SPATIOTEMPORAL ANALYSIS OF EASTERN EQUINE ENCEPHALITIS HUMAN INCIDENCE

by

Jessika L. Ava

A Thesis Submitted to the Faculty of the

MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

In Partial Fulfilment of the Requirements

For the Degree of

MASTER OF SCIENCE WITH A MAJOR IN BIOSTATISTICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

STATEMENT BY AUTHOR

The thesis titled Spatiotemporal Analysis of Easter Equine Encephalitis Human Incidence has been submitted in partial fulfillment of the requirements for a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained by the author.

Jessika L. Ava

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

_____________________________ _____May 2, 2017______

Dr. Chengcheng Hu, Professor of Biostatistics Date

ACKNOWLEDGMENTS

I would first like to thank my thesis committee chair Dr. Heidi Brown of the Mel and Enid Zuckerman College of Public Health at University of Arizona. Dr. Brown has offered help, support, and patience; she consistently allowed this project to be my own work, but offered guidance in the right direction when necessary.

I would also like to thank thesis director, Dr. Chengcheng Hu and committee member, Dr. Melanie Bell of the Mel and Enid Zuckerman College of Public Health at the University of Arizona. Dr. Hu and Dr. Bell have offered guidance and support throughout this Master's degree.

Lastly, I would like to thank my fellow classmates, friends, and family for all the support over the last two years.

TABLE OF CONTENTS

ABSTRACT

Spatial and temporal components play a critical role in explaining variability across geographic regions and time, and are necessary components to space-time epidemiological research.

Until recent years, most spatial epidemiological studies have used simple space-time analyses, but the continuous advancements in statistical modeling software and geographic information systems have made more complex spatial analyses readily available. However, methods may be problematic and several ongoing statistical weaknesses have been documented, including failing to account for three significant correlative factors - spatial, temporal, and spatiotemporal autocorrelations.

Using Eastern Equine Encephalitis (EEE) human incidence data, this Master's thesis aimed to answer the research question, *is there a northeastern shift in human EEE incidence within the United States,* by identifying a statistical model that adjusts for spatial, temporal, and spatiotemporal autocorrelations.

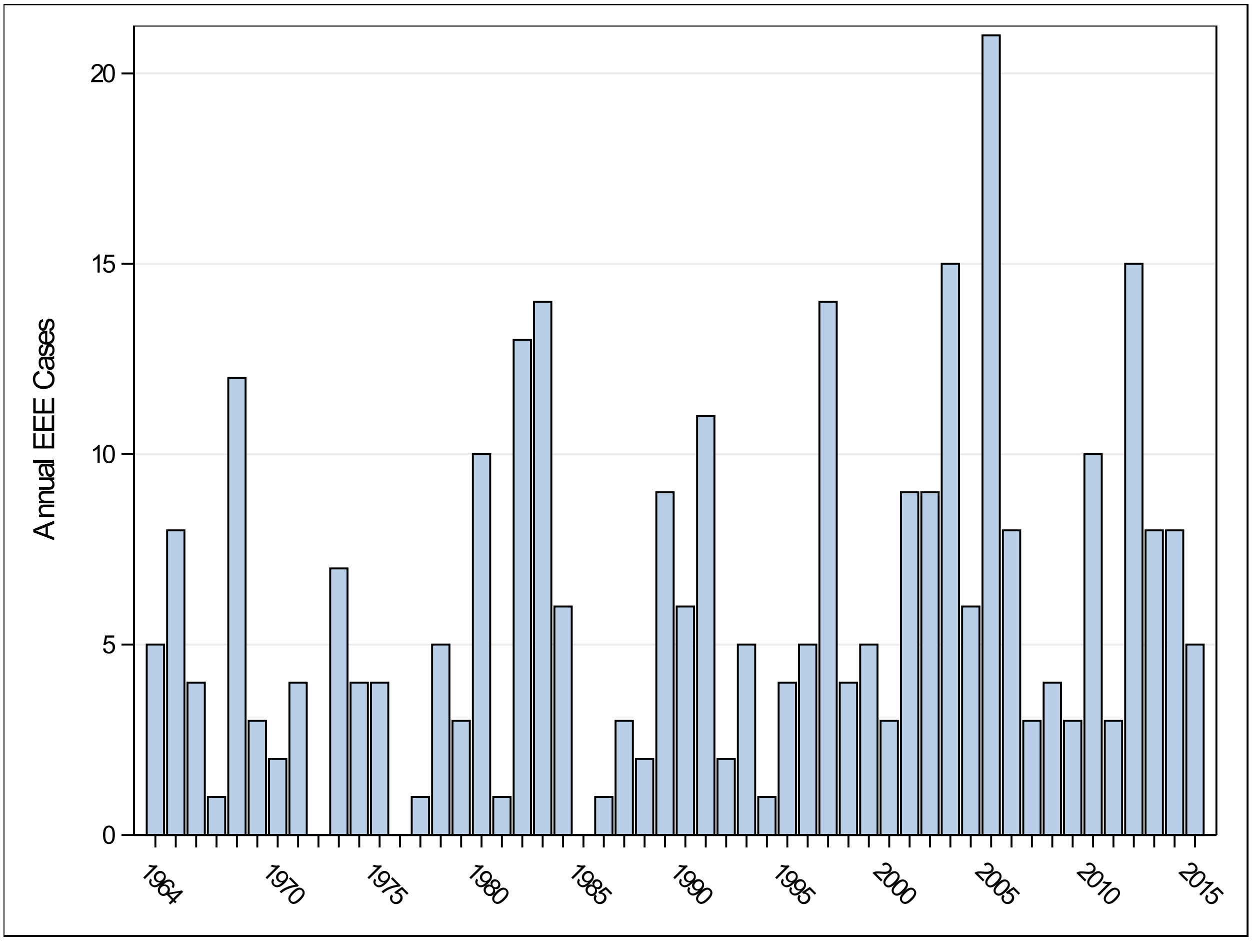
This thesis introduced the *spatial autoregressive distributed lag* (SADL) model, a model that adjusts for spatial, temporal, and spatiotemporal autocorrelations. However, results demonstrated that EEE is too rare an event for the SADL model to be appropriate, and a non-autocorrelation model was used as the final model. Results showed that EEE incidence is significantly increasing over time for all infected regions of the United States, with a significant difference of 1.4 cases/10 million between 1964 and 2015. Results did not demonstrate a northeastern shift in EEE incidence as the northeastern US had the highest expected incidence across the entire study period (1964-1967: 2.9/10 million; 2012-2015: 6.8/10 million), but results did demonstrate that the northeastern US had the quickest increasing risk for EEE as compared to other infected regions of the US with an increase in expected incidence of 3.9/10 million between 1964 and 2015.

CHAPTER 1: BACKGROUND AND INTRODUCTION

1.1 Eastern Equine Encephalitis Background

Discovered in 1933 and predominantly affecting equines, Eastern Equine Encephalitis (EEE) is a *Culesita melanura* mosquito-borne viral disease endemic to the eastern United States (Armstrong and Andreadis 2013). Until the last decade, human outbreaks were isolated, rare, and restricted primarily to the Southeast and Mid-Atlantic. However, in 2005, a human outbreak occurred in the northeastern state of New Hampshire, followed by ongoing increasing frequency in novel northeastern regions. Proposed reasons for this seemingly northward expansion are climate change leading to more vector-favorable seasonal temperatures, suburban development nearer to vector habitats, or wetland restoration increasing habitat for the primary vector (Armstrong and Andreadis 2013). Figure 1 shows human EEE cases as reported to the Centers for Disease Control and Prevention between 1964 and 2015.

Figure 1: Annual Eastern Equine Encephalitis Human Cases as reported to the CDC, 1964-2015.



EEE virus results in neurological impairment and mortality in both humans and equines (Armstrong and Andreadis 2013). Although human cases are rare, ranging from 0 to 21 annual US cases between the years 1964 and 2015, the EEE virus has the highest mortality of all mosquito-borne pathogens in North America (CDC 2015). The human case fatality rate is 35 to 75 percent, with approximately half of survivors experiencing permanent neurologic sequelae (Deresiewicz et al. 1997). Equine case fatality rate is 75 to 90 percent (APHIS 2008). Veterinary vaccines exist for horses, but currently no vaccine or effective treatment exists for humans (Armstrong and Andreadis 2013).

1.2 Sample Population and Data Sources

Human EEE data were extracted from the Centers for Disease Control and Prevention (2015). A total of 309 new cases occurred within 24 infected states between the years of 1964 and 2015 (mean annual incidence within United States,1.9/100million; mean annual incidence within infected states, 3.2/100million). Cases occur in the southeastern, east central, and northeastern United States and included the states of Texas, Arkansas, Louisiana, Mississippi, Florida, Georgia, Alabama, Virginia, North Carolina, South Carolina, Wisconsin, Michigan, Indiana, Delaware, Connecticut, Maryland, New Jersey, New York, Pennsylvania, Vermont, Rhode Island, Maine, Massachusetts, and New Hampshire. See summary Table 1.

Table 1: Human EEE annual cases and incidence per infected state and region, 1964-2015

Total Cases, 1964-2015	309
No. Infected States	24
	mean (sd, variance)
Annual Cases per Infected States	0.25 (0.82, 0.67)
Annual Cases per Region*	0.98 (1.81, 3.28)
Annual Incidence per Infected States	3.2 /100 million
Annual Incidence per Region*	4.6/100 million

Regions defined as: 1-TX, AR, LA, MS; 2-FL, GA, AL; 3- VA, NC, SC; 4- WI, MI, IN; 5- DE, CT, MD, NY, NY, PA; 6- VT, RI, ME, MA,NH. See Chapter 2 for more information on regions.

1.3 Adjusting for High Zero Outcomes

Approximately 86 percent of the outcome data are 0, indicating no new cases. The high degree of zero outcomes poses a challenge to this analysis as it can lead to data overdispersion (mean<variance) and potentially biased results (Rose et al. 2006). Two methods are utilized to adjust for the high number of zero outcomes, a negative binomial distributed model and aggregating data.

Negative Binomial Distribution

A Poisson or negative binomial distribution is appropriate for count outcome data, which is the case in this analysis (Rose et al. 2006). However, the data overdispersion that might result from a high degree of zero outcomes violates the Poisson mean/variance equality assumption and may potentially underestimate standard errors. The problem of overdispersion may be overcome using a negative binomial distribution, as it includes a dispersion parameter that accounts for overdispersion (Rose et al. 2006).

Aggregating Data

To further overcome issues that might arise from the high number of zero outcomes, data are collapsed into geographic regions and/or into uniform time intervals rather than individual states and individual years. By aggregating data, the additional non-zero observations that become available when outcomes are grouped can now be analyzed. This provides a more robust analysis while also maintaining the observed trends of the original data (see Chapter 4: Discussion).

1.4 Prior Research and Common Weaknesses in Space-Time Analysis

Among the most common weaknesses in epidemiology space-time analyses are oversight of spatial and/or temporal and/or spatiotemporal autocorrelations (defined in Section 1.6) (Ocana-Riola 2010; Beale et al. 2008; Elliot et al. 2001; Abellan et al. 2008). Neglecting to adjust for these autocorrelations violates the independently identically distributed random observation assumption of linear regression statistical models and can result in inaccurate and biased estimates.

1.5 Purpose of Study and Research Question

Using EEE human incidence data, this Master's thesis aims to overcome commonly seen weakness of failing to adjust for autocorrelations while answering the research question, *is there a northeastern shift in human EEE incidence within the United States*? To achieve this, a statistical model that merges spatial and temporal factors into a single analysis while adjusting for spatial, temporal, and spatiotemporal autocorrelations (as defined in Section 1.6) is identified. By aggregating the data in space and time, the proposed model also overcomes the limitation of rare and excessive zero outcomes, common to diseases with low incidence rates such as EEE. Since this thesis' main focus is to identify a model that adjusts for autocorrelation factors, potential environmental predictors of EEE incidence are omitted and are suggested for future research.

1.6 Adjusting for Correlations

Defining Correlations Relevant to Current Research

The following describes relevant correlations as they are defined for this project; however, varying terms are used within the literature. *Correlation* describes the degree of relationship between a pair of variables (Fitzmaurice et al. 2011, 27). *Autocorrelation* describes the correlation of values of the same variable at different time points or different spatial units. *Spatial autocorrelation* describes the correlation of values of the same variable measured in adjacent spatial units; as an example, the measure of correlation between observations in region A and the same variable observed in adjacent region B at the same time point. *Temporal autocorrelation* describes the correlation between values of the same variable at different times; as an example, the correlation between subject A's measurement at time 1 and subject A's measurement at time 2 (De Smith 2015, ch.9). *Spatiotemporal autocorrelation* describes the correlation of observations within adjacent spatial units at different times; as an example, the correlation between observations in region A at time 1 and observations in adjacent region B at time 2 (Beale et al. 2008). Models that address the correlations are discussed below.

General First Order Spatial Autoregressive Distributed Lag Model

This section describes the space-time analysis model known as the *general first order spatial autoregressive distributed lag* (general SADL) model developed by Elhorst (2001). A general SADL model is applicable to examining the relationship between a count or density outcome and predictors as they are observed over both space and time, while adjusting for spatial, temporal, and spatiotemporal autocorrelations (Elhorst 2001). Thus, the model overcomes the commonly seen weakness of failing to adjust for autocorrelations. The autocorrelations are explicitly expressed through modeling response variables as covariates, including a temporal lag response variable (response variable at previous time measurement), an adjacent region response variable, and an adjacent region temporal lag response variable (adjacent region, previous time measurement).

The general first order SADL model is defined as:

$$\mathbf{Y}_t = \beta_1\mathbf{Y}_{t-1} + \beta_2\mathbf{WY}_t + \beta_3\mathbf{WY}_{t-1} + \beta_4\mathbf{X}_t + \beta_5\mathbf{X}_{t-1} + \beta_6\mathbf{WX}_t + \beta_7\mathbf{WX}_{t-1} + \boldsymbol{\varepsilon}_1$$

where $\mathbf{Y}_t$ denotes an nx1 vector consisting of one observation for every spatial unit of the outcome variable in the t^{th} time period (t=1,…n). $\mathbf{X}_t$ denotes an nx1 vector of non-outcome predictors. **W** denotes an nxn weight matrix describing the geographical units. t-1 denotes the first order temporal lag, a variable multiplied by **W** denotes its spatially lagged value. $\boldsymbol{\varepsilon}_1$ is an nx1 vector containing error terms and $\sim N(0, \sigma^2)$, where $\sigma^2 =$

variance. The null hypothesis states $\beta_1=\ldots=\beta_{7=}0$, no spatial or temporal effects on the outcome (Elhorst 2001).

GEE and GLM

While attempting to utilize the Elhorst SADL model's explicit temporal autocorrelation adjustment methods, this project will also analyze both negative binomial distributed Generalized Linear Models (GLM) and Generalized Estimating Equations (GEE). GLM and GEE differ in that a GEE accounts for potential covariation that may occur from repeated measurement through the modeling of random effects, while GLM does not account for any potential covariation from repeated measurements.

1.7 Potential Predictor Variables

Table 2 lists the description of each potential autocorrelation predictor variable and how the predictor variable relates to the SADL model. The outcome variable is cases aggregated by region for Analysis 1 and cased aggregated by region and time for Analysis 2 (see Chapter 3 for Analyses 1 and 2 discussions). Temporal variable is years 1964-2015 modeled individually for Analysis 1 or, due to the excessive number of zero outcomes, as 4-year intervals for Analysis 2. Northern and southern adjacent region cases, as well as cases lagged one year are also included as potential covariates. Lagged variables are one year lag for Analysis 1 and 4-year interval lags for Analysis 2.

Table 2: Potential Predictor Variables for Analysis 1 and 2 and the Autocorrelation as it Relates to the SADL Model

Potential Predictor Variable	Variable Type	Description
Year	Temporal Explanatory Var.	1964-2015
Aggregated Year	Temporal Explanatory Var.	Years 1964-2015 aggregated into 4 year intervals
Region	Spatial Correlation Var.	Region 1-6
Regional Cases Lag	Spatial Correlation Var.	Regional Cases, Previous Year or Previous Year aggregate
North and South Adjacent Regions Cases	Spatial Correlation Var.	Regional Cases for Adjacent Regions
North and South Adjacent Regions Cases Lag	Spatiotemporal Correlation Var.	Regional Cases for Adjacent Regions, Previous Year or Previous Year aggregate
Log(regional cases)	Dependent var.	Response Var.
Log(region population)	Offset	Offset to account for rate dependent var.

CHAPTER 2: METHODOLOGY

2.1 Regional Aggregation: Aggregating States

The data were first collapsed into 6 regions. Regions shift northeastern as numerical designation increase, with Region 6 being the most northeastern. Regional incidence was defined as number of new cases per region population per year. Regions were selected due to geographic proximity and approximately similar area. Regions were defined as follows, Region 1: Texas, Arkansas, Louisiana, Mississippi; Region 2: Florida, Georgia, Alabama; Region 3: Virginia, North Carolina, South Carolina; Region 4: Wisconsin, Michigan, Indiana; Region 5: Delaware, Connecticut, Maryland, New Jersey, New York, Pennsylvania; Region 6: Vermont, Rhode Island, Maine, Massachusetts, New Hampshire.

2.2 Temporal Aggregation: Aggregating Years

To further reduce the number of zero outcomes, the data were also aggregated by time in Analysis 2. First, to assess the maintenance of the trend of zero outcomes over time, four aggregates were observed, 4 year, 6 year, 8 year, and 10 year aggregates. The optimum dataset used an aggregate that balanced between the shortest interval possible (thus maintaining the highest number of data points) and preserving the general trend observed over time, while also providing a viable reduction from individual year zero outcome percentages.

2.3 Analyses

Two analyses were conducted, Analysis 1: Regional Aggregate and Analysis 2: Regional Aggregate and Time Aggregate. Due to data overdispersion and count outcome data, a negative binomial distribution was used. Due to a rate outcome, cases per region, an offset term was used and defined as log(regional population).

Model Development

For Analysis 1, a regional aggregate outcome variable was used and time was measured as individual years (1964-2015). For Analysis 2, a regional and time aggregated outcome was used and time was measured in 4 year intervals. For both analyses, first, eight negative binomial General Linear Models (GLM) adjusted for linear time were considered as potential models, with each model increasing in the number of potential autocorrelation covariates. Next, eight negative binomial GLMs adjusted for categorical time and the same potential covariates were considered as potential models. Next, similar to the GLM, eight negative binomial Generalized Estimating Equations (GEE) using linear time and the variable, *state,* as repeated random effects and the variable, *time,* as within subjects random effects were considered as potential models, with each model increasing in the number of potential autocorrelation covariates. Followed by the same eight negative binomial GEEs but adjusted for categorical time. Table 3 lists each compared model.

Model Selection

Only models that converged and demonstrated all significant covariates were considered for model selection. Models were compared using the lowest goodness of fit statistic (AIC for GLM or QIC for GEE), reasonable parameter estimates, parsimony to achieve the simplest model, and the model's ability to account for potential covariation from repeated time measurements through modeling random effects. AIC and QIC are goodness of fit measurements used to compare similar models, with the lowest value of compared models indicating the best fit model (Rosner 2011).

Assessing Goodness of Fit

Lastly, the final model for Analysis 1 and 2 were each assessed for goodness of fit, including a residual versus observed plot, and if appropriate, a Chi Square goodness of fit test. A residual plot displays a model's Pearson residual and observed data points. A Pearson residual is defined as the observed values minus the predicted values (Rosner 2011). Each point's distance from the zero line shows the difference between the predicted and observed values. A residual vs. predicted plot is used to provide evidence as to how well the model fits the data and is used to detect outliers, non-linearity, and unequal error variance. A good fit model will have a residual vs observed plot that is randomly distributed around the zero value and fairly symmetrical.

A Chi Square goodness of fit test is used to test the association between the modeled data and observed data (Rosner 2011). Currently, no statistical test for overall goodness of fit is available for GEE and a Chi square test is not appropriate for GEE. (SAS 2016).

Table 3: Potential Models for Analyses

Model	Variables	Time
NB GLM	Year, Region	Categorical
NB GLM	Year, Region, Region Lag	Categorical
NB GLM	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases	Categorical
NB GLM	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases N Neighbor Region Cases Lag, S Neighbor Region Cases Lag	Categorical
NB GLM	Year, Region	Linear
NB GLM	Year, Region, Region Lag	Linear

NB GLM	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases	Linear
NB GLM	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases N Neighbor Region Cases Lag, S Neighbor Region Cases Lag	Linear
NB GEE	Year, Region	Categorical
NB GEE	Year, Region, Region Lag	Categorical
NB GEE	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases	Categorical
NB GEE	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases N Neighbor Region Cases Lag, S Neighbor Region Cases Lag	Categorical
NB GEE	Year, Region	Linear
NB GEE	Year, Region, Region Lag	Linear
NB GEE	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases	Linear
NB GEE	Year, Region, Region Lag N Neighbor Region Cases, S Neighbor Region Cases N Neighbor Region Cases Lag, S Neighbor Region Cases Lag	Linear

NB: negative binomial. GLM: General Linear Model. GEE: Generalized Estimating Equations. Year is individual year for Analysis 1 and aggregated years for Analysis 2. Lin represents linear time, Cat represents categorical time. Covariates are defined in Table 2.

2.4 Final Model Selection

Lastly, the best fit models of Analysis 1 and Analysis 2 were compared to determine the overall final best fit model that was used to answer the thesis research question. Models were compared using the residual vs. observed plot, reasonableness of parameter estimates, parsimony, the additional data made available by not aggregating time, and the model’s ability to account for potential covariation from repeated time measurements through modeling random effects.

Although AIC or QIC is used to compare between similar models, it is not relevant when comparing models of differing outcome data. Since Analysis 1 consisted of outcome data aggregated by region and Analysis 2 consistsed of outcome data aggregated by region and time, the AIC or QIC was not an appropriate method of comparison between Analysis 1 and 2.

2.5 ANCOVA to Answer Research Question

An ANVOCA was conducted (H_0: equal E(Y) for all regions and times) to determine if significant differences exist between annual regional incidences. This will assist in answering the research question, *Is there a northeastern shift in EEE cases over time?* (See Chapter 4 Discussion.)

2.6 Statistical Software

SAS statistical software and the GENMOD procedure were used. See Appendix for relevant SAS code.

CHAPTER 3: RESULTS

3.1 Analysis 1: Regional Aggregate

Collapsing states into 6 regions resulted in a reduction of zero outcomes from 86 to 57 percent.

Model Comparison Results

Table 4 shows each model compared in Analysis 1, including the generalized linear model (GLM) or general estimating equation (GEE), the covariates, linear or categorical time, model convergence, if all covariates were significant, and values of AIC for GLMs or QIC for GEEs. The best fit model is highlighted (see Chapter 4 for discussion of model selection).

Table 4: Summary of Analysis 1 Models with Best Fit Model Highlighted

Model	Variables	Linear Time(Y/N)	Model Converge(Y/N)	All Sign. Covariates	AIC or QIC
NB GLM	Year Region	N	N	N	2956
NB GLM	Year Region RegionLag	N	N	N	2910
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	N	N	N	2780
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	N	N	N	2781
NB GLM	Year Region	Y	Y	Y	3213
NB GLM	Year Region RegionLag	Y	Y	Y	3171
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	Y	Y	N	3169

NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	Y	Y	N	3165
NB GEE	Year Region	N	N	N	.
NB GEE	Year Region RegionLag	N	N	N	.
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	N	N	N	.
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	N	N	N	.
NB GEE	Year Region	Y	Y	Y	1108
NB GEE	Year Region RegionLag	Y	Y	N	1128
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	Y	Y	N	1151
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	Y	Y	N	1180

NB:Negative Binomial; GLM: General Linear Model; GEE: Generalized Estimating Equation, N:no, Y:yes. Covariates are defined in Table 2.

3.2 Analysis 2: Regional Aggregate, Time Aggregate

Table 5 demonstrates that all time aggregates showed the same general trend towards an overall decrease in percentage of zero outcomes over time, and all demonstrated a viable reduction in the percentages of zero outcomes that are seen in individual years. The 4-year interval was chosen due to its aggregation of the fewest years, thus allowing the dataset to maintain the highest number of data points and thereby providing the most information for more valid results. Additionally, 4 is divisible by the total 52 years, providing for an equal number of years in each time interval.

Further collapsing the regional data by time resulted in a reduction in zero outcomes from 57.5 to 21 percent.

Table 5: Percentage of Regional Zeros Outcome by Aggregated Years and Individual Years, for Select Years This table shows that the overall trends in percentage of zero outcomes remains the same for data aggregated into 4 year, 6 year, 8 year, and 10 year intervals, where the percentage of regions zero outcomes tends to demonstrate a general decrease over time.

10 Year Aggregates	**1964-1973**	**1974-1983**	**1984-1993**	**1994-2003**	**2004-2013**	**2014-2015**			
%0 outcomes	73	62	62	51	45	33			
8 Year Aggregates	**1964-1971**	**1972-1979**	**1980 – 1987**	**1988-1995**	**1996-2003**	**2004-2011**	**2012-2015**		
%0 Outcomes	73	72	60	58	47	47	33		
6 Year Aggregate	**1964-1969**	**1970-1975**	**1976-1981**	**1982-1987**	**1988-1993**	**1994-1999**	**2000-2005**	**2006-2011**	**2012-2105**
% 0 outcomes	75	63	77	58	55	55	48	45	33
4 Year Aggregates	**1964-1967**	**1972-1975**	**1976-1979**	**1984-1987**	**1988-1991**	**1996-1999**	**2000-2003**	**2004-2007**	**2012-2015**
% 0 outcome	75	60	83.3	73	56.2	48	45.8	52	33.3
Individual Years	**1964**	**1972**	**1976**	**1984**	**1988**	**1996**	**2000**	**2004**	**2012**
% 0 outcome	88	100	100	29	75	88	67	67	54

For space efficiency purposes, only select years are shown.

Model Comparison Results

Table 6 shows each model compared in Analysis 2, including the GLM or GEE, the covariates, linear or categorical time, model convergence, covariate significance, and the AIC or QIC value. The best fit model is highlighted (see Chapter 4 for discussion of model selection).

Table 6: Summary of Analysis 2 Models with the Best Fit Model Highlighted

Model	**Variables**	**Linear Time(Y/N)**	**Model Converge(Y/N)**	**All Sign. Covariates**	**AIC or QIC**
NB GLM	Year Region	N	Y	N	5605
NB GLM	Year Region RegionLag	N	Y	N	4826
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	N	Y	N	2781.6
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	N	Y	N	2781.2
NB GLM	Year Region	Y	Y	Y	5837
NB GLM	Year Region RegionLag	Y	Y	Y	5206

NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	Y	Y	Y	3423
NB GLM	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	Y	Y	N	3143
NB GEE	Year Region	N	Y	N	-6877
NB GEE	Year Region RegionLag	N	Y	N	-8375
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	N	Y	N	-5115
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	N	Y	N	-4998
NB GEE	Year Region	Y	Y	Y	-6185
NB GEE	Year Region RegionLag	Y	Y	Y	-5985
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases	Y	Y	N	-3009
NB GEE	Year Region RegionLag Nneighbor_Region_Cases SNeighbor_Region_Cases N_Reg_CasesLag S_Reg_CasesLag	Y	Y	N	-3347

NB:Negative Binomial; GLM: General Linear Model; GEE: Generalized Estimating Equation, N:no, Y:yes. Covariates are defined in Table 2.

3.3 ANCOVA

The ANCOVA model rejected the H_0 hypothesis of equal expected incidence for all regions and times, and demonstrated a significant difference exists in expected incidences between all regions and time points.

CHAPTER 4: DISCUSSION

4.1 Analysis 1: Regional Aggregate

Model Selection

Of the models compared for Analysis 1, three models converged and contained only significant covariates, 1 GLM and 2 GEEs. The linear time GLM using year and region as covariates, show the best fit of the GLMs since it was the only model to fit the criteria of convergence and significant covariate; the linear time GEE using year and time as covariates is the best fit of the GEEs as demonstrated by the lowest QIC.

Between the best fit GLM and GEE, the GEE is chosen due to its ability to measure potential covariance from repeated time measurements through the use of modeling a random effect. The best fit model from Analysis 1 is the negative binomial GEE with linear time and categorical region covariates.

The model equation is:
$Log(E(Y_{ij})/p) = -47.53X_1 + 0.016X_{j2} + -2.101X_3 + -0.523X_4 + -1.270X_5 + -1.792X_6 + -1.665X_7 + 0X_8 + Zu$

where Y_{ij} represents case count for ith region at j time, p represents regional population, X_1 represents intercept and equals 1 for all subjects and occasions, X_{j2} represents years at j time, X_3= 1 if i=Region 1 and equals 0 otherwise, X_4= 1 if i=Region 2 and equals 0 otherwise, X_5=1 if i=Region 3 and equals 0 otherwise, X_6=1 if i=Region 4 and equals 0 otherwise, X_7=1 if i=Region 5 and equals 0 otherwise, X_8=1 if i=Region 6 and equals 0 otherwise, Zu represents random effect, i=1,….,6 region, j= 1964,…..,.2015.

Goodness of Fit

No covariates show high (>0.6) collinearity. Distribution of outcomes show a negative binomial distribution for each compared model. The negative binomial dispersion parameter value was significantly greater than 0 for all models, indicating the negative binomial distribution is an appropriate choice.

Figure 2: Residual vs. Observed Plot for Analysis 1 Best Fit Model

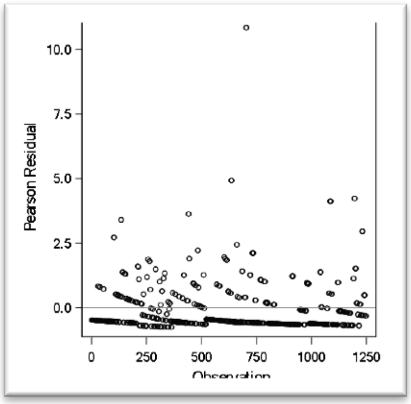


Figure 2 shows the Pearson residual versus observed scatter plot for the best fit Analysis 1 model, the negative binomial GEE using linear time and categorical region as covariates. The residual plot demonstrates a nonrandom distribution with most points falling near the y-axis, known as a y-axis unbalance pattern (Statwing 2016). A y-axis unbalance is often due to a negative binomial distribution with a high number 0/low values, or it could indicate a missing predictor, both elements which are present in this analysis. This plot demonstrates that the model may not fit the data extremely well, however, this can be expected since environmental predictors were intentionally omitted for these analyses as it is beyond the scope of this research project.

4.2 Analysis 2: Regional Aggregate, Time Aggregate

Model Selection

Of the models compared for Analysis 2, five models converged and contained only significant covariates, 3 GLMs and 2 GEEs. Of the GLMs, the GLM using categorical year, region, and lag as covariates, shows the best fit due to its lowest AIC; and of the GEEs the GEE using linear time and categorical region as covariates is the best fit of the GEE as demonstrated by the lowest QIC.

Between the best fit GLM and GEE, the GEE is chosen due to its ability to measure potential covariance from repeated time measurements through the use of modeling a random effect. The best fit model from Analysis 2 is the negative binomial GEE with linear time and categorical region covariates.

The model equation is:

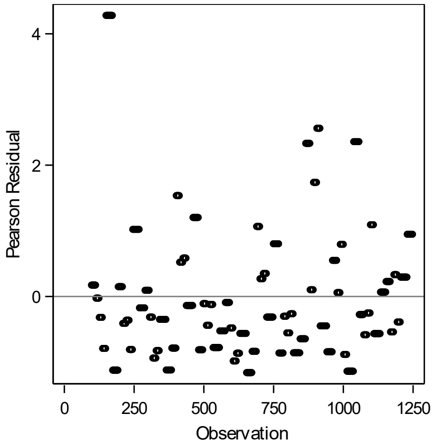
$$Log(E(Y_{ij})/p) = -50.0784X_1 + 0.0178X_{j2} + -2.8054X_3 + -0.4825X_4 + -1.6357X_5 + -2.1079X_6 + -1.5632X_7 + 0.00X_8 + Zu$$

where Y_{ij} represents case count for ith region at j time, p represents regional population, X_1 represents intercept and equals 1 for all subjects and occasions, X_{j2} represents years at j time, X_3= 1 if i=Region 1 and equals 0 otherwise, X_4= 1 if i=Region 2 and equals 0 otherwise, X_5=1 if i=Region 3 and equals 0 otherwise, X_6=1 if i=Region 4 and equals 0 otherwise, X_7=1 if i=Region 5 and equals 0 otherwise, X_8=1 if i=Region 6 and equals 0 otherwise, Zu represents random effect, i=1,….,6 region, j= [1964-67],….,[2012-15] 4-year aggregates.

Goodness of Fit

None of the covariates show high (>0.6) collinearity. Distribution of outcomes shows a negative binomial distribution for each compared model. The negative binomial dispersion parameter value is significantly greater than 0 for all models, indicating the negative binomial distribution was an appropriate choice. Figure 3 shows the Pearson residual versus observed scatter plot for Analysis 2. The residual plot demonstrates that the model provides a somewhat decent fit, as can be seen by the approximately random and symmetrical distribution around the zero line, with the exception of outliers. However, the plot demonstrates the model is still not an ideal fit as can be seen by the slight y-axis unbalance and the majority of points below the zero line demonstrating that the majority of predicted values are greater than observed values.

Figure 3: Residual vs. Observed Plot for Analysis 1 Best Fit Model



4.3: Final Model Selection

In both analyses, the negative binomial GEE is shown to be the best fit model, this is somewhat expected as the GEE takes into account the potential covariance that could occur with repeated time measurements; the observations within this dataset were taken over a period of 52 years (1964-2015), thus a model that accounts for this repeated measurement could be hypothesized to be the best fit model.

The best fit model of Analysis 1, provides more data for which to analyze since it did not include the additional time aggregation as seen in Analysis 2 data, thus Analysis 1 has the potential to provide more robust results that better represents the observed outcomes. The best fit model from Analysis 2 has an inherent weakness in that it further condenses the data into time aggregates, thus having a lower number of data to analyze than does the model from Analysis 1.

The residual vs observed plot for Analysis 1 (Figure 2) demonstrates y-axis unbalance, as characterized by data points congested towards the y-axis and with residuals becoming larger as the prediction moves from smaller to larger. A y-axis unbalance is often due to a missing predictor (Statwing 2016). This may be the case for this analysis as it is likely that the inclusion of environmental predictors such as weather, geographic ecosystem and others might improve the model, but environmental predictors are outside the scope of the analyses performed here. The residual vs observed plot for Analysis 2 (Figure 3) demonstrates a decently fit model, although not a superb fit model due to the presence of outliers, non-random clumping, and a slight y-axis unbalance. Furthermore, although the Analysis 2 residual plot does not demonstrate a strong y-axis unbalance, it hints at a y-axis unbalance that suggests a missing predictor.

Parameter estimates for best fit models of Analysis 1 and 2 are presented in Tables 7 and 8. Parameter estimates, standard errors, and confidence limit were very similar for both models and the difference between estimated incidences of the models may be negligible. Both models show a similar parameter estimate trend with a negative intercept, positive time association, and negative regional association. Overall, the differences in parameter estimates does not provide strong evidence for the best fit model.

Table 7: Analysis 1 Parameter Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-47.5298	12.1196	-71.2838	-23.7757	-3.92	<.0001
Year	0.0158	0.0061	0.0039	0.0278	2.60	0.0093
Region 1	-2.1011	0.0046	-2.1101	-2.0921	-458.37	<.0001
Region 2	-0.5249	0.0027	-0.5302	-0.5196	-192.82	<.0001
Region 3	-1.2703	0.0028	-1.2757	-1.2648	-458.92	<.0001
Region 4	-1.7922	0.0057	-1.8034	-1.7811	-315.84	<.0001
Region 5	-1.6646	0.0019	-1.6682	-1.6609	-891.60	<.0001
Region 6	0.0000	0.0000	0.0000	0.0000	.	.

Table 8: Analysis 2 Parameter Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-50.0784	12.3531	-74.2900	-25.8668	-4.05	<.0001
Year Aggrt.	0.0178	0.0062	0.0057	0.0300	2.87	0.0041
Region 1	-2.8054	0.0089	-2.8229	-2.7879	-314.78	<.0001
Region 2	-0.4825	0.0025	-0.4874	-0.4776	-192.81	<.0001
Region 3	-1.6357	0.0048	-1.6452	-1.6262	-337.65	<.0001
Region 4	-2.1079	0.0079	-2.1234	-2.0924	-266.91	<.0001
Region 5	-1.5632	0.0013	-1.5657	-1.5607	-1226.1	<.0001
Region 6	0.0000	0.0000	0.0000	0.0000	.	.

Conclusion for Final Model

The Analysis 2 model, the time aggregated negative binomial GEE using linear time and categorical regions as covariates, is chosen as best fit model and is used for the final analysis. Although Analysis 1 allows for additional data to be analyzed due to the additional observations that can be extracted by not aggregating time, Analysis 2 was chosen as the best fit model due to its residual plot that suggests a better fit.

Defining the Final Model

The overall final model is a generalized estimating equation mixed effects model using continuous time and categorical region as predictors, with time aggregated into 4 year intervals, and the variable, *state,* as repeated random effects and the variable, *time,* as within subject random effects. The final model equation is as follows:

$$Log(E(Y_{ij})/p) = -50.0784X_1 + 0.0178X_{j2} + -2.8054X_3 + -0.4825X_4 + -1.6357X_5 + -2.1079X_6 + -1.5632X_7 + 0.00X_8 + Zu$$

where Y_{ij} represents case count for ith region at j time, p represents regional population, X_1 represents intercept and equals 1 for all subjects and occasions, X_{j2} represents years at j time, X_3= 1 if i=Region 1 and equals 0 otherwise, X_4= 1 if i=Region 2 and equals 0 otherwise, X_5=1 if i=Region 3 and equals 0 otherwise, X_6=1 if i=Region 4 and equals 0 otherwise, X_7=1 if i=Region 5 and equals 0 otherwise, X_8=1 if i=Region 6 and equals 0 otherwise, Zu represents random effect, i=1,….,6 region, j= [1964-67],….,[2012-15] 4-year aggregates.

4.4 Final Model Goodness of Fit

No test for overall goodness of fit is currently available for GEE (SAS 2016). However, the Pearson residuals vs. observed plot can be used to provide evidence for goodness of fit. Figure 4 shows the residual plot for the final model. The plot demonstrates the model provides a decent, but not superb fit as the plot hints as y-axis unbalance that may suggest a missing predictor (Statwing 2016) and slight clustering below the zero line that reflects a majority of observed values are lower than predicted values. A more detailed discussion of Figure has previously been presented in Section 3.2.

Figure 4: Pearson Residual versus Observed Scatter Plot

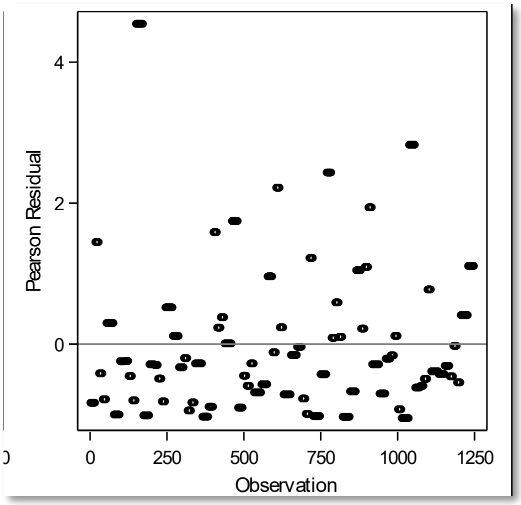


Table 9 compares the expected EEE incidence to the observed incidence for select years and regions. The table demonstrates that, generally, the model predicts the observed incidence fairly closely with some expectations falling slightly below or slightly higher than the observed incidence.

Table 9: Observed and Predicted Incidence for Select Year Intervals and Regions

Year, Region	Predicted Incidence	Observed Incidence
1964-7 reg1	1.75 x 10 ^-8	0
1964-7 reg2	1.79 x 10^-7	4.01x10^-7
1968-71 reg2	1.92 x 10^-7	1.44 x 10^-7
1968-71 reg3	6.06 x 10^-8	4.33x10^-8
1980-83 reg2	2.34 x 10^-7	5.44x10-7
1980-83 reg3	7.50 x 10^-8	1.29x10^-7
1988-91 reg3	2.35 x 10^-7	3.89x10^-7
1988-91 reg4	5.40 x 10^-8	8.98 x 10^-8
1996-99 reg4	6.22 x 10^-8	8.98x10^-8
1996-99 reg5	1.07 x 10^-7	0
2004-07 reg5	1.24 x 10^-7	0
2004-07 reg6	5.91 x 10^-7	1.98 x 10^-6

For space efficiently purposes, only selected years and regions are shown.

4.5 Answering Research Question

SADL Model

The final model failed the thesis aim of identifying a Spatial Autocorrelation Distributed Lag (SADL) model that overcomes the commonly seen epidemiological space-time analysis weakness of neglecting spatial, temporal, and spatiotemporal autocorrelations. However, the author concludes that the SADL model may only be appropriate for events less rare than EEE, as such rare events will likely not be impacted by previously occurring or neighboring events, and that the final non-SADL model remains relevant to answering the research question.

Model Interpretation

Table 10 provides expected incidence for all regions and aggregated years. Recall, ANCOVA results demonstrate each regional incidence is statistically significant from one other across time. Table 10 demonstrates that EEE appears to be increasing over time for all infected regions of the United States.

This thesis aims to answer the research question, *is there a northeastern shift in EEE incidence over time*? (Recall that regions shift northeast with increasing numerical designation, and Region 6 is the most northeastern region.) For all time points, Region 6 demonstrates the highest EEE incidence of all regions. Moreover, with the exception of Region 2, expected incidence appears to be increasing as regions shift from central US (Region 1) towards northeastern US (Region 6) for all time points. These patterns demonstrate that EEE does not appear to be shifting northeastern over time, but has always had a higher incidence in the northeast and prevalence appears to have always followed an increasing trend from central US to northeastern US (with exception of some Mid-Atlantic states that compose Region 2).

However, the magnitude of increase over time is highest in Region 6 as compared to other regions. This demonstrates that although EEE does not appear to be shifting northeastern over time, the northeastern US does appear to have the quickest rate of increase in risk as compared to other regions of the US. Overall, this shows that although EEE is not emerging into NE regions of the US, EEE incidence is increasing the quickest in the northeast. See Table 10.

Furthermore, as shown in Table 10, the model is used to predict EEE incidence until 2047. This model predicts a continued increase over time until approximately year 2040, where EEE decreases, followed by a continued increase in rates. The pattern of highest rates of incidence increase for the northeastern Region 6 remains the same in predicted future time points.

To conclude, EEE does not appear to be emerging or shifting into more northeastern regions. However, EEE incidence is increasing at the highest rate in the northeastern US, demonstrating that northeastern US has historically had and continues to hold the highest risk of EEE infection. The mid-Atlantic region holds the second highest risk for human EEE infection. EEE incidence appears to be increasing for all infected regions of the US over time. Predicted values over the next 30 years, as shown on Table 10, demonstrate a continuation of this trend until around 2040, where expected incidence may dramatically decrease followed by another rise.

However, it must be noted that the expected incidence is very minimal for most regions. Table 10, for ease of interpretation, provides results as new cases per 100 million individuals. However, the entire population of infected states is approximately 170 million and each region's population is within the 10 millions. Thus, for most regions and time points, the expected incidence is less than 1 person per entire region. (The population for each region is as follows: Region 1: 37,570,000, Region 2: 34,839,000, Region 3: 23,102,000, Region 4: 22,265,000, Region 5: 40,642,614, Region 6: 11,093,562 (US Census Bureau).)

Table 10: Expected Regional Incidence [E(Y)] per 100 million people 1964-2047

Year Aggregates	Reg1 E(Y)	Reg2 E(Y)	Reg3 E(Y)	Reg4 E(Y)	Reg5 E(Y)	Reg6 E(Y)
1964-67	1.75	17.87	5.64	3.52	6.06	28.95
1968-71	1.88	19.19	6.06	3.78	6.51	31.09
1972-75	2.02	20.61	6.51	4.06	6.99	33.39
1976-79	2.17	22.14	6.99	4.36	7.51	35.86
1980-83	2.33	23.77	7.50	4.68	8.07	38.51
1984-87	2.50	25.53	8.06	5.02	8.66	41.36
1988-92	2.69	27.42	8.65	5.40	9.30	44.42
1992-95	2.89	29.44	9.29	5.80	9.99	47.70
1996-99	3.10	31.62	9.98	6.22	10.73	51.23
2000-03	3.33	33.96	10.72	6.68	11.52	55.01
2004-07	3.57	36.47	11.51	7.18	12.38	59.09
2008-11	3.84	39.16	12.36	7.71	13.29	63.45
2012-15	4.12	42.06	13.27	8.28	14.27	68.14
2016-19	*4.43*	*45.17*	*14.26*	*8.89*	*15.33*	*73.18*
2024-27	*5.10*	*52.10*	*16.44*	*10.25*	*17.68*	*84.40*
2032-35	*5.89*	*60.09*	*18.96*	*11.83*	*20.39*	*97.35*
2040-43	*1.14*	*11.65*	*3.68*	*2.29*	*3.95*	*18.87*
2044-47	*1.23*	*12.51*	*3.95*	*2.46*	*4.25*	*20.27*

[

E(Y) for region and aggregated years. Results are per 100,000,000 people. Predictions for future years are italicized. Recall that regions shift north-eastern with higher numerical designation, and Region 6 is the most north-eastern, New England, region. Regions are defined as: 1-TX, AR, LA, MS; 2-FL, GA, AL; 3- VA, NC, SC; 4- WI,MI, IN; 5- DE, CT, MD, NY, NY, PA; 6- VT, RI, ME, MA,NH.

CHAPTER 5: LIMITATIONS, FUTURE RESEARCH, CONCLUSIONS

5.1: Research Limitations

Environmental predictors were outside the scope of this analysis, thus the final model that intentionally omitted environmental predictors may not be the best fit model. Moreover, rare events and excess of zero outcomes, although adjusted for, may still result in somewhat biased results for the human incidence analysis (Rose et al. 2006). The spatial autoregressive distributed lag (SADL) model that accounts for the commonly seen weakness of epidemiology space time analysis, failure to adjust for spatial, temporal, and spatiotemporal autocorrelations is not applicable to extremely rare events such as EEE.

5.2 Future Research

Future research that builds upon the current analysis by including environmental predictors could better determine the relationship between space, time, and human EEE incidence.

5.3 Conclusions

EEE incidence appears to be increasing within all infected regions of the United States over time (Table 10). EEE does not appear to be emerging or shifting into more northeastern regions of the US. However, EEE incidence is increasing at the highest rate in the northeastern US, demonstrating that northeastern US has historically had and continues to hold the highest risk of EEE infection. Overall, the New England region of the United States not only holds the highest risk for human EEE infection, but also the quickest rate of increase in risk for human EEE infection over time. Other than the New England region of the US, the mid-Atlantic states hold the second highest risk of human EEE infection. Values predicted until 2047 demonstrate a continuation of this trend until around 2040, where expected incidence may dramatically decrease followed by another rise.

Furthermore, it should be reiterated that this model may be missing significant environmental predictors since environmental predictors were outside the scope of this analysis. Therefore, although this analysis provides evidence for the aforementioned results, future research that builds upon this model with environmental predictors should be conducted before definitive conclusions are made.

APPENDIX: RELEVANT AND CONDENSED SAS CODE

```
*NEGATIVE BINOMIAL RREGRESSION FOR ANALYSIS 1;
*CATEGORICAL TIME;
proc genmod data = human2V2 plots=predicted;
        class year region;
   model Region_Cases= year region /dist=negbin link=log offset=Pop_Region_Log;
run;

proc genmod data=human2V2  plots=predicted;
     class year region;
     model Region_Cases= Year Region Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
run;

proc genmod data=human2V2  plots=predicted;
     class year region;
     model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases / dist=nb link=log offset=Pop_Region_Log;
run;
```

```
proc genmod data=human2V2  plots=predicted;
   class year region;
   model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases
   Nneighbor__Region_Cases_Lag Sneighbor_Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
run;

*LINEAR TIME;
proc genmod data = human2V2  plots =(predicted reschi resdev) ;
      class region;
  model Region_Cases= year region /dist=negbin link=log offset=Pop_Region_Log;
run;

proc genmod data=human2V2  plots=predicted;
   class region;
   model Region_Cases= Year Region Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
run;

proc genmod data=human2V2  plots=predicted;
   class region;
   model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases / dist=nb link=log offset=Pop_Region_Log;
run;

proc genmod data=human2V2  plots=predicted;
   class region;
   model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases
   Nneighbor__Region_Cases_Lag Sneighbor_Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
run;


*NEGATIVE BINOMIAL GEE FOR ANALYSIS 1;

*CATEGORICAL TIME;
proc genmod data = human2V2 plots=predicted;
      class year state region t;
  model Region_Cases= year region /dist=negbin link=log offset=Pop_Region_Log;
       repeated subject=state / withinsubject=t ;
run;


proc genmod data=human2V2  plots=predicted;
   class t year state region;
   model Region_Cases= Year Region Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
       repeated subject=state / withinsubject=t ;
run;
```

```
proc genmod data=human2V2  plots=predicted;
     class t state year region;
     model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t ;
run;


proc genmod data=human2V2  plots=predicted;
     class t state year region;
     model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases
     Nneighbor__Region_Cases_Lag Sneighbor_Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t ;
run;


*CONTINUOUS TIME;
proc genmod data=human2V2 plots =(predicted reschi resdev);
     class region state t;
     model Region_Cases= Year Region  / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t ;
run;

proc genmod data=human2V2 plots =(predicted reschi resdev);
     class region state t;
     model Region_Cases= Year Region Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

proc genmod data=human2V2 plots =(predicted reschi resdev);
     class region state t;
     model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t ;
run;

proc genmod data=human2V2 plots =(predicted reschi resdev);
     class region state t;
     model Region_Cases= Year Region Region_Cases_Lag Nneighbor_Region_Cases
Sneighbor_Region_Cases
     Nneighbor__Region_Cases_Lag Sneighbor_Region_Cases_Lag / dist=nb link=log
offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t ;
run;
```

```
*ANALYSIS 2;

*NEGATIVE BINOMIAL GLM FOR ANALYSIS 2;
*CATEGORICAL TIME;
proc genmod data = agg2d plots=(predicted reschi resdev);
        class year_agg region;
   model timeagg_region_cases= year_agg region /dist=negbin link=log
offset=Pop_Region_Log;
run;

proc genmod data=agg2d  plots=(predicted reschi resdev);
    class year_agg region;
    model timeagg_region_cases= Year_agg Region lag_timeagg_region_cases / dist=nb
link=log offset=Pop_Region_Log;
run;

proc genmod data=agg2d  plots=(predicted reschi resdev);
    class year_agg region;
    model timeagg_region_cases= Year_agg Region
    lag_timeagg_region_cases Nneighbor_timeagg_region_cases
Sneighbor_timeagg_region_cases / dist=nb link=log offset=Pop_Region_Log;
run;

proc genmod data=agg2d  plots=(predicted reschi resdev);
    class year_agg region;
    model timeagg_region_cases= Year_agg Region
    lag_timeagg_region_cases Nneighbor_timeagg_region_cases Nneighbor_lag_timeagg
Sneighbor_timeagg_region_cases Sneighbor_lag_timeagg / dist=nb link=log
offset=Pop_Region_Log;
run;


*CONTINUOUS TIME;
proc genmod data = agg2d plots=(predicted reschi resdev);
        class  region;
   model timeagg_region_cases= year_agg region /dist=negbin link=log
offset=Pop_Region_Log;
run;

proc genmod data=agg2d  plots=(predicted reschi resdev);
    class region;
    model timeagg_region_cases= Year_agg Region lag_timeagg_region_cases / dist=nb
link=log offset=Pop_Region_Log;
run;

proc genmod data=agg2d  plots=(predicted reschi resdev);
    class region;
    model timeagg_region_cases= Year_agg lag_timeagg_region_cases
Nneighbor_timeagg_region_cases Sneighbor_timeagg_region_cases / dist=nb link=log
offset=Pop_Region_Log;
run;
```

```
proc genmod data=agg2d  plots=(predicted reschi resdev);
     class region;
     model timeagg_region_cases= Year_agg Region lag_timeagg_region_cases
Nneighbor_timeagg_region_cases Nneighbor_lag_timeagg Sneighbor_timeagg_region_cases
Sneighbor_lag_timeagg / dist=nb link=log offset=Pop_Region_Log;
run;



* NEGATIVE BINOMIAL GEE FOR ANALYSIS 2;

*CATEGORICAL TIME;
proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg year_agg t;
     model timeagg_region_cases= Year_agg Region  / dist=nb link=log
offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t year_agg;
     model timeagg_region_cases= Year_agg Region
     lag_timeagg_region_cases/ dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t year_agg;
     model timeagg_region_cases= Year_agg Region
     lag_timeagg_region_cases Nneighbor_timeagg_region_cases
Sneighbor_timeagg_region_cases / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t year_agg;
     model timeagg_region_cases= Year_agg lag_timeagg_region_cases
Nneighbor_timeagg_region_cases Nneighbor_lag_timeagg Sneighbor_timeagg_region_cases
Sneighbor_lag_timeagg / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

*CONTINUOUS TIME;
proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t;
     model timeagg_region_cases= Year_agg Region  / dist=nb link=log
offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;
```

```
proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t;
     model timeagg_region_cases= Year_agg region lag_timeagg_region_cases / dist=nb
link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t;
     model timeagg_region_cases= Year_agg Region lag_timeagg_region_cases
Nneighbor_timeagg_region_cases Sneighbor_timeagg_region_cases / dist=nb link=log
offset=Pop_Region_Log;
     repeated subject=state / withinsubject=t;
run;

proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t;
     model timeagg_region_cases= Year_agg Region lag_timeagg_region_cases
Nneighbor_timeagg_region_cases Nneighbor_lag_timeagg Sneighbor_timeagg_region_cases
Sneighbor_lag_timeagg / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
run;

*FINAL MODEL ESTIMATES;
*(Condensed);
proc genmod data=agg2d plots =(predicted reschi resdev);
     class region state t_agg t;
     model timeagg_region_cases= Year_agg Region  / dist=nb link=log
offset=Pop_Region_Log;
 repeated subject=state / withinsubject=t;
estimate "1964-7 reg1" intercept 1 year_agg 1964 region 1 0 0 0 0 0 / e exp;
estimate "1964-7 reg2" intercept 1 year_agg 1964 region 0 1 0 0 0 0 / e exp;
estimate "1964-7 reg3" intercept 1 year_agg 1964 region 0 0 1 0 0 0 / e exp;
estimate "1964-7 reg4" intercept 1 year_agg 1964 region 0 0 0 1 0 0 / e exp;
estimate "1964-7 reg5" intercept 1 year_agg 1964 region 0 0 0 0 1 0 / e exp;
estimate "1964-7 reg6" intercept 1 year_agg 1964 region 0 0 0 0 0 1 / e exp;
run;


*ANCOVA;
PROC genmod DATA=agg2d;
  CLASS region state t_agg t;
  model timeagg_region_cases= Year_agg Region  / dist=nb link=log offset=Pop_Region_Log;
           repeated subject=state / withinsubject=t;
 lsmeans region/pdiff at year_agg=1964;
 lsmeans region/pdiff at year_agg=1968;
 lsmeans region/pdiff at year_agg=1972;
 lsmeans region/pdiff at year_agg=1976;
 lsmeans region/pdiff at year_agg=1980;
 lsmeans region/pdiff at year_agg=1984;
 lsmeans region/pdiff at year_agg=1988;
 lsmeans region/pdiff at year_agg=1992;
 lsmeans region/pdiff at year_agg=1996;
 lsmeans region/pdiff at year_agg=2000;
 lsmeans region/pdiff at year_agg=2000;
 lsmeans region/pdiff at year_agg=2003;
 lsmeans region/pdiff at year_agg=2008;
 lsmeans region/pdiff at year_agg=2012;
RUN;
```

REFERENCES

Abellan, J.J., Richardson, S., and Best, N. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives*, 116(8).

APHIS. (2008). United States Department of Agriculture. Eastern Equine Encephalomyelitis. APHIS Veterinary Service Factsheet.

Armstrong, P. and Andreadis, T. (2013). Eastern equine encephalitis virus – old enemy, new threat. *New England Journal of Medicine*, 368(18).

Beale, L., Abellan, J., Hodgson, S., and Jarup, L. (2008). Methodologic issues and approaches to spatial epidemiology. *Environmental Health Perspectives*, 116(8).

Centers for Disease Control and Prevention (CDC). (2015). Eastern Equine Encephalitis. http://www.cdc.gov/easternequineencephalitis/ <accessed January 30, 2016>

Derby, N. (2011). An introduction to the analysis of rare events. Stakana Analytics. http://www.wuss.org/proceedings11/Papers_Derby_N_76404.pdf <accessed June 23, 2016>

Deresiewicz, R., Thaler, S., and Zhamani, A. (1997). Clinical and neuroradiographic manifestations of Eastern Equine Encephalitis. *New England Journal of Medicine*, 336, 1867-74.

De Smith, M.J. (2015). Statistical Analysis Handbook: Chapter 9. Autocorrelation. Web-based. http://www.statsref.com/HTML/?introduction.html <accessed June 23, 2016>

Elliot, P., Wakefield, J., Best, N., and Briggs, D. (2001). *Spatial epidemiology: methods and applications*. Bias and confounding in spatial epidemiology. Oxford Scholarship Online: September 2009. http://www.oxfordscholarship.com/view/101093/acprof:oso/9780198515326.001.0001/acprof-9780198515326-chapter-5 <accessed June 5, 2016>

Elhorst, J. (2001). Dynamic models in space and time. *Geographic Analysis,* 33(2), 119-140.

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*. 2nd edition. Hoboken, New Jersey: Wiley Publication.

Ocana-Riola, R. (2010). Common errors in disease mapping. *Geospatial Health*, 4(2),129-154.

Rose, C.E., Martin, S.W., Wannemuehler, K., Plikaytis, B.D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*,16(4),463-481.

Rosner, B. (2011). *Fundamentals of Biostatistics*. 7th edition. Boston, Massachusettes: Books/Cole Publishing.

SAS (2016). Assessing fit, correcting overdispersion in generalized linear models. http://support.sas.com/kb/22/630.html

Statwing (2016). Statwing Documentation. Interpeting Residual Plots: Y-Axis Unbalanced http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#y-unbalanced-header <accessed March 14, 2017>

United States Census Bureau. (2010). United States Census 2010. http://www.census.gov/2010census/ <accessed May 30, 2016>