International Scholarly Research Network
ISRN Biomathematics
Volume 2012, Article ID 132342, 12 pages
doi:10.5402/2012/132342

Research Article

Spatially Explicit Nonlinear Models for Explaining the Occurrence of Infectious Zoonotic Diseases

Stephen Jones,[1] William Conner,[2] and Bo Song[2]

[1] *BlueCross BlueShield of Tennessee, Department of Medical Informatics, 1 Cameron Hill Circle, Building 2.1, Chattanooga, TN 37402, USA*
[2] *Forestry and Natural Resources Department, Clemson University, Georgetown, SC, USA*

Correspondence should be addressed to Stephen Jones, stephen_jones@bcbst.com

Received 2 August 2012; Accepted 24 September 2012

Academic Editors: H. Ishikawa, M. Jose, Y. Pan, and W. Raffelsberger

Zoonotic diseases can be transmitted via an arthropod vector, and disease risk maps are often created based on underlying associative factors within the surrounding landscape of known occurrences. A limitation however is the ability to map disease risk at a meaningful geographic scale, and traditional regression modeling approaches may not always be appropriate. Our objective was to determine if nonlinear modeling could improve explanatory power in describing the occurrence of 2 tick-borne diseases (Lyme disease (LD) and Rocky Mountain spotted fever (RMSF)) known to occur in Tennessee. Medically diagnosed cases of LD (ICD-9: 088.81) and RMSF (ICD-9: 082.0) were extracted from a managed care organization data warehouse for the 2000–2009 time period. Four separate modeling techniques were constructed (logistic regression, classification and regression tree (CART), gradient boosted tree (GBT), and neural network (NNET)) and compared for accuracy. Results suggest that areas higher in disease prevalence were not necessarily the same areas having high predicted disease risk. GBT best explained LD occurrence (misclassification rate: 0.232; ROC: 0.789). RMSF prevalence was best explained with an NNET algorithm (misclassification rate: 0.288; ROC: 0.696). Covariates explaining disease risk included forested wetlands, urbanization, and median income. Nonlinear modeling may provide better results than traditional regression-based approaches.

1. Introduction

Because zoonotic diseases are transmitted via an arthropod vector, it is often of interest to understand vector habitat in the epidemiologic study of diseases. It is common in spatial epidemiology to describe vector habitat and then create causal inference risk maps of potentially high-risk areas based on habitat preferences [1, 2]. These geospatial mapping exercises outline areas having high probabilities of vector prevalence and then infer disease risk based on probable presence or absence. For example, abundance of the tick genus *Ixodes*, one of which is the vector primarily responsible for the transmission of Lyme disease (LD), is associated with temperature, landscape slope [3], forested areas with sandy soils [4], and increasing residential development [5]. Tularemia prevalence is positively associated with dry forested habitat areas [6]. Human populations living within forested areas and on specific soils are at higher risk of contracting LD [7, 8]. Human monocytic ehrlichiosis (HME or *Ehrlichia chaffeensis*) is more associated with wooded habitats compared to neighboring grassy areas [9].

A major limitation in the study of such diseases, however, is the ability to comprehensively track disease prevalence at a meaningful geographic scale [8]. Data aggregations and disease prevalence rates are most often presented at the county level [1, 10, 11]. Unfortunately, county level assessments compared to ZIP code level analyses may mask smaller isolated high-risk areas as well as obscure within county variability [12–15]. In 2007, the Centers for Disease Control and Prevention (CDC) called for a means to improve data collection methods to determine probable pathogen exposure sites based specifically on patient activity spatial patterns [10]. This suggests geocoding the residential location (street address or ZIP code) of the infected patient and conducting a radial search around that point to examine the underlying landscape [16]. However, data describing possible

pathogen exposure sites are limited [7, 10], and means to collect this information can be very costly. Therefore, studies within the wildlife and ecological sciences are often limited in predictive power due to the inability to generate large sample sizes, either because of costs, data availability, or both [17].

The usefulness of administrative medical claims data in the study of infectious diseases has been previously discussed [14, 15, 18]. Administrative medical claims data contain, among other things, a patient's ZIP code at the time of service, date of medical service, and medical diagnoses describing the reason(s) why the patient is seeking medical care. The use of this data is relatively easy and inexpensive to work with and could represent a volume rich source of persons diagnosed with zoonotic diseases. The geographic element of a patient's residence location combined with the diagnosis provides spatially explicit information regarding what the patient was exposed to, and potentially where the exposure may have occurred. Spatially explicit disease case models using data from managed care organizations (MCO) do not exist. It is the purpose of this study to determine if meaningful exploratory spatial models can be constructed at the ZIP code level to help describe the occurrence of 2 tick-borne zoonotic diseases known to occur in Tennessee (LD and Rocky Mountain spotted fever (RMSF)) by comparing nonlinear modeling approaches to traditional regression analyses.

2. Methods

2.1. Study Area. The study area for this project included the state of Tennessee, USA, a southeastern state and is approximately bounded within the southernmost west coordinate (−90.309200, 34.995800) to the northern most east coordinate (−81.646900, 36.611900). Estimated land cover percentages for the state are as follows: open water (2.7%), forested wetland (3.0%), nonforested wetland (0.4%), grassland/pasture (37.2%), cropland (5.8%), upland deciduous forest (40.6%), upland mixed forest (4.4%), upland coniferous forest (3.6%), urban/developed (1.9%), and nonvegetated (0.2%) [19].

2.2. Disease Case Data. Medically diagnosed cases of LD and RMSF from January 1, 2000 to December 31, 2009 were collected from the electronic data warehouse system of a large MCO located in Tennessee. These diseases were selected because they occurred in at least 20% of the sample units (i.e., ZIP codes) and therefore would not be potentially biased by issues related to rare event modeling. The process of data collection was described in detail elsewhere [18], but, briefly, zoonotic disease cases within the study area of Tennessee were extracted from MCO claims data warehouse if they had any of following diagnosis codes: for LD (ICD-9 code: 088.81) and RMSF (ICD-9 diagnosis code: 082.0). Disease cases without at least 3 separate line items in the claims system were removed. Any patient receiving medical services for one of the selected diseases prior to the start of the study period or after the study period was removed from the analysis.

2.3. Spatial Sample Unit. This study uses two types of spatial data: (1) disease occurrence data at the ZIP code level extracted from medical claims and (2) underlying spatial data to describe the sociodemographic, geographic, and habitat characteristics surrounding the ZIP code centroid. ZIP codes can have either a geographic centroid or population-weighted centroid. A geographic centroid is defined by the US Census Bureau as the center of the tabulation area as it relates to the geographic extremes of the physical boundaries of the polygon. A population-weighted centroid is the center of the tabulation area as determined by where the majority of the population is located within the polygon. For this study, the geographic centroid was converted to a population weighted ZIP code centroid to create the spatial sample units. This weighted-average transformation was accomplished using the underlying inscribed census block population counts within the enclosing ZIP code to calculate an adjusted longitude (x_z) and latitude (y_z); see the following formula:

$$x_z = \frac{\sum_{i=1}^{z} p_i x_i}{\sum_{i=1}^{z} p_i}, \qquad y_z = \frac{\sum_{i=1}^{z} p_i y_i}{\sum_{i=1}^{z} p_i}, \tag{1}$$

where x_z is transformed population-weighted x-coordinate for ZIP code z, p_i is the population of the ith census block within ZIP code z, and x_i is the x-coordinate value of the ith census block repeat for the y-coordinate.

2.4. Dependent (Response) Variable. For the purposes of this study, spatial models are considered to be exploratory models at the ZIP code level, and separate models were built for each of the 2 studied diseases. Two separate modeling exercises were conducted across the 2 diseases using different dichotomous (i.e., binary) response variables. The first approach assigned a value of 1 to all ZIP codes if the disease in question was present at any time during the study period; otherwise the ZIP is assigned a value of 0. ZIP codes with a value of 1 are hereafter considered "case" sites.

The second modeling approach assigned a 1 to only those ZIP codes with a z score greater than zero. This was done to explain characteristics of ZIP codes having an observed disease case volume above expectation relative to all other ZIP codes. The observed number of cases in a ZIP code was the per 100 k rate averaged over the study period, and the expected number of cases within a ZIP code was derived from the statewide prevalence rate averaged over the entire study period. Thus ZIP code rates were proportional to the member population within that ZIP code. A z score was calculated for each ZIP code using the standard formula:

$$z_i = \frac{y_i - \overline{y_j}}{\sigma_j}, \tag{2}$$

where z_i is z score for ZIP code i, y_i is observed per 100 k rate of cases in ZIP code i averaged over the entire study period, $\overline{y_j}$ is mean of the disease rate cases averaged across the set of j ZIP codes, and σ_j is standard deviation of the disease rate cases across the set of j ZIP codes.

Table 1: Top 11 wetland types by area in Tennessee and selected for study.

Wetland type	Description	Area (km^2)	Percent of total area	Cumulative* %
L1UBHh	Lacustrine limnetic unconsolidated bottom permanently flooded dike/impounded	2726.4	30.9%	30.9%
PFO1A	Palustrine forested broad-leaved deciduous temporary flooded	1812.1	20.5%	72.0%
R2UBH	Riverine lower perennial unconsolidated bottom permanently flooded	1347.9	15.3%	82.0%
PFO1C	Palustrine forested broad-leaved deciduous permanently flooded	1061.7	12.0%	90.8%
PUBHh	Palustrine unconsolidated bottom permanently flooded dike/impounded	254.2	2.9%	84.6%
PFO6F	Palustrine forested deciduous semipermanently flooded	205.3	2.3%	86.3%
PFO1F	Palustrine forested broad-leaved deciduous semipermanently flooded	109.6	1.2%	86.5%
PUBHx	Palustrine unconsolidated bottom permanently flooded excavated	96.7	1.1%	87.4%
R2UB3H	Riverine lower perennial unconsolidated bottom mud permanently flooded	84.9	1.0%	88.3%
PEM1A	Palustrine emergent persistent temporary flooded	79.5	0.9%	89.1%
PEM1C	Palustrine emergent persistent seasonally flooded	68.5	0.8%	89.8%

* When land use categories are ranked in descending order relative to percent of total area, cumulative percent is calculated in this descending order.

2.5. Independent Variables. Underlying sociodemographic, geographic, and habitat characteristics of the landscape surrounding the population-weighted ZIP code centroid served as explanatory variables. Clinical variables representing the per 100 k rate of LD and RMSF as well as other zoonotic diseases (human monocytic ehrlichiosis, babesiosis, tularemia, La Crosse viral encephalitis, and West Nile virus) within the ZIP code were also included. Independent variables in the model are considered multilevel because data aggregations were done at 2 spatial scales, 1.6 km and 8 km. Sociodemographic factors included total population count and median income from the 2000 US Census Bureau estimates within 1.6 km and 8 km of the ZIP centroid. Geographic factors included continuous distance (km) to the nearest river/stream and the number of river kilometers within the 2 radial aggregation bands. Habitat characteristics included the amount (km^2) of land use type and wetland type (described below) within the 2 radial aggregation bands.

Land use data was downloaded from the Tennessee Spatial Data Server (TSDS) and is a generalized version of the detailed vegetation map that was prepared in compliance with the National Gap Analysis Program effort. The 10 land cover types were derived from classification techniques performed on Landsat Thematic Mapper imagery and included open water, forested wetland, nonforested wetland, pasture/grassland, cropland, upland deciduous forest, upland mixed forest, upland coniferous forest, urban/developed, and nonvegetated (barren land and strip mines/rock quarries/gravel pits). The strip mines/rock quarries/gravel pits class were taken from ancillary data sets and added to the classification file. The forest classes were extracted from satellite imagery and reclassified. Forest communities were interpreted from aerial videography acquired in April 1995 and correlated to the satellite imagery [19].

Digital wetland areal data was downloaded from the TSDS and is sourced from the National Wetlands Inventory (NWI) database. The US Fish and Wildlife Service (USFWS) and the US Geological Survey (USGS) are the federal agencies primarily responsible for providing geospatial information relative to the Nation's wetlands. This data layer represents the extent, approximate location, and type of wetlands and deepwater habitats in the conterminous United States. These data delineate the areal extent of wetlands and surface waters as defined elsewhere [20]. Certain wetland habitats are excluded from the National mapping program because of the limitations of aerial imagery as the primary data source used to detect wetlands. This data layer was digitized from USGS topographic base maps. Alpha-numeric codes describing the type of wetland are attributed to each digitized polygon and correspond to the wetland and deepwater classifications. For example, "L1UB1Hx" indicates the delineated area as

(i) L: lacustrine (system),

(ii) 1: Limnetic (subsystem),

(iii) UB: unconsolidated bottom (class),

(iv) 1: Cobble-Gravel (subclass),

(v) H: permanently flooded (water regime modifier),

(vi) X: excavated (modifier).

There were a total of 567 different described wetland types in the Tennessee NWI wetlands data layer. To reduce the amount of potential explanatory variables, the top 11 wetland types by area were selected (Table 1). This reduced set of wetland areas account for approximately 90% of the entire landscape, so little information was lost and provided a refined basis for predictive modeling.

All continuous independent variables (i.e., covariates) were transformed using a quantitative binning procedure. This was done to improve model performance so as not to restrict the relationships between covariates and response to only linear interpretations. For each covariate, 4 bins were created using quartiles to generate groups by splitting the data into bins having approximately the same frequency of observations. For example, the covariate "median income" could be separated into 4 bins, where INCOME_BIN_1 has all observations with an income less than \$29,000, INCOME_BIN_2 (\$29–\$33,000), INCOME_BIN_3 (\$33–\$39,000), and INCOME_BIN_4 (> \$39.000). These transformed variables are then treated as ordinal dummy variables

in the modeling procedures. When modeling a particular disease, geographic cooccurrence of all other diseases was included as a binary indicator (0, 1 where 1 indicates that another disease was also recorded in the ZIP code).

Patient level characteristics (e.g., age, gender, and comorbidities) were excluded from analyses because the intent of this study was to determine what geographically based risk factors could explain disease occurrence. Additionally, we aimed to produce risk factors that could be replicated in other environments without requiring known case/patient level information.

2.6. Analytical Modeling Techniques. Four separate modeling techniques were compared (stepwise logistic regression, classification decision tree, gradient boosted tree, and neural network) to determine which model type performs best (i.e., champion model). The modeling dataset consisted of 615 ZIP code records with 2 different binary response variables (evidence of disease, above average prevalence according to z score) and all aforementioned explanatory variables. The dataset was partitioned into two mutually exclusive data sets, a training data set, and a validation data set. The training data set was used for preliminary model fitting, and then once the model was built, the validation data set was used to fine tune (to help prevent over-fitting) and assess the final adequacy of the model. The data partitions were created using stratified sampling (stratified by the binary response variable), and the training data set included approximately 80% ($n = 490$) of the observations, and the validation set contained the remaining 20% ($n = 125$).

Stepwise logistic regression (SLR) is a variable selection algorithm that begins with no candidate variables in the model and then systematically adds effects that are significantly associated with the response variable [21]. Effects can be subsequently removed if it is not significantly associated with the response once another variable enters the model. This selection process continues until either (1) no other effect in the model meets the "stay significance level" or (2) the user-defined number of iterations criterion is met. The entry significance level value was set to 0.5 to ensure that effects with potential were considered, while stay significance was set to a more conservative 0.05 to guard against type I errors (concluding that a factor was influential when in fact it was not).

A classification and regression (CART) decision tree [22] is a commonly used algorithm in data-mining and machine-learning techniques. Classifications are used with nominal targets, while regression trees are used with continuous targets. A tree is created by applying a series of simple interpretable rules to the data in a recursive partitioning factor using a splitting criterion. These rules are then used to classify new observations into a series of tree nodes. One of the major benefits of a decision tree is its ability to use missing data which can often be as informative as known data, unlike regression techniques which cannot process this information directly. A classification tree was created using the Pearson Chi-square P value statistic as a splitting criterion. Maximum threshold P values for variable consideration in the splitting criterion were set to 0.2 with a Bonferroni adjustment (to account for multiple comparisons), and the minimum number of acceptable observations for a categorical value was set at 15.

Gradient boosting within classification and regression trees (GBT) is an emerging technique in data-mining algorithms which has been shown to outperform traditional decision tree approaches [23, 24]. Boosting is an adaptive method designed to improve predictive performance by combining multiple simple models into one overall "ensemble" model [25, 26]. Boosting is described in detail elsewhere [25], but, briefly, this approach recursively resamples the data to generate results that form a weighted average of the resampled data set. The successive samples are adjusted to accommodate previously computed inaccuracies. This continues until a user-defined limit is reached, and then each tree within the series is combined to form a single final algorithm explaining the response variable.

A neural network (NNET) is a type of model that is designed to mimic the neurophysiology of the human brain, in that it attempts to "learn" as it moves along the data and examines it. These types of models are referred to as feed-forward backpropagation networks [27]. As with the gradient boosting technique, they are typically used when understanding that the effects of the model are less important compared to model performance. That is, the output of the model cannot be readily interpreted as the aforementioned SLR and CART techniques can. In a neural network, there are three kinds of units in the modeling procedure.

(1) Input units obtain the values of covariates and standardize those values.

(2) Hidden units perform internal computations, providing the nonlinearity that makes neural networks powerful.

(3) Output units compute predicted values and compare those predicted values with the values of the response variable.

Each unit produces a single computed value, and this computed value is passed along the connections to other hidden or output units. Output units (i.e., predicted values) are compared with the response variable value to compute the error function in an attempt to minimize the error. For this project, the multilayer perceptron (MLP) method which is the most common network technique was utilized. The MLP was leveraged because they are best used when prior knowledge of the relationship between inputs and targets is unknown.

2.7. Model Comparisons. All models were built using SAS Enterprise Miner [28]. A model champion was chosen using the lowest overall misclassification rate applied to the validation dataset, which represents the percentage of all incorrectly predicted observations. This metric was chosen because we are simultaneously interested in minimizing false positive rates (proportion of ZIP codes predicted to have a disease case but did not) and false negative rates (proportion of ZIP codes predicted to not have a disease case but did) because both carry a public health liability. In addition,

the following model fit statistics were examined: receiver operator characteristic (ROC) curves, averaged squared error, sensitivity, specificity, and positive predictive values (PPVs). ROC curves plot sensitivity (true positive) on the y-axis and 1 specificity (false positive) on the x-axis, which can be used to visually interpret how well models perform relative to one another. To provide an interpretation for GBT and NNET models, the original complete data set ($n = 615$) was scored with the predictive algorithms produced by the final GBT and NNET models. This scoring calculated a predictive probability ranging from 0 to1 for each observation (i.e., ZIP code), detailing the likelihood that the disease in question would be present in the ZIP code. We then applied an explanatory CART model to the data to determine which independent variables were most associated with predicted probabilities greater than 0.5 [29].

3. Results

Of the 615 ZIP codes modeled, LD occurred in 49.9% ($n = 307$), RMSF occurred in 46.8% ($n = 288$), and LD or RMSF occurred in 97% ($n = 595$) of the ZIP codes. Approximately 33% ($n = 204$) of the ZIP codes had at least one case of LD and one case of RMSF. Of the 307 ZIP codes with LD, 51 had above average prevalence rates of LD (i.e., z score >0). Of the 288 ZIP codes with RMSF, 48 had above average prevalence rates (i.e., z score >0). Lastly, 2% ($n = 12$) of all ZIP codes had above average prevalence rates for both RMSF and LD.

The average LD rate across all ZIP codes and the entire study period was 4.56 per 100 k (SD: 9.46). The highest average LD rate (81.3 per 100 k; $n = 2$) occurred in ZIP code 38564 within the Knoxville region of Jackson County. The highest raw count of LD cases ($n = 29$) occurred in ZIP code 37830 of Anderson County (Knoxville region). The average RMSF rate across all counties and the entire study period was 4.05 per 100 k (SD: 9.32). The highest average RMSF rate (98.1 per 100 k; $n = 1$) occurred in ZIP code 37140 within the Nashville region of Hickman County. The highest raw count of RMSF cases ($n = 28$) occurred in ZIP code 38401 of Maury County (Nashville region). Approximately 38% of the LD cases occurred in the Nashville regional area (middle of state), and only 5% occurred in the Johnson City area (northeast portion of state). Similarly, 45% of the RMSF cases occurred in the Nashville regional area, and only 3% occurred in the Johnson City (Table 2, Figures 1 and 2).

Exploratory models examining ZIP codes having at least one occurrence of LD or RMSF successfully converged across all 4 modeling procedures. For the LD models, the GBT outperformed all others with a misclassification rate of 0.232, average squared error of 0.187, and ROC value of 0.789 (Table 3, Figure 3) using misclassification rate as the champion model selection criterion. Covariates most useful in explaining LD occurrence within the GBT model were cooccurrences of RMSF, amount of forested and nonforested wetlands, upland deciduous forests and urbanized/developed lands, population counts, median income, and wetland type palustrine unconsolidated bottom permanently flooded dike/impounded (PUBHh). Occurrence of RMSF was best explained using a neural network algorithm (misclassification rate = 0.288; average square error = 0.232; ROC = 0.696) (Table 3, Figure 4). Similar to the LD model, covariates most useful in explaining RMSF occurrence within the NNET model were cooccurrences of LD, amount of forested and nonforested wetlands, pasture/grasslands, and urbanized/developed lands, and population counts.

Table 2: Regional summary of disease distribution (Lyme disease and Rocky Mountain spotted fever) for the 2000–09 study period within Tennessee according to medically diagnosed claims data.

	Lyme disease N (%)	Rocky Mountain spotted fever N (%)	Total (%)
Nashville	343 (38%)	296 (45%)	639 (41%)
Knoxville	271 (30%)	149 (23%)	420 (27%)
Chattanooga	96 (11%)	87 (13%)	183 (12%)
Jackson	80 (9%)	80 (12%)	160 (10%)
Memphis	69 (8%)	26 (4%)	95 (6%)
Johnson City	44 (5%)	23 (3%)	67 (4%)
Totals	903	661	1,564

The algorithms from the champion models were used to score the validation data set ($n = 125$). Areas higher in disease prevalence were not necessarily the same areas having high predicted risk of disease infection (Figures 5 and 6). Table 3 provides a comprehensive assessment of all modeling outcomes for LD and RMSF and details covariates useful in explaining the variability in disease occurrence. A ZIP code was predicted to be a "case" site if the posterior probability was greater than or equal to 0.50, and therefore all model fit statistics are based on this predicted probability threshold. The symbols denote the general direction of the data, where a "+" indicates a positive relationship between the covariate and the response (i.e., as the covariate increases, the likelihood of a disease case occurring also increases), a "−" indicates a negative relationship between the covariate and the response (i.e., as the covariate increases, the likelihood of a disease case occurring decreases), and a "+/−" indicates a nonlinear relationship (i.e., in some ranges of the covariate, the likelihood of a disease case occurring decreases while, in other ranges, likelihood of disease increases). Note that the interpretations of the signs are only generalizations for two reasons: first, not all modeling procedures can be directly interpreted, and second, raw data were transformed using the binning procedure to segment each variable into groups thus allowing for non-liner interpretations. Additionally, P values for covariates are not reported because only the SLR procedure produces this type of interpretable statistic.

Model fit was adequate for both LD and RMSF. Figure 7 displays the performance of each model against the posterior probability predictions. The dotted 45° line represents a perfect model fit based on the predictions from the algorithm. For example, within the posterior probability range of 0.50-0.60 one would expect from a perfect model that approximately 50–60% of the ZIP codes actually had a disease case. Additionally, this chart can be used to determine

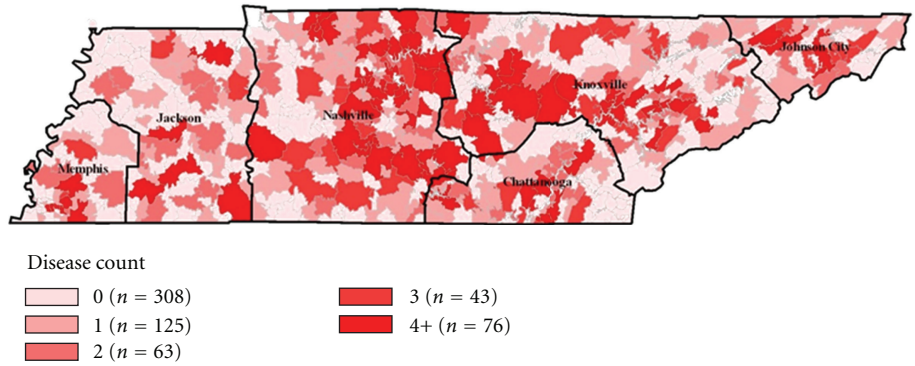

Figure 1: Spatial distribution of medically diagnosed Lyme disease cases (raw count) within Tennessee ZIP codes during the 2000–09 study period: dark black outlines define regional areas.

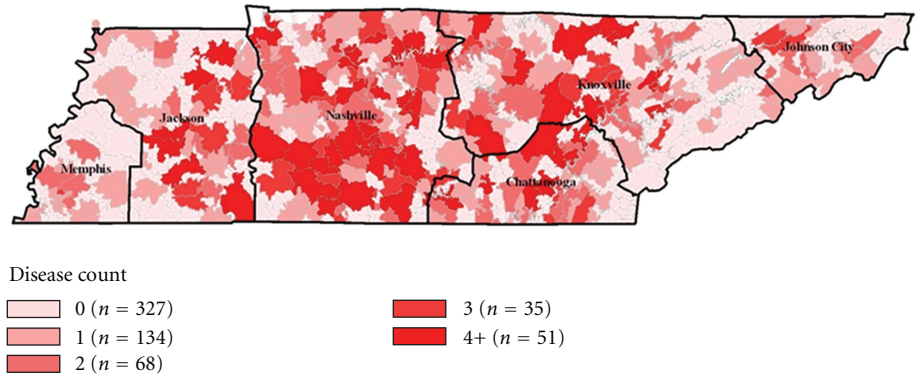

Figure 2: Spatial distribution of medically diagnosed Rocky Mountain spotted fever cases (raw count) within Tennessee ZIP codes during the 2000–09 study period: dark black outlines define regional areas.

the optimal posterior probability that should be used as a threshold to assign a predicted classification of "case" to the ZIP code. Moving the threshold value of the prediction can thus alter the model fit statistics because the model evaluation is based in part on the ability to predict a "case."

Exploratory models using the z score to define ZIP codes with above average prevalence rates were unsuccessful across all modeling types. The models did not pick any successful covariates to explain the above average prevalence rates, and therefore each algorithm simply predicted all observations to have below average prevalence rates. No other results are reported for these models.

4. Discussion

Results from this study suggest that LD and RMSF prevalence rates are associated with varying landscape characteristics. Disease prevalence was explained reasonably well within the spatially explicit models at the ZIP code level using administrative medical claims data as a source for diagnosed cases. It is believed that this is the first study that has attempted to use claims data for modeling the spatial characteristics of zoonotic diseases. This work also supports, at least in part, the viability of collecting and studying disease prevalence at the ZIP code level.

Three out of the four models suggested that LD prevalence increased with increasing urbanization. Two different covariates reflect urbanization in this study: urbanization as a land use type and population counts. Both covariates indicated a consistently positive relationship with disease risk across the 4 models. Assuming that urbanization is indicative of residential habitation, others have also suggested that residential factors were associated with increased risk of LD [30, 31]. Others found LD risk to be reduced in highly developed areas [7]. It is likely that land use types between studies are different and therefore produce different findings. Others [7] specifically described highly developed areas as multiunit residential neighborhoods and found these areas to be negatively associated with risk of LD. The urbanization variable used in our study is defined in terms of land use type, not actual physical representations of housing structures. Further, others [7] report an adjusted odd ratio upper confidence limit equal to 1 for this urbanization variable, which denotes the possibility that no significant association exists (i.e., in statistics, an odds ratio of 1 indicates that the independent variable does not have any statistical influence on the outcome variable).

LD prevalence was significantly associated with both forested and nonforested wetland areas. In a comprehensive

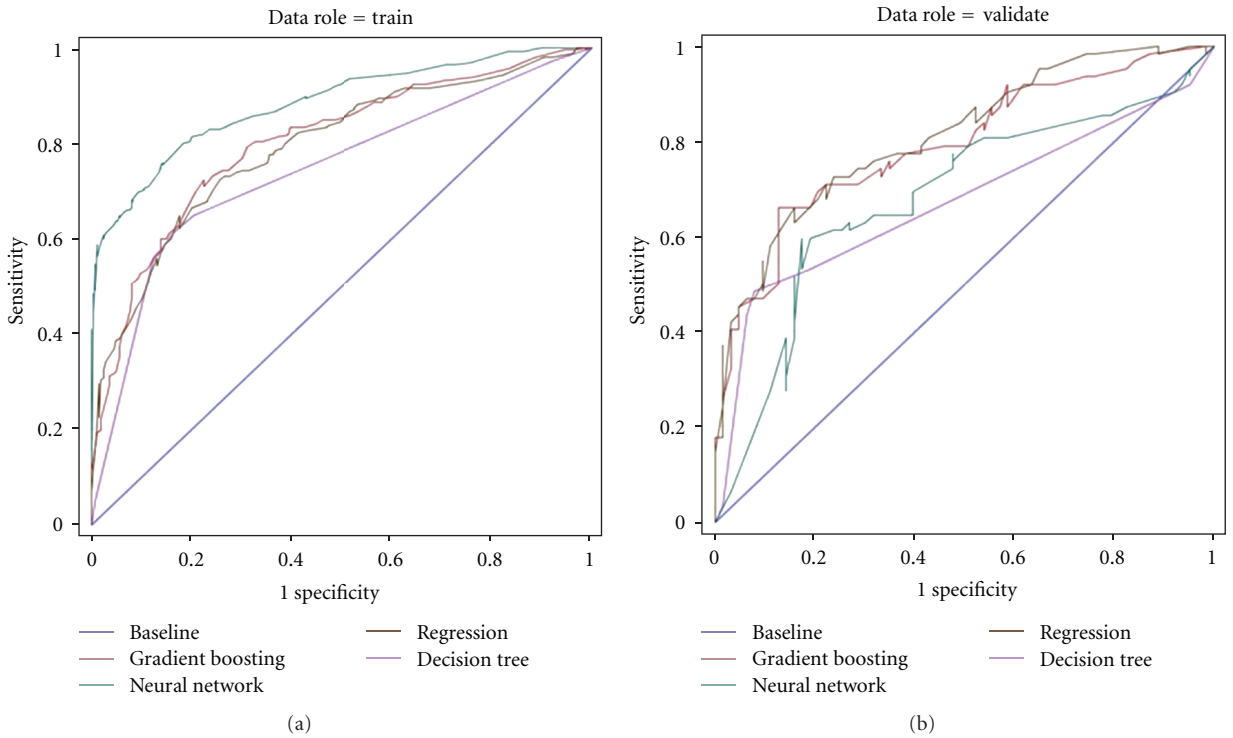

Figure 3: Receiver operator characteristic (ROC) curves for spatial models explaining occurrence of medically diagnosed cases of Lyme disease for the 2000–09 study period within Tennessee.

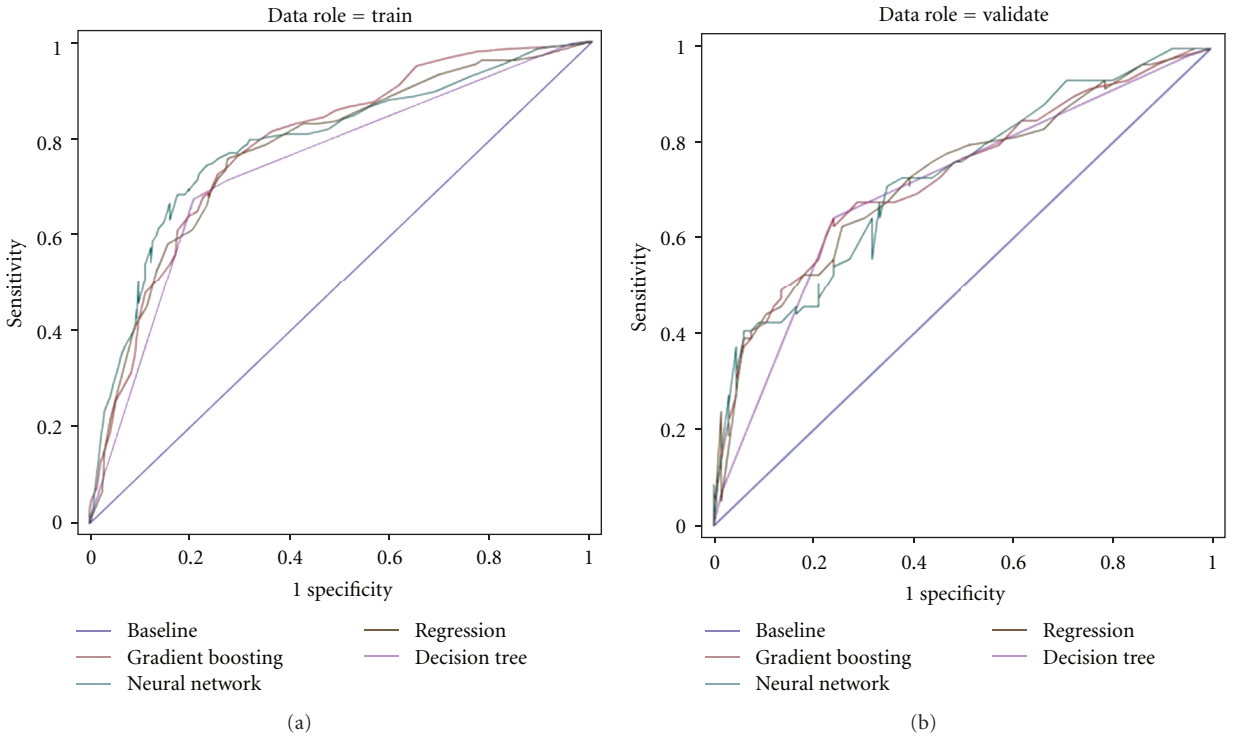

Figure 4: Receiver operator characteristic (ROC) curves for spatial models explaining occurrence of medically diagnosed cases of Rocky Mountain spotted fever for the 2000–09 study period within Tennessee.

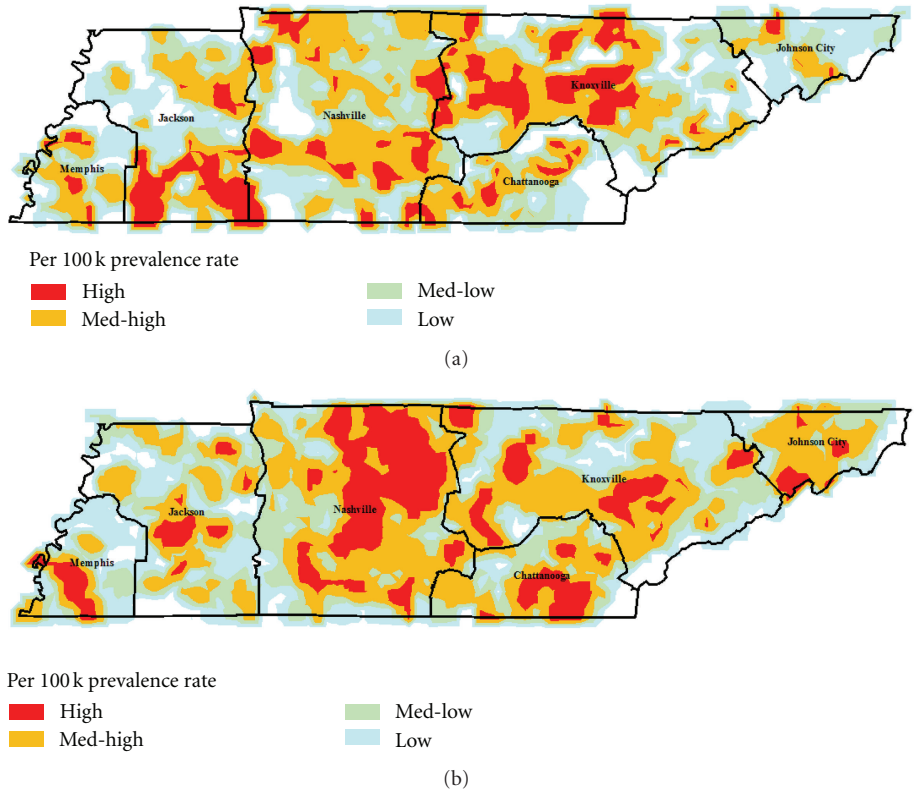

Figure 5: Delineated risk areas for Lyme disease according to raw disease prevalence per 100 k rates (a) and predicted probabilities from spatial predictive models (b).

review of literature related to LD risk [8], LD was consistently associated with forested areas. A probable explanation is that these land use types provide valuable habitat for host abundance [32–34]. A crude analysis between disease prevalence and forested wetland area suggest a positive correlation when forested wetlands account for up to 2.5% of the surrounding sample area. However, disease prevalence declines when the amount of forested wetlands is above this amount. Similarly, a positive correlation exists between disease prevalence and upland deciduous forests when this land use type accounts for up to 24% of the surrounding sample area. Above this amount, the relationship becomes negative. Others [7] reported that persons living in forested areas had elevated risk (OR: 3.7; 95% CI: 1.2–11.8) of LD exposure. This nonlinear relationship between disease prevalence within deciduous forests and nonforested wetlands may result from the complex vector-host interaction. For example, an area that is 100% forested may not be inhabited by humans and, therefore, reduces the possibility of disease transmission from vector to host. Consequently, an area that is 100% urbanized may eliminate vector habitat, thus removing all chances of a vector-host interaction. *Borrelia burgdorferi*, the causative agent of LD, may occur in urban and suburban development areas as well as in isolated park/forest preserves where deer, rodents, and birds can thrive [35]. Others [4] reported that *I. scapularis* were most abundant on sandy soils with deciduous forests.

The positive association between LD occurrence and median incomes may be more an artifact of the data source rather than an actual correlation. The data source is from persons with health insurance, both commercially insured and government subsidized programs for those who cannot afford coverage (i.e., Medicaid). Relatively wealthier persons have more access to care and tend to disproportionately utilize medical services compared to lower income persons [36, 37].

Covariates explaining RMSF prevalence were mostly similar to LD, and thus similar interpretation of results is assumed. However, one notable difference was RMSF that was significantly associated with the amount of pasture/grassland within all 4 models. The American Dog tick (*Dermacentor variabilis*) is the most commonly identified species responsible for transmitting the *Rickettsia rickettsii* bacterial organism that causes RMSF in humans. *D. variabilis* is considered an ixodid tick (hard-shell tick), and these are commonly found in grassland areas including pastures, old fields, clearings around homes, and brushy habitats [38, 39].

When evaluating either LD or RMSF, the cooccurrence of each other was significant throughout all 8 models. There are two, though possibly more, likely explanations for this relationship. As previously mentioned, significant explanatory covariates were similar for each disease. Therefore, it is plausible that suitable habitat features are overlapping for the tick vectors [39]. Another possible reason for this

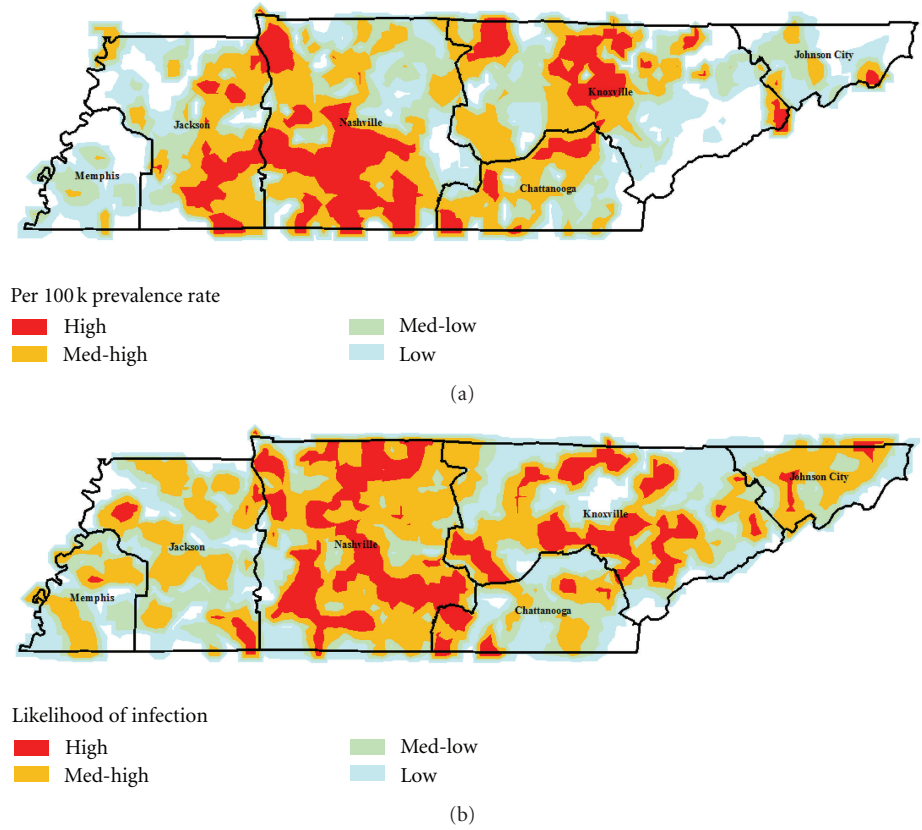

Figure 6: Delineated risk areas for RMSF according to disease prevalence per 100 k rates (a) and predicted probabilities from spatial predictive models (b).

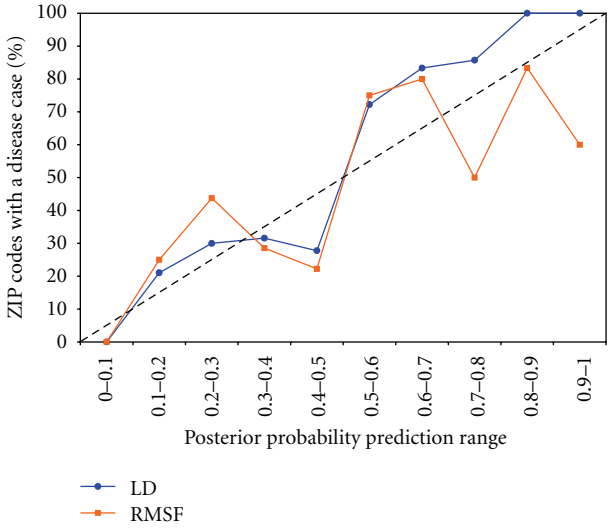

Figure 7: Performance of champion models as a function of the posterior probability predictions on the validation datasets.

interaction is that both diseases have similar clinical presentations; thus cases may be misdiagnosed between the two diseases [40]. In highly endemic areas within the US where awareness of RMSF is high, many patients receive an alternate diagnosis when initially seeking medical attention. Cases not laboratory confirmed are frequently not RMSF, and laboratory confirmation using weak diagnostic criteria may lead to false positives [41]. Because of the possibility of misdiagnoses, clinicians should receive confirmatory laboratory results prior to making a definitive clinical diagnosis.

Areas higher in disease prevalence were not necessarily the same areas having high predicted risk of disease infection. This supports our original project intent to illustrate the need to build spatially explicit models. Traditional risk maps can highlight temporally static areas where case volumes are high relative to other spatial units. This approach benefits from its simplicity; however, it lacks statistical validation and does not account for other influencing factors and is influenced by population.

Limitations in this present study include the inability to definitively confirm a diagnosed case of LD and/or RMSF as such. Although we tested for spatial autocorrelation using variograms and found none to be present, it is still possible that findings could be influenced by autocorrelation if measured differently. Land use and wetlands data do not necessarily reflect the same temporal period as the diagnosed disease case. The champion models for LD and RMSF were the GBT and NNET, respectively. Although they performed well, these modeling procedures do not produce directly interpretable results. Therefore, the ability to describe the

Table 3: Model summary statistics for spatially explicit models describing the occurrence of medically diagnosed cases of Lyme disease and Rocky Mountain spotted fever for the 2000–09 study period within Tennessee.

Model type	Lyme disease (LD)				Rocky mountain spotted fever (RMSF)			
	GBT	SLR	NNET	CART	GBT	SLR	NNET	CART
Model performance								
Misclassification rate	0.232*	0.272	0.288	0.296	0.304	0.312	0.288*	0.296
Average square error	0.187	0.182	0.253	0.206	0.230	0.210	0.232	0.213
ROC	0.789	0.812	0.688	0.674	0.702	0.727	0.696	0.712
PPV	83.7%	75.0%	77.1%	85.7%	69.8%	68.5%	72.5%	70.4%
Sensitivity	66.1%	67.7%	59.7%	48.4%	62.7%	62.7%	62.7%	64.4%
Specificity	87.3%	77.8%	82.5%	92.1%	75.8%	74.2%	78.8%	75.8%
Input variables**								
Land cover								
Forested wetland	+/−		−				+	+/−
Nonforested wetland	+/−						−	
Pasture/grassland					+	+/−	+/−	+
Upland deciduous forest	+/−	+/−			−			
Urban/developed	+	+	+		+		+/−	
Wetland type								
PUBHh	−							
Geographic								
Distance to river			+/−					
Demographic								
Population counts	+	+	+	+	+	+	+	+/−
Median income	+	+	+/−	+				
Clinical								
Lyme Dis. cooccurrence					+	+	+	+
RMSF cooccurrence	+	+	+	+				

* Best model chosen using lowest misclassification rate on validation dataset.
** Denotes that aggregations were made at 1.6 and 8 km where applicable.
Variables missing from this table indicate nonsignificance across all models, and plus and minus signs indicate direction of relationship.

quantitative impact of the covariates without deriving them from the SLR or CART results is limited.

5. Conclusions

Findings from this study suggest that administrative medical claims data is a viable source to study and map disease risk for LD and RMSF. Spatial models predicting disease risk are favorable to defining risk by mapping areas of high prevalence. Spatial factors associated with medically diagnosed cases of zoonoses agree with other literature using actual CDC reported cases. Little work exists using more advanced nonlinear modeling techniques like those used in this study, and it is recommended to explore these options as they may provide better results than traditional regression-based approaches. Administrative medical claims data is relatively easy to access given the appropriate permissions; relatively no cost once access is granted and provides the researcher with a volume rich dataset from which to study.

References

[1] M. C. Wimberly, A. D. Baer, and M. J. Yabsley, "Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens," *International Journal of Health Geographics*, vol. 7, p. 15, 2008.

[2] A. M. Winters, R. J. Eisen, S. Lozano-Fuentes, C. G. Moore, W. J. Pape, and L. Eisen, "Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado," *American Journal of Tropical Medicine and Hygiene*, vol. 79, no. 4, pp. 581–590, 2008.

[3] R. S. Lane and H. A. Stubbs, "Host-seeking behavior of adult Ixodes pacificus (Acari: *Ixodidae*) as determined by flagging vegetation," *Journal of Medical Entomology*, vol. 27, no. 3, pp. 282–287, 1990.

[4] U. Kitron, C. J. Jones, J. K. Bouseman, J. A. Nelson, and D. L. Baumgartner, "Spatial analysis of the distribution of Ixodes dammini (Acari: *Ixodidae*) on white-tailed deer in Ogle County, Illinois," *Journal of Medical Entomology*, vol. 29, no. 2, pp. 259–266, 1992.

[5] S. Aronoff, *Geographic Information Systems: A Management Perspective*, WDL Publications, Ottawa, Canada, 1989.

[6] R. J. Eisen, P. S. Mead, A. M. Meyer, L. E. Pfaff, K. K. Bradley, and L. Eisen, "Ecoepidemiology of tularemia in the South-central United States," *American Journal of Tropical Medicine and Hygiene*, vol. 78, no. 4, pp. 586–594, 2008.

[7] G. E. Glass, B. S. Schwartz, J. M. Morgan, D. T. Johnson, P. M. Noy, and E. Israel, "Environmental risk factors for Lyme disease identified with geographic information systems," *American Journal of Public Health*, vol. 85, no. 7, pp. 944–948, 1995.

[8] M. E. Killilea, A. Swei, R. S. Lane, C. J. Briggs, and R. S. Ostfeld, "Spatial dynamics of lyme disease: a review," *EcoHealth*, vol. 5, no. 2, pp. 167–195, 2008.

[9] H. Gaff and E. Schaefer, "Metapopulation models in tick-borne disease transmission modelling," *Advances in Experimental Medicine and Biology*, vol. 673, pp. 51–65, 2010.

[10] L. Eisen and R. J. Eisen, "Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases," *Emerging Infectious Diseases*, vol. 13, no. 12, pp. 1816–1820, 2007.

[11] R. Sugumaran, S. R. Larson, and J. P. DeGroote, "Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States," *International Journal of Health Geographics*, vol. 8, no. 1, p. 43, 2009.

[12] F. Mostashari, M. Kulldorff, J. J. Hartman, J. R. Miller, and V. Kulasekera, "Dead bird clusters as an early warning system for West Nile virus activity," *Emerging Infectious Diseases*, vol. 9, no. 6, pp. 641–646, 2003.

[13] R. J. Eisen, R. S. Lane, C. L. Fritz, and L. Eisen, "Spatial patterns of lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure," *American Journal of Tropical Medicine and Hygiene*, vol. 75, no. 4, pp. 669–676, 2006.

[14] S. G. Jones, W. Conner, B. Song, D. Gordon, and A. Jayakaran, "Comparing spatio-temporal clusters of arthropod-borne infections using administrative medical claims and state reported surveillance data," *Spatial and Spatio-Temporal Epidemiology*, vol. 3, no. 3, pp. 205–213, 2012.

[15] S. G. Jones and M. Kulldorff, "Influence of spatial resolution on space-time disease cluster detection," *PLoS ONE*. In press, http://dx.plos.org/10.1371/journal.pone.0048036. 2012.

[16] J. Wieczorek, Q. Guo, and R. J. Hijmans, "The point-radius method for georeferencing locality descriptions and calculating associated uncertainty," *International Journal of Geographical Information Science*, vol. 18, no. 8, pp. 745–767, 2004.

[17] J. A. Bissonette, "Small sample size problems in wildlife ecology: a contingent analytical approach," *Wildlife Biology*, vol. 5, no. 2, pp. 65–71, 1999.

[18] S. G. Jones, S. Coulter, and W. Conner, "Using administrative medical claims data to supplement state disease registry systems for reporting zoonotic infections," *Journal of American Medical Informatics Association*. In press.

[19] Tennessee Wildlife Resources Agency, Tennessee Land Use/Land Cover Landsat TM imagery, Tennessee Spatial Data Service metadata files, http://www.tngis.org/frequently_accessed_data.html, 1997.

[20] L. Cowardin, V. Carter, E. Golet, and E. LaRoe, "Classification of wetlands and deepwater habitats of the United States," US Fish and Wildlife Service FWS/OBS 79/31, 1979.

[21] M. Efroymson, "Multiple regression analysis," in *Mathematical Methods for Digital Computers*, A. Ralston and H. S. Wilf, Eds., chapter 17, Wiley, New York, NY, USA, 1960.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, Calif, USA, 1984.

[23] B. De Ville, *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*, SAS Publishing, Cary, NC, USA, 2006.

[24] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[25] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[26] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[27] A. Lapedes and R. Farber, "Nonlinear signal processing using neural networks: prediction and system modeling," Tech. Rep. LA-UR87-2662, Los Alamos National Laboratory, Los Alamos, NM, USA, 1987.

[28] SAS Institute Inc, *SAS Enterprise Miner 6.1: Single-User Installation Guide*, SAS Institute Inc, Cary, NC, USA, 2009.

[29] R. Wall and P. Cunningham, "Exploring the potential for rule extraction from ensembles of neural networks," in *Proceedings of the 11th Irish Conference on Artificial Intelligence and Cognitive Science*, J. Griffith and C. O'Riordan, Eds., Computer Science Technical Report TCD-CS-2000-24, pp. 52–68, Trinity College, Dublin, Ireland, 2000.

[30] A. C. Steere, S. E. Malawista, D. R. Snydman et al., "Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three connecticut communities.," *Arthritis and Rheumatism*, vol. 20, no. 1, pp. 7–17, 1977.

[31] G. O. Maupin, D. Fish, J. Zultowsky, E. G. Campos, and J. Piesman, "Landscape ecology of Lyme disease in a residential area of Westchester County, New York," *American Journal of Epidemiology*, vol. 133, no. 11, pp. 1105–1113, 1991.

[32] R. G. McLean, S. R. Ubico, C. A. N. Hughes, S. M. Engstrom, and R. C. Johnson, "Isolation and characterization of *Borrelia burgdorferi* from blood of a bird captured in the Saint Croix River Valley," *Journal of Clinical Microbiology*, vol. 31, no. 8, pp. 2038–2043, 1993.

[33] H. S. Ginsberg, P. A. Buckley, M. G. Balmforth, E. Zhioua, S. Mitra, and F. G. Buckley, "Reservoir competence of native North American birds for the lyme disease spirochete, *Borrelia burgdorferi*," *Journal of Medical Entomology*, vol. 42, no. 3, pp. 445–449, 2005.

[34] N. H. Ogden, R. L. Lindsay, K. Hanincová et al., "Role of migratory birds in introduction and range expansion of *Ixodes scapularis* ticks and of *Borrelia burgdorferi* and *Anaplasma phagocytophilum* in Canada," *Applied and Environmental Microbiology*, vol. 74, no. 12, pp. 3919–3919, 2008.

[35] L. A. Magnarelli, A. Denicola, K. C. Stafford, and J. F. Anderson, "*Borrelia burgdorferi* in an urban environment: white-tailed deer with infected ticks and antibodies," *Journal of Clinical Microbiology*, vol. 33, no. 3, pp. 541–544, 1995.

[36] R. G. Wilkinson and K. E. Pickett, "Income inequality and population health: a review and explanation of the evidence," *Social Science and Medicine*, vol. 62, no. 7, pp. 1768–1784, 2006.

[37] A. Lusardi, D. Schneider, and P. Tufano, "The economic crisis and medical care usage. Harvard business school," Working Paper 10-079, 2010.

[38] Q. H. Liu, G. Y. Chen, Y. Jin et al., "Evidence for a high prevalence of spotted fever group rickettsial infections in diverse ecologic zones of Inner Mongolia," *Epidemiology and Infection*, vol. 115, no. 1, pp. 177–183, 1995.

[39] P. Parola and D. Raoult, "Ticks and tickborne bacterial diseases in humans: an emerging infectious threat," *Clinical Infectious Diseases*, vol. 32, no. 6, pp. 897–928, 2001.

[40] E. J. Masters, G. S. Olson, S. J. Weiner, and C. D. Paddock, "Rocky Mountain spotted fever: a clinician's dilemma," *Archives of Internal Medicine*, vol. 163, no. 7, pp. 769–774, 2003.

[41] C. G. Helmick, K. W. Bernard, and L. J. D'Angelo, "Rocky Mountain spotted fever: clinical, laboratory, and epidemiological features of 262 cases," *Journal of Infectious Diseases*, vol. 150, no. 4, pp. 480–488, 1984.