A BAYESIAN HIERARCHICAL SPATIAL MODEL FOR WEST NILE VIRUS IN NEW YORK CITY: EVALUATING AN APPROACH TO HANDLE LARGE SPATIAL DATA SETS

by

JAMIAN KRISHNA PACIFICI

(Under the Direction of Nicole Lazar)

ABSTRACT

In 1999 West Nile Virus (WNV) was first detected in the Western Hemisphere in New York City, NY and is now responsible for approximately 1.8 million human infections, 360000 illnesses, and 1308 deaths. Our objectives are to: 1) develop a geostatistical model to assess which covariates influence WNV distribution, 2) assess the predictive performance of the developed model, and 3) explore approaches to reduce the computational burden associated with fitting spatial models to large data sets. We found a positive effect of land cover type, housing density, and distance to wetlands, and a negative effect of slope on the prevalence of WNV in New York City. We found predicted WNV occurrence to be highest in Staten Island. We believe the use of hierarchical modeling provides a useful approach to exploring the prevalence and distribution of WNV, but does not come without a cost regarding the difficulty of optimal implementation.

INDEX WORDS: Bayesian modeling, hierarchical modeling, infectious diseases, large spatial data sets, predictive process models, spatial modeling, West Nile virus

A BAYESIAN HIERARCHICAL SPATIAL MODEL FOR WEST NILE VIRUS IN NEW YORK CITY: EVALUATING AN APPROACH TO HANDLE LARGE SPATIAL DATA SETS

by

JAMIAN KRISHNA PACIFICI

B.S., North Carolina State University, 2003

M.S., North Carolina State University, 2007

Ph.D., University of Georgia, 2011

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

A BAYESIAN HIERARCHICAL SPATIAL MODEL FOR WEST NILE VIRUS IN NEW YORK CITY: EVALUATING AN APPROACH TO HANDLE LARGE SPATIAL DATA SETS

by

JAMIAN KRISHNA PACIFICI

Major Professor: Nicole Lazar

Committee: John Drake
Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

I dedicate this dissertation to my son, Samson.

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor Nicole Lazar for all of her support and guidance throughout this process. I would especially like to acknowledge her patience and availability to answer all of my questions. I would also like to acknowledge my committee members, John Drake and Lynne Seymour for all their support and patience. I believe they provided me with a diverse set of skills which continually nurtured my development. I owe a special thanks to John Drake for his financial support stemming from this project.

I would like to thank all of my fellow graduate students for their support. I have thoroughly enjoyed my time at UGA and will remember all of the laughs we have had. I would like to give a special recognition to my wife Lara, and for everything she does to make our lives special.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Disease ecologists are interested in understanding the factors that drive the transmission of infectious diseases in order to prevent costly outbreaks. Such outbreaks, particularly from new or reoccurring diseases, have been estimated to cost approximately 7.5 billion dollars annually (Pimentel et al. 2005) and are estimated to be responsible for about 15 million annual deaths worldwide not including the additional millions of deaths that occur as a consequence of past infections or complications with chronic infections (World Health Organization 2004). Emerging diseases or re-emerging diseases have been defined as "infections that have newly appeared in a population or have existed previously but are rapidly increasing in incidence or geographic range" (Morse 1995). The classification of infectious diseases as emerging or re-emerging is helpful because the control measures to prevent outbreaks differ between the groups (Morens et al. 2004). Interest often lies in understanding the underlying causes of emergence or re-emergence and how the disease evolves in relation to its environment and hosts. This evolution is complex as it depends on how the pathogen interacts, not only with its host and possibly new hosts, but with the ecosystem surrounding it. Not only can pathogen-induced changes occur within host individuals, but these effects can be propagated through to ecosystem processes (e.g., productivity, nutrient cycling) and landscape level effects (e.g., land use). The interplay between hosts and pathogens and their surrounding environment is an important avenue of infectious disease ecology research (Ostfeld et al. 2008).

An additional area of research is to understand how abiotic and biotic factors influence the severity and frequency of disease outbreaks. Unfortunately, the prediction of the magnitude and duration of these impacts for a single pathogen is very difficult. To make such predictions accurately requires the integration of different scientific tools with the ultimate goal of incorporating the role of disease into conceptual models of community, ecosystem, and landscape ecology.

It has been recognized that there is often a form of spatial dependency in disease outbreaks (Chuang et al. 2012, Dowdy et al. 2012, Jones et al. 2008, Young and Jensen 2012). Often specific locations tend to have higher re-occurrences of certain diseases and once an outbreak occurs these disease "hotspots" propagate the transmission of the disease in a spatially dependent manner (Dowdy et al. 2012, Jones et al. 2008). It is therefore critical to be able to identify these "hotspots" as well as to understand what factors are responsible for the spatial dependence of disease occurrences. The field of spatial statistics is an area rich in techniques to effectively explore these types of relevant questions in disease ecology.

Spatial Statistics

One of the key features of spatial data is the autocorrelation of observations in space. Observations in close spatial proximity tend to be more similar than observations that are farther apart. Although this is not the defining feature of spatial data, the characterization of the spatial correlation is a common interest for many applications of spatial statistics. Often the scientific question of interest evolves from simply "How many?" to "How many and where?" thus necessitating the use of specific methods. Classical inference and estimation including estimating the parameters of the data-generating mechanism, testing hypotheses about these

parameters, estimating the mean vector, and predicting the response at unobserved locations, are still valid except that the location of individual observations is now important and comparisons among different locations in the study area can now be validly assessed.

Traditionally, the appropriate statistical modeling framework is dependent upon the type of data collected. Broadly speaking, three different types of spatial data have been classified based on the characteristics of their domain (Cressie 1993): geostatistical data or point-referenced data, wherein the domain is a continuous fixed set such that the attributes of interest can be observed everywhere within the domain (i.e. theoretically an infinite number of samples can be placed within the fixed domain); lattice data or areal data, wherein the domain is fixed and discrete such that the number of locations can be enumerated; and point pattern data, wherein the variable of interest is the location of events and thus the domain is stochastic (i.e., changes with each realization of the spatial process). Although it is helpful to classify the different types of spatial data, they are not mutually exclusive and there are often cases that feature more than one type, for example, we may have ozone levels at specific sites within a region on a specific day. These observations would come from fixed monitoring stations for which exact spatial coordinates are known. A second component of this data set could be the number of children in the area's zip codes that were reported at local emergency rooms with acute asthma symptoms on the following day. Now we have both point-reference data (ozone levels) and lattice data (number of children at local emergency rooms) and we are interested in establishing a connection between high ozone and subsequent high pediatric emergency room asthma visits.

The set of statistical tools and forms of analysis differ depending on the type of data collected as do the types of questions that can be answered. Here we will focus on geostatistical

data because the data in our case study was collected in this fashion. There are three important elements of geostatistical modeling: stationarity, variograms, and isotropy. It is important to note the fundamental concept underlying the theory is a stochastic process $\{Y(s): s \in D\}$ where *D* is a fixed subset of *r*-dimensional Euclidean space. It is also important to recognize that the continuity refers to the domain, not necessarily to the attribute being measured. Whether the attribute is continuous or discrete has no bearing on whether the data are geostatistical or not. For a more complete description of spatial statistical methods including underlying theory, proofs and derivation, please consult one of the texts devoted to the subject (e.g., Banerjee et al. 2004, Cressie 1993, Schabenberger and Gotway 2005).

Stationarity

First let us assume that our spatial process has a mean, $\mu(\boldsymbol{s}) = E[Y(\boldsymbol{s})]$, and that the variance of $Y(\boldsymbol{s})$ exists for all $s \in D$. The process $Y(\boldsymbol{s})$ is said to be Gaussian if, for any $n \geq 1$ and any set of sites $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $\boldsymbol{Y} = [Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)]^T$ has a multivariate normal distribution. Strictly stationary is then defined for any given $n \geq 1$, any set of *n* sites $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and any $\mathbf{h} \in \mathbb{R}^d$, such that the distribution of $[Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)]$ is the same as that of $[Y(\boldsymbol{s}_1 + \mathbf{h}), \ldots Y(\boldsymbol{s}_n + \mathbf{h})]$. Weak stationarity (or second-order stationarity) is defined as the process having a constant mean, $\mu(\boldsymbol{s}) \equiv \mu$, and $Cov[Y(\boldsymbol{s}), Y(\boldsymbol{s} + \mathbf{h})] = C(\mathbf{h})$ for all $\mathbf{h} \in \mathbb{R}^d$ such that $\boldsymbol{s}$ and $\boldsymbol{s}$ + **h** both lie within *D*. Weak stationarity implies that the covariance relationship between the values of the process at any two locations can be summarized by a covariance function $C(\mathbf{h})$, and this function depends only on the separation vector **h**. Strong stationarity implies weak stationarity when all variances are assumed to exist, but the converse is not necessarily true (it does hold for Gaussian processes).

Variograms

A third type of stationarity called intrinsic stationarity exists when we assume $E[Y(\boldsymbol{s}+\mathbf{h})-Y(\boldsymbol{s})]=0$ and define $E[Y(\boldsymbol{s}+\mathbf{h})-Y(\boldsymbol{s})]^2=Var[Y(\boldsymbol{s}+\mathbf{h})-Y(\boldsymbol{s})]=2\gamma(\mathbf{h})$ where the left-hand side depends only on **h** and not the particular choice of $\boldsymbol{s}$. $2\gamma(\mathbf{h})$ is then called the variogram, and $\gamma(\mathbf{h})$ is called the semivariogram. Note that intrinsic stationarity says nothing about the joint distribution of a collection of variables, $[Y(\boldsymbol{s}_1),\dots,Y(\boldsymbol{s}_n)]$, and therefore provides no likelihood. The relationship between the variogram and covariance function is: $\gamma(\mathbf{h})=C(\mathbf{0})-C(\mathbf{h})$. A valid variogram satisfies a negative definiteness condition such that for any set of locations $\{\boldsymbol{s}_1,\dots,\boldsymbol{s}_n\}$ and any set of constants $\{a_1,\dots,a_n\}$ where $\sum_i a_i=0$ then $\sum_i\sum_j a_i a_j\gamma(\boldsymbol{s}_i-\boldsymbol{s}_j)\leq 0$.

Isotropy

If the semivariogram function $\gamma(\mathbf{h})$ depends upon the separation vector only through its length $\|\mathbf{h}\|$, then we say that the process is isotropic and anisotropic if this condition does not hold. The isotropic semivariogram rises from the origin and if *C*(**h**) decreases monotonically with increasing **h**, then $\gamma(\mathbf{h})$ will approach $Var\big(Y(\boldsymbol{s})\big)=\sigma^2$ either asymptotically or exactly at a particular lag **h***. The asymptote itself is termed the sill and the lag at which the sill is reached is called the range. Observations between locations for which the distance exceeds the lag **h*** are uncorrelated. If the semivariogram does not pass through the origin the intercept is termed the nugget effect. Figure 1.1 provides an example of an exponential semivariogram with sill = 1.2 and nugget = 0.2 (technically speaking the range is infinite). Isotropic processes are popular because of their simplicity, interpretability and because there are a number of relatively simple parametric forms available as candidates for semivariograms. Common covariance functions for

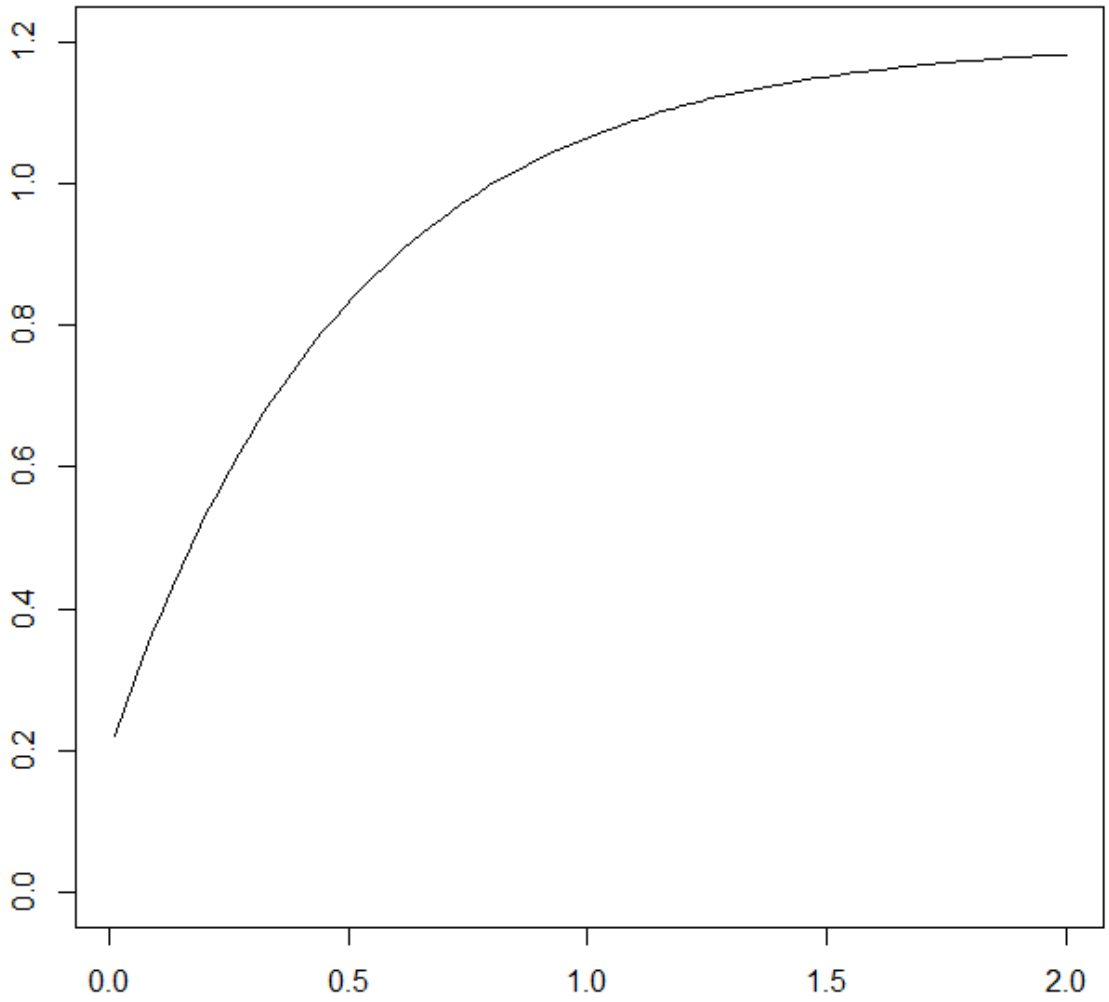
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0

Figure 1.1. Semivariogram from an exponential correlation model with parameters $\sigma^2 = 1, \tau^2 = 0.2, \phi = 2$, sill of 1.2 ($\sigma^2 + \tau^2$) and nugget of 0.2.

parametric isotropic models include spherical, exponential, powered exponential, Gaussian, and Matérn (see Banerjee et al. 2004, Cressie 1993, Schabenberger and Gotway 2005 for a complete description). In our case study we used the exponential covariance function:

$$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) \text{ if } t > 0 \\ \tau^2 + \sigma^2 \text{ otherwise} \end{cases}$$

where $t = \|\mathbf{h}\|$. Technically the sill is only reached asymptotically such that the range is infinite. In this case interest lies in the effective range, i.e., the distance at which there is essentially no lingering spatial correlation (≤ 0.05). Estimating the three parameters: ϕ, σ^2, τ^2 is usually of interest.

Bayesian Hierarchical Modeling

Estimation and model fitting of variograms has traditionally been done "by eye" or using trial and error to choose values of the nugget, sill, and range parameters that provide a good match to the empirical semivariogram. There have been other attempts to find parameter values (e.g., method of moments, estimators based on order statistics and quantiles) as well as nonlinear maximization routines to find the parameter values that minimize some goodness-of-fit criterion (Banerjee et al. 2004, Schabenberger and Gotway 2005). More recently, parametric approaches have been used such as least squares, maximum likelihood, restricted maximum likelihood, composite likelihood or generalized estimating equations (e.g., see Schabenberger and Gotway 2005).

Although frequentist methods exist, maximum likelihood is usually no longer feasible when complex dependencies or hierarchical structure within the data exist (Cressie and Wikle 2011). It is often very natural to view data exhibiting spatial dependency in a hierarchical or

multi-level framework (Banerjee et al. 2004) which makes them particularly suitable for Bayesian modeling using Markov Chain Monte Carlo (MCMC) algorithms.

In Bayesian modeling, the goal is to compute the form of the posterior distribution of the parameters, given the data and prior information. The posterior distribution thus contains all of the relevant information for making inference including predictions at unobserved locations. The posterior distribution can be obtained by solving the following equation

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior density, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of the data and $p(\boldsymbol{\theta})$ is a prior density through which previous knowledge about $\boldsymbol{\theta}$ can be incorporated. Unfortunately, the solution to the above equation for most spatial models does not usually have a closed form and instead we rely on MCMC algorithms to generate samples from the posterior distributions of the parameters. Fortunately there has been rapid development in the area of MCMC methods and once these algorithms have converged, with sufficiently large samples we may compute any statistic from the posterior distribution such as posterior means, standard deviations, quantiles, or correlations between the parameters.

We must take care to ensure that the algorithms converge and that we only analyze samples generated after convergence. Convergence itself can be assessed by generating samples from multiple Markov chains, begun with widely dispersed initial values, and comparing the behavior of the chains. We can do this visually by making simultaneous plots of the values of the multiple chains for each parameter, with good mixing of the chains implying convergence. In addition we can also compute the Gelman-Rubin scale-reduction factor (Gelman et al. 2004),

which compares variation within the chains to variation across the chains. Convergence will lead to values of this factor close to 1.

There is an important practical limitation in using geostatistical models within Bayesian MCMC modeling. At each step of the algorithm we must sample from the full conditional distribution which requires computing the inverse of the N x N covariance matrix where N is the number of unique locations. In practice, we can actually solve a system of N equations, but this still becomes an enormous drain on computer time when N is large and the operation must be repeated for thousands of MCMC iterations. Things look even worse when our goal is spatial prediction ("kriging", Cressie 1993, Diggle et al. 1998) at a large number of unsampled locations. Fortunately, there has been an increase in the number of available methods to handle large spatial data sets (e.g., Latimer et al. 2006, Illian et al. 2007, Banerjee et al. 2008, Latimer et al. 2009, Sang and Huang 2012). Different approaches provide different techniques to reduce complexity and the computational burden. Several authors have explored methods that are more efficient by approximating complex likelihoods instead of dealing with the full likelihood directly. Vecchia (1988) suggested approximating the likelihood through products of conditional distributions while Fuentes (2007) used spectral representations of the spatial process to approximate the likelihood. A second approach is to modify the correlation structure to enable more efficient computation. For example, Sang and Huang (2012) developed an approach that takes advantage of more efficient sparse matrix solvers by using compacted correlation functions that yield sparse correlation structures. A third approach is to represent the spatial process in a lower-dimensional subspace (Stein 2007, Cressie and Johannesson 2008, Banerjee et al. 2008). The general idea behind these methods is to express the spatial process realization over a small set of locations (i.e. "knots") where the number of "knots" is much smaller than the number of

observed sites. Ideally there will be insignificant loss of spatial information when using the smaller set of knots given that there is adequate domain coverage. The critical issue for these types of models is selecting the number and location of the knots, which is a challenging problem.

Chapter Description

Chapter 2 –

In late summer 1999 the first known case of West Nile Virus (WNV) was detected in the Western Hemisphere, specifically New York City, NY in the northeastern United States (Anderson et al. 1999, Lanciotti et al. 1999). WNV is a mosquito-borne flavivirus native to Europe, Asia, and Africa (Lanciotti et al. 1999). Birds serve as the vertebrate reservoir hosts in the transmission cycle of WNV, while humans and other mammals are incidental hosts (Lanciotti et al. 1999). The initial exposure resulted in an outbreak of human encephalitis with 589 patients suspected of having WNV in New York City, four of which were fatal (Nash et al. 2001). In only 4 years WNV reached the west coast and is now responsible for approximately 1.8 million human infections, 360000 illnesses, and 1308 deaths (Kilpatrick 2011).

Our objective in Chapter 2 is to develop a geostatistical model to effectively assess which covariates influence WNV distribution and occurrence while accounting for spatial dependence in the data, to assess the predictive performance of the developed model, and to explore approaches to reduce the computational burden associated with fitting spatial models to large data sets.

Chapter 3 –

I provide a synthesis and conclusion of the previous chapter. In addition I highlight major contributions and discuss future research needs.

Literature Cited

Anderson, J. F., T. G. Andreadis, C. R. Vossbrinck, S. Tirrell, E. M. Wakem, R. A. French, A. E. Garmendia, and H. J. Van Kruiningen. 1999. Isolation of West Nile Virus from mosquitoes, crows, and a cooper's hawk in Connecticut. Science 286:2331-2333.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall. Boca Raton, FL, USA.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society, Series B 70:825-848.

Chuang, T., C.W. Hockett, L. Kightlinger, and M.C. Wimberly. 2012. Landscape-level spatial patterns of West Nile Virus risk in the northern Great Plains. American Journal of Tropical Medicine and Hygiene 86:724-731.

Cressie, N. A. C. 1993. Statistics for Spatial Ddata. Wiley. New York, New York, USA.

Cressie, N. A. C. and G. Johannesson. 2008. Fixed rank kriging for large spatial datasets. Journal of the Royal Statistical Society, Series B 70:209-226.

Cressie, N. A. C. and C. K. Wikle. 2011. Statistics for Spatio-Temporal Data. Wiley. Hoboken, New Jersey, USA.

Diggle, P.J., J.A. Tawn, and R.A. Moyeed. 1998. Model-based geostatistics (with discussion).

Applied Statistics 47:299-350.

Dowdy, D.W., J.E. Golub, R.E. Chaisson, and V. Saraceni. 2012. Heterogeneity in tuberculosis and the role of geographic hotspots in propagating epidemics. Proceedings of the National Academy of Sciences 109: 9557 – 9562.

Fuentes, M. 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102:32-331.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis. Chapman and Hall. Boca Raton, FL, USA.

Illian, J. B., J. Møller, and R. P. Waagepetersen. 2007. Spatial point process analysis for a plant community with high biodiversity. Environmental and Ecological Statistics 16:389-405.

Jones, K.E., N.G. Patel, M.A. Levy, A. Storeygard, D. Balk, J.L. Gittleman, and P. Daszak. 2008. Global trends in emerging infectious diseases. Nature 451:990-994.

Kilpatrick, A. M. 2011. Globalization, land use, and the invasion of West Nile Virus. Science 334:323-327.

Lanciotti, R. S. et al. 1999. Origin of the West Nile Virus responsible for an outbreak of encephalitis in the northeastern United States. Science 286:2333-2337.

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33-50.

Latimer, A. M., S. Banerjee, H. Sang, E. S. Mosher, and J. A. Silander Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters 12:144-154.

Morens, D. M., G. K. Folkers, and A. S. Fauci. 2004. The challenge of emerging and re-emerging infectious diseases. Nature 430:242-249.

Morse, S.S. 1995. Factors in the emergence of infectious diseases. Emerging Infectious Disease 1:7-15.

Nash, D. et al. 2001. The outbreak of West Nile Virus infection in the New York City area in 1999. The New England Journal of Medicine 344:1807-1814.

Ostfeld, R.S., F. Keesing, and V.T. Eviner. 2008. Infectious Disease Ecology. Princeton University Press. Princeton, NJ, USA.

Pimentel, D., R. Zuniga, and D. Morrison. 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics 52:273-288.

Sang H. and J. Z. Huang. 2012. A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society, Series B 74:111-132.

Schabenberger, O., and C.A. Gotway. 2005. Statistical Methods for Spatial Data Analysis. Chapman and Hall. Boca Raton, FL, USA.

Stein, M. L. 2007. Spatial variation of total column ozone on a global scale. The Annals of Applied Statistics 1:191-200.

Vecchia, A. V. 1988. Estimation and model identification for continuous spatial processes.

Journal of the Royal Statistical Society, Series B 50:297-312.

World Health Organization. 2004. The World Health Report. World Health Organization, Genève.

Young, S.G., and R.R. Jensen. 2012. Statistical and visual analysis of human West Nile Virus infection in the United States, 1999-2008. Applied Geography 34:425-431.

CHAPTER 2

A BAYESIAN HIERARCHICAL SPATIAL MODEL FOR WEST NILE VIRUS IN NEW YORK CITY: EVALUATING AN APPROACH TO HANDLE LARGE SPATIAL DATA SETS[1]

[1]Pacifici, K., J.M. Drake, and W.I. Bajwa. To be submitted to *Ecological Applications.*

Abstract

Infectious diseases are responsible for extensive costs to society both in terms of dollars and human losses. In 1999 West Nile Virus (WNV) was first detected in the Western Hemisphere in New York City, NY and reached the west coast in 4 years. It is now responsible for approximately 1.8 million human infections, 360000 illnesses, and 1308 deaths. The WNV epidemic was unprecedented and underscored the ease with which pathogens can move among population centers. In order to fully understand the dynamics of WNV and prevent transmission through control measures, it is important to identify covariates that drive the spatial distribution and occurrence of WNV through the use of statistical spatial models. Computational challenges become highly relevant when fitting spatially explicit models to large data sets. Fortunately, there has been a rise in the development of approaches to handle large spatial data sets including the use of predictive process models to reduce the dimension of the problem. Our objectives are to: 1) develop a geostatistical model to effectively assess which covariates influence WNV distribution and occurrence while accounting for spatial dependence in the data, 2) assess the predictive performance of the developed model, and 3) explore approaches to reduce the computational burden associated with fitting spatial models to large data sets. We found a positive effect of wetlands land cover, housing density, and distance to wetlands, and a negative effect of slope on the prevalence of WNV in New York City. The effective range of correlation between two locations is approximately 560 meters suggesting small-scale local dependence. We found predicted WNV occurrence to be highest in Staten Island and that this was highly dependent on land cover type specifically areas with wetlands. Model validation for our model resulted in a correct classification rate of 68% when using a cut-point of 0.5. The ability to predict disease outbreaks as a function of covariates can be an important step to reducing the

spread of diseases and lessening the associated costs. Furthermore, the identification of specific covariates that have a positive effect on the occurrence of WNV (i.e. land cover type, distance to wetlands, and housing density) allow land managers to focus control efforts on areas that are more likely to propagate disease transmission. We believe the use of hierarchical modeling provides a useful approach to exploring the prevalence and distribution of WNV, but does not come without a cost regarding the difficulty of optimal implementation.

INTRODUCTION

Understanding the environmental factors that help explain and ultimately drive disease dynamics and transmission is an important step in reducing the negative costs associated with many infectious diseases. Human diseases have been estimated to have an associated cost of 7.5 billion dollars annually (Pimentel et al. 2005) and an estimate exceeding 25% of all annual human deaths worldwide have been attributed to infectious diseases (Morens et al. 2004). In late summer 1999 the first known case of West Nile Virus (WNV) was detected in the Western Hemisphere, specifically New York City, NY in the northeastern United States (Anderson et al. 1999, Lanciotti et al. 1999). WNV is a mosquito-borne flavivirus native to Europe, Asia, and Africa (Lanciotti et al. 1999). Birds serve as the vertebrate reservoir hosts in the transmission cycle of WNV, while humans and other mammals are incidental hosts (Lanciotti et al. 1999). The initial exposure resulted in an outbreak of human encephalitis with 589 patients suspected of having WNV in New York City, four of which were fatal (Nash et al. 2001). In only 4 years WNV reached the west coast and is now responsible for approximately 1.8 million human infections, 360000 illnesses, and 1308 deaths (Kilpatrick 2011).

The WNV epidemic was unprecedented and underscored the ease with which pathogens can move among population centers. Because of the extreme health risks associated with the virus and the lack of understanding related to how the virus was introduced, the New York City Department of Health and Mental Hygiene (NYCDOHMH) established a multi-faceted program to detect and prevent the spread of WNV in New York City and the surrounding metropolitan area (NYCDOHMH 2009). As part of the effort a vector surveillance system for monitoring mosquitoes was implemented consisting of trapping and testing mosquitoes at numerous sites throughout the five boroughs of the city. The number of sites, the frequency of trap placement,

and the number of traps per site varied depending on where a pathogen was most likely to be present based on avian and mosquito surveillance data from 1999 onward. Although establishing a rigorous surveillance program is the first step in combatting WNV, it is essential for the data to directly inform control actions in order to effectively prevent future outbreaks. It is therefore necessary to examine past data in order to precisely understand what covariates affect the occurrence of WNV and are accurate predictors for the spatial distribution of the disease.

The field of spatial statistics provides a vast array of approaches for effectively exploring the spatial dynamics of WNV. Traditionally, the appropriate statistical modeling framework is dependent upon the type of data collected. Broadly speaking, three different types of spatial data have been classified based on the characteristics of their domain (Cressie 1993): geostatistical data or point-referenced data, wherein the domain is a continuous fixed set such that the attributes of interest can be observed everywhere within the domain (i.e. theoretically an infinite number of samples can be placed within the fixed domain); lattice data or areal data, wherein the domain is fixed and discrete such that the number of locations can be enumerated; and point pattern or point process data, wherein the variable of interest is the location of events and thus the domain is stochastic (i.e., changes with each realization of the spatial process). Although it is helpful to classify the different types of spatial data, they are not mutually exclusive and there are often cases that feature more than one type of spatial data.

Regardless of the nature of the collection of spatial data, ecological studies often neglect to model spatial relationships (Beale et al. 2007) and we can assume that this is often due to the computational burden of running spatially explicit models for large data sets (Banerjee et al. 2004, Christman 2008). Recent advances in remote sensing technology (e.g., Geographical Information Systems and Global Positioning Systems) and an interest in large-scale long-term

ecological inventories (e.g., National Ecological Observatory Network; NEON, Inc.) have resulted in an increase in the number of large spatially-referenced data sets. The challenge lies in dealing with the increased number of matrix decompositions (increases as $O(n^3)$ in the number of locations, n) that are required and this is often referred to as the "big n" problem for large data sets. Fortunately, there has been an increase in the number of available methods to handle large spatial data sets (e.g., Latimer et al. 2006, Illian et al. 2007, Banerjee et al. 2008, Latimer et al. 2009, Sang and Huang 2012). Different approaches provide different techniques to reduce complexity and the computational burden. Several authors have explored methods that are more efficient by approximating complex likelihoods instead of dealing with the full likelihood directly. Vecchia (1988) suggested approximating the likelihood through products of conditional distributions while Fuentes (2007) used spectral representations of the spatial process to approximate the likelihood. A second approach is to modify the correlation structure to enable more efficient computation. For example, Sang and Huang (2012) developed an approach that takes advantage of more efficient sparse matrix solvers by using compacted correlation functions that yield sparse correlation structures. A third approach is to represent the spatial process in a lower-dimensional subspace (Stein 2007, Cressie and Johannesson 2008, Banerjee et al. 2008). The general idea behind these methods is to express the spatial process realization over a small set of locations (i.e. "knots") where the number of "knots" is much smaller than the number of observed sites. Ideally there will be insignificant loss of spatial information when using the smaller set of knots given that there is adequate domain coverage. The critical issue for these types of models is selecting the number and location of the knots, which is a challenging problem.

Given the large scale nature of many infectious diseases including West Nile Virus, and the increasing availability of data, our goal is to explore and evaluate one methodological approach to model spatial data efficiently. Banerjee et al. (2008) proposed an approach, predictive process modeling, which attempts to express the original spatial process (i.e. parent process) over a lower-dimensional subspace ("knots"). This approach has many advantages (e.g., requires no additional tuning parameters in the model) and we believe an illustration of the approach will provide researchers with a valuable tool. Specifically, our objectives are to: 1) develop a geostatistical model to effectively assess which covariates are responsible for influencing WNV distribution and occurrence, 2) assess the predictive performance of the developed model, 3) explore the use of predictive process models to reduce the computational burden associated with fitting spatial models to large data sets, and 4) evaluate four alternative ways to select knots in order to maximize the accuracy and precision of the predictive process models. We illustrate this approach with one year of data from New York City, NY, USA.

METHODS

The data were collected by the New York City Department of Health and Mental Hygiene (NYCDHMH) in 2008 across the five boroughs of New York City, New York, USA (Bronx, Brooklyn, Manhattan, Queens, and Staten Island, Fig. 2.1). Mosquitoes were collected weekly in CDC light and Reiter's gravid traps at 225 sites from May 2008 – September 2008 (Fig. 2.1). Trap catch was separated in the lab to species, and grouped into pools of up to 50 individuals from the same species, on the same date and collected from the same trap. These pools were then tested for WNV using PCR. The results of the test (positive or negative for WNV) were used as the binary response variable for the statistical models described below. We only used results from *Culex pipiens* and *Culex pipiens-restuans* as most virus isolations have

been made from various *Culex* species mosquitoes (Turell et al. 2001). Each trap location was mapped according to the geographic coded record from the NYCDOHMH and converted to the NAD 1983 State Plane New York Long Island FIPS 3104 coordinate system for further analysis. Eleven covariates thought to influence WNV distribution and variability were collected. The covariates were land cover type (Land cover), distance to water (Distance water), distance to open space (Distance openspace), distance to building (Distance building), population density (Population density), lot density (Lot density), housing density (House density), distance to wetlands (Distance wetlands), aspect (Aspect), slope (Slope), and land surface temperature (Land surface temp). The covariates were either provided by the NYCDOHMH or obtained using the National Land Cover Database 2006 (http://www.mrlc.gov/nlcd2006.php). We treated Land Cover type as a factor variable and used six levels because our study area was predominantly covered by these six land classifications: Developed open space, Developed Low Intensity, Developed Medium Intensity, Developed High Intensity, "Forest" (Deciduous Forest and Evergreen Forest combined), and "Wetlands" (Woody Wetlands and Emergent Herbaceous Wetlands combined).

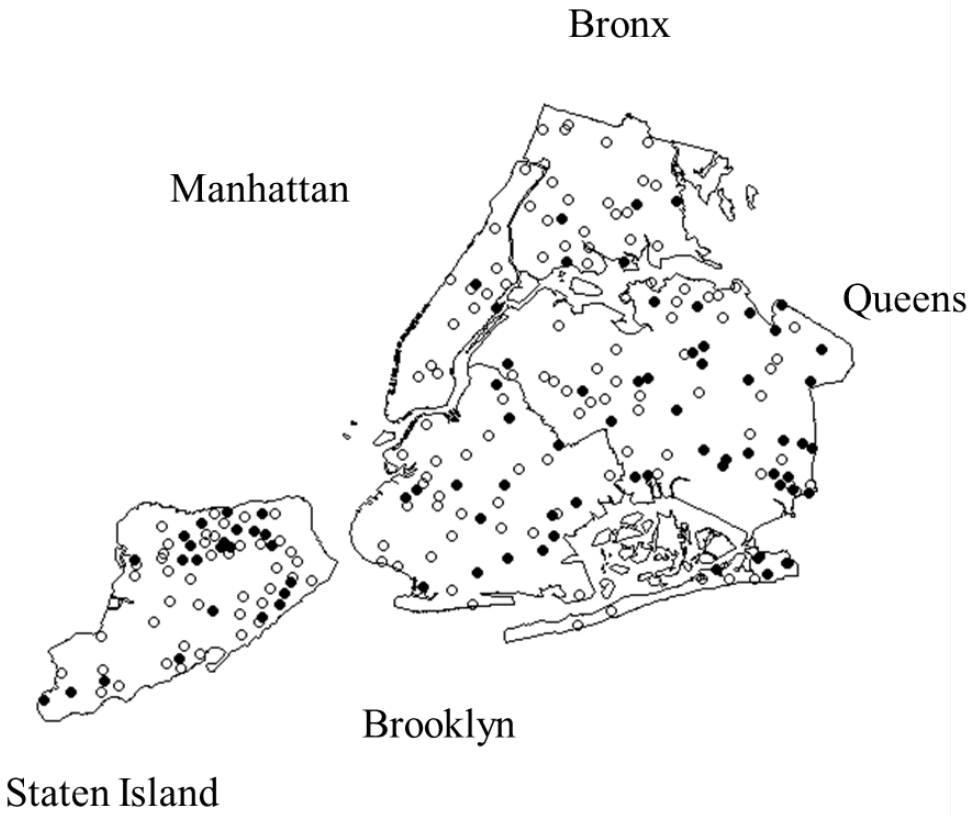
Bronx
Manhattan
Queens
Brooklyn
Staten Island

Figure 2.1. Mosquito trap locations in New York City, NY, USA with five boroughs identified. Black circles represent cases where a positive test for WNV occurred and white circles represent cases where a negative test result occurred during mosquito sampling from May – September 2008. See text for more description.

Statistical Model

We were interested in exploring the influence of the eleven covariates on the occurrence and distribution of WNV in New York City, NY, USA. To do so we developed a multi-stage hierarchical geostatistical model that allows for spatial correlation among observations. We first discuss a basic logistic model that can accommodate the non-normal distribution of WNV (binary response) and then add in spatial random effects. Suppose we have $i = 1,\ldots, n$ locations. We set $y_i = 1$ if location i tested positive for WNV and $y_i = 0$ otherwise. Conditional upon our m predictor variables, $\mathbf{x_i}$ for site i, we assume that the y_i's follow a Bernoulli distribution, $y_i \overset{i.i.d.}{\sim} Bernoulli(p(\mathbf{x}_i))$ with $P(y_i = 1|\mathbf{x_i}) = p(\mathbf{x_i})$. A logistic link is used to model the relationship between the response data vector $\mathbf{y} = (y_1, \ldots, y_n)$, and the matrix of predictor variables $\mathbf{X} = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T]$, where $\boldsymbol{x}_i^T$ is the 1 x m vector of covariates for the i-th point,

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is the vector of parameters to be estimated. Classical inference would typically use iterative methods to obtain estimates of the parameters, $\widehat{\boldsymbol{\theta}}$, by maximizing the likelihood function,

$$\prod_{i=1}^{n} \frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\theta})}. \tag{2}$$

Spatial Random Effects

It has been suggested that West Nile Virus outbreaks occur with some form of spatial dependence (e.g., Chuang et al. 2012, Young and Jensen 2012); unfortunately the mean function in (1) does not accommodate spatial correlation among observations. Not accounting for

existing spatial dependence can impair the estimates of precision and often times leads to a false sense of confidence when evaluating the significance of important predictors. Therefore we used a multi-stage hierarchical model that explicitly incorporates spatial structure to ensure our estimates of uncertainty were accurate. Multi-stage models provide more modeling flexibility which often results in a better model fit compared to simpler single-stage models. In our case the first stage of the hierarchical model assumes that the observed responses over the locations are conditionally independent given the spatial effects and adds spatially correlated random effects to the mean structure in (1). The second stage specifically models the nature of the association between the random effects. We adopt a Bayesian approach to estimation and inference (e.g., Gelman et al. 2004) therefore we require prior distributions of the model parameters for full specification of the hierarchical model.

Specifically, suppose the trap collection sites are spatially referenced as $S = \{s_1, \ldots, s_N\}$. We can envision the response as $y(s_i) = 1$ or 0 depending on whether the test for WNV was positive or not at site s_i. The probability that $y(s_i) = 1$ depends on spatially-referenced predictor variables, $\mathbf{x}(s_i)$ for site s_i, the regression slope parameters $\boldsymbol{\theta}$, and the location-specific random effects $w(s_i)$to yield:

$$p(s_i) = \frac{\exp(\mathbf{x}(s_i)^T\boldsymbol{\theta} + w(s_i))}{1+\exp(\mathbf{x}(s_i)^T\boldsymbol{\theta} + w(s_i))}. \quad (3)$$

Here we assume $s \in D$, where D is a fixed subset of $\Re^2$.

The second stage of the hierarchical model specifies the association in the random effects. We chose to use a Gaussian Process to specify the random effect, denoted by $w(s) \sim GP(\mu(s), K(\cdot))$ where $\mu(s)$ is the process mean and $K(\cdot)$ is a positive definite covariance

function (see Cressie 1993 or Banerjee et al. 2004 for more details). The Gaussian Process provides several advantages, most notably very convenient distribution theory (Banerjee et al. 2004). By specifying the mean and covariance structure we can obtain the joint, conditional and all other distributions with little work (Banerjee et al. 2004). The Gaussian Process is also a popular choice for modeling spatial variation due to its ability to directly model spatial correlation.

We assume $w(s) \sim GP(0, K(\phi))$, where $K(s - s'; \phi) = \sigma^2 \rho(s - s'; \phi)$ such that the process realization $\mathbf{w} = [w(\boldsymbol{s}_i)]_{i=1}^{N} \sim MVN(\mathbf{0}, \sigma^2 H(\phi))$, where $H(\phi) = \left[\rho(s_i - s_j; \phi)\right]_{i,j=1}^{N}$ is the N x N spatial correlation matrix. If the $w(s_i)$ were i.i.d., then we would have the usual generalized linear mixed effects model, but instead we have a generalized linear mixed model with spatial structure in the random effects (Banerjee et al. 2004). The spatial correlation function, $\rho(s_i - s_j; \phi)$, captures the strength of spatial association and is also known as a positive definite function because it must ensure that the matrix $H(\phi)$ is positive definite for any collection of sites (Banerjee et al. 2004). Here we use the exponential correlation function, $\rho(s - s'; \phi) = \exp(-\phi \| s - s' \|)$ with two parameters, the spatial decay parameter ϕ and the spatial effect variance σ^2. We also describe the effective range r of the spatial process by solving $\exp(-\phi r) = 0.05$ (i.e., $r \approx 3/\phi$) which describes the range at which the spatial correlation drops to 0.05 (Banerjee et al. 2004).

To complete the full specification of the model we need to assign prior distributions for the parameters. We chose a non-informative flat prior for the fixed effects $\boldsymbol{\theta}$ (i.e., $p(\theta) \propto 1$). Following Gelman et al. (2004) we assumed the spatial effect variance parameter, σ^2 follows an inverse-Gamma distribution, $\sigma^2 \sim IG(a, b)$. We fixed a = 2 such that the distribution has mean b

and an infinite variance. The large variance should allow for the data to overwhelm the prior beliefs and dominate the inference. The spatial decay parameter was distributed as Uniform, $\phi \sim U(c, d)$. As other authors have noted (Berger et al. 2001; Banerjee et al. 2004), we required a fairly informative prior to ensure proper and well-identified posteriors and sufficient convergence of the algorithm. We set *d* as the maximum intersite distance between locations in our data set and specified *c* as a minimal distance that we believed would always be exceeded.

Predictive Process Models

Estimation and prediction of the above parameters involves evaluating the likelihood and the *N* x *N* matrix. Although there are fast linear solvers to conduct such explicit inversions it is still expensive for big *N*. This problem becomes exacerbated for Bayesian inference which requires repeated evaluation as needed in Markov Chain Monte Carlo (MCMC) algorithms.

Recently, Banerjee et al. (2008) suggested a class of models known as predictive process models that operate on an identified lower-dimensional subspace. By projecting the original process to the lower-dimensional subspace the dimension of the matrix which needs to be evaluated is reduced thus decreasing computational time. Following Banerjee et al. (2008) a set of "knots" are specified $S^* = \{s_1^*, \ldots, s_m^*\}$, which may or may not form a subset of the entire collection of observed locations S. The Gaussian Process is now defined as $\boldsymbol{w}^* = [w(\boldsymbol{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, \sigma^2 H^*(\phi))$, where $H^*(\phi) = \left[\rho(s_i^* - s_j^*; \phi)\right]_{i,j=1}^m$ is the corresponding *m* x *m* spatial correlation matrix. The predictive process $\widetilde{w}(\boldsymbol{s})$ derived from the parent process $w(\boldsymbol{s})$ is defined as $\widetilde{w}(\boldsymbol{s}) = E[w(\boldsymbol{s})|\boldsymbol{w}^*] = \boldsymbol{h}^T(\boldsymbol{s}; \phi) H^{*-1}(\phi) \boldsymbol{w}^*$, where $H^*(s - s'; \phi) = \sigma^2 \rho(s - s'; \phi) = H^*(\phi)$ and $\boldsymbol{w}^*$ is defined above. The process is completely specified

given the covariance function of the parent process and S^*. The probability of a positive WNV detection at a site now depends on the predictive process model,

$$p(s_i) = \frac{\exp(\mathbf{x}(s_i)^T\boldsymbol{\theta}+\tilde{w}(s_i))}{1+\exp(\mathbf{x}(s_i)^T\boldsymbol{\theta}+\tilde{w}(s_i))}. \quad (4)$$

The dimension reduction is seen by replacing the *N* random effects $[w(\boldsymbol{s}_i)]_{i=1}^N$, with *m* random effects in $\boldsymbol{w}^*$ where we work with an *m*-dimensional joint distribution involving only *m* x *m* matrices. Finley et al. (2009a) and Finley et al. (2009b) point out the predictive process systematically underestimates the variance of the parent process *w*(**s**) at any location **s** and this is observed in the results from Banerjee et al. (2008). To fix this problem Finley et al. (2009b) propose a modified predictive process, defined as $\ddot{w}(\boldsymbol{s}) = \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s})$, where $\tilde{\epsilon}(\boldsymbol{s}) \overset{indep}{\sim} N(0, H(\boldsymbol{s},\boldsymbol{s}) - \boldsymbol{h}^T(\boldsymbol{s};\phi)H^{*-1}(\phi)\boldsymbol{h}(\boldsymbol{s};\phi)$ is a process of independent variables with spatially adaptive variances. This allows the modified predictive process $\ddot{w}(\boldsymbol{s})$, to have the same properties of the predictive process $\tilde{w}(\boldsymbol{s})$, while reducing the systematic bias in estimating the variance of the parent process *w*(**s**). More details of the predictive process model including derivations and theoretical results can be found in Banerjee et al. (2008), while details of the modified predictive process can be found in Finley et al. (2009a) and Finley et al. (2009b).

Selection of knots

The selection of knots for the modified predictive process model is an important and difficult step. Essentially the problem comes down to a spatial design problem except that we already have samples collected at *N* locations. Two common approaches for obtaining optimal spatial designs are design-based approaches and model-based approaches (see Xia et al. 2006 for review). The design-based approaches attempt to obtain optimal placement of locations and are

evaluated by how well the given set of points covers the study region irrespective of the assumed covariance function. Common design-based approaches include using a regular grid of locations that are spatially balanced (Stevens and Olsen 2004), space-filling knot selection based on the designs of Nychka and Saltzman (1998) or modifying a regular grid by augmenting the lattice with close pairs or infill (e.g., Diggle and Lophaven 2006). Alternatively, model-based approaches rely on an assumed spatial model and either focus on optimizing the placement of locations in regards to the ability to make predictions (e.g., Zhu 2002), to estimate model parameters (e.g., Müller and Zimmerman 1999), or a mixture of both (e.g., Zhu and Stein 2005).

The selection of the number of knots is governed by computational cost regardless of the choice of approach for optimizing knot-selection. This requires implementing the analysis over different choices for the number of knots and must include run time along with the stability of estimated parameters. We chose to explore four different design-based approaches and conducted each approach at eight different numbers of knots ranging from 15% - 60% of the total number of observed locations. The four different design-based approaches were: (1) a systematic grid covering the entire study area, (2) random placement of knots throughout the study area, (3) proportional placement of knots in relation to observed locations, i.e. areas with more observed locations had more knots, and (4) an optimal spatial design which minimized a geometric space-filling design (Nychka and Saltzman 1998). We ensured the same number of total knots were used for each of the four approaches and increased the number of knots from a minimum of 34 (15% of total number of observed locations) to a maximum of 141 (63% of total number of observed locations) for each run.

Model validation and prediction

In addition to fitting the model to the full data set, we randomly sampled 20% of the data and excluded them in order to validate the model (holdout data set). The remaining 80% of the data were then used for model construction. Using the model constructed with 80% of the data we predicted the response (probability of positive infection) at the locations in the holdout data set and evaluated the model's performance.

We were interested in predicting WNV prevalence throughout the entire study area as well as using predicted responses to compare knot-placement and intensity approaches. To facilitate this we required predictions at new unsampled locations based on the model fit to the full data. We obtained new covariate information at 434 sites within the study area (Fig. 2.2) and predicted probability of infection using the posterior predictive distribution (e.g., Gelman et al. 2004) at each new location.

Convergence diagnostics

All analyses were conducted using the software program R (R Development Core Team 2011). MCMC was implemented via the R package spBayes (Finley et al. 2007) using adaptive MCMC (Roberts and Rosenthal 2009). Three independent chains were run for 250,000 iterations for each trial. Each chain was given dispersed initial values for each parameter in the model and the posterior samples were thinned to keep every 20th iteration. Convergence was assessed by examining posterior samples using the software package CODA (Plummer et al. 2006) available in R and consisted of graphing trace plots, assessing autocorrelation and computing Gelman-Rubin diagnostics (Gelman et al. 2004).

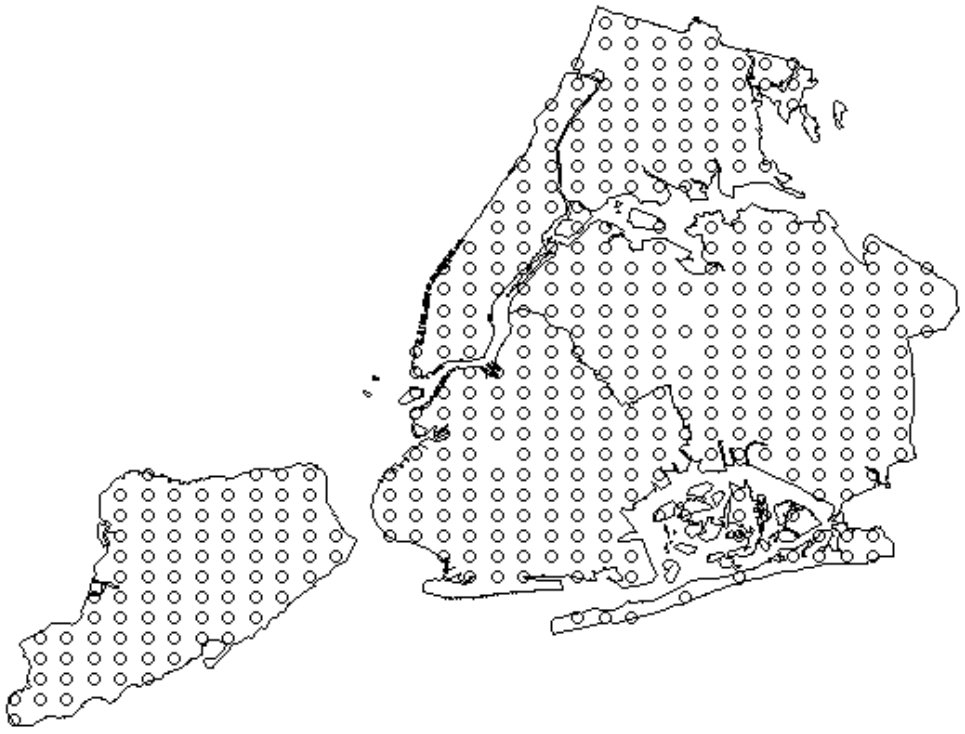

Figure 2.2. New locations for prediction of West Nile virus from Bayesian hierarchical spatial model in New York City, NY, USA.

RESULTS

The average distance between sites is approximately 19780 meters with a maximum distance of 54711 meters and a minimum distance of 21 meters. Table 2.1 provides parameter estimates (50, 2.5, 97.5 percentiles) for the multi-stage hierarchical spatial model fit to the full data set which consisted of 225 sites. The credible intervals suggest that there is a positive effect of land cover type (specifically wetlands), housing density, and distance to wetlands and a negative effect of slope on the occurrence of WNV in New York City (Table 2.1). The medians of $\sigma^2$and ϕ are 242.19 and 602.84, respectively. The effective range ($\approx 3/\phi$) for the spatial model is approximately 560 meters suggesting that at this distance the correlation between two locations is 0.05. There is a large amount of variability around the estimate of ϕ and the associated effective range, however, suggesting that there is quite a bit of uncertainty regarding the strength of spatial dependence among sites. Furthermore the upper estimate of the credible interval (1499.56 meters) suggests relatively small-scale local dependence given the average distance between sites and distribution of sites.

Figures 2.3 and 2.4 provide the probability map and associated error for the predicted probability of WNV in New York City based on the multi-stage hierarchical model using all of the data. It is apparent that the model predicts a high probability of WNV in Staten Island which corresponds to the large number of observed cases found in Staten Island. It is also evident that uncertainty is lowest in Staten Island due in part to the large amount of information collected in this borough. The model does a poor job of predicting observed cases in the middle of Queens and the northern region of the Bronx and Manhattan. There are also high levels of uncertainty with the predictions in the middle of Queens, but the uncertainty is lower in the northern region of the Bronx. Examining maps of the distribution of land cover type (Fig. 2.5) and housing unit

Table 2.1. Parameter estimates (median and 95% credible intervals) from Bayesian hierarchical spatial model fit to the full data set exploring the influence of eleven covariates on the distribution and occurrence of WNV in New York City, NY. Effective range is defined as distance at which the correlation among locations drops to 0.05. Italicized variables represent covariates that have support and are considered important to understanding the distribution or prediction of WNV occurrence.

Parameter	50%	2.50%	97.50%
σ^2	242.19	116.09	451.50
ϕ	602.84	221.76	979.72
Effective range (meters)	560.33	343.58	1499.56
Developed Low Intensity	3.90	-4.84	13.67
Developed Med. Intensity	1.22	-8.96	11.57
Developed High Intensity	0.35	-10.54	11.79
Forest	1.69	-10.52	14.06
Wetlands	20.08	5.59	36.74
Distance water	-1.06	-4.57	2.07
Distance open space	0.14	-3.12	3.37
Distance building	2.10	-1.07	5.56
Aspect	-1.07	-4.04	1.73
Slope	-2.20	-5.79	0.92
Population density	-1.38	-8.34	4.64
Lot density	-4.35	-14.98	3.38
House density	6.03	1.57	11.70

Distance wetlands	3.70	-0.65	8.74
Land surface temp	-0.27	-3.49	2.92

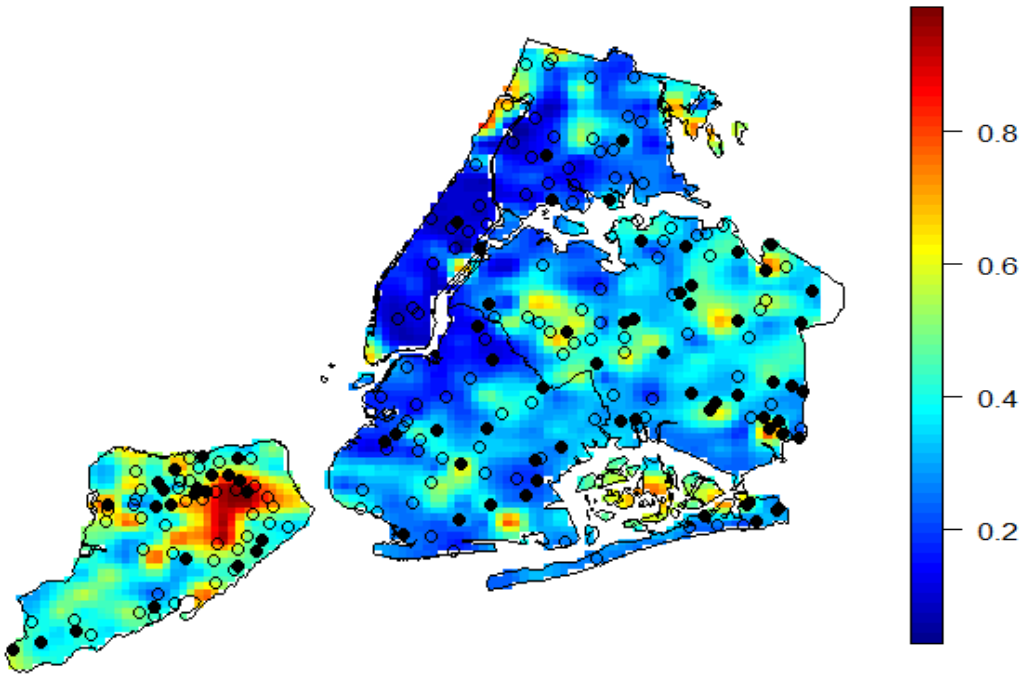
0.8
0.6
0.4
0.2

Figure 2.3. Mean predicted occurrence of WNV in New York City, NY, USA. Predictions are made from a Bayesian hierarchical spatial model fit to the full data set of detected and nondetected cases of WNV collected from May – September 2008. Black circles represent cases where a positive test for WNV occurred and white circles represent cases where a negative test result occurred during mosquito sampling from May – September 2008.

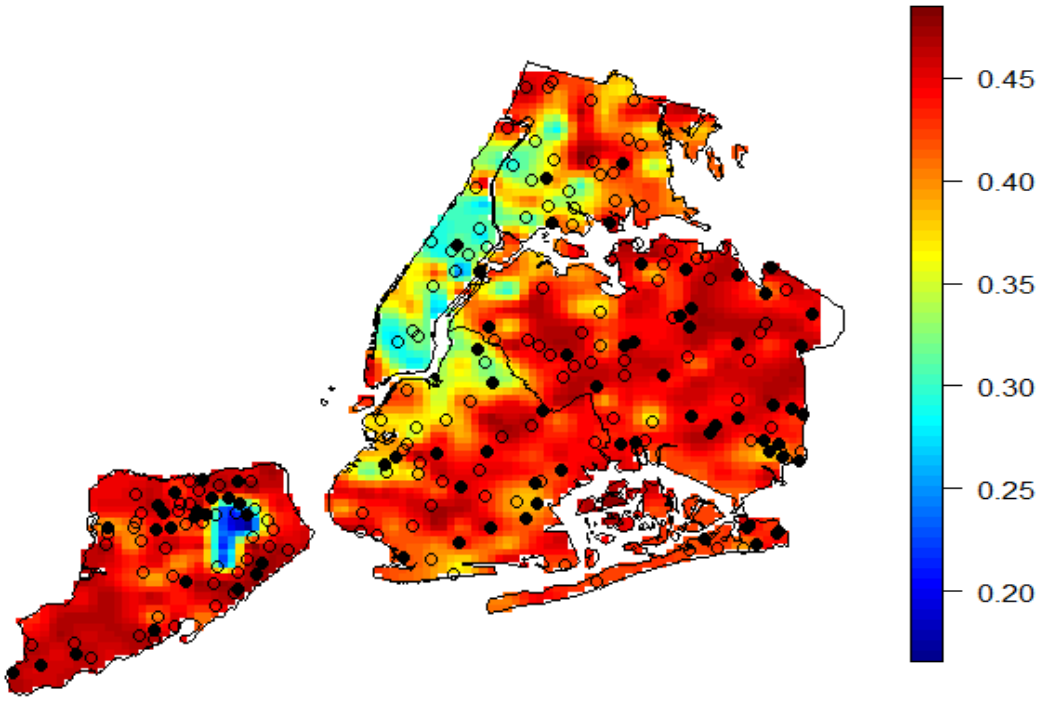
0.45
0.40
0.35
0.30
0.25
0.20

Figure 2.4. Estimated uncertainty from mean predictions of WNV occurrence in New York City, NY, USA. Predictions are made from a Bayesian hierarchical spatial model fit to the full data set of detected and nondetected cases of WNV collected from May – September 2008. Black circles represent cases where a positive test for WNV occurred and white circles represent cases where a negative test result occurred during mosquito sampling from May – September 2008.

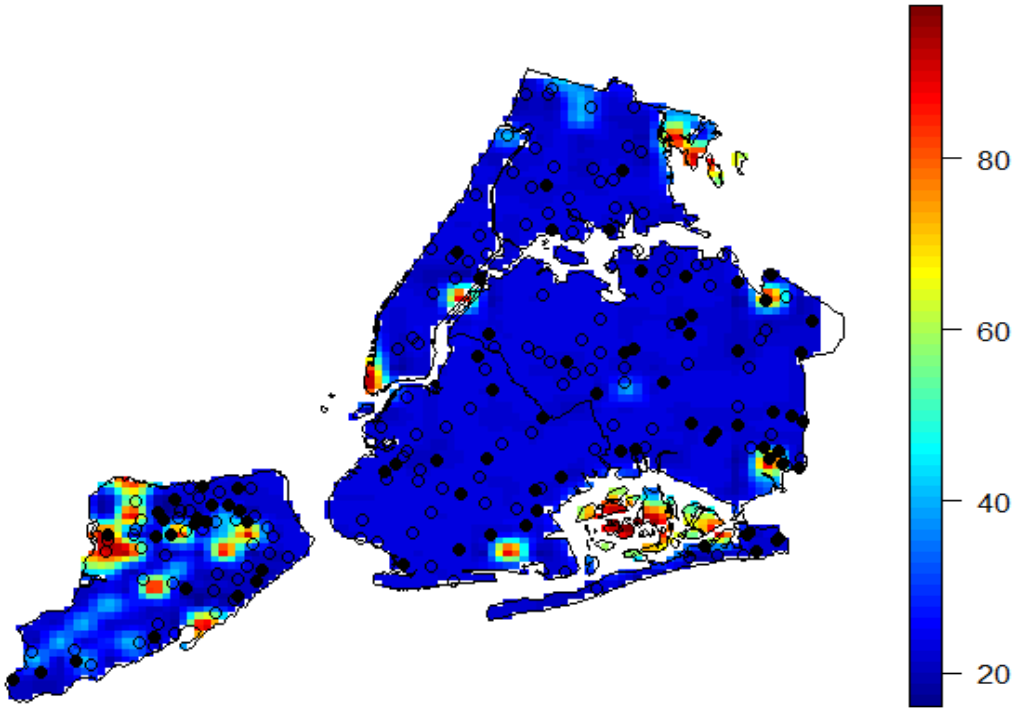
80
60
40
20

Figure 2.5. Distribution of six land cover types in New York City, NY, USA based on the National Land Cover Database 2006. The six land cover types are: 21 – Developed open space, 22 – Developed low intensity, 23 – Developed medium intensity, 24 – Developed high intensity, 41 and 42 – Deciduous forest and Evergreen forest, referred to as “Forest”, and 90 and 95 – Woody wetland and Emergent herbaceous wetland, referred to as “Wetlands”.

density (Fig. 2.6) suggests the model's predictions are highly dependent on these two covariates; high predicted probability in areas classified as wetlands and with high housing unit density. The majority of locations with high predicted probability of WNV correspond to areas that have a higher number of observed positive cases relative to negative cases, for instance in Brooklyn and Queens. This result makes intuitive sense because the ratio of positive to negative cases is the dominant type of information to inform the model. Model validation for the multi-stage hierarchical model fit to the full data set resulted in a correct classification rate of 68% when using a cut-point of 0.5. Model construction for validation used 185 sites to estimate parameters and predict to the 46 sites in the hold-out data set.

Banerjee et al. (2008) and Finley et al. (2009b) suggest that the best way to evaluate knot performance is by comparing the covariance function of the parent process to the predictive process. Table 2.2 displays estimates of $\sigma^2$and ϕ for a combination of knot intensities and placement strategies for the multi-stage hierarchical model fit using a modified predictive process. We only report levels of knot intensity when the total computing time was less than that of the model fit to the full data set (i.e., number of knots < 40% of full data set). Three of the four approaches (random allocation, proportional allocation, and optimal design) did a sufficient job of estimating σ^2 regardless of the number of knots used while the systematic grid did a poor job of estimation for all of the levels of knots. The systematic grid routinely underestimated the uncertainty in σ^2 and often the estimate of the median was significantly biased low. There is very little variation among the other three approaches (random, proportional, and optimal) regarding the ability to accurately represent σ^2 with a variety of knot intensities. Based on computing time it appears as though choosing any of the three approaches (random, proportional,

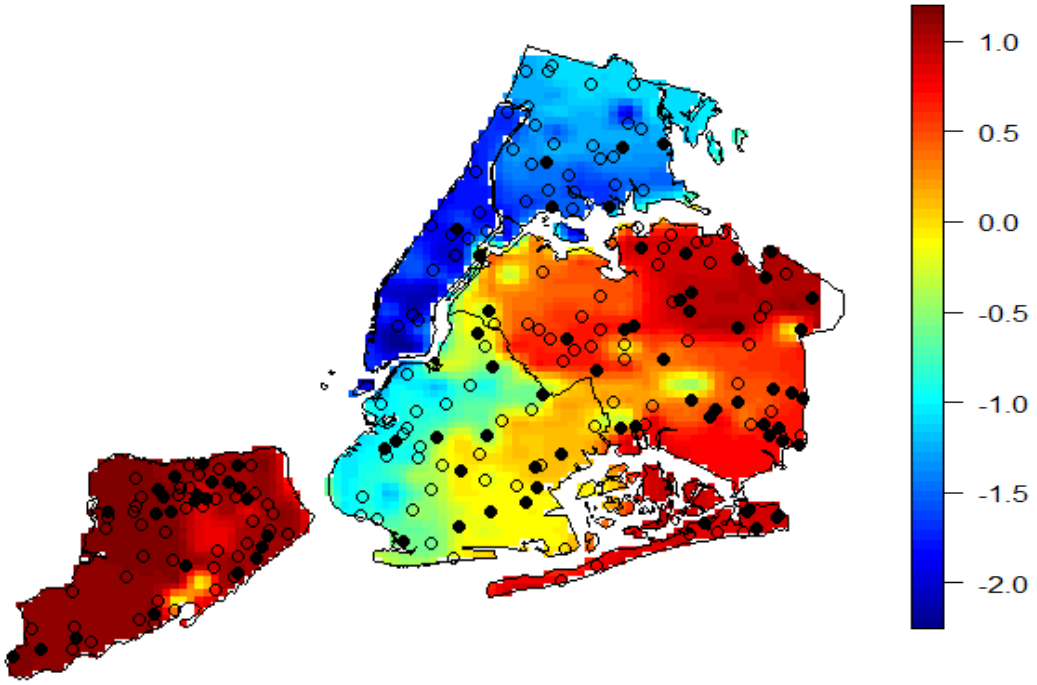
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0

Figure 2.6. Distribution of housing unit density in New York City, NY, USA. The measurement unit has been centered and re-scaled to have mean equal to zero and variance equal to one with higher numbers representing higher housing unit density.

Table 2.2. Median and 95% credible intervals for estimates of the parameters of the covariance function from a Bayesian hierarchical spatial model for the full data set (in italics) and predictive process model using different levels of knots and knot placement approaches. Note that the numbers of knots are in percentages of the total number of sites (225 total sites): 15% - 34 knots, 20% - 46 knots, 25% - 58 knots, 28% - 65 knots, 34% - 79 knots, 37% - 86 knots. The four knot placement approaches are systematic grid (Grid), randomly placed (Random), proportional to observed locations (Proportional), and an optimal space-filling design (Optimal). Computing time represents the total amount of time to complete estimation of parameters in hours.

		σ^2			ϕ			Computing
Model	Knots % of Total	Median	2.50%	97.50%	Median	2.50%	97.50%	Time (hours)
Full	-	242.19	116.09	451.50	602.84	221.76	979.72	***14.02***
Grid	15	194.97	194.96	195.23	465.01	77.08	974.48	4.05
	20	252.02	234.83	281.46	449.25	11.14	973.98	5.56
	25	219.28	181.30	249.91	446.79	32.97	972.66	6.75
	28	196.02	64.69	378.52	552.21	128.03	978.31	8.71
	34	207.10	202.98	210.02	489.19	96.35	974.67	10.05
	37	164.47	134.02	168.18	481.49	89.18	974.39	11.44
Random	15	233.39	104.34	450.01	542.89	89.84	975.09	3.65
	20	237.84	107.37	443.88	559.85	149.44	976.24	5.71
	25	236.85	110.65	436.89	573.87	123.28	978.08	7.28
	28	236.74	102.76	451.50	553.68	110.28	977.28	8.48
	34	234.83	103.82	444.17	575.28	146.97	977.65	10.19
	37	236.03	111.22	426.24	554.84	124.96	978.94	10.64

Proportional	15	235.62	108.15	440.58	437.18	48.67	971.08	4.24
	20	231.61	103.95	439.51	540.02	57.02	975.77	5.99
	25	230.89	103.91	439.82	539.07	71.79	975.08	7.14
	28	239.39	103.80	450.27	542.75	74.81	979.14	8.42
	34	245.44	114.08	445.49	473.56	67.57	975.39	10.16
	37	240.34	111.85	449.51	475.43	78.58	972.53	11.70
Optimal	15	235.07	103.24	438.17	349.27	51.05	964.72	3.50
	20	238.23	103.30	447.57	518.03	114.68	975.36	5.55
	25	234.49	103.94	424.21	579.25	118.88	979.77	7.16
	28	240.00	111.64	444.93	451.13	71.42	973.27	7.45
	34	242.71	113.59	458.10	609.21	159.59	981.18	10.30
	37	236.68	102.30	454.75	548.05	101.65	980.15	12.37

and optimal) with 34 knots should be the optimal balance of accuracy and efficiency because very little is gained by increasing the number of knots.

Estimation of ϕ appeared to pose a greater challenge for all four approaches. All of the approaches did a poor job of capturing the estimated variability around σ^2 (using the full data set) and instead overestimated the uncertainty. The performance did not appear to improve with any of the four approaches when increasing knot intensity. It appears as though there were different optimal knot intensities for the four approaches. The systematic grid and proportional allocation approach performed the best when using 65 knots, while the random allocation of knots and optimal design allocation of knots performed best when using 79 knots. Overall the optimal design allocation with 79 knots was closest to representing the estimate and associated uncertainty from the full data set. This ambiguity suggests that the modified predictive process model had a difficult time estimating σ^2 regardless of the approach or number of knots and there was not an increase in performance as knot intensity was increased as would be expected.

In addition we can examine the performance of the modified predictive process to estimate the influence of covariates. Table 2.3 provides parameter estimates for the four covariates with evidence of support for all four approaches and knot intensities. It is clear that performance does not vary much among the random allocation, proportional allocation and optimal design allocation while the systematic grid performed much worse when knot intensity was low.

DISCUSSION

Our goal was to explore the influence of several covariates on the prevalence of WNV in New York City while evaluating a recently developed approach to handle large spatial datasets.

Table 2.3. Parameter estimates (median and 95% credible intervals) for the four covariates with support from the Bayesian hierarchical spatial model fit to the full data set (in italics) and predictive process model using different levels of knots and knot placement approaches. Note that the numbers of knots are in percentages of the total number of sites (225 total sites): 15% - 34 knots, 20% - 46 knots, 25% - 58 knots, 28% - 65 knots, 34% - 79 knots, 37% - 86 knots. The four knot placement approaches are systematic grid (Grid), randomly placed (Random), proportional to observed locations (Proportional), and an optimal space-filling design (Optimal).

	Knots % of Total	Wetlands land cover			Slope			House density			Distance wetlands		
		50%	2.50%	97.50%	50%	2.50%	97.50%	50%	2.50%	97.50%	50%	2.50%	97.50%
Full		*20.08*	*5.59*	*36.74*	*-2.20*	*-5.79*	*0.92*	*6.03*	*1.57*	*11.70*	*3.70*	*-0.65*	*8.74*
Grid	15	19.23	5.41	33.55	-2.11	-4.99	0.73	5.60	1.43	9.95	3.43	-0.51	7.52
	20	3.91	1.40	7.66	-0.32	-0.76	0.07	0.79	0.23	1.39	0.46	-0.08	1.00
	25	4.20	1.45	8.15	-0.36	-0.85	0.08	0.91	0.27	1.61	0.54	-0.08	1.17
	28	5.96	2.14	10.83	-0.54	-1.30	0.15	1.38	0.39	2.39	0.82	-0.12	1.78
	34	19.03	5.44	32.94	-2.26	-5.30	0.71	5.58	1.45	9.95	3.34	-0.63	7.59
	37	18.21	5.41	31.27	-1.91	-4.67	0.69	5.28	1.31	9.21	3.23	-0.50	6.93
Random	15	19.67	5.36	36.83	-2.21	-5.77	0.82	5.91	1.47	11.50	3.50	-0.83	8.75

	20	19.80	4.78	35.99	-2.22	-5.81	0.91	6.21	1.86	11.83	3.78	-0.19	8.92
	25	19.64	4.70	36.31	-2.20	-5.71	0.87	5.89	1.40	11.25	3.59	-0.53	8.48
	28	20.06	5.62	36.17	-2.15	-5.71	0.87	5.94	1.59	11.51	3.56	-0.57	8.47
	34	20.04	5.46	36.55	-2.11	-5.77	1.00	5.83	1.43	11.79	3.54	-0.56	8.57
	37	19.97	5.63	35.83	-2.18	-5.62	0.99	6.08	1.62	11.58	3.67	-0.56	8.76
Proportional	15	20.14	5.16	36.25	-2.23	-5.82	0.91	6.11	1.65	11.81	3.65	-0.51	8.61
	20	19.65	5.12	35.68	-2.20	-5.81	0.74	5.84	1.65	11.61	3.48	-0.58	8.55
	25	19.80	5.53	35.25	-2.14	-5.76	0.84	5.88	1.50	11.58	3.50	-0.48	8.53
	28	20.09	5.32	36.65	-2.10	-5.93	0.97	6.24	1.65	11.50	3.77	-0.50	8.60
	34	20.08	5.20	37.07	-2.28	-6.15	0.81	6.18	1.63	11.51	3.73	-0.76	8.48
	37	20.22	5.67	37.11	-2.11	-5.64	0.91	6.25	1.70	12.06	3.74	-0.57	9.02
Optimal	15	19.55	5.58	35.73	-2.20	-5.72	0.89	5.90	1.59	11.28	3.52	-0.53	8.67
	20	20.23	5.68	36.20	-2.08	-5.59	0.95	5.91	1.51	11.71	3.60	-0.66	8.47
	25	19.46	5.64	34.92	-2.13	-5.69	0.88	5.95	1.64	11.00	3.59	-0.56	8.29
	28	19.50	5.91	35.42	-2.11	-5.60	1.09	6.17	1.67	11.96	3.73	-0.46	8.96
	34	20.44	5.91	36.61	-2.14	-5.82	0.95	6.01	1.51	11.67	3.58	-0.67	8.68

37	19.93	5.12	36.59	-2.15	-5.88	1.00	6.02	1.57	11.72	3.69	-0.66	8.74

In order to do so we developed a hierarchical Bayesian model that incorporated spatial structure in the random effects and permitted accurate estimates of the associated uncertainty. We found four covariates (land cover type, distance to wetlands, slope, and housing density) to have a significant influence on the distribution of WNV. The hierarchical model using the full data set did a sufficient job of predicting WNV throughout the study area with a few exceptions. The most notable exceptions were in Staten Island and northern regions of the Bronx and Manhattan where the model appeared to over predict WNV occurrence. The model's predictions do, however, align with other authors' assertion that disease transmission and occurrence are driven by the distribution of land cover type (Bowden et al. 2011, Magori et al. 2011) and urban density (Gibbs et al. 2006, Ruiz et al. 2007). Not surprisingly, Staten Island and the northern region of the Bronx have high levels of both of these types of land cover (Figures 2.5 and 2.6).

The model also appeared to underrepresent disease prevalence in some interior portions of Brooklyn and Queens. We believe that this was a result of very small scale spatial structure that may not have been fully captured by the model. The average inter-site distance was large (19780 m) while the spatial range was estimated to be relatively small (560 m) suggesting that the spatial dependence occurs at a small scale, but the observed locations are more often than not very far away. With enough data this would not be problematic as the correlation could be estimated with few data points existing close together, unfortunately, with only 225 sites this may not have been the case in our study. If identifying small scale spatial structure is the primary focus, for instance because control measures may be implemented at a neighborhood or smaller scale wherein understanding individual transmission is important (e.g., Salje et al. 2012), then other approaches such as statistical agent-based models (Hooten and Wikle 2010) may characterize the process better and provide greater insight. In our study we now have evidence

to suggest that small scale spatial structure is important so perhaps a future step would be to explore models that scale up starting with a mechanistic model that describes the small-scale processes.

An additional limitation that we observed was that the data were not collected according to a probability-based sampling survey and were instead collected in a haphazard manner resulting in large variation in the placement of sampling locations. Ideally, an optimal spatial sampling design could be implemented that would prioritize areas of concern and allow for smaller scale spatial structure to be accurately estimated (e.g., Xia et al. 2006). In addition, we did notice a small level of discrepancy between the sampling coverage allocated to different boroughs relative to the amount of area in the borough. Although this was not a prevalent issue it was most evident in Staten Island which makes up roughly 22% of the land area in New York City (265.5 km^2/1213 km^2), but approximately 29% of the sampling locations occurred in this borough (66/225 total sites). We could envision a spatial sampling design that allocates sampling locations proportional to land area as one approach that may improve model performance. Often it is not possible to redesign surveys for various reasons (i.e. too costly, logistical constraints, project already initiated) and so a second suggestion would be to explicitly account for the preferential sampling in the model directly. Diggle et al. (2010) develop a geostatistical model that accounts for the stochastic dependence between the process that determines the data locations and the response being modeled. This could potentially alleviate some of the difficulties that we observed and perhaps improve overall model performance.

Our hierarchical model introduced spatial effects into the transformed mean allowing for similar mean responses at proximate locations after adjusting for covariates. An alternative direct or first-stage model could have been considered that would only focus on the spatial

relatedness of observations, but we were interested in spatial explanation in the mean response as a function of covariates. An additional constraint was that data had been collected and aggregated over a relatively large window of time (May – Sept.) so that the actual observations of proximate locations may really occur weeks apart. Therefore we believed it was necessary to restrict our interest to the mean response and add in spatial dependence at this stage of the model. A more complex model could be fit that addressed the temporal nature of the data and attempted to explore the spatial and temporal dynamics of WNV (e.g. Cressie and Wikle 2011), but that was not the focus here. If data were available for multiple years it would be possible to investigate the temporal variability occurring at multiple scales: within a single year, perhaps using monthly or weekly slices of the data, and among years exploring the spatial dynamics across different years. One can imagine that the data requirements for such an analysis would increase dramatically emphasizing the value of an approach like predictive process modeling. Sahu and Bakar (in press) provide a useful framework and example of using predictive process modeling for a large spatiotemporal data set.

Overall we found the modified predictive process models to adequately capture most attributes of the parent process. All of the models did a decent job of estimating σ^2 and the coefficients for the covariates when the number of knots was greater than 20% of the total number of locations, but had difficulties in estimating ϕ regardless of the number of knots. Interestingly, there was little evidence of increased performance in estimating parameters of the covariance function as the number of knots increased above 20% of the total number of sites. Banerjee et al. (2008) showed that the accuracy of the predictive process models gets worse as the separation between knots increases. Finley et al. (2009b) also found that predictive process models fail to capture local, small-scale dependence accurately and that the smaller the spatial

range the denser the locations of knots need to be to pick up the spatial effect. In our setting we had a small spatial range (560 m) and even with high numbers of knots relative to the number of locations (> 40%) there was poor coverage of the entire study region. So although we were increasing knot intensity, in relation to the study region it was not enough to sufficiently capture the fine scale spatial dependency. We therefore observed an increase in uncertainty surrounding the estimation of ϕ even as the number of knots increased.

The difficulty with estimating ϕ could also simply be due to lack of information in the sample. We believe two sources contribute to the lack of information in the sample: 1) lack of data given the large number of covariates and spatial effects we are trying to estimate, and 2) spatial signal weakened because of control efforts by NYCDOHMH. Although our sample size was not small there may not have been sufficient information available to estimate ϕ accurately along with the associated covariates and spatial effects.

A more serious consideration is that the NYCDOHMH is interested in minimizing the impact of mosquito-borne diseases. They implement an aggressive program to eliminate potential breeding sites via the elimination of standing water and the application of larvicide, and the curtailment of adult mosquitos when necessary, through the use of adulticide (NYCDOHMH 2009). These direct control measures are based on surveillance data and are used to specifically target neighborhoods and communities where the virus is reappearing. Although these approaches are undoubtedly beneficial, they can present a problem when trying to model the spatial distribution and dynamics of WNV. Because areas with reappearing WNV outbreaks are treated quickly it becomes difficult to collect data that contain enough information about WNV distribution and the spatial signal necessary for accurate estimation. If there is strong dependence among locations or specific site-level covariates that are driving WNV occurrence

some of the information to estimate those effects accurately is lost when sites are sprayed or otherwise treated. Ultimately, it is important to understand these geographic hotspots as they are often most responsible for propagating disease epidemics (Dowdy et al. 2012). Overall immediate response control measures are an effective approach to reducing the human related risk of WNV, but the recognition of the challenges it creates for modeling is worth mentioning.

We found very little variation in the performance of three of the four different approaches to selecting knot locations. Overall the random, proportional and optimal design approaches to selecting knot locations performed better than the systematic grid approach and this difference was most evident when using a specified number of knots that was less than 30% of the total number of locations. Although we found very little discrepancy among the three approaches we evaluated, which all performed sufficiently well, there are other approaches to selecting knots that may provide more accurate estimates given the nature of this study. We only chose to focus on design-based approaches, but a mixed approach that maximizes study coverage while taking estimation and/or prediction of the covariance parameters into account may be optimal. An additional approach that could provide improvements is to not select the knots prior to estimation. Instead a form of adaptive knot selection where the knot locations are not fixed, but are a stochastic component of the model and can be added in an adaptive fashion as estimation takes place (Guhaniyogi et al. 2011), may be useful. For this study such an approach could allow for a more directed effort in coverage of the study area and potentially provide better estimates of the fine scale spatial structure given the ad-hoc placement of observations. For example, as Guhaniyogi et al. (2011) found, knot locations can be adaptively placed such that the knots divide the domain of interest into equal portions to maximize the ability to estimate the covariance parameters. Issues surrounding knot placement are critical to effectively using

predictive process models and should be considered carefully. Ultimately, the decision for where to place knots may depend on the objectives of the modeling (estimation/prediction vs. full study area coverage), but it appears as though when there is fine scale spatial structure and interest lies in estimating that spatial structure then the four design-based approaches we explored could be improved upon.

The ability to predict disease outbreaks as a function of covariates can be an important step to reducing the spread of diseases and lessening the associated costs. Furthermore, the identification of specific covariates that have a positive effect on the prevalence of WNV (i.e. land cover type, distance to wetlands, and housing density) allows land managers to recognize areas that can potentially be disease hotspots and to focus control efforts on these specific locations. This type of information permits more effective and efficient measures to combat the spread of WNV. We believe the use of hierarchical modeling which permits accurate estimation of uncertainty while incorporating important covariates of interest and spatial dependence provides a useful approach to exploring the prevalence and distribution of WNV. The use of predictive process modeling offers a useful approach to reduce computational burdens associated with large spatial data sets. The use of these models, however, does not come without a cost regarding the selection of the number of knots and knot locations. We have shown several of these approaches perform sufficiently well and hope that the illustration of predictive process modeling enables other researchers to explore the use of these methods for their own work.

LITERATURE CITED

Anderson, J. F., T. G. Andreadis, C. R. Vossbrinck, S. Tirrell, E. M. Wakem, R. A. French, A. E. Garmendia, and H. J. Van Kruiningen. 1999. Isolation of West Nile Virus from mosquitoes, crows, and a cooper's hawk in Connecticut. Science 286:2331-2333.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall. Boca Raton, FL, USA.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society, Series B 70:825-848.

Beale, C. M., J. J. Lennon, D. A. Elston, M. J. Brewer, and J. M. Yearsley. 2007. Red herrings remain in geographical ecology: A reply to Hawkins et al. (2007). Ecography 30:845-847.

Berger, J., V. De Oliveira, and B. Sanso. 2001. Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association 96:1361-1374.

Bowden, S. E., K. Magori, and J. M. Drake. 2011. Regional differences in the association between land cover and West Nile Virus disease incidence in humans in the United States. American Journal of Tropical Medicine and Hygiene 84:234-238.

Christman, M. C. 2008. Statistical modeling of observational data with spatial dependencies.

Journal of Wildlife Management 72:23-33.

Chuang, T., C.W. Hockett, L. Kightlinger, and M.C. Wimberly. 2012. Landscape-level spatial patterns of West Nile Virus risk in the northern Great Plains. American Journal of Tropical Medicine and Hygiene 86:724-731.

Cressie, N. A. C. 1993. Statistics for Spatial Data. Wiley. New York, New York, USA.

Cressie, N. A. C. and G. Johannesson. 2008. Fixed rank kriging for large spatial datasets. Journal of the Royal Statistical Society, Series B 70:209-226.

Cressie, N. A. C. and C. K. Wikle. 2011. Statistics for Spatio-Temporal Data. Wiley. Hoboken, New Jersey, USA.

Diggle, P. J. and S. Lophaven. 2006. Bayesian geostatistical design. Scandinavian Journal of Statistics 33:53-64.

Diggle, P. J., R. Menezes, and T. L. Su. 2010. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society, Series C 59:191-232.

Dowdy, D.W., J.E. Golub, R.E. Chaisson, and V. Saraceni. 2012. Heterogeneity in tuberculosis and the role of geographic hotspots in propagating epidemics. Proceedings of the National Academy of Sciences 109: 9557 – 9562.

Finley, A. O., S. Banerjee, and B. P. Carlin. 2007. spBayes: An R package for univariate and

multivariate hierarchical point-referenced spatial models. Journal of Statistical Software 19:1-24.

Finley, A. O., S. Banerjee, P. Waldmann, and T. Ericsson. 2009a. Hierarchical spatial modeling of additive and dominance genetic variation for large spatial trial datasets. Biometrics 65:441-451.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand. 2009b. Improving the performance of predictive process models for large datasets. Computational Statistics and Data Analysis 53:2873-2884.

Fuentes, M. 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102:32-331.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis. Chapman and Hall. Boca Raton, FL, USA.

Gibbs, S.E.J., M.C. Wimberly, M. Madden, J. Masour, M.J. Yabsley, and D.E. Stallknecht. 2006. Factors affecting the geographic distribution of West Nile Virus in Georgia, USA: 2002 – 2004. Vector-Borne and Zoonotic Diseases 6: 73 – 82.

Guhaniyogi, R., A. O. Finley, S. Banerjee, and A. E. Gelfand. 2011. Adaptive Gaussian predictive process models for large spatial datasets. Environmetrics 22:997-1007.

Hooten, M. B. and C. K. Wikle. 2010. Statistical agent-based models for discrete spatio-temporal systems. Journal of the American Statistical Association 105:236-248.

Illian, J. B., J. Møller, and R. P. Waagepetersen. 2007. Spatial point process analysis for a plant community with high biodiversity. Environmental and Ecological Statistics 16:389-405.

Kilpatrick, A. M. 2011. Globalization, land use, and the invasion of West Nile Virus. Science 334:323-327.

Lanciotti, R. S. et al. 1999. Origin of the West Nile Virus responsible for an outbreak of encephalitis in the northeastern United States. Science 286:2333-2337.

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33-50.

Latimer, A. M., S. Banerjee, H. Sang, E. S. Mosher, and J. A. Silander Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters 12:144-154.

Magori, K., W. I. Bajwa, S. Bowden, and J. M. Drake. 2011. Decelerating spread of West Nile Virus by percolation in a heterogeneous urban landscape. PLOS Computational Biology 7:1-13.

Morens, D. M., G. K. Folkers, and A. S. Fauci. 2004. The challenge of emerging and re-emerging infectious diseases. Nature 430:242-249.

Müller, W. G., and D. L. Zimmerman. 1999. Optimal designs for variogram estimation.

Environmetrics 10:23-37.

Nash, D. et al. 2001. The outbreak of West Nile Virus infection in the New York City area in 1999. The New England Journal of Medicine 344:1807-1814.

New York City Department of Health and Mental Hygiene. 2009. Comprehensive mosquito surveillance and control plan.

Nychka, D. and N. Saltzman. 1998. Design of air-quality monitoring networks. In "Case Studies in Environmental Statistics, Lecture Notes in Statistics". Editors Nychka, D. L. Cox, W. Piegorsch. Springer Verlag. New York, NY, USA.

Pimentel, D., R. Zuniga, and D. Morrison. 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics 52:273-288.

Plummer, M., N. Best, K. Cowles and K. Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC, R News 6:7-11.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roberts, G. O. and J. S. Rosenthal. 2009. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics 18:349-367.

Ruiz, M.O., E.D. Walker, E.S. Foster, L.D. Haramis, and U.D. Kitron. 2007. Association of West Nile Virus illness and urban landscapes in Chicago and Detroit. International Journal of Health Geographics 6: 1 – 11.

Sahu, S.K. and K. S. Bakar. In Press. Hierarchical Bayesian auto-regressive models for large

space time data with applications to ozone concentration modelling. Applied Stochastic Models in Business and Industry, In Press.

Salje, H., J. Lessler, T.P. Endy, F.C. Curriero, R.V. Gibbons, A. Nisalak, S. Nimmannitya, S. Kalayanarooj, R.G. Jarman, S.J. Thomas, D.S. Burke, and D.A.T. Cummings. 2012. Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. Proceedings of the National Academy of Science 109: 9535 – 9538.

Sang H. and J. Z. Huang. 2012. A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society, Series B 74:111-132.

Stein, M. L. 2007. Spatial variation of total column ozone on a global scale. The Annals of Applied Statistics 1:191-200.

Stevens D. L. Jr. and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99:262-278.

Vecchia, A. V. 1988. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society, Series B 50:297-312.

Xia, G., M. L. Miranda, and A. E. Gelfand. 2006. Approximately optimal spatial design approaches for environmental health data. Environmetrics 17:363-385.

Young, S.G., and R.R. Jensen. 2012. Statistical and visual analysis of human West Nile virus infection in the United States, 1999-2008. Applied Geography 34:425-431.

Zhu, Z. 2002. Optimal sampling design and parameter estimation of Gaussian random fields. Ph.D. Thesis. University of Chicago. Department of Statistics.

Zhu, Z. and M. Stein. 2005. Spatial sampling design for parameter estimation of the covariance

function. Journal of Statistical Planning and Inference 134:583-603.

CHAPTER 3

CONCLUSION

Disease ecologists are faced with the difficult task of understanding the proximate and ultimate factors influencing disease transmission and outbreaks. This problem becomes exacerbated as the number of recognized infectious diseases increases and the potential for perilous outcomes becomes more probable (Daszak et al. 2000). It is necessary that scientists have a suite of sophisticated and effective tools to aid in disease prevention and control. It is often the case that substantial data exist for many diseases resulting in very large data sets which increase the computational burden even when fitting the simplest models. Scientific interest, however, often requires exploring more complicated models that recognize the inherent dependencies and hierarchical structure in the data. In this thesis I have evaluated a tool that permits accurate predictions of disease outbreaks while evaluating the influence of important covariates on the occurrence of a disease. I have used a case study of West Nile Virus in New York City to explore the use of hierarchical multi-stage models to address scientifically valid questions about disease ecology while overcoming complications of model fitting with large spatial data sets.

In chapter 2 I developed a multi-stage hierarchical spatial model that explored the influence of several covariates on the prevalence of West Nile Virus in New York City while evaluating a recently developed approach to handle large spatial data sets. I found that four covariates (land cover type, distance to wetlands, slope and housing density) had a significant

influence on the distribution of West Nile Virus. I found that the effective range of spatial correlation among sites is approximately 560 meters although there is substantial uncertainty around this parameter and the 95% credible interval ranges from 344 meters to 1500 meters. This still suggests that there is significant small scale spatial correlation and sites that are separated by a distance greater than 560 meters can be considered independent if one site becomes infected. Staten Island had the highest levels of predicted probability of occurrence of West Nile Virus with select other areas having predicted probabilities exceeding 0.6. The uncertainty in these estimates was also lowest in Staten Island suggesting that control efforts should be focused in these areas and that Staten Island can be potentially regarded as a West Nile Virus hotspot. I found that with a relatively low number of knots (<30% of total number of observations) and using one of three knot allocation approaches (random, proportional, or optimal) the spatial process can be estimated accurately. The systematic grid allocation approach performed poorly under almost all scenarios and is not recommended.

Chapter 2 highlighted the advantages of using a hierarchical modeling framework to incorporate spatial dependence and propagate uncertainty into the estimates of interest. In our case study data were collected over a large window of time (May – Sept.) so that the actual observations of proximate locations may really occur weeks apart. A traditional first-stage model that only focuses on the spatial relatedness of the observations would have falsely assumed that the observations were collected at a single instance resulting in an upwardly biased estimate of spatial correlation. By including spatial dependence at the second-stage of the model (the mean response included spatial dependence) we overcame this difficulty and therefore accounted for the underlying spatial dependence in the data.

I found several key limitations in this study that influenced the model's ability to accurately predict West Nile Virus occurrence. The most relevant limitation was that the NYCDOHMH actively implements a program to eliminate potential mosquito breeding sites and adult mosquitoes (NYCDOHMH 2009). These actions inhibit the estimation of the spatial signal because information is lost when one site tests positive and surrounding locations are subsequently sprayed with larvicide. Of course these measures are appropriate and necessary to prevent outbreaks, but it requires the collection of more data to fully capture the spatial process. Future research should be focused on determining the effectiveness of the control measures.

A second major avenue of future research should be the incorporation of time into the spatial model. The statistical analysis of spatial data recognizes the fact that the configuration of observations carries important information about the relationship of the data points. By this same token the argument can be made that the temporal relationship of the data is equally important to acknowledge and provides useful information regarding the underlying processes of interest. It is therefore not a surprise that recent advances have been made to accommodate both spatial and temporal variation in data acquisition and modeling. This is also supported by the increase in data that are indexed by both time and space. Examples include environmental monitoring such as wind monitoring (Cressie and Huang 1999) and climate data (Fasso and Cameletti 2009, Wikle et al. 1998) where data are collected over a spatial domain and customarily monitored over long periods of time. Other specific examples include changes in real estate markets (Gelfand et al. 2003) over space and time or the spatial relationship of brain activity monitored via multiple brain scans (Bowman 2007) thus producing a dynamic spatio-temporal process. The West Nile Virus case study contains data for nine years (1999-2008) and this information should be included in the analysis. It is therefore necessary to develop some of

the ideas regarding space-time modeling and how they are relevant for West Nile Virus prevention in New York City.

The introduction of time into spatial modeling brings a substantial increase in the scope of work forcing the analyst to make several critical decisions regarding spatial correlation, temporal correlation, and how space and time interact in the data. First, one must make a distinction between the types of data collected, as in any spatial analysis. This requires determining if the spatial process and spatial data are to be viewed as geostatistical data (point level), lattice data (areal level) or point process data. A parallel distinction must be made for temporal data to determine if the temporal scale is to be viewed as continuous or discrete. If it is viewed as being discrete the analyst must further determine if the measurements are to be viewed as a block average over the time interval or whether they should be viewed merely as a count attached to an associated time interval. Furthermore if time is discretized it is important to distinguish between observing a time series of spatial data (e.g., the same points are being sampled in each time unit) and a cross-section of the data where the locations change with each time step. Additional complications involve missing data and the prediction of new locations both spatially and temporally as some time points will be missing spatial information and spatial points will be missing temporal information. Worse is the case when predictions are being made without either spatial or temporal information. For the West Nile Virus case study, data can be thought of as geostatistical where the temporal component represents a cross-section of the data (i.e., the locations change with each time step) although some of the same locations are used throughout the duration of the study.

Current approaches to modeling spatio-temporal data can be broken into three main categories: (1) conditional approaches, (2) methods for random fields in $\mathbf{R}^{d+1}$, and (3) joint

analysis of spatio-temporal data. The third option contains the majority of the current research including the development of separable and non-separable covariance functions along with the use of hierarchical modeling. I will briefly outline the first two approaches and the associated difficulties and then spend the majority of the time discussing the third and most useful approach.

Conditional approaches specifically isolate a particular time point (or location) and apply standard techniques for the type of data that results. Essentially it is a separate analysis for time at different spatial locations or separate analysis of spatial data at different time points. These sorts of analyses are obviously conditional on the approach taken and have been used mainly because of their simplicity. Often a two stage approach is taken wherein the second stage is to combine the results from the conditional analyses in the first stage. This is a somewhat common practice in statistics where multiple sources of variation are at work but methodology and/or software are unavailable for joint modeling. An example is nonlinear mixed model applications for clustered data (Schabenberger and Gotway 2005). Two important sources of variation are the changes in response as a function of covariates for each cluster and cluster-to-cluster heterogeneity. The first source is captured by a nonlinear regression model and the second source is expressed by using random effects among clusters. The two-stage approach here is to model each cluster separately and then to combine the regression coefficients into a set of overall coefficients in the second stage (Davidian and Giltinan 1995).

Although conditional approaches are appealing because of their simplicity several serious shortcomings arise that further limit their application. Data that are sparse in time or space can present difficulties. If it is not possible to analyze the data spatially for a time point then data collected at that time will not contribute in the second stage when the spatial analyses are

combined. The integration of observations over time or space (as a solution to this problem) can confound the spatial or temporal effects making it difficult to understand the process at hand. Alternatively too many spatial or temporal locations may result in an unwieldy number of analyses. If the locations are unique (cross-sectional data) then there is no suitable way to conduct a conditional analysis. An additional difficulty is that there may be multiple ways to summarize the results from the first-stage analysis and often an accurate summary would require the temporal or spatial correlation between the statistics of interest. Finally there is no approach to model space-time interactions and allow for predictions in time and space. Separate analyses in time (space) allow predictions in time (space) only. As we will see this is also a problem when using separable covariance functions in joint space-time analyses.

To treat spatio-temporal data as spatial data with an extra dimension ($\mathbf{R}^{d+1}$), the second approach, is not encouraged specifically because time and space are not directly comparable. Spatial coordinates are not directly comparable to temporal units because space has no past, present, or future. It is possible to circumvent this problem by using valid separable covariance structures as Gneiting (2002) points out. Strictly from a mathematical point of view the space-time domain $\mathbf{R}^d$ x $\mathbf{R}$ is no different than the spatial domain $\mathbf{R}^{d+1}$. This is not the same, however, as modeling spatio-temporal data as "3-D"data. A spatio-temporal process with a spatial component in $\mathbf{R}^2$ may be separable, but it is not a process in $\mathbf{R}^3$. Gneiting (2002) further argues a valid spatio-temporal covariance function is also a valid spatial covariance function although the difference between space and time must be acknowledged. Specifically there are two distances between points in a space-time process: the Euclidean distance between points in space and the Euclidean distance between points in time. If time is considered an "added" dimension, then the anisotropy must be taken into account. An example of a valid separable covariance function is if

the time points are evenly spaced, then the temporal process has an AR(1) structure and the spatial component has an exponential structure. Mitchell and Gumpertz (2003) used this to model CO_2 released over time on rice plots.

The joint modeling of spatial and temporal data has seen a recent influx of methods over the last decade. All of the approaches can be based on the basic necessities needed to describe a spatio-temporal random field; three simple expressions are needed: (1) a simple stochastic representation, (2) a closed-form spectral density function and (3) a closed form of the covariance function. Based on one of these three forms, statistical inference and prediction of the model can be made. Unfortunately only one or two simple forms may be available for a spatio-temporal stationary random field. Often the case is that for a particular situation only one of these expressions can be formulated. Specific methods are developed loosely based on one or more of these simple components to characterize the spatio-temporal random field. A further division is that of hierarchical modeling which is used characterize a spatio-temporal random field and often bypasses limitations or constraints of other approaches. This approach appears to hold the most promise for the West Nile Virus case study. The division between frequentist and Bayesian inference is often blurred among these approaches although the majority of hierarchical modeling approaches utilize Bayesian inference while the other approaches often resort to frequentist estimation techniques (e.g., Restricted Maximum Likelihood Estimation). I will further make the distinction between estimation techniques as I discuss the specific models. It should be reiterated that none of these approaches based on the components of a spatio-temporal random field are *necessarily* mutually exclusive, for example several hierarchical models are used to describe covariance functions as well as in the representation of spectral density

functions, but the separation of these techniques in this fashion helps to illuminate the approaches and their subtle differences.

Space-time Covariance Functions

Let $\{Z(s;t): s \in D \subset \mathbb{R}^d; t \in [0,\infty)\}$ denote a spatio-temporal random process observed at *N* space-time coordinates ($\mathbf{s}_1;t_1$),…($\mathbf{s}_N;t_N$). Optimal prediction (in space and time) of the unobserved parts of the process, based on the observations is often the ultimate goal, but to achieve this goal, a model is needed for how various parts of the process co-vary in space and time. If we assume that variance is finite (regularity condition) then we can define the mean function as $\mu(\mathbf{s};t) = E(Z(\mathbf{s};t))$ and the covariance function as Cov(Z($\mathbf{s}$;*t*),Z($\mathbf{r}$;*q*)); $s, r \in D$, t>0, q>0. It is then necessary (with the exception of some of the hierarchical models) to assume second-order stationary in space and time such that

$$Cov(Z(s;t), Z(r;q)) = C(s-r;t-q)$$

for certain functions C. This assumption is often made so that the covariance function can be estimated from data. The function C must satisfy a positive-definiteness condition to be a valid covariance function; that is for any real $a_1,\ldots,a_m$ and any positive *m*, C must satisfy

$$\sum_{i-1}^{m}\sum_{j=1}^{m} a_i a_j C(r_i - r_j; q_i - q_j) \geq 0 .$$

Then and only then is the covariance function valid to be used in spatio-temporal prediction. We will further assume that C is continuous, therefore positive-definiteness is equivalent to the process having a spectral distribution function (Cressie 1993). To ensure positive-definiteness it

is common to specify a parametric family of distributions for C whose members are known to be positive definite.

One common class of covariance functions is separable covariance functions where a valid spatial covariance and a valid temporal covariance function are combined in product form. A separable spatio-temporal covariance function decomposes the covariance into purely spatial and purely temporal components. The components usually have different parameters to allow for space-time anisotropy. The final spatio-temporal covariance function is then the result of addition and multiplication operators. Separable covariance functions are easy to work with and valid, provided the components are valid covariance functions. However the class of separable covariance functions is severely limited because they do not model space-time interaction.

The other class of covariance functions is that of non-separable covariance functions. Typically non-separable covariance functions are more complicated than separable models but incorporate space-time interactions. Some non-separable models reduce to separable ones for particular values of model parameters and allow a test of separability (also see Li et al. 2007 and Crujeiras et al. 2010 for two different nonparametric tests of separability). Methods for constructing valid covariance functions with spatio-temporal interactions include the monotone function approach of Gneiting (2002), the spectral method (Cressie and Huang 1999; Fuentes 2002), the mixture approach (Ma 2002), partial differential equation approach (Jones and Zhang 1997), and several other approaches (Ma 2008). I will give a brief overview of each of these approaches including the method of estimation.

Monotone function approach (Gneiting 2002)

This method is powerful because it builds valid covariance functions from elementary components whose validity is easily checked. It is also useful because it avoids operations in the spectral domain and does not rely on Fourier inversions (such as Cressie and Huang 1999). The general idea is to choose two functions such that one is completely monotone and the other is positive with a completely monotone derivative (Tables 1 and 2 in Gneiting 2002 provide a list of such functions). A valid non-separable covariance function can then be built and a test of separability can also be conducted. Estimation of the parameters in the model can be conducted via restricted maximum likelihood estimation.

Spectral density approach (Cressie and Huang 1999; Fuentes 2002)

Earlier we mentioned that valid covariance functions (continuous and positive-definite) have a spectral representation. This is confirmed by Bochner's theorem (see Cressie 1993) and suggests the following method for constructing valid covariance functions: determine a valid spectral density and take its inverse Fourier transform. Unfortunately, this approach can be limited if there is no explicit expression for the covariance function. Cressie and Huang (1999) apply the following method: determine a valid spatio-temporal spectral density and integrate. But they note that because the covariance function and spectral density are a Fourier transform pair, integration of the spatial and temporal components can be separated in the frequency domains and thus the spatio-temporal spectral density is obtained with a one-dimensional Fourier transform although complicated integration is still needed from there. Estimation can be conducted by likelihood methods although the complicated integration introduces difficulties. Fuentes (2002) uses the convolution of local stationary processes to propose various parametric

approaches for estimating the spectral density of a nonstationary spatial process and applies this to air pollution data. Several nonparametric approaches are suggested including a new algorithm for fitting such models.

Mixture approaches (Ma 2002)

Instead of integration in the frequency domain, non-separable covariance functions can also be constructed by summation or integration in the spatio-temporal domain. Space-time interactions can be incorporated by "mixing" the product of the covariance functions or correlation functions. Ma (2002) goes on to further describe several approaches (positive power mixture and scale mixture) using bivariate random vectors with a joint distribution function independent of both the spatial and temporal covariance functions.

Stochastic partial differential equations (Jones and Zhang 1997)

A dynamic or stochastic model may be formally described by a stochastic (partial) differential equation. Jones and Zhang (1997) consider a separable spatio-temporal random field as

$$\left(\frac{\partial^2}{\partial {s_1}^2}+\frac{\partial^2}{\partial {s_2}^2}-\phi^2\right)\left(\frac{\partial}{\partial t}+\alpha\right)Z(s;t)=\varepsilon(s;t), s\in\mathbb{R}^2, t\in\mathbb{R},$$

where $\varepsilon(s;t)$ is a zero mean Gaussian white noise field with variance σ^2. This results in a theoretically valid correlation function, but is limited by the difficulty in imagining a physical mechanism which would lead to such a relation. In addition the correlation function does not have the partial derivative with respect to *t* or **s**, so that the partial derivatives in the above equation do not exist in the usual sense. Alternatively, Brown et al. (2000) use a "blur" approach

where the model operates successively in time; the spatial field at time $t + 1$ is obtained by "blurring" the field at time t and adding a spatial random field.

Although both of these approaches can be useful the complexity in the analysis specifically in the setup of the stochastic partial differential equation to represent a physical (or biological) process limits their utility.

Hierarchical Models

Although I have outlined several approaches to calculate valid covariance functions for a spatio-temporal random field these approaches are limited by the need to completely specify the joint space-time covariance structure. In addition such covariance functions are often not realistic for complicated dynamical processes and dimensionality can prohibit practical implementation. Alternatives to this problem exist and rely heavily on hierarchical modeling. Hierarchical modeling is a broad class of models in which the components of the model are defined by submodels at different hierarchical stages. The advantage is mainly in flexibility and allowing for allocation of the total uncertainty to the various components or levels.

A natural approach to spatio-temporal hierarchical modeling for complex dynamical processes is a combination of spatial and time series techniques, which is accomplished by a spatio-temporal dynamic (hierarchical) model formulation. However, estimation in this context can still be problematic due to the high dimensionality of the state process. Several modeling strategies have been proposed to address this problem. One approach is to reduce the dimensionality by projecting the state-process on some set of spectral basis functions (Wikle and Cressie 1999) or by projecting onto a lower subspace (Banerjee et al. 2008; Sahu and Bakar in press). Alternatively, one might specify very simple, random walk dynamics or incorporate

dynamical biological models directly into the parameterization (Wikle et al. 2001). Estimation in hierarchical models uses either the Expectation-Maximization algorithm (Bolin et al. 2009; Fasso and Cameletti 2009; Xu and Wikle 2007) or a Bayesian approach that allows for the complicated structure to be modeled in terms of means at various stages, rather than a model for a massive joint covariance matrix (Cressie and Wikle 2002; Huang and Cressie 1996; Wikle et al. 1998; Wikle et al. 2001). Furthermore, hierarchical models offer the analyst various opportunities to explore and employ trade-offs between large and small scale dynamics, and space-time dynamics.

Numerous examples exist in the literature that rely on hierarchical modeling of spatio-temporal processes including spatially varying coefficients (time series of spatial processes; Banerjee et al. 2004; Gelfand et al. 2003), estimating vegetation trends using the EM algorithm (Bolin et al. 2009), estimation of snow melt using Kalman filtering (Cressie and Wikle 2002; Huang and Cressie 1996), modeling the dependence in fMRI data by using a two stage autoregressive approach (Derado et al. 2009), agent based models (Hooten and Wikle 2010), and a regime switching model for wind energy (Gneiting et al. 2006). Although all of these approaches vary slightly in their model formulation and each have attributes specific to their problem at hand the basic hierarchical formulation is as follows with levels being added or subtracted as needed:

i. Data: specifies a measurement error model for the observation data
ii. Process: specifies models for the response and the incorporation of space-time dynamics
iii. Large and small scale and space-time dynamics: provides parameters for the process models that generate the spatial dependence structure and includes dynamical terms as well

iv. Model parameters (measurement variances, model variances, dynamical parameters): setting priors for the parameters

v. Hyperparameters: any additional parameters for prior distributions

Moving forward with the West Nile Virus case study I envision that a hierarchical space-time model using Bayesian inference would be most appropriate. This framework would allow for a flexible approach to model the spatio-temporal dynamics of West Nile Virus while accounting for the various sources of uncertainty. It will be necessary and relevant to explore approaches that can efficiently handle large amounts of data. I believe the research contained within this thesis provides a stepping stone for more sophisticated analyses while also providing a useful tool to help disease ecologists understand and control costly disease outbreaks.

Literature Cited

Banerjee, S, B. Carlin, and A. Gelfand. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall, Boca Raton, Fl.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society, Series B 70:825-848.

Bolin, D., J. Lindstrom, L. Eklundh, and F. Lindgren. 2009. Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields. Computational Statistics and Data Analysis 53: 2885-2896.

Bowman, F. 2007. Spatiotemporal models for region of interest analyses of functional neuroimaging data. Journal of the American Statistical Association 102: 442-453.

Brown, P. K. Karesen, G. Roberts, S. Tonellato. 2000. Blur-generated non-separable space-time models. Journal of the Royal Statistical Society B 62: 847-860.

Cressie, N. 1993. Statistics for Spatial Data. Wiley Interscience, New York, NY.

Cressie, N. and H. Huang. 1999. Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of the American Statistical Association 94: 1330-1340.

Cressie, N. and C. Wikle. 2002. Space-time Kalman filter. Encyclopedia of Environmetrics 4: 2045-2049.

Crujeiras, R., R. Fernandez-Casal, W. Gonzalez-Manteiga. 2010. Nonparametric test for separability of spatio-temporal processes. Environmetrics 21: 382-399.

Daszak, P., A.A. Cunningham, and A.D. Hyatt. 2000. Emerging infectious diseases of wildlife - Threats to biodiversity and human health. Science 287: 443 – 449.

Davidian, M. and D.M. Giltinan. 1995. Nonlinear Models for Repeated Measurement Data. Chapman and Hall. New York, NY.

Derado, G., F. Bowman, and C. Kilts. 2009. Modeling the spatial and temporal dependence in fMRI data. Biometrics 49: 1-9.

Fasso, A. and M. Cameletti. 2009. The EM algorithm in a distributed computing environment for modeling environmental space-time data. Environmental Modelling and Software 24: 1027-1035.

Fuentes, M. 2002. Spectral methods for nonstationary spatial processes. Biometrika 89: 197 210.

Gelfand, A., H. Kim, C. Sirmans, and S. Banerjee. 2003. Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association 98: 387 396.

Gneiting, T. 2002. Nonseparable, stationary covariance functions for space-time data. Journal

of the American Statistical Association 97: 590-600.

Gneiting, T., K. Larson, K. Westrick, M. Genton, and E. Aldrich. 2006. Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method. Journal of the American Statistical Association 101: 968-979.

Hooten, M. and C. Wikle 2010. Statistical agent-based models for discrete spatio-temporal systems. Journal of the American Statistical Association 105: 236-248.

Huang, H. and N. Cressie. 1996. Spatio-temporal prediction of snow water equivalent using the Kalman filter. Computational Statistics and Data Analysis 22: 159-175.

Jones, R. and Y. Zhang. 1997. Models for continuous stationary space-time processes. In: T.G. Gregoire, D.R. Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren, and R.D. Wolfinger, (eds) Modeling Longitudinal and Spatially Correlated Data, Lecture Notes in Statistics 122: 289-298. Springer, Heidelberg.

Li, B. M. Genton, and M. Sherman. 2007. A nonparametric assessment of properties of space time covariance functions. Journal of the American Statistical Association 102: 736-744.

Ma, C. 2002. Spatio-temporal covariance functions generated by mixtures. Mathematical Geology 34: 965-975.

Ma, C. 2008. Recent developments on the construction of spatio-temporal covariance models.

Stochastic Environmental Research and Risk Assessment 22: S39-S47.

Mitchell, M. and M. Gumpertz. 2003. Spatio-temporal prediction inside a free-air CO_2 enrichment system. Journal of Agricultural, Biological, and Environmental Statistics 8: 310-327.

Schabenberger, O. and C. Gotway. 2005. Statistical Method for Spatial Data Analysis. Chapman and Hall, Boca Raton, Fl.

Wikle, C. and N. Cressie. 1999. A dimension-reduced approach to space-time Kalman filtering. Biometrika 86: 815-829.

Wikle, C., L. Berliner, N. Cressie. 1998. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 5: 117-154.

Wikle, C., R. Milliff, D. Nychka, L. Berliner. 2001. Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. Journal of the American Statistical Association 96: 382-397.

Xu, K. and C. Wikle. 2007. Estimation of parameterized spatio-temporal dynamic models. Journal of Statistical Planning and Inference 137: 567-588.