure at the individual level. An overlap in explanatory power between the regional geology variable and the regional water quality variable was identified by sequentially adding the two variables to a regional model to see which variable was retained. Because the regional water quality variable forced the regional geology variable out of the model when both were included, it can be concluded that these variables explain the same portion of variability in the dataset.

Surface water quality In the individual models distance to *E. coli* impaired stream was not an important risk factor for disease, which contrasts the findings of O'Brien et al. (2001) who found an association between wading in surface water streams and increased risk for *E. coli* O157 in the individual. It is possible that, given the practice of 'posting' impaired streams with signs to discourage body contact with surface water in the United States, this result may indicate that the signage is effective in reducing risk for disease. However, given that the distance to impaired streams is a risk factor for cryptosporidiosis, it is more likely that the distance variable does not effectively capture the potential for bodily contact with surface water.

In contrast, the regional surface water quality variable (ratio of *E. coli* impaired to assessed stream miles) was positively associated with *E. coli* O157 infection in the population. More research is needed to understand the role of regional water quality on risk for this disease.

Poverty rate Poverty rate was not retained in the *E. coli* individual models which may indicate that poverty rate at the block group level is not a good indicator of risk for disease in the individual. A better variable to include in the individual model may be the socioeconomic status of the individual, rather than the individual's block group. The significance of poverty rate as a risk factor for disease in the regional models, however, agrees with the findings of Sakuma (2006) and Hasin (2007) who associated increased socioeconomic status with lowered disease, and contrasts with the findings of Chang et al. (2009) who found a negative association between poverty and *E. coli* O157 infection, which could be related to decreased access to healthcare and underdiagnosis.

Well permits In both the individual and regional models, well permit density was positively related to *E. coli* O157. While not indicative of drinking water source for a given case of disease, well permit density can indicate whether private water supply is more or less common in a county, so it is a measure of the availability of private water supply. The finding that increased disease risk in the individual and in the population is associated with decreased well density indicates that the availability of private water supplies have a protective effect at both scales. This result contrasts with findings by Strachan et al. (2006) who associated private water supply with increased risk for *E. coli* O157 infection.

The advantage of generating multiple models is that together they can be used to assess consistency. If all models have similar coefficient values, then higher confidence

can be placed on the model results. Although the magnitude of the coefficients can be somewhat variable between models, if the sign is generally consistent one can conclude that the variable has a consistent effect of increased (positive sign) or decreased (negative sign) risk as the variable increased in value and that this is independent of the model.

6.3 Importance of scale in identification of risk factors

The purpose of this section is to discuss the insights that can be gained by examining the environmental risk factors for cryptosporidiosis and *E. coli* O157 infection at two different scales, with case-control models of individual risk (individual models) and ecological models of population level risk (regional models). It is important to note that the two modeling approaches ask two different questions; the individual models assess risk in the individual, while the regional models assess risk in the population.

6.3.1 Methodological insights

Two well-known fallacies, the ecological fallacy and the atomistic fallacy, demonstrate the importance of drawing inferences at a scale appropriate to the data. The ecological fallacy, which can be defined as making inferences about the individual based on group-level data, is illustrated in this research by identifying risk factors that are significant in the regional models, but not significant in the individual models. For example, poverty rate at the zip code level was important in the regional level models,

but the block group poverty rate was not a significant risk factor in the individual level models for both cryptosporidiosis and *E. coli* O157. The atomistic fallacy is defined as drawing inferences about the group based on individual-level data, and can be illustrated by the identification of geology as a significant risk factor for both cryptosporidiosis and *E. coli* O157 in the individual. The geology variable in the individual models was represented by proximity of the patient's residence to karst geologic areas. When this variable was converted to a regional variable by calculating the percent karst geology within the aggregate area, it was not significant in the regional models for either disease. It is therefore important to recognize that variables can be significant risk factors at one level, but their analogues (scaled up or scaled down) may not be significant at another level.

In the research presented here, some of the variables included in the individual models, are necessarily measured at the group level because they were unavailable at the individual level. This can further complicate the scale effect. These variables include socioeconomic status, represented by poverty rate (block group level), and the agricultural variables (population densities at the zip code level). Inclusion of group level, or contextual, variables can provide important additional information that is not captured by individual variables because group level variable may constrain the choices available to the individual. Ignoring the relevant group-level variables in an individual model can lead to an incomplete understanding of the risk factors for disease in the individual, and has been labeled the "psychologistic fallacy" (Diez-Roux 1998).

The advantage of employing multiple models and multiple approaches is the ability to compare statistical regression modeling methods and results to understand the relationship between risk factors and disease. Risk factors that are associated with disease at the individual level and at the regional level are important across multiple spatial scales, while risk factors that are statistically significant only at the individual level or at the regional level should be incorporated in models for disease at those scales. This has implications for epidemiological studies; variables that are significant only at the individual level should be incorporated in a case-control model at the individual scale, yet variables that are significant at both the individual and regional scales may behave linearly such that scale is not important.

Risk factors for disease in the individual include behavioral or lifestyle factors specific to the individual that can influence exposure. These variables may include consumption of unpasteurized dairy products, visits to farms, and playing or swimming in contaminated surface water (Dietz and Roberts 2000, Suresh et al. 2012, Chalmers et al. 2000, Strachan et al. 2006, Robertson and Yasvinski 2012, Benham et al. 2006, Lake et al. 2007, Hunter et al. 2007). The individual models developed here do not capture this individual behavior, but instead can capture potential exposure to risk for the individual by inclusion of environmental variables at the local scale. Proximity to contaminated surface water or to agricultural activity can increase the likelihood that a person will come in contact with an environmental reservoir of cryptosporidiosis or *E. coli* O157, and in this way can increase background risk. The individual model, therefore, can be thought of as the environmental background risk at the individual

level (large-scale environmental risk). Similarly, the aggregated models capture the environmental and socioeconomic risk factors at a more regional level (zip code level), but do not address behavioral or lifestyle risk factors in the population.

6.3.2 Epidemiological insights

This research has shown that environmental variables are important risk factors for cryptosporidiosis and *E. coli* O157 in the individual and in the population.

Surface water streams that are recognized by the state and by the Environmental Protection Agency as impaired due to the presence of *E. coli* are considered a public health threat because *E. coli* is an indicator for the presence of both cryptosporidiosis and *E. coli* O157 in surface water. So far surface water quality spatial databases have been ignored in environmental disease modeling yet in this research proximity to an *E. coli* impaired surface water stream was positively associated with cryptosporidiosis at the individual scale. Likewise, the regional surface water quality was positively associated with *E. coli* O157 infection in the population. More research is needed to understand the role of *E. coli* in surface water as an indicator for public health threat for both of these diseases.

Geology, specifically limestone and dolomite formations that are prone to karst landforms, should be incorporated at the individual level as a proximity measure, not an abundance measure, as the distance to limestone areas was a significant predictor for both diseases.

Agricultural animal population densities are best incorporated at the individual level, and it is important to separate beef cattle populations from dairy cattle populations. Because land use is important at both scales, and both developed land and pasture land are associated with increased risk for cryptosporidiosis, future studies should partition the cryptosporidiosis cases by species (*c. hominis* and *c. parvum*) to examine the role of pasture and land use. This means that diagnostic methods must go beyond identification of probable cases, to identification of *cryptosporidium* at the species level.

6.4 Other considerations

While geospatial analysis and the explicit inclusion of space when assessing disease risk can be a valuable tool to identify populations at risk, any model is only as good as the data and assumptions used to generate the model. Non-reporting of diseases because a patient does not seek medical attention, because a sample is not taken during the medical examination, or due to incomplete reporting of disease incidence, can erode the quality of a dataset. Patterns in non-reporting can also affect results if members of one socioeconomic group are less likely (or more likely) to be exposed or to seek treatment (Lake et al. 2007). Care must be taken, therefore, in interpretation of model results because spatial differences in reporting rates may introduce bias into the modeled relationship between disease and environmental risk factors.

Selection of a "best" model is secondary to identifying risk factors for *E. coli* O157 and cryptosporidiosis from a set of potential environmental and socioeconomic explanatory variables. An understanding of the environmental and socioeconomic variables that emerge as significant risk factors for disease can help to inform policies to combat disease prevalence. Often these relationships can be discerned by exploratory mapping of disease cases overlayed with risk factors however, the modeling process is important to statistically quantify the risk and the important risk factors. Risk maps such as those displayed in Figure 29 and Figure 42 can visually display the environmental background risk for disease within the study area. Care must be taken in interpreting a map such as this because individual behavior has not been factored in. Background risk maps may therefore be most useful in developing the framework for a public education campaign to combat a disease endemic within a population, or in developing a more in depth epidemiological study. Knowledge of the explanatory variables used by the model to generate the risk map can be useful to identify the behaviors to target in such a campaign, for example warning residents of the risks associated with agricultural animal contact or playing in or around an impaired stream. In this way, limited public health resources can be targeted at the locations and the behaviors most associated with disease.

Revisiting the four analytical methods for modeling disease as described by Strachan and others (Strachan et al. 2006) (Section 2.3), this research uses the best available data to assess the importance of risk factors in disease by developing multiple regression models, employing Strachan's second and third analytical methods. The

remaining two methods (epidemiological studies of individual behavior and microbiological studies of the risk from a given vector) could not be employed using the data and resources available, however this research can be used as a springboard to develop a more in depth and focused epidemiological or microbiological study of the most important risk factors and their relationship with disease on an individual and microbiological level.

CHAPTER 7 CONCLUSION

For cryptosporidiosis, beef cow population density was positively associated with disease in individual. Land use variables were important in both the individual models (percent agricultural or developed land) and regional models (percent agricultural, developed or forested land), suggesting multiple transmission sources. Proximity to karst geology and *E. coli* impaired streams were important risk factors at the individual level, yet the equivalent aggregated variables were not important in the regional models. Poverty rate was important only at the regional level, suggesting that the socioeconomic conditions of the larger neighborhood (zip code level) are more important as a risk factor for disease than the very local neighborhood (block group level). Increasing well permit density was associated with decreasing disease in both the regional and individual models.

For *E. coli* O157, agricultural variables beef cow and hog population density were important at the individual level as were land use variables percent developed or percent agricultural land. Residence within an area close to both developed and agricultural land was also associated with increased risk in the individual models, which was confirmed in the regional models suggesting that land use is important at both the local and regional scale. Proximity to karst geology was associated with disease in the individual models, but not the regional models. In contrast, proximity to *E. coli* impaired streams was not important at the local scale, but the proportion of *E. coli* impaired streams was important at the regional scale. As with cryptosporidiosis,

poverty rate was important at only the regional level. Well permit density was associated with reduced risk for disease in the population and the individual.

The scale at which data are collected and aggregated is important. The relative importance of different risk factors can depend on the scale at which they are measured and modeled, and aggregation of disease case data to rate data can impact the risk factors that emerge as important in a regression model because the research question changes from an examination of risk factors for the individual to risk factors for the population. Development of both individual and regional models can help to distinguish whether risk factors are important on an individual level, on a regional level, or not important at all.

Further research is needed into the role of surface water impairment and risk for cryptosporidiosis and *E. coli* O157 in the individual and the population. Proximity to karst geology is important at the individual level and should be included in studies of risk factors for waterborne pathogens. Cattle populations should be partitioned into dairy and beef cattle when included as risk factors in a model for cryptosporidiosis and *E. coli* O157.

Models such as these can be useful to identify important risk factors for disease and to generate a background risk model. These results can then be used to develop a public education campaign to target limited public health resources to address behavior associated with the most important risk factors in the communities where opportunities for such behavior are most likely. These results can also be used as the framework upon

which to develop a more comprehensive epidemiological or microbiological study to examine specific pathways for disease that focus on individual level risk factors.

LIST OF REFERENCES

Adams, D. A., K. M. Gallagher, R. A. Jajosky, J. Ward, P. Sharp, W. J. Anderson, J. P. Abellera, A. E. Aranas, M. Mayes, M. S. Wodajo, D. H. Onweh & M. Park (2010) Summary of Notifiable Diseases - United States, 2010. *Morbidity and Mortality Weekly Report,* 59**,** 1-111.

Agterberg, F. P. (1974) Automatic Contouring of Geological Maps to Detect Target Areas for Mineral Exploration. *Mathematical Geology,* 6**,** 373-395.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19**,** 716-723.

Anselin, L. (2002) Under the Hood. Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics,* 27**,** 247-267.

Anselin, L., Y. W. Kim & I. Syabri (2004) Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems,* 6**,** 197-218.

Anselin, L., I. Syabri & Y. Kho (2006) GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis,* 38**,** 5-22.

Baddeley, A., M. Berman, F. N.I., A. Hardegen, R. K. Milne, D. Schuhmacher, R. Shah & R. Turner (2010) Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics,* 4**,** 1151-1201.

Baldursson, S. & P. Karanis (2011) Waterborne transmission of protozoan parasites: Review of worldwide outbreaks - An update 2004-2010. *Water Research,* 45**,** 6603-6614.

Beach, M. J. 2008. Waterborne: Recreational Water. In *Cryptosporidium and cryptosporidiosis, 2nd edition.,* eds. R. Fayer & L. Xiao, 329-362. Boca Raton, FL: CRC Press.

Bel, L., D. Allard, J. M. Laurent, R. Cheddadi & A. Bar-Hen (2009) CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics and Data Analysis,* 53**,** 3082-3093.

Benham, B. L., C. Baffaut, R. W. Zeckoski, K. R. Mankin, Y. A. Pachepsky, A. M. Sadeghi, K. M. Brannan, M. L. Soupir & M. J. Habersack (2006) Modeling bacteria fate and transport in watersheds to support TMDLs. *American Society of Agricultural and Biological Engineers,* 49**,** 987-1002.

Bivand, R. S., E. J. Pebesma & V. Gomez-Rubio. 2008. *Applied Spatial Data Analysis with R*. New York: Springer.

Bradley, K. & R. M. Atkinson. 2011. Update to Cryptosporidiosis Case Definition. 7. Centers for Disease Control and Prevention, Infectious Disease Committee.

Center for Food Security and Public Health. 2005. Cryptosporidiosis. ed. College of Veterinary Medicine, Iowa State University. Ames, Iowa.

Centers for Disease Control and Prevention (2011a) Cryptosporidiosis outbreak at a summer camp - North Carolina, 2009. *Morbidity and Mortality Weekly Report,* 60**,** 918-922.

---. 2011b. General Information *Escherichia coli* O157:H7 and other Shiga toxin-producing *Escherichia coli* (STEC). Atlanta, GA.

Chalmers, R. M., H. Aird & F. J. Bolton (2000) Waterborne Escherichia coli 0157. *Symposium Series in Social and Applied Microbiology,* 29**,** 124S-132S.

Chang, M., S. L. Groseclose, A. A. Zaidi & C. R. Braden (2009) An ecological analysis of sociodemographic factors associated with the incidence of salmonellosis, shigellosis, and *E. coli* O157:H7 infections in US counties. *Epidemiology and Infection,* 137**,** 810-820.

Chelghoum, N., K. Zeitouni & A. Boulmakoul. 2002. A Decision Tree for Multi-Layered Spatial Data. In *Symposium on Geospatial Theory, Processing and Applications*. Ottawa, Canada.

de Smith, M., M. Goodchild & P. Longley. 2009. *Geospatial Analysis - a comprehensive guide to principles, techniques and software tools*. www.spatialanalysisonline.com (last accessed 02-03-2011).

Denton, G. M., C. J. Brame, D. H. Arnwine, L. K. Cartwright & M. H. Graf. 2008. 2008 305(b) Report. The Status of Water Quality in Tennessee. Tennessee Department of Environment and Conservation, Division of Water Pollution Control.

Dietz, V. J. & J. M. Roberts (2000) National surveillance for infection with Cryptosporidium parvum, 1995-1998: what have we learned? *Public Health Reports,* 115**,** 358-363.

Diez-Roux, A. V. (1998) Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health,* 88**,** 216-222.

Dohoo, I. W., W. Martin & H. Stryhn. 2003. *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC.

Environmental Systems Research Institute, ArcGIS Desktop: Release 10. Redlands, CA.

Fayer, R. 2008. General Biology. In *Cryptosporidium and Cryptosporidiosis,* eds. R. Fayer & L. Xiao, 1-42. Boca Raton, Florida, USA: CRC Press.

Friesema, I. H. M., J. Van de Kassteele, C. M. De Jager, A. E. Heuvelink & W. Van Pelt (2010) Geographical Association Between Livestock Density and Human Shiga toxin-producing *Escherichia coli* O157 Infections. *Epidemiology and Infection,* in press.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold & J. Wickam (2011) Completion of the 2006 National Land Cover Database for the Conterminous United States. *PE&RS*, 77**,** 858-864.

Goldberg, D. W. & J. P. Wilson. 2011. USC WebGIS Services.

Gordis, L. 2009. *Epidemiology*. Philadelphia: Saunders Elsevier.

Gould, L. H., R. K. Mody, K. L. Ong, P. Clogher, A. B. Cronquist, K. N. Garman, S. Lathrop, C. Medus, N. L. Spina, T. H. Webb, P. L. White, K. Wymore, R. E. Gierke, B. E. Mahon & P. M. Griffin (2013) Increased Recognition of Non-O157 Shiga Toxin-Producing *Escherichia coli* Infections in the United States During 2000-2010: Epidemiologic Features and Comparison with *E. coli* 0157 Infections. *Foodborne Pathogens and Disease*, 10**,** 453-460.

Greene, D. C. & W. J. Wolfe. 2000. Superfund GIS - 1:250,000 Geology of Tennessee. Nashville, Tennessee: U.S. Geological Survey accessed from Tennessee Spatial Data Server http://www.tngis.org/frequently_accessed_data.html.

Hanna, S. S., R. M. Atkinson, T. F. Jones & J. R. Dunn. 2010. Shiga Toxin-Producing *Escherichia coli* (STEC) in Tennessee: Surveillance and Current Clinical Laboratory Practices. In *Tennessee Medecine*, 39-42.

Hasin, T., R. Dagan, G. Boutboul, E. Derazne, O. Atias & D. Cohen (2007) Socioeconomic correlates of antibody levels to enteric pathogens among Israeli adolescents. *Epidemiology and Infection,* 135**,** 118-125.

Hlavsa, M. C., V. A. Roberts, A. R. Anderson, V. R. Hill, A. M. Kahler, M. Orr, L. E. Garrison, L. A. Hicks, A. Newton, E. D. Hilborn, T. J. Wade, M. J. Beach & J. S. Yoder (2011) Surveillance of Waterborne Disease Outbreaks and Other Health Events Associated with Recreational Water --- United States, 2007 -- 2008. *Morbidity and Mortality Weekly Report,* 60**,** 1-32.

Hoek, R., I. Oliver, M. Barlow, L. Heard, R. Chalmers & S. Paynter (2008) Outbreak of *Cryptosporidium parvum* among children after a school excursion to an adventure farm, South West England. *Journal of Water and Health,* 6.

Hu, W., K. Mengersen & S. Tong (2009) Spatial Analysis of Notified Cryptosporidiosis Infections in Brisbane, Australia. *Annals of Epidemiology,* 19**,** 900-907.

Hunter, P. R., S. J. Hadfield, D. Wilkinson, I. R. Lake, F. C. D. Harrison & R. M. Chalmers (2007) Suptypes of *Cryptosporidium parvum* in humans and disease risk. *Emerging Infectious Diseases,* 13**,** 82-88.

IBM SPSS for Windows, Release 20 (2011) SPSS, Inc., Chicago.

Jagai, J. S., J. K. Griffiths, P. H. Kirshen, P. Webb & E. N. Naumova (2010) Patterns of protozoan infections: spatiotemporal associations with cattle density. *Ecohealth,* 7**,** 33-46.

Jamieson, R., R. Gordon, D. Joy & H. Lee (2004) Assessing microbial pollution of ruralsurface waters: A review of current watershed modeling approaches. *Agricultural Water Management,* 70**,** 1-17.

Jarup, L. 2000. The role of geographical studies in risk assessment. In *Spatial Epidemiology: Methods and Applications,* eds. P. Elliott, J. Wakefield, N. G. Best & D. Briggs. New York: Oxford University Press.

Kammann, E. E. & M. P. Wand (2003) Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 52**,** 1-18.

Kassenborg, H. D., C. W. Hedberg, M. Hoekstra, M. C. Evans, A. E. Chin, R. Marcus, D. J. Vugia, K. Smith, S. D. Ahuja, L. Slutsker & P. M. Griffin (2004) Farm visits and undercooked hamburgers as major risk factors for sporadic *Escherichia coli* O157:H7 infection: Data from a case-control study in 5 FoodNet sites. *Clinical Infectious Diseases,* 38 (Suppl 3)**,** S271-8.

King, G. & L. Zeng (2001) Logistic Regression in Rare Events. *Political Analysis,* 9**,** 137-163.

Kistemann, T., S. Zimmer, I. Vagsholm & Y. Andersson (2004) GIS-supported investigation of human EHEC and cattle VTEC O157 infections in Sweden: Geographical distribution, spatial variation and possible risk factors. *Epidemiology and Infection,* 132**,** 495-505.

Klumb, C., T. Robinson, E. Cebelinski, B. H. Alexander & K. Smith. 2010. Evaluation of Potential Relationships Between Well Water, Geologic Sensitivity, Cattle Density, and *Cryptosporidium parvum* Infections in Minnesota, 2000-2008 [abstract]. In *International Conference on Emerging Infectious Diseases 2010 Slide Session Abstracts*.

Krieger, N., P. Waterman, J. Chen, M.-J. Soobader, S. V. Subramanian & R. Carson (2002) Zip Code caveat: Bias due to spatiotemporal mismatches between Zip Codes and US Census-defined geographic areas -- The Public Health Disparities Geocoding Project. *American Journal of Public Health,* 92**,** 1100-1102.

Lake, I. R., F. C. D. Harrison, R. M. Chalmers, G. Bentham, G. Nichols, P. R. Hunter, R. S. Kovats & C. Grundy (2007) Case-control study of environmental and social factors influencing cryptosporidiosis. *European Journal of Epidemiology,* 22**,** 805-811.

Lu, Z., D. Lapen, A. Scott, A. Dang & E. Topp (2005) Identifying Host Sources of Fecal Pollution: Diversity of *Escherichia coli* in Confined Dairy and Swine Production Systems. *Applied and Environmental Microbiology,* 71**,** 5592-5998.

Michel, P., J. B. Wilson, S. W. Martin, R. C. Clarke, S. A. McEwen & C. L. Gyles (1999) Temporal and geographical distributions of reported cases of *Escherichia coli* O157:H7 infection in Ontario. *Epidemiology and Infection,* 122**,** 193-200.

Naumova, E. N., E. O'Neil & I. MacNeil (2005) INFERNO: a system for early outbreak detection and signature forecasting. *Morbidity and Mortality Weekly Report,* 54**,** 77-83.

O'Brien, S. J., K. A. Goutam & C. Gilham (2001) Contact with Farming Environment as a Major Risk Factor for Shiga Toxin (Vero Cytotoxin)-Producing *Escherichia coli* O157 Infection in Humans. *Emerging Infectious Diseases,* 7**,** 1049-1051.

Olsen, S. J., G. Miller, T. Breuer, M. Kennedy, C. Higgins, J. Walford, G. McKee, K. Fox, W. Bibb & P. Mead (2002) A Waterborne Outbreak of *Escherichia coli* O157:H7 Infections and Hemolytic Uremic Syndrome: Implications for Rural Water Systems. *Emerging Infectious Diseases,* 8**,** 370-375.

Parajuli, P. B., K. R. Mankin & P. L. Barnes (2009) Source specific fecal bacteria modeling using soil and water assessment tool model. *Bioresource Technology,* 100**,** 953-963.

Pohlmann, J. T. & D. W. Leitner (2003) A Comparison of Ordinary Least Squares and Logistic Regression. *Ohio J. Sci.,* 103**,** 118-125.

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

R: A language and environment for statistical computing. 2.15.1, R Foundation for Statistical Computing, Vienna, Austria.

Rangel, J. M., P. H. Sparling, C. Crowe, P. M. Griffin & D. L. Swerdlow (2005) Epidemiology of *Escherichia coli* O157:H7 Outbreaks, United States, 1982-2002. *Emerging Infectious Diseases,* 11**,** 603-609.

Robertson, W. & G. Yasvinski. 2012. Exposure. In *World Health Organization (WHO). Animal Waste, Water Quality and Human Health,* eds. A. Dufour, J. Bartram, R. Bos & V. Gannon. London, UK: IWA Publishing.

Robinson, T. J., E. A. Cebelinski, C. Taylor & K. Smith, E. 2010. Evaluation of the positive predictive value of rapid assays used by clinical laboratories in Minnesota for the diagnosis of cryptosporidiosis. e53-e55. Clinical Infectious Diseases.

Said, B., F. Wright, G. L. Nichols, M. Reacher & M. Rutter (2003) Outbreaks of infectious disease associated with private drinking water supplies in England and Wales 1970 - 2000. *Epidemiology and Infection,* 130**,** 469-479.

Sakuma, M., U. Mitsuyoshi & O. Nobuhiko (2006) Verocytotoxin-producing *Escherichia coli*, Japan, 1999-2004. *Emerging Infectious Diseases,* 12**,** 323-325.

Scallan, E., R. M. Hoekstra, F. J. Angulo, R. V. Tauxe, M.-A. Widdowson, S. L. Roy, J. L. Jones & P. M. Griffin (2011) Foodborne Illness Acquired in the United States - Major Pathogens. *Emerging Infectious Diseases,* 17**,** 7-15.

Scallan, E., B. E. Mahon, R. M. Hoekstra & P. M. Griffin (2013) Estimates of Illnesses, Hospitalizations and Deaths Caused by Major Bacterial Enteric Pathogens in Young Children in the United States. *The Pediatric Infectious Disease Journal,* 32**,** 217-221.

Sehgal, R., Y. Kumar & S. Kumar (2008) Prevalence and geographical distribution of *Escherichia coli* 0157 in India: a 10-year survey. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 102**,** 380-383.

Strachan, N. J. C., G. M. Dunn, M. E. Locking, T. M. S. Reid & I. D. Ogden (2006) *Escherichia coli* O157: Burger bug or environmental pathogen? *International Journal of Food Microbiology,* 112**,** 129-137.

Suresh, K. G., G. A. Toranzos, R. Fayer, V. Nissaparton, R. Olveda, N. Ashbolt & V. Gannon. 2012. Assessing the importance of zoonotic waterborne pathogens. In *World Health Organization (WHO). Animal Waste, Water Quality and Human Health,* eds. A. Dufour, J. Bartram, R. Bos & V. Gannon. London, UK: IWA Publishing.

Tennessee Department of Environment and Conservation. 2008. Year 2008 303(d) List. Tennessee Department of Environment and Conservation, Division of Water Pollution Control.

Tikkinen, K. A. O., A. Auvinen, T. M. Johnson, II, J. P. Weiss, K. Tapani, A. Tlitinen, O. Polo, M. Partinen & T. L. J. Tammela (2009) A systematic Evaluation of Factors Associated With Nocturia - The Population-based FINNO Study. *American Journal of Epidemiology,* 170.

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

United States Census Bureau. 2012. TIGER/Line Shapefiles and TIGER/Line Files.

United States Department of Agriculture (2009) 2007 Census of Agriculture, Tennessee State and County Data Volume 1, Chapter 2: County level data. http://www.agcensus.usda.gov/Publications/2007/Full_Report/Volume_1,_Chapter_2_County_Level/Tennessee/tnv1.pdf (last accessed 02-03-09).

United States Environmental Protection Agency. 2009. National Summary of Impaired Waters and TMDL Information. In *Watershed Assessment, Tracking and Environmental Results*. United States Environmental Protection Agency.

Valcour, J. E., M. Pascal, S. A. McEwen & J. B. Wilson (2002) Associations between indicators of livestock farming intensity and incidence of human Shiga toxin-producing *Escherichia coli* infection. *Emerging Infectious Diseases,* 8**,** 252-257.

Yoder, J. S. & M. J. Beach (2009) *Cryptosporidium* surveillance and risk factors in the United States. *Experimental Parasitology,* 124**,** 31-39.

Yoder, J. S., R. M. Wallace, S. A. Collier, M. J. Beach & M. C. Hlavsa (2012) Cryptosporidiosis Surveillance - United States, 2009-2010. *Morbidity and Mortality Weekly Report,* 61**,** 1-12.

Young, L. J., C. A. Gotway, J. Yang, G. Kearney & C. DuClos (2009) Linking Health and Environmental Data in Geographic Analysis: It's so Much More than Centroids. *Spatial and Spatio-temporal Epidemiology 1***,** 73-84.

VITA

Ingrid E. Luffman earned a Bachelor of Science degree in Honors Mathematics-Science from the University of Ottawa, Canada in 1993. She received a Master of Science degree from the Department of Earth Sciences at the University of Ottawa in 1997. Her thesis research investigated the effect of spatial statistics of water flow parameters on the flow of solutes in groundwater. In 1997, she began teaching as adjunct faculty at East Tennessee State University in the then Department of Geography, Geology, and Geomatics. In 2005, she accepted a position as Lecturer at East Tennessee State University in the Department of Economics, Finance, Geography and Urban Studies. While pursuing the doctorate, she worked as a Lecturer in the Department of Economics, Finance, Geography and Urban Studies, which in 2008 was restructured into the Department of Geosciences. In 2013, Ingrid was awarded the Doctorate of Philosophy in Geography by the University of Tennessee.