CHAPTER 1

WEST NILE VIRUS MOSQUITO ABUNDANCE MODELING USING A NON-STATIONARY SPATIO-TEMPORAL GEOSTATISTICS

EUN-HYE YOO[1], DONGMEI CHEN[2], CURTIS RUSSEL[3]

[1]Department of Geography, University at Buffalo, SUNY, Buffalo, NY, USA
[2]Department of Geography, QueenÕs University, Kingston, Ontario, Canada
[3]Enteric, Zoonotic and Vector-Borne Diseases, Public Health Ontario, Canada

Abstract

The lack of spatial coverage and missing observations of adult mosquito surveillance data challenges the quantitative assessment of human exposure to West Nile virus (WNv). We developed a geostatistical spatio-temporal prediction model for missing WNv mosquito data. In the proposed Poisson generalized linear mixed model, the effects of meteorological and physiographic conditions on mosquito abundance are modeled as a drift, and the spatio-temporal variations around the drift, possibly correlated, are captured by a spatio-temporal residual random field. The proposed model accounts for discrete counts of the mosquito surveillance data within a generalized linear mixed model, and tackles the non-stationarity in WNv mosquito abundance data by restricting the decision of stationarity to a local neighborhood surrounding the target prediction point.

key words: Geostatistical space-time model; West Nile Virus (WNv); Moving local neighborhoods; Poisson Generalized Linear Mixed Model (GLMM); Non-stationarity

Please enter \offprintinfo{(Title, Edition)}{(Author)}
at the beginning of your document.

1.1 Introduction

West Nile virus (WNv) has been recognized as a globally distributed disease since it's first outbreak in New York City in 1999, and it is the fast growing mosquito-borne health threat in the US with 3,545 known cases and 147 deaths (Centers for Disease Control and Prevention, as of September, 2012). Both scientists and vector control practitioners have been exploiting various ways to assess the spatio-temporal human risk of transmission and respond adequately to reduce potential health threats (Theophilides et al. 2003, Johnson et al. 2006, Bolling et al. 2009). Some studies have used entomological risk of vector exposure as a key determinant of WNv disease risk in humans, while others focused on disease risk based on avian and equine surveillance or mandatory human case reports (Griffith 2005, Ward et al. 2006, Beroll et al. 2007, Carney et al. 2011).

Entomological risk measures can be used for a direct assessment and prediction of human WNv infection risk (Kilpatrick et al. 2006), while their effectiveness depends on the quality of mosquito surveillance data. Mosquito data collected at a set of trap sites have been used to identify site-specific meteorologic conditions and local environmental factors that account for the variation of mosquito abundance (Soverow et al. 2009). However, mosquito surveillance data are available only from a sparse monitoring network due to the labor intensive and costly data collection procedures, and technical challenges in equipment maintenance often result in unintentional missing observations. Building a spatio-temporal entomological risk model using sparse monitored data is challenging, but the presence of missing observations imposes further difficulties to build a spatio-temporal risk assessment. That is, mosquito data are missing at a specific site for a time series, but also missing observations constitute a subset of a total number of sites for a specific trap night.

In the current paper, we aim to develop a geostatistical spatio-temporal model to predict missing values of the mosquito surveillance data. The proposed prediction model (i) explicitly takes into account the discrete nature of observed mosquito counts within a framework of generalized linear models (McCullagh & Nelder 1989); (ii) incorporates a spatio-temporal correlated error structure by extending a generalized linear model for poisson data to a generalized linear mixed effect model (GLMM); and (iii) accommodates the non-stationarity in the mean of latent mosquito abundance process using a moving local neighborhood approach.

Despite the lack of surveillance mosquito data, some studies (Diuk-Wasser et al. 2006, Reisen et al. 2006, Liu & Weng 2009, Soverow et al. 2009, Morin & Comrie 2010, Chuang et al. 2011, 2012) have successfully shown the effects of weather and environmental conditions on the mosquito abundance. Both Shaman & Day (2007) and Ruiz et al. (2010) pointed out that increased temperature has a direct effect on the spread of WNv mosquito infection, and Diuk-Wasser et al. (2006) have identified key environmental predictors of mosquito abundance using remote sensing data and Geographical Information Systems (GIS). Most studies mentioned above, however, used a log transformed mosquito counts as a response variable to achieve a linear relationship between environmental/climate conditions and mean mosquito counts. Mosquito surveillance data, on the other hand, contain excessive number of zeros

(30% in the current study) and the correlated error structure, which, in consequence, leads to a poor performance of a linear regression with log transformed count data (O'Hara & Kotze 2010). We will investigate the associations of environmental and weather conditions with the underlying (latent) process that is assumed to generate the observed mosquito counts using a poisson regression model.

We will further extend the generalized linear model for poisson data to a generalized linear mixed effect model (GLMM) to introduce a spatio-temporal correlation. We hypothesize that the latent process has a spatio-temporal correlated error structure because of the mosquitoes' biological behavior, but also from the lack surveillance data and missing covariates (Zuur et al. 2009). The causal effects of poor match between the spatial extent of the phenomenon of interest and the units for which data are available on the spatial error autocorrelation have been frequently documented in the literature (Arbia 1989, Anselin & Rey 1991, Goodchild 2001). The proposed spatio-temporal predictive model of WNv vector mosquito abundance shares similar problems because the spatial coverage/temporal intervals for which surveillance data are collected may not reflect the spatial extent and temporal duration of the true mosquito abundance patterns. The covariates used in regression models, particularly, landscape elements that are known to influence abundance of vector populations, are typically defined over certain areal units, such as buffer zones or administrative units whose delineation is rather subjective and arbitrary (Liu & Weng 2009).

The discrepancy between the spatial (or temporal) scale of the analysis and that of the process underlying the observed data may cause correlated error structure. Although often ignored, such discrepancy may conceal some associations between mosquito abundance and environmental conditions and yield non-stationary residuals. As an attempt to alleviate such non-stationarity problems while accommodating a joint spatio-temporal error structure within a poisson regression, we use a spatio-temporal moving neighborhood approach (Haas 1990). The implementation of the local spatio-temporal prediction model, however, requires the determination of spatio-temporal cylinder size, which could be different from prediction point to point. In the current study, we determine the optimal size of spatio-temporal cylinder size based on the sensitivity analysis, where the effect of the cylinder size on the model prediction accuracy is examined using the leave-one out cross-validation method. Based on the optimal size, i.e., the minimum number of spatio-temporal data points adjacent the prediction point, we obtain the geostatistical spatio-temporal model prediction of missing values and the corresponding prediction error variances. For a model assessment, we use a cross-validation technique where the data set is split into training and the prediction data set.

1.2 Methods

1.2.1 Spatial-Temporal Process Modeling

Mosquito abundance data, i.e., adult mosquito counts captured at a specific trap site $s_n, n = 1, \ldots, N$ and a trap night $t_p, p = 1, \ldots, P$, are viewed as a realization of a spatio-temporal Poisson random variable (RV). We assume that the average intensity process of such a spatio-temporal poisson RV is influenced by local climate conditions, physiographic characteristics, and the timing when the trapping efforts were made. More specifically, we consider weekly average temperature and precipitation as climate predictors and included two land-use variables, i.e., the proportions of residential area and water body within a buffer of radius 200 m centered at each trap site and the areal average of 30 × 30 Normalized Difference Vegetation Index (NDVI) within 1 km centered at each trap site as environmental covariates.

Modeling the effects of environmental factors on the dynamic changes of mosquito population is not always straightforward (Diuk-Wasser et al. 2006), because each species prefers a certain habitat and accordingly thrives in different landscapes with features essential to its life history. A further challenge exists in modeling the association of environmental condition with the behavior and life cycle of each mosquito species because most spatial covariates are measured on spatial units (supports or neighborhoods) whose size, shape, and orientation are rather arbitrarily determined. In the current study, our decision on the buffer distance used to determine the neighborhoods of each trap site and our selection of environmental covariates are based on our preliminary analyses and the literature review (Diuk-Wasser et al. 2006), as well as our understanding of species specific behavior, such as the short flight range of *Cx. pipiens-restuans*. The spatio-temporal poisson model for mosquito abundance is defined as

$$\begin{aligned} Y(s_n,t_p)|\lambda(s_n,t_p) &\sim Poisson(\lambda(s_n,t_p)) \qquad (1.1)\\ Z(s_n,t_p) &= \log\lambda(s_n,t_p) = \mu(s_n,t_p) + R(s_n,t_p)\\ &= \beta_0 + \sum_{i=1}^{7}\beta_i x_i(s_n,t_p) + R(s_n,t_p) \end{aligned}$$

where $Y(s_n,t_p)$ denotes the poisson RV whose realization is associated with observed mosquito counts at the n-th trap site on the p-th week. The underlying average intensity of mosquito abundance is denoted by $\lambda(s_n,t_p)$, where the log transformed intensity $Z(s_n,t_p)$ is spatially and temporally varying as a linear function $\mu(s_n,t_p)$ of predictors $x_i, i = 1,\dots,7$. Here, x_1, x_2 denote the year and the week of trap night, and x_3 is the population density, x_4, x_5, x_6 are the proportion of residential land-use (1 km buffer), waterbody (0.5 km buffer), and the average NDVI within 1 km centered at the n-th trap site, respectively. The average temperature and precipitation of the p-th week are denoted by x_7 and x_8. The spatio-temporal variation unexplained by the selected covariates is modeled by a stochastic residual component $R(s_n,t_p)$ whose a stationary covariance function $C_R(\mathbf{h},\tau)$ can be identified under the model decision of a space-time stationary mean $\mu(s_n,t_p)$ as

$$\begin{aligned} C_R(\mathbf{h},\tau) &= E\{R(s,t)\cdot R(s',t')\}\\ &= E\{[Z(s,t)-\mu(s,t)][Z(s+\mathbf{h},t+\tau)-\mu(s+\mathbf{h},t+\tau)]\}\\ &= C_Z(\mathbf{h},\tau) \qquad (1.2) \end{aligned}$$

where $\mathbf{h}$ denotes the difference between any two trap sites s and $s' = s + \mathbf{h}$ and τ denotes the time lag between any pair of weekly observations $t, t' = t + \tau$.

Recognizing the differences between space and time, we further assume that the spatio-temporal covariance can be decomposed into a purely spatial covariance $C_1(\mathbf{h})$ and a purely temporal covariance $C_2(\tau)$ as (Cressie 1993)

$$C_R(\mathbf{h}, \tau) = C_1(\mathbf{h}) \cdot C_2(\tau). \tag{1.3}$$

1.2.2 Moving-Cylinder Spatio-Temporal Kriging

In theory, the separable spatio-temporal covariance in Eq (1.3) should be derived from residual data $r(s_n, t_p) = Z(s_n, t_p) - \mu(s_n, t_p)$, which are not directly available in most practical applications. This is problematic in the inference of the residual covariance model, because the residual covariance estimate $C_{\hat{R}}(\mathbf{h}, \tau)$, which is inferred from the estimated residuals $\hat{r}(s_n, t_p) = Z(s_n, t_p) - \hat{\mu}(s_n, t_p)$, depends on the filtering algorithm, such as poisson regression, used to evaluate the trend estimate $\hat{\mu}(s_n, t_p)$ (Kyriakidis & Journel 1999). In addition, the behavior of *Cx. pipiens-restuans*, such as a short distance flying range and their preference for certain habitats, strongly indicates the presence of non-stationarity of mosquito abundance. To overcome such practical limitations, we restrict our decision of space-time stationarity to a local neighborhood around a prediction point. That is, the spatio-temporal local neighborhood for each target prediction point is specified by the spatial window (the radius of the spatio-temporal cylinder) and the time duration (the height of spatio-temporal cylinder), which vary from a prediction point to another depending on the neighboring data availability. Using the data found within the spatio-temporal neighborhood, the drift and the residual estimates of the average intensity process $\lambda(s_0, t_0)$ at each prediction point are estimated using a poisson regression and simple kriging in sequence. The prediction process is summarized as follows:

step 1. Construct a local spatio-temporal cylinder per prediction point. The goal is to search for a circular neighborhood with minimum window size surrounding the prediction point to achieve the local stationarity of the underlying process. On the other hand, the window size should be large enough to contain a number of data pairs to ensure the accuracy of variogram estimate. To balance between the local stationarity and the accuracy of variogram estimate, we adopt an automatic window sizing approach (Haas 1990). That is, the window radius of spatio-temporal cylinder is varying from 1 to 7 km and the height of the cylinder varies from -5 to 5 weeks (previous 5 weeks to the next 5 weeks) from the week of prediction point until at least 45 data points are included.

step 2. Once a local spatio-temporal neighborhood is determined at each prediction point, a poisson generalized regression is performed using the set of covariates mentioned above. The regression result is recorded, i.e., a spatio-temporal drift estimate at the prediction point and the deviance residual data within its local neighborhoods.

step 3. Model the spatial structure of the stochastic spatio-temporal process by pooling the residual data of a prediction point $(\mathbf{u}_0, t_0)$ obtained from step 2. The spatio-temporal sample variogram is calculated as

$$\hat{\gamma}(\mathbf{h}_k, \tau_l) = \frac{1}{2N_{kl}} \sum_{i=1}^{N_{kl}} [\hat{r}(\mathbf{u}, t) - \hat{r}(\mathbf{u}', t')]_i^2 \tag{1.4}$$

where $\mathbf{h}_k, \tau_l$ denote the spatial and temporal lag class indexed by $k = 1, \ldots, m_S$ and $l = 1, \ldots, m_T$, respectively. For a consistent modeling, the same number (m_S, m_T) of the spatio-temporal variography is calculated at each prediction point. The (k, l)-th variogram value $\hat{\gamma}(\mathbf{h}_k, \tau_l)$ is calculated using N_{kl} pairs of residual data points $\hat{r}(\mathbf{u}, t), \hat{r}(\mathbf{u}', t')$ whose separation vector belongs to the k-th spatial lag class $|\mathbf{u} - \mathbf{u}'| \in \mathbf{h}_k$ and the l-th temporal lag class $|t - t'| \in \tau_l$. The spatio-temporal covariogram in Eq (1.3) can be rewritten using variogram as

$$\gamma_R(\mathbf{h}, \tau; \boldsymbol{\theta}) = (a_1 + s_1)\gamma_2(\tau) + (a_2 + s_2)\gamma_1(\mathbf{h}) - \gamma_1(\mathbf{h})\gamma_2(\tau) \tag{1.5}$$

where (a_1, s_1) denote the nugget and the partial sill of spatial variogram $\gamma_1(\mathbf{h})$ with the spatial range r_1, and (a_2, s_2) denote the nugget and the partial sill of temporal variogram $\gamma_2(\tau)$ with the temporal range τ_2. Using the empirical variogram $\hat{\gamma}(\mathbf{h}_k, \tau_l)$ obtained at each prediction point, we fit a spatio-temporal variogram model using a weighted least square approach. That is, the vector of spatio-temporal variogram parameters $\boldsymbol{\theta} = [a_1, s_1, r_1, a_2, s_2, r_2]$ are estimated using the number of observations within the space-time lag class as a weight.

step 4. Obtain simple kriging (SK) prediction and prediction error variance of residual at a prediction point using the deviance residual data and the fitted spatio-temporal variogram model as:

$$\hat{r}(s_0, t_0) = \mathbf{w}_0^T \mathbf{r}, \qquad \hat{\sigma}^2(s_0, t_0) = \sigma_0^2 - \mathbf{w}_0^T \mathbf{c}_0 \tag{1.6}$$

where the kriging weights $\mathbf{w}_0$ are obtained by solving the following spatio-temporal kriging system

$$\begin{bmatrix} \mathbf{C}_R \end{bmatrix} \begin{bmatrix} \mathbf{w}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{c}_0 \end{bmatrix} \tag{1.7}$$

where $\mathbf{C}_R$ is a data-to-data spatio-temporal covariance matrix within the moving cylinder and $\mathbf{c}_0$ is the data to the target covariance vector. The simple kriging weights $\mathbf{w}_0$ are obtained by solving the kriging system in Eq (1.7).

step 5. Combine the spatio-temporal drift estimate obtained at the step 2 and the stochastic residual value estimated from the step 4 to obtain the prediction of the log transformed spatio-temporal process $\hat{\lambda}(s_0, t_0)$.

1.3 Data Analysis and Results

1.3.1 Mosquito Surveillance Data

The Greater Toronto Area (GTA) is the largest urban agglomeration in Canada with diverse land use, including urban, suburban, rural, and agricultural areas. The study area consists of 4 health units (Hamilton, Peel, City of Toronto, and York) and 949 census tracts with a population size of 3,328,590 (2006 Census).

We focus on one of the most important vectors of WNv in the northeastern U.S. and Canada, *Culex Pipiens (Cx.pipiens)*, which has a short flight range with a maximum of 2 km and usually stays within 200 m of the area of larval emergence. They prefer human settlements and stagnant water for larval habitats (Turell et al. 2005, Kilpatrick et al. 2006). *Culex restuans* is another competent vector of WNv that is almost indistinguishable from *Cx.pipiens* adults (Degaetano 2005, Diuk-Wasser et al. 2006). We grouped these two species together into the *Culex pipiens-restuans* complex (Kilpatrick et al. 2006, Bolling et al. 2009). A total of 4,040 mosquito observations are collected over two years (2007-2008), which consist of the weekly surveillance records from fixed sites and a small number of observations from temporary sites. Typically the mosquito trapping season lasts about 18 weeks per year; roughly from early June through mid-October.

We consider any trap site with more than 15 weeks of surveillance records a permanent trap site, and a temporary site, otherwise. A total of 141 sites, corresponding to 73% of the entire trap sites in the study area, are permanent with varying number of missing observations. It is important to distinguish permanent trap sites from temporary ones, because our prediction efforts will be made only on permanent sites. The spatial and temporal patterns of trapping frequency allow us to better understand missing data, which are clustered in their spatial and temporal configuration. Particularly, the spatial variation of the trapping frequency informs us how many values need to be predicted per site and where they are located.

The surveillance data availability (or trapping frequency) is spatially and temporally varying. The surveillance data are collected for a total of 36 weeks over two year long surveillance period across 194 trap sites. Missing data are commonly encountered at the beginning (week 23-25) and at the end of the season (week 41), and temporary trap sites are spatially clustered in the Hamilton region and most trap sites with one year record only are placed in the Peel region (see Figure 1.1). Most trap sites located in the region of Peel has incomplete one-year record (a total of 18 weeks), i.e., a total of 63 trap sites have 2 to 3 missing values. Permanent trap sites located in other regions have the two year long record with a different number of missing values. For example, the City of Toronto have 43 sites with 5 to 6 missing values and 32 trap sites in York region have on average 3 to 4 missing values. Only three trap sites in the Hamilton region have one missing value. In summary, a total of 87 sites (45%) have two year long records (more than 30 weeks of observations) and 66 sites (34%) have a full year long record. residuals.

Yearly variation in the observed mosquito population is summarized in Table 1.1. Despite the similar trapping efforts made in 2007 (a total of 2,007 records) and 2008

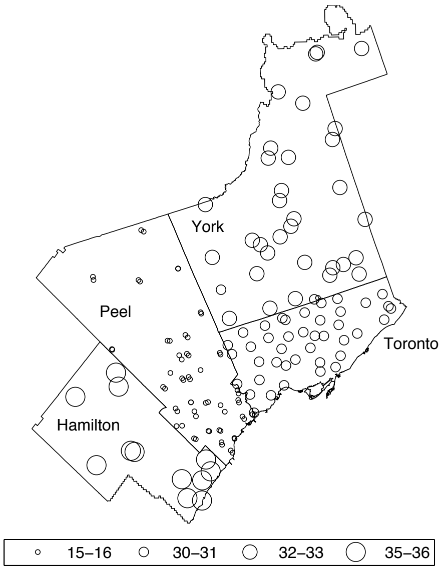

Figure 1.1 Regional variation of trapping frequency.

year	# of obs.	min	mean	max	sum	st.dev
2007	2,007	0	6.83	139	13,713	14.15
2008	2,033	0	11.27	181	22,921	21.97

Table 1.1 Summary of *Cx. pipiens-restuans* data

(a total of 2,033 records), the total of 22,921 mosquitoes captured in 2008 is almost doubled the mosquito counts (13,713) of 2007. The difference is also found in the highest record of mosquito population: the maximum mosquito count captured in 2007 was 139 and 181 in 2008, which amounts to 30% increases.

The weekly variation of mosquito counts is summarized by the boxplot in Figure 1.2. Each box contains the mosquito counts collected from all trap sites where trapping efforts were made on a week over two years. The weekly summary of mosquito counts may not reveal the true weekly variations as shown in the boxplot due to large variation. For example, some trap sites may capture over 100 mosquitoes during the peak of summer, while other trap sites may catch one or two mosquitoes due to disturbance made around trap sites at the same week. On the other hand, the variation of data, i.e., the range between the upper and lower quartile (the height of each box) and outliers (denoted by symbol (+) beyond the whisker), are substantially different from week to week. For instance, the maximum number of mosquito counts per week and the variance of mosquito counts show a parabolic behavior of mosquito abundance as a function of the time of a trap night. That is, the total number of mosquitoes captured at the beginning and the end of the trapping season is small, but they soon increase as summer begins for the following seven to nine weeks.

The variation of mosquito counts per week, in fact, represents their spatial variability across the study area. Most weekly observations except the last two weeks at the end of the trap season have many outliers, i.e., the unusually high mosquito counts, particularly, during week 27 - 35. Unless the environmental conditions, such as residential area, waterbody, and NDVI, change weekly, it is less likely that the traditional poisson regression can capture the extreme spatial heterogeneity present in data (the difference between minimum and maximum observation per week). In addition, the weekly variation of mosquito abundance in space is greater than what predictors, such as weather conditions and the time of trap night, can explain. One may argue that the differences in the trapping efforts made in the beginning and the end of the trap season versus the mid-summer are responsible for such differences in spatial heterogeneity, but it is clear that there is a need for a flexible and general modeling approach to accommodate such high spatial and temporal variabilities in data. In the next subsection, we will illustrate the application of the global poisson regression to the mosquito data and demonstrate the limitations due to the extreme spatial and temporal heterogeneity of the mosquito data.

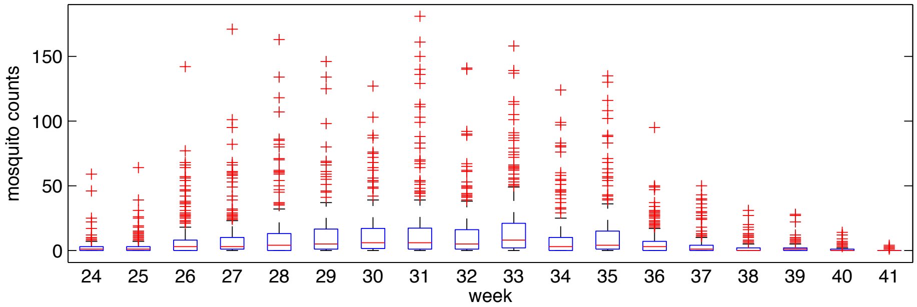

Figure 1.2 Weekly variation of mosquito counts

1.3.2 Global space-time poisson regression

A global poisson generalized linear model in Eq (1.1) is applied to the mosquito surveillance data, and the resulting spatio-temporal drift model parameter estimates are summarized in Table 1.2. Except the intercept, all other predictors are statistically significant at the significance level 0.05. The increase of mosquito population in 2008 compared to the year 2007 is substantial as shown in the model coefficient estimate $\hat{\beta}_1$. The effects of the population density $\hat{\beta}_3$ is minimal, however, due to the coarse resolution of census data used. Among environmental covariates, the coefficient estimates $\hat{\beta}_4, \hat{\beta}_5$ for the residential and water body proportion within the buffer zone and the estimate $\hat{\beta}_6$ of the spatial average of NDVI index are worth noting. The higher value of NDVI (approaching to 1) implies more green vegetation present in the area where the trap site is placed. As the unit of greenness index (NDVI) increases, less ground for *Cx. pipiens-restuans* habitats such as human settlements remains. The negative relationship is also shown in the negative coefficient of NDVI index, i.e., $\hat{\beta}_6 = -1.658$. While the effect of residential area $\hat{\beta}_4 = 0.004$ is significant but not as strong as that of NDVI, probably because most trap sites are placed near residential areas and the difference in the proportions of residential area is trivial. Last, it is clearly shown that the average temperature is a major driver of *Cx. pipiens-restuans* abudance in line with the previous studies (Diuk-Wasser et al. 2006, Brown et al. 2008), where the habitat preference of *Cx. pipiens-restuans* for urban and highly populated area in relation to forest and green land is empirically demonstrated.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
estimate	0.523	-0.011	0.013	0.004	-0.076	-1.658	0.106	-0.001
Pr(<z)	0.012	0.001	0.003	0.000	0.002	0.050	0.001	0.000

Table 1.2 Spatio-temporal drift model coefficient estimates

The global poisson regression reveals interesting associations of mosquito abundance to the selected covariates, weather and environmental conditions as well as the time of the trap nights. However, residual analyses in Figures 1.3 indicate that the substantial variation of mosquito abundance remains to be further explored. Both the histogram of the deviance residuals Figure 1.3(a) and the scatter plot between fitted values and residuals Figure 1.3(b) show the non-normality and the heterogeneity in residuals, respectively. Highly skewed residual distribution also indicates that the global model did not successfully capture the high variability in mosquito abundance, in addition to the inhomogeneous variance of residuals across the fitted values. Not only the non-stationarity of the trend in the average intensity process, but both the spatial and temporal variograms of residuals in Figures 1.3(c) and (d) suggest the presence of a strong spatial and temporal autocorrelation in residuals. The spatial variogram in Figure 1.3(c) is calculated from weekly residual data. A separate spatial variogram is calculated each week using the residual data at separation vector $|\mathbf{h}| = 1$ km. The multiple dots in each separation vector distance denote the sample variogram value, i.e., the spatial variability of residual data at a separation vector distance $\mathbf{h}_k, k = 1, \ldots, 15$. Similarly, the temporal variability of the global poisson model residuals are calculated at each trap site with temporal lag $\tau = 1$ over up to 6 weeks $\tau_l, l = 1, \ldots, 6$. As shown in both sample variograms, the spatial variability per week is different from week to week and the temporal variability per site is varying site to site, i.e., the wide range of dots per lag distance in both spatial and temporal lags. This result confirms our hypothesis that a substantial amount of covariance heterogeneity remains in the mosquito abundance data. The solid line in both variographies, that is, the arithmetic average of sample variogram values at each lag, however, shows the need of spatial and temporal covariance structure models. The spatial variability is higher than temporal variability, where there is a trend of the variability which increases up to 4 km and two weeks until the variability gets stabilized.

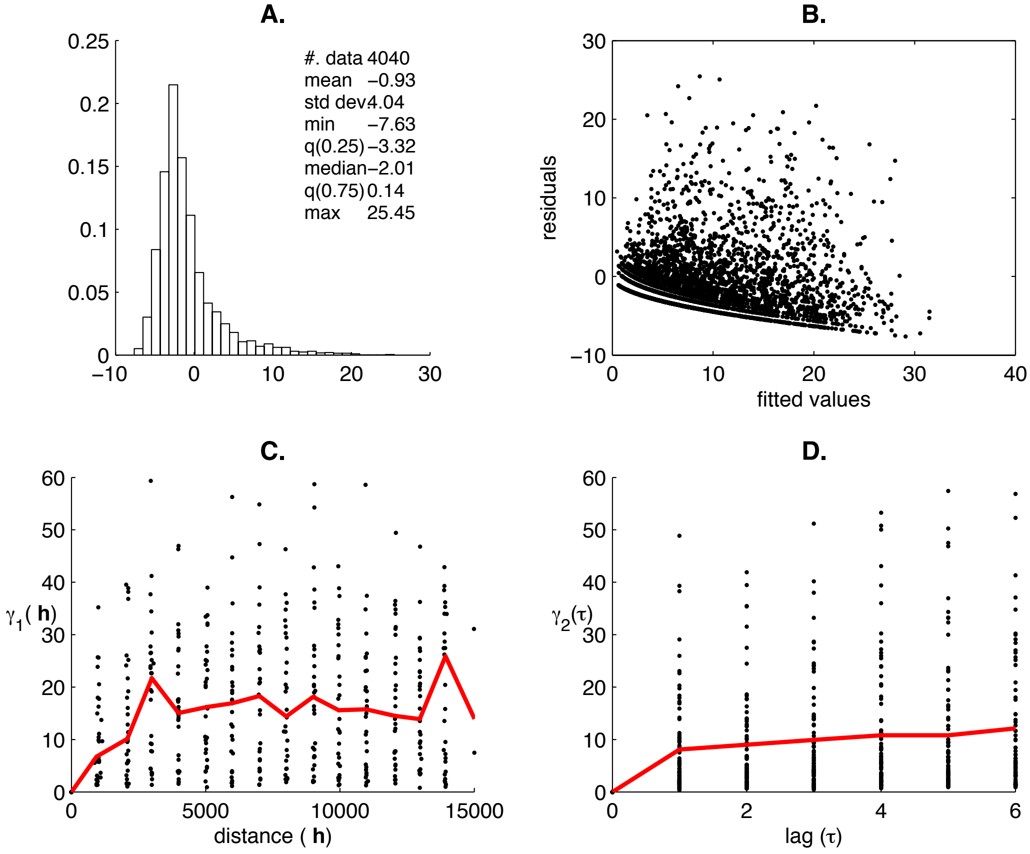

Figure 1.3 Residual analysis of global GLM residuals: (a) Histogram of residuals, (b) Scatter plot of residual deviance vs. fitted values, (c) and (d) Spatial and temporal variography of residuals, respectively.

The non-stationarity in the trend and spatial and temporal covariance structure calls for a model that addresses several issues mentioned above. In the following subsection, we will introduce a local spatio-temporal predictive mosquito abundance model, which takes into account the nature of mosquito count data using a poisson regression model and accommodates the non-stationarity of the underlying process by building multiple models over subregions. Our goal is to achieve a quasi-stationarity, in which the trend and the covariances are stationary, by performing a local poisson regression model and specifying a joint space-time covariance inside the subregion using local data only.

1.3.3 Local spatio-temporal model calibration

We adopt a local spatio-temporal geostatistical model to address the non-stationarity of residuals. The drift of the log transformed spatio-temporal average intensity process is locally estimated per prediction point with a poisson regression model using local data alone. All the local data fall within the moving spatio-temporal cylinder. The spatio-temporal correlation present in residuals are explicitly taken into account in the joint space-time covariance model. While the proposed local spatio-temporal poisson model is a promising alternative to the global poisson regression model, a set of key parameters needs to be determined before any model inference and prediction are made. The key parameters, such as the size of spatio-temporal moving cylinder and the hyperparameters for spatial and temporal variogram models, such as, range, sill, and nugget, play an important role to control the quality of predictions. Various statistical approaches, such as a Bayesian approach and E-M algorithm, can be used to tackle such problems, but it may intensify the complexity of the computation. In the current paper, we use a cross-validation method to identify the optimal moving cylinder size combined with our prior knowledge and understanding of mosquito behaviors and regional landscape to determine optimal parameters of model variograms.

The implementation of the local spatio-temporal prediction model requires the determination of spatio-temporal cylinder size, which could be different from prediction point to point. Unlike the global poisson regression, the quality of model prediction and inference are spatially and temporally varying and they are highly dependent on the neighboring data. In the current paper, the size of the spatio-temporal cylinder centered at a prediction point is determined with respect to the availability of neighboring data. The quality of local prediction depends on the values of the local neighborhoods. Homogeneous local neighboring data will yield less biased model parameter estimates and more accurate prediction outcomes. It is important to identify an optimal cylinder size that is just large enough to contain sampling points to estimate local poisson regression model coefficients and the spatial and temporal variogram variance function with accuracy sufficient for the prediction (Haas 1990). We also want to make the prediction procedures automatic, since there are many number of missing values in the data, i.e., a total of 141 prediction points. We use a programmable approach to determine the cylinder size per prediction point by recursively increasing the radius of spatio-temporal cylinder until enough number of data points are included.

In order to identify the optimal cylinder size, we conducted a sensitivity analysis where the effect of the cylinder size on the model prediction accuracy is examined using the leave-one out cross-validation method. We consider a range of minimum data points between 40 and 120 to be sufficient to perform a poisson regression model and to calculate a reliable sample variogram, while avoiding the potential nonstationarity in the trend and covariance. At each validation point, the posterior distribution of the missing value is obtained using the neighboring sample data located within the spatio-temporal cylinder that is parameterized by the number of minimum data points. The predicted value are compared with the true value, i.e., the observed mosquito count, and the discrepancy between the estimates and the observed value is compared. Most discrepancy values are within ± one standard deviation (15.412) of the mean discrepancy (1.142), but extremely high or low discrepancy is found especially when the observed count is high. Here, we used an empirical cut off values -50 and 50 to examine the portion of biased predictions per the minimum number of data points. That is, if the difference between predicted mosquito count and the observed count is either above 50 or below -50, the prediction is considered biased. Table 1.3 summarizes the sensitivity of prediction accuracy with respect to the minimum number of sample data used to define a moving spatio-temporal cylinder. The result shows that the chances of obtaining biased prediction increase, as the minimum number of data points is above 70. On the other hand, the prediction errors are minimum when the neighboring data criterion is set between 40 and 60 data points. Based on our finding, we use the minimum number of data points to 45 in the following local space-time predictions.

# min. data	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**110**	**120**
% bias	**1.932**	1.982	1.976	2.158	2.154	2.158	2.158	2.146	2.156

Table 1.3 The proportion of biased prediction under different cylinder sizes

1.3.4 Local space-time poisson regression

The goal is to make an inference about the unobserved spatio-temporal mean process $\lambda(s_0, t_0)$ underlying missing observations. We further assume that $\log \lambda(s_0, t_0)$ consists of a spatially and temporally varying drift $\mu(s_0, t_0)$ and a Gaussian stochastic process $R(s_0, t_0)$ with a zero mean, variance $C_R(\mathbf{0}, 0)$, and a stationary spatio-temporal correlation function $C_R(\mathbf{h}, \tau)$. The prediction at any missing data point (s_0, t_0) requires the search of local data, a local poisson regression model estimation, spatio-temporal variogram modeling, and the simple kriging prediction of the spatio-temporal random effect. In the previous section, we have shown that the quality of prediction highly depends on the design of the cylinder, i.e., the sensitivity of the model prediction accuracy to the number of minimum data points used to determine the moving spatio-temporal cylinder. In the current application, we further refined the cylinder size (i.e., 45 data points as an optimal size) to be within 1-7 km from the prediction location (target trap site) and within 6 weeks prior and after the trap-night. This decision is based on our understanding of the *Cx. pipens-restuans*' short distance migration and their variation over time during trap season.

The number of missing observations are significantly different from one region to another. We focus only on the prediction of missing values at permanent trap sites during the study period (June through mid-October each year), because of the sparse network of trap sites and the coarse resolution of explanatory variables. Typically, the number of missing values per trap site ranges between 1 - 6 points at the beginning or the end of the trapping season. Most trap sites located in the region of Peel (100%) and the City of Toronto (98%) have a small number of missing values.

For illustration, we randomly select a trap site located in the City of Toronto. This trap site, denoted by a square symbol in Figure 1.4, has five missing values. The prediction goal is to infer the underlying spatio-temporal average intensity process for the missing mosquito counts. The temporal profile of mosquito abundance over 36 weeks (entire study period) are shown in Figure 1.5(a), where the observed weekly mosquito counts are denoted by stems terminated with circles, and the prediction for the unobserved spatio-temporal mean counts underlying five missing values are denoted by stems terminated with stars. The target prediction of the spatio-temporal average count, i.e., $\log \hat{\lambda}(s_n, t_p)$ at the p-th week $p = 1, 2, 17, 18, 36$ is obtained by a linear combination of the estimated drift $\hat{\mu}(s_n, t_p) = \hat{\beta}_0 + \sum_{i=1}^{7} \hat{\beta}_i x_i(s_n, t_p)$ and a realization $\hat{r}(s_n, t_p)$ of the stochastic residual process obtained from simple kriging. When a multivariate Gaussian distribution is assumed for residual random variables, a simple kriging prediction is equivalent to the conditional expectation of the random process and simple kriging variance corresponds to the conditional variance of such multivariate Gaussian random field (Deutsch & Journel 1998). Therefore, we can easily derive the predictive distribution of the stochastic residual process instead of the summary statistics, such as mean and variance, using a stochastic simulation.

The predicted spatio-temporal mean value and the associated measure of uncertainty presented in Figures 1.5(a) and (b) correspond to the conditional expectation and conditional variance of a realization of spatio-temporal mean process $\hat{\lambda}(s_n, t_p)$ at the n-th trap site and the p-th week. The predicted mean values are similar to the previous or following weeks' observations at the same trap site, but also their predicted values are affected by the observation at neighboring trap sites. As a measure of uncertainty, the prediction error variance in Figure 1.5(b) is not complete, because it is a function of the spatial-temporal covariance model and the configuration of observed spatio-temporal data and the target prediction point (Goovaerts 1997). This measure of uncertainty is independent of the attribute of surveillance data, i.e., mosquito counts, whereas the higher predicted mean values are likely to have higher prediction errors.

Validation is critical in a generalized linear mixed modeling (GLMM) process, because the complexity of the higher dimensional GLMM makes it more difficult to get statistical goodness-of-fit measures. We adopt a cross-validation method to assess the prediction accuracy, where the data set is split into training and the prediction data set. It is highly expected that the quality of model prediction has a strong relationship with the unknown (or missing) value itself. The more extreme target values are, the more biased the accuracy of model prediction is. To facilitate the model performance assessment, we conducted a cross-validation study across four different prediction groups whose members are selected based on the rank of observed values

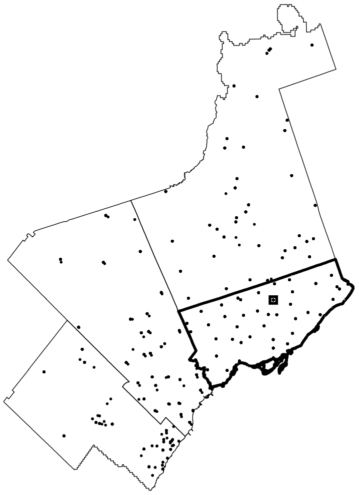

Figure 1.4 A randomly selected trap site in the City of Toronto.

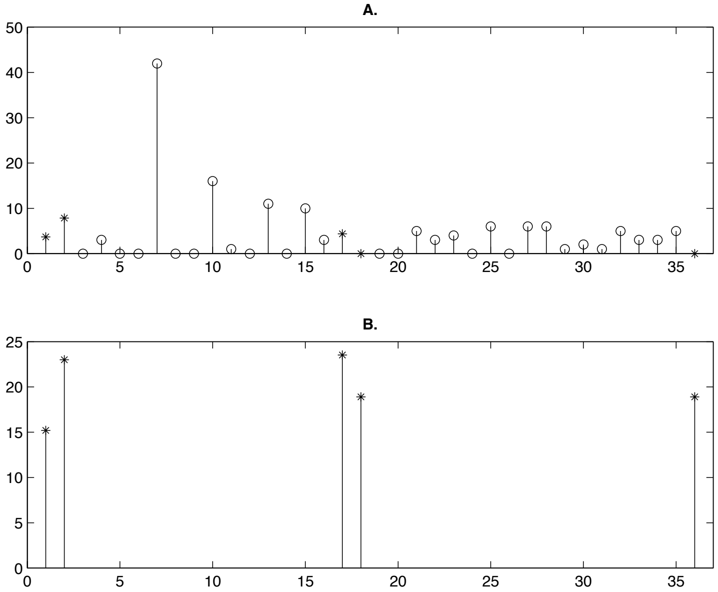

Figure 1.5 (a) A complete temporal profile of mean counts at the selected trap site in Figure 1.4 complemented by the model prediction for missing values, denoted by stems terminated with star (*) symbol. (b) The prediction error variances associated with local spatio-temporal predictions for missing values.

(y_L, y_U). The definition of group is summarized in Table 1.4. Approximately, 70% of data belong to Group 1, and the other data are evenly divided into Group 2 - 4 (approximately 10% each).

	Group 1	Group 2	Group 3	Group 4
(y_L, y_U)	(0, 7)	(7, 12)	(7, 25)	(25, 181)
# prediction points	2816	411	394	419

Table 1.4 Prediction Groups

Using the observed mosquito count and the seven covariates, we obtained mosquito count predictions at prediction points and assessed the prediction accuracy by comparing the predicted values to the observed count. The bias of the predicted values, i.e., the difference between the observed mosquito count and the predicted per prediction group, is summarized by boxplot in Figure 1.6(a). As expected, Group 4 has the highest bias both in the magnitude (-100 to 150) and in the proportion ($\approx$ 20%). We further investigated the predictions with substantial bias, in both positive (under-estimation) and negative (over-estimation), which is denoted by dashed line in the box plot. Only Group 4 (higher number of weekly mosquito observations) has substantially large positive bias. That is, the model underestimates true values when the observation is extremely high. This result comes with no surprise because the proposed model prediction is the spatial and temporal average of neighboring observation. Not only positive, but also substantially high magnitude of negative bias is obtained in Group 4, which is also explained by examining the spatial pattern of trap sites and the corresponding week with high biases. Model predictions at a number of trap sites that belong to Group 4 are either over- or under-estimated, which might be due to the fact that their observed mosquito counts in consecutive weeks were varying significantly and the proposed model could not capture such dynamic changes.

Both Figures 1.6(b) & 1.7 show the week and the location of trap sites with substantial under- or over-predictions, respectively. The over-estimation is obtained across a wide range of week 26 to the week 36, while the under-estimation is obtained in the middle of summer (weeks 29 - 34). This poor performance of the model might be due to spatially and temporally local variations occurred during the mosquito trapping.

1.4 Summary & Conclusions

We proposed a geostatistical spatio-temporal model to predict missing observations by combining the effects of meteorological and environmental conditions on WNv mosquito abundance with a stochastic spatio-temporal random field within a poisson generalized linear mixed model (Diggle & Ribeiro 2007). The proposed geostatistical spatio-temporal model takes into account the discrete counts in nature of the

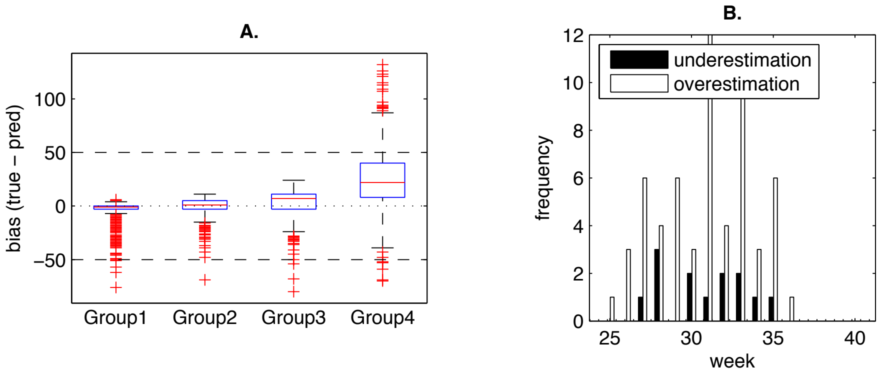

Figure 1.6 (a) Prediction bias per Group. (b) Weekly distribution of under- and over-predictions.

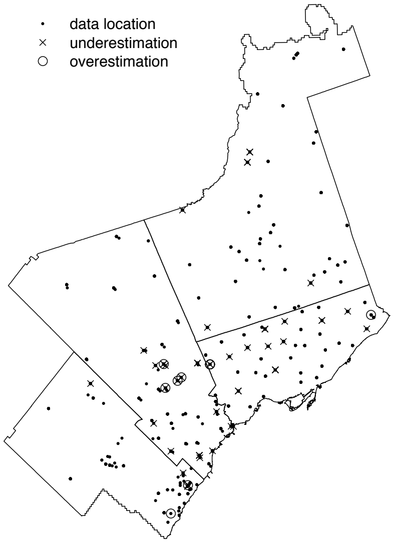

Figure 1.7 Spatial distribution of under- and over-predictions.

mosquito surveillance data and accommodates the spatially and temporally correlated error structure.

We identified factors that account for the variation in WNv vector population, including average temperature, NDVI index, and the dummy variable for year, and modeled the spatio-temporal fluctuations remaining in residuals by a stochastic spatio-temporal random field. One of the major issues to apply the proposed geostatistical space-time model to mosquito surveilance data is the presence of non-stationarity, which affects the predictive power of the proposed model. To address the non-stationarity problem in the current study, we used a moving local neighborhood where the study region and time domains are divided into homogeneous sub-units. Within a spatio-temporal neighborhood whose boundary is drawn on the basis of trapping frequency and the marginal distribution of the observed data to meet a quasi-stationarity decision, we derived a space-time covariance model and obtained the conditional expectation of the unobserved spatio-temporal mean process underlying mosquito count data at any permanent trap site with missing values. A log transformed predicted mean count of the latent spatio-temporal field was then obtained by combining a drift estimate and the local simple kriging residual prediction using spatially and temporally adjacent neighboring residual data. The predicted mosquito counts are accompanied with the measure of uncertainty associated with prediction based on spatial and temporal data configuration with respect to the target prediction point as well as the joint space-time covariance structure.

The model assessment is conducted using a cross-validation technique, which shows that the model prediction of extreme values is likely to involve high biases compared to the prediction of average count. We expect the results of bias analysis and uncertainty associated with spatio-temporal prediction of mosquito counts will be used to guide the sampling design in mosquito control programs, which will facilitate control efforts and reduce transmission risk to humans.

In summary, we demonstrated the application of the proposed geostatistical spatio-temporal model to missing data prediction problem and reported the prediction accuracy associated with predicted values. While we limited our goal to site-specific predictions in the current study, the proposed model could be further extended to estimate a entomological risk surface. To maintain the predictive power, however, it is necessary to use predictors at finer spatial scales and temporal intervals. As discussed in the result section, the uncertainty measure, i.e., prediction error variance, reported in the current study can be further explored, since they are independent of the attribute value and depend only on the spatial and temporal configuration and the joint spatio-temporal correlation error structure. The potential uses of uncertainty measures associated with predicted values are diverse: they can be used as a means of assessing the reliability of the predicted mosquito counts, but also they can be used as a guideline design the following surveillance, i.e., where to place a trap site to improve the risk estimates of exposure to WNv infection or to close an existing trap site to minimize redundancy, and how often the trapping efforts should be made during trap season.

In future work, we will conduct a systematic evaluation of the uncertainty associated with predicted missing observations, for example, using a spatial stochastic simulation in a Monte Carlo framework, but also we will assess the risk of exposure to WNv by combining entomological risk map with other sources of surveillance data, e.g., avian surveillance data and human case data. The inherent weakness of different types of surveillance data, i.e., mosquitoes, dead birds reports and testing, and the mandatory human case reports, perhaps, can be accounted for by taking an hybrid approach (Winters et al. 2008).

Bibliography

Anselin, L. & Rey, S. (1991), 'Properties of tests for spatial dependence in linear regression models', *Geographical Analysis* **23**(2), 112–131.

Arbia, G. (1989), *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Kluwer Academic Publisher, Dordrecht, The Netherlands.

Beroll, H., Berke, O., Wilson, J. & Barker, I. K. (2007), 'Investigating the spatial risk distribution of West Nile virus disease in birds and humans in southern Ontario from 2002 to 2005', *Population Health Metrics* **5**, 3.

Bolling, B. G., Barker, C. M., Moore, C. G., Pape, W. J. & Eisen, L. (2009), 'Seasonal patterns for entomological measures of risk for exposure to *Culex* vectors and West Nile virus in relation to human disease cases in northeastern Colorado', *J Med Entomol* **46**(6), 1519–31.

Brown, H., Childs, J., Diuk-Wasser, M. & Fish, D. (2008), ‘Ecological factors associated with West Nile virus transmission’, *Emerging Infectious Disease* **14**(10), 1539–1545.

Carney, R. M., Ahearn, S. C., McConchie, A., Glasner, C., Jean, C., Barker, C., Park, B., Padgett, K., Parker, E., Aquino, E. & Kramer, V. (2011), ‘Early warning system for West Nile virus risk areas, California, USA’, *Emerging Infectious Diseases* **17**(8), 1445–54.

Chuang, T.-W., Henebry, G. M., Kimball, J. S., VanRoekel-Patton, D. L., Hildreth, M. B. & Wimberly, M. C. (2012), ‘Satellite microwave remote sensing for environmental modeling of mosquito population dynamics’, *Remote Sensing of Environment* **125**(0), 147 – 156.

Chuang, T.-W., Hildreth, M. B., Vanroekel, D. L. & Wimberly, M. C. (2011), ‘Weather and land cover influences on mosquito populations in weather and land cover influences on mosquito populations in Sioux Falls, South Dakota’, *Journal of Medical Entomology* **48**(3), 669–679.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, New York.

Degaetano, A. T. (2005), ‘Meteorological effects on adult mosquito (*Culex*) populations in metropolitan New Jersey’, *International journal of biometeorology* **49**(5), 345–53.

Deutsch, C. V. & Journel, A. G. (1998), *GSLIB: Geostatistical Software Library and User’s Guide*, 2nd edn, Oxford University Press, New York.

Diggle, P. & Ribeiro, P. (2007), *Model-based Geostatistics*, Springer Series in Statistics, Springer, New York.

Diuk-Wasser, M., Brown, H., Andreadis, T. & Fish, D. (2006), ‘Modeling the spatial distribution of mosquito vectors for West Nile virus in Connecticut, USA’, *Vector-Borne and Zoonotic Diseases* **6**(3), 283–295.

Goodchild, M. F. (2001), Models of scale and scales of modeling, *in* N. J. Tate & P. M. Atkinson, eds, ‘Modeling Scale in Geographical Information Science’, John Wiley & Sons, New York, pp. 3–10.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York.

Griffith, D. A. (2005), ‘A comparison of six analytical disease mapping techniques as applied to West Nile virus in the coterminous United States’, *International Journal of Health Geographics* **4**, 18.

Haas, T. (1990), ‘Kriging and automated variogram modeling within a moving window’, *Atmospheric Environment* **24A**(7), 1759–1769.

Johnson, G. D., Eidson, M., Schmit, K., Ellis, A. & Kulldorff, M. (2006), 'Geographic prediction of human onset of West Nile virus using dead crow clusters: an evaluation of year 2002 data in New York State', *American journal of epidemiology* **163**(2), 171–80.

Kilpatrick, A., Kramer, L., Campbell, S., Alleyne, E., Dobson, A. & Daszak, P. (2006), 'West Nile virus risk assessment and the bridge vector paradigm', *Emerging Infectious Disease* **11**(3), 425–429.

Kyriakidis, P. & Journel, A. (1999), 'Geostatistical space-time models: a review', *Mathematical Geology* **31**(6), 651–684.

Liu, H. & Weng, Q. (2009), 'An examination of the effect of landscape pattern, land surface temperature, and socioeconomic conditions on WNv dissemination in Chicago', *Environmental Monitoring and Assessment* **159**, 143–161.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman & Hall/CRC.

Morin, C. W. & Comrie, A. C. (2010), 'Modeled response of the West Nile virus vector *Culex quinquefasciatus* to changing climate using the dynamic mosquito simulation model', *International Journal of Biometeorology* **54**(5), 517–529.

O'Hara, R. B. & Kotze, D. J. (2010), 'Do not log-transform count data', *Methods in Ecology and Evolution* **1**(2), 118–122.

Reisen, W., Fang, Y. & Martinez, V. (2006), 'Effects of temperature on the transmission of West Nile virus by *Culex tarsalis* (diptera: *Culicidae*)', *Journal of Medical Entomology* **43**, 309–317.

Ruiz, M., Chaves, L., Hamer, G., Sun, T., Brown, W., Walker, E., Haramis, L., Goldberg, T. & Kitron, U. (2010), 'Local impact of temperature and precipitation on West Nile virus infection in *Culex* species mosquitoes in northeast Illinois, USA', *Parasites & Vectors* **3**(1), 19.

Shaman, J. & Day, J. (2007), 'Reproductive phase locking of mosquito populations in response to rainfall frequency', *PLoS One* **2**(3), e331.

Soverow, J., Wellenius, G., Fisman, D. & Mittleman, M. (2009), 'Infectious disease in a warming world: how weather influenced West Nile virus in the United States (2001–2005)', *Environ Health Perspect* **117**(7), 1049–1052.

Theophilides, C. N., Ahearn, S. C., Grady, S. & Merlino, M. (2003), 'Identifying West Nile virus risk areas: the dynamic continuous-area space-time system', *Am J Epidemiol* **157**(9), 843–54.

Turell, M., Dohm, D., Sardelis, M., O'guinn, M. & G.A (2005), 'An update on the potential of north american mosquitoes (diptera: *Culicidae*) to transmit West Nile virus', *Journal of Medical Entomology* **42**(1), 57–62.

Ward, M. P., Schuermann, J. A., Highfield, L. D. & Murray, K. O. (2006), 'Characteristics of an outbreak of West Nile virus encephalomyelitis in a previously uninfected population of horses', *Veterinary Microbiology* **118**(3–4), 255 – 259.

Winters, A. M., Bolling, B. G., Beaty, B. J., Blair, C. D., Eisen, R. J., Meyer, A. M., Pape, W. J., Moore, C. G. & Eisen, L. (2008), 'Combining mosquito vector and human disease data for improved assessment of spatial West Nile virus disease risk', *The American journal of tropical medicine and hygiene* **78**(4), 654–65.

Zuur, A., Leno, E., Walker, N., Saveliev, A. & Smith, G. (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer.