Biostatistics (2018), **0**, 0, *pp.* 1–14
doi:10.1093/biostatistics/samplebibtex

Application of generalized additive models in Lyme disease epidemiology

Yang Yang

Department of Statistics, Kansas State University, Manhattan, KS, US

sheepyang@ksu.edu

Summary

Lyme disease is the most frequently diagnosed vector-borne disease and the sixth most commonly reported notifiable infectious disease in the United States. The incidence of Lyme disease in human is caused by the bite of ticks infected by spirochete bacteria named $Borrelia\ burgdorferi$. Over the last two decades, the geographic distribution of high incidence areas appears to be expanding rapidly, which highlights the urgent need for surveillance and prevention. In this study, we propose three generalized additive models (GAM) for Lyme disease epidemic data. Both interpolation and extrapolation performances of proposed models are evaluated and compared by mean absolute errors (MAE), R-square, percent of deviance explained, Akaike information criterion (AIC) and graphical method. Estimates by a robust statistical model could act as a tool to guide surveillance, control, and prevention efforts and act as a baseline for studies tracking the spread of infection.

Key words: Lyme disease; Generalized Additive Models; Akaike information criterion (AIC).

[*] To whom correspondence should be addressed.

1. Introduction

Lyme disease is the most commonly reported zoonotic tick-borne diseases worldwide. In the United States, infection of Lyme disease is mainly caused by spirochete bacteria, *Borrelia burgdorferi* (Burgdorfer *and others* (1982)). The pathogen is transmitted to humans mostly by the bite of a tiny tick named *Ixodes scapularis*, which is commonly called the black-legged or deer tick.

Clinical manifestations of Lyme disease are divided into three stages (Steere *and others* (1986)). Typically, the illness begins with the characteristic skin lesion, erythema chronicum migrans (ECM), which are bulls-eye-shaped rash with central clearing and slowly expands. Together with the featured rash, patients may also experience associated flu-like symptoms, such as fever, headache, fatigue, muscle and joint aches, and swollen lymph nodes (Steere *and others* (1983)). A few weeks to months later, untreated individuals may develop manifestations involving nervous system, heart, and joints (Pachner and Steere (1985)). In late stage, months to years after disease onset, the arthritis as well as neurological and cardiac symptoms increase and become intermittent or chronic (Logigian *and others* (1990)). With prompt and appropriate antibiotic treatment, most cases in early stage of Lyme disease recover rapidly and completely. In its later stage, some severe cases can be difficult to treat and may require intravenous treatment. Even after appropriate antibiotic treatment, patients may still have persistent or recurrent symptoms called post-treatment Lyme disease syndrome (PTLDS) and need long-term antibiotic treatment (Rahn (1991)). Figure 1 shows the classic symptoms of Lyme disease.

Like most communicable diseases, Lyme disease is largely preventable. Tick control including burning or removing vegetation, especially acaricide application to wildlife, reduces ixodid vector populations by up to 94% (Poland (2001)). Avoiding exposure to ticks by personal precautions is the best defense against Lyme disease. Effective protection includes wearing protective clothing with permethrin, avoiding contact with ticks, and using insect repellent. However, a Lyme disease

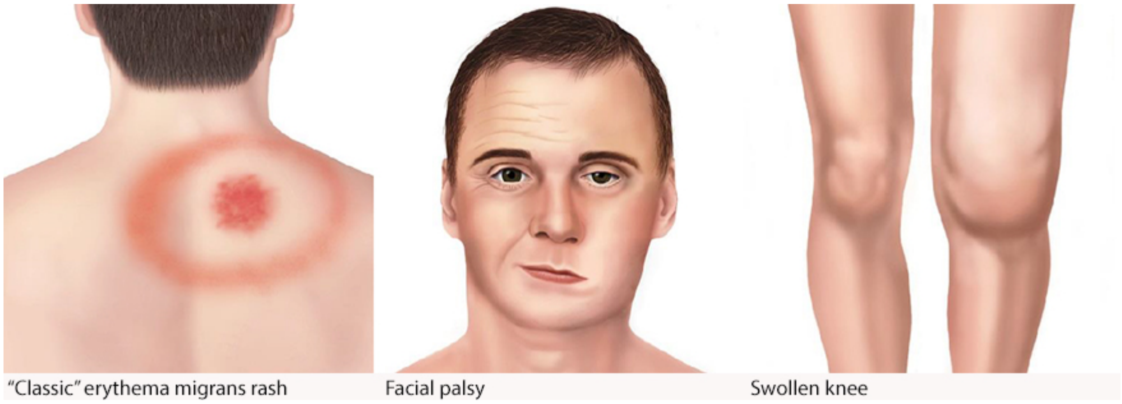


Fig. 1: Lyme disease symptoms. This figure is obtained from CDC website.[†]

vaccine is no longer available, and it is unlikely that any will be developed and licensed in the near future (Aronowitz (2012)).

The name of Lyme disease comes from the first occurrence in the United States in 1975 at Lyme, Connecticut. Over the last two decades, its incidence and geographic expansion have been rapidly increasing making it the most frequently reported vector-borne disease. According to the U.S. Centers for Disease Control and Prevention (CDC), there are 42,743 confirmed and probable cases of Lyme disease reported in 2017, about 9% more than in 2016. Although more than 30,000 cases are reported each year, estimates using other methods suggest that the actual number of diagnosed cases may be ten times, approximately 300,000, as many cases go unreported. Ninety-five percent of all confirmed Lyme disease cases stemmed from high-risk areas in the northeast, midwest, and the west coast, including 14 states. However, the number of counties with an incidence which is greater than 10 per 100,000 persons, increased from 324 in 2008 to 454 in 2017. Also, the two tick species known to carry the pathogen of human Lyme disease were reported to have spread into half of U.S. counties. The geographic distribution changes of Lyme disease shown in Figure 2, also indicate the potential explanation of the pathogen inhabits.

Since Lyme disease is a tremendous public health problem in the United States, which highlights the urgent need for effective surveillance and prevention. Improvement of control measures

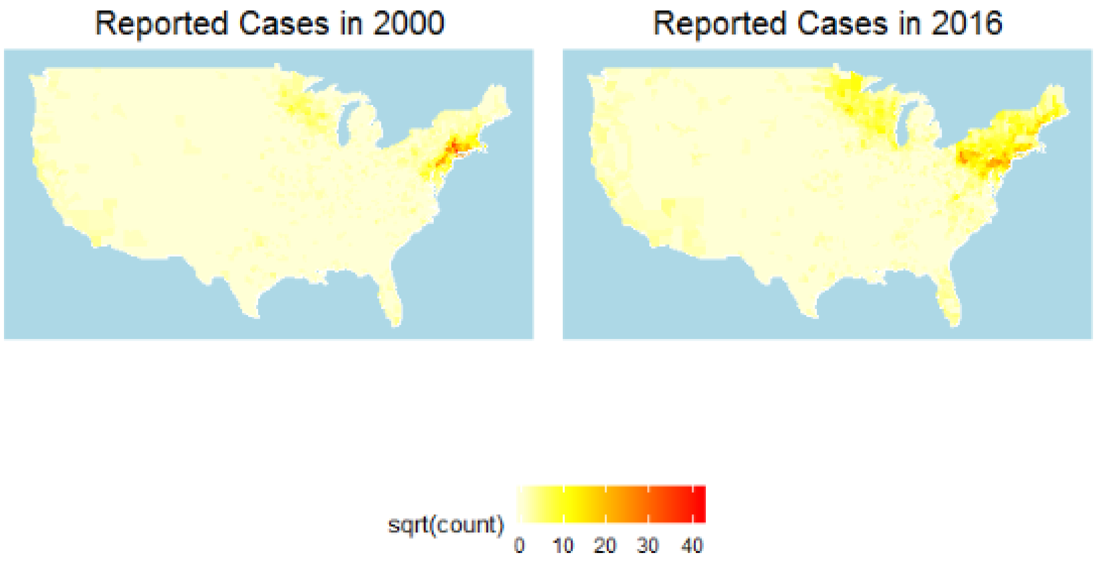


Fig. 2: Geographic expansion of confirmed cases from 2000 (left) to 2016 (right). Square root of case count is adopted to amplify the differences.

requires geographic distribution of the incidence of Lyme disease to follow disease trends and raise public awareness. In this study, we propose three generalized additive models for Lyme disease epidemic data. Both interpolation and extrapolation performances of proposed models are evaluated and compared by mean absolute errors (MAE), R-square, percent of deviance explained, Akaike information criterion (AIC) and graphical method. Estimates by a robust statistical model would facilitate Lyme disease surveillance and prevention by identifying risk factors, predicting spread trend and shaping public policy.

2. Methods

2.1 *Data description*

In this study, data is obtained from the surveillance report of Lyme disease provided by CDC. This case report includes case counts in all 3,193 counties in the U.S. from 2010 to 2016. After rebuilding the data, reported case counts from 3,075 counties during 2010 to 2016 are used for further analysis. This yields a total of $n = 52,428$ observations with corresponding spatial location and time.

2.2 *Generalized additive model*

Three generalized additive models with spatial and temporal random effect are developed and described.

The data model from the first generalized additive model can be written as

$$\mathbf{z} \sim NB(\mu, \phi) \tag{2.1}$$

$$\log(\mu) = \beta_0 + \alpha_s + \alpha_t \tag{2.2}$$

where $\mathbf{z}$ is a set of n observations at fixed spatial locations and time, which follows a negative binomial distribution with parameters μ and ϕ. The parameters μ is related to intercept β_0, spatial random effect α_s and temporal random effect α_t with a link function.

The corresponding process model is written as

$$\alpha|\sigma^2_\alpha, \phi \sim N(0, \sigma^2_\alpha C(\phi)) \tag{2.3}$$

α, referred to α_s or α_t, is a spatial or temporal random effect which follows a zero-mean Gaussian distribution with variance σ^2_α and correlation matrix $C(\phi)$ with parameter ϕ. For element of i^{th} row and j^{th} column in matrix $C(\phi)$,

$$C_{ij} = (1 + d_{ij}/\phi + \frac{2}{5}(d_{ij}/\phi)^2 + \frac{1}{15}(d_{ij}/\phi)^3)e^{-d_{ij}/\phi} \tag{2.4}$$

where d_{ij} is the spatial distance or temporal difference between i^{th} and j^{th} observations.

In the second generalized additive model, the data model is shown below

$$\mathbf{z} \sim Poisson(\lambda) \tag{2.5}$$

$$\lambda = e^{\beta_0 + \alpha_s + \alpha_t} \tag{2.6}$$

where $\mathbf{z}$ is assumed to follow a Poisson distribution with parameter λ which is related to intercept β_0, spatial random effect α_s and temporal random effect α_t with a link function.

The process model is the same as the first model in 2.3 and 2.4.

For the third data model, $\mathbf{z}$ follows a zero-inflated Poisson distribution with parameter λ which also link to a linear combination of predictors, the same as in 2.6. The data model is described as,

$$\mathbf{z} = \begin{cases} Poisson(\lambda) & y = 1 \\ 0 & y = 0 \end{cases} \tag{2.7}$$

Where $\mathbf{y}$ follow a Bernoulli distribution with parameter p,

$$\mathbf{y} \sim Bernoulli(p) \tag{2.8}$$

For spatial random effect α_s and temporal random effect α_t, the process model is the same as the first two models in 2.3, but the element in correlation matrix $C(\phi)$ changes to,

$$C_{ij} = 1 - 1.5 d_{ij}/\phi + o.5(d_{ij}/\phi)^3 \; if \; d_{ij} < \phi, \; otherwise \; 0 \tag{2.9}$$

d_{ij} refers to the spatial distance or temporal difference between i^{th} and j^{th} observations.

2.3 *Experimental set-up and model evaluation*

Model fitting is conducted and evaluated in five experimental set-ups.

i)The data are randomly split into training data and test data with percentage of 81% and 19% respectively.

ii)To assess the robustness of model in interpolation, observations in 2004 are removed and used as test set. The model is then trained with the rest data. The expected value generated by model prediction is compared with real data in 2004.

iii)Model performance of extrapolation is evaluated by training without using observations in 2016. The predicted expected value is compared with reported counts in 2016.

iv)Extrapolation ability is also assessed by forecast for 2017 by model trained with total data set. The result is compared with density and distribution of reported counts in 2016.

v)Similarly, in areas with high incidence, the local forecast for 2017 is conducted and evaluated as countrywide.

In each experimental set-up with three models, mean absolute errors (MAE) and R-square values are employed to assess the goodness-of-fit of the models. Model fitting is implied by gam function in R package mgcv.

3. Results

To examine the accuracy of the proposed models for prediction and forecast, five experimental set-ups are tested by examining the MAE, R^2 and percent of deviance explained. The results are shown in Table 1.

Experimental set-ups are designed for performance on interpolation and extrapolation. Interpolation performances, which are the first two experiment set-ups includes predict based on independent training and test set(Random) and prediction for 2004. While extrapolation is tested based on prediction for 2016, as well as 2017 in aspects of the country and high incidence areas where data is not available to estimate MAE. For the zero-inflated Poisson model, a pseudo-R^2 is given in term of a percent of deviance explained.

Experimental set-up	NB		Poisson		0-inflated Pois	
	MAE	R^2	MAE	R^2	MAE	Deviance explained
Random	7.283	0.376	7.028	0.460	7.079	65.2 %
Predict 2004	5.518	0.392	5.903	0.465	5.912	65.4 %
Forecast 2016	10.764	0.409	10.293	0.470	8.72	66.1 %
Forecast 2017	NA	0.389	NA	0.463	NA	65.3 %
Local Forecast 2017	NA	0.165	NA	0.459	NA	63.1 %

Table 1: MAE, R^2 values and percent of deviance explained of GAMs with data model from a negative binomial distribution(NB), Poisson distribution(Poisson), zero-inflated Poisson distribution(0-inflated Pois). MAE values of the countrywide or local forecast for 2017 are denoted as NA due to data not available for 2017. Results are generated based on random training and test dataset(Random), prediction for 2004(Predict 2004), prediction for 2016(Forecast 2016), prediction for 2017 in the U.S.(Forecast 2017) and high incidence areas(Local Forecast 2017).

From the results in Table 1, three proposed model performed relatively similarly in each experimental set-up. The interpolation performances of three models are better compared with those of extrapolation, which is reasonable.

To select the model with the best fit, the Akaike information criterion (AIC) is adopted to estimate the relative quality of the proposed models. The AIC scores are shown in Table 2. Compared with the other two models, the negative binomial model with lowest AIC value in all experimental set-ups turn to be the best model on Lyme disease data.

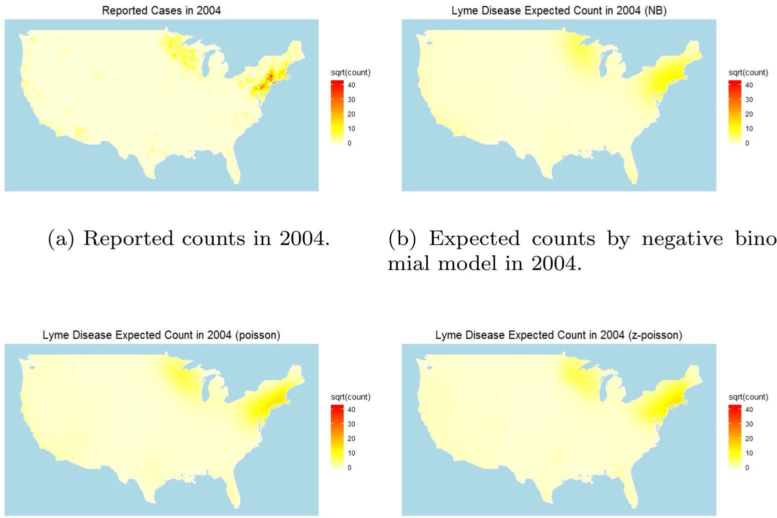
AIC	NB	Poisson	0-inflated Pois
Random	103,991	488,844	458,427
Predict 2004	121,750	564,487	531,287
Forecast 2016	118,230	538,053	509,876
Forecast 2017	128,017	591,751	557,595
Local Forecast 2017	90,812	503,387	475,633

Table 2: The Akaike information criterion (AIC) results from the data model using a negative binomial distribution(NB), Poisson distribution(Poisson), zero-inflated Poisson distribution(0-inflated Pois). Results are generated based on random training and test dataset(Random), prediction for 2004(Predict 2004), prediction for 2016(Forecast 2016), prediction for 2017 in the U.S.(Forecast 2017) and high incidence areas(Local Forecast 2017).

To compare the prediction with the real data, expected values predicted from the three proposed models, together with the reported counts in 2004 and 2016 are displayed in Figure 3 and

Figure 4.

In general, the distributions of expected counts predicted by the negative binomial model (Figure 3b and Figure 4b), the Poisson model (Figure 3c and Figure 4c) and the zero-inflated Poisson model (Figure 3d and Figure 4d) all appear to be similar to the real data (Figure 3a and Figure 4a).


Reported Cases in 2004
Lyme Disease Expected Count in 2004 (NB)
Lyme Disease Expected Count in 2004 (poisson)
Lyme Disease Expected Count in 2004 (z-poisson)
sqrt(count)
40
30
20
10
0

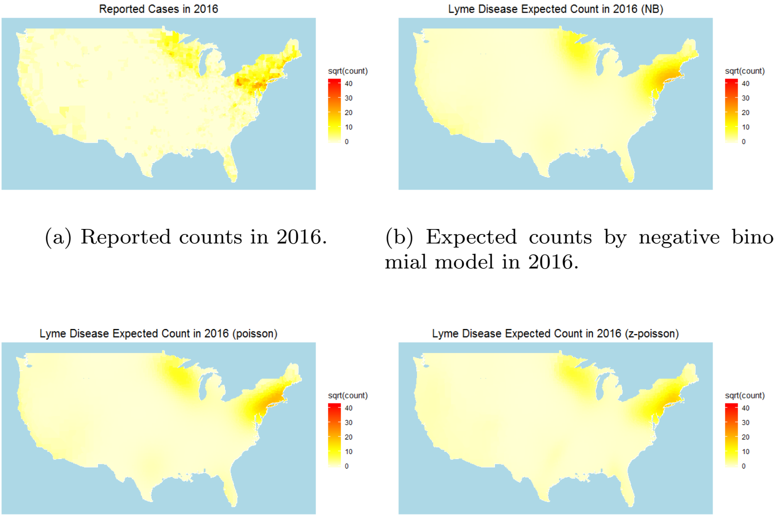
(a) Reported counts in 2004.

(b) Expected counts by negative binomial model in 2004.

(c) Expected counts by Poisson model in 2004.

(d) Expected counts by zero-inflated Poisson model in 2004.

Fig. 3: Reported and expected counts in 2004 in U.S.. Square root of count is adopted to amplify the differences.


Reported Cases in 2016
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2016 (NB)
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2016 (poisson)
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2016 (z-poisson)
sqrt(count)
40
30
20
10
0

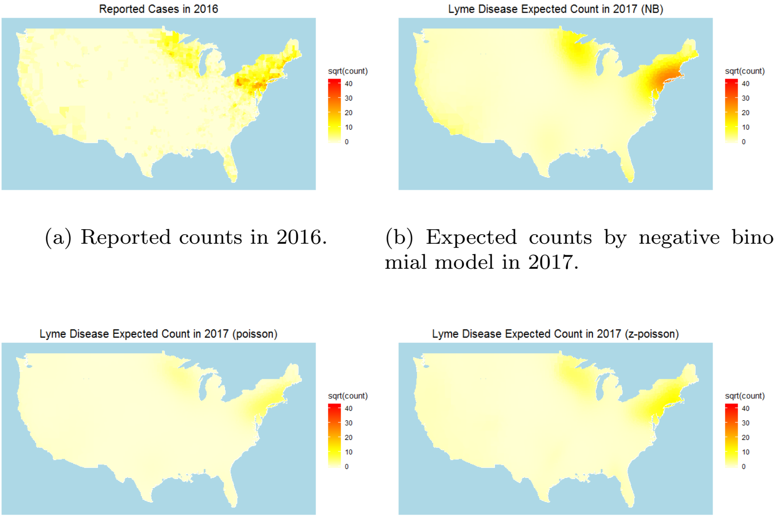
(a) Reported counts in 2016.

(b) Expected counts by negative binomial model in 2016.

(c) Expected counts by Poisson model in 2016.

(d) Expected counts by zero-inflated Poisson model in 2016.

Fig. 4: Reported and expected counts in 2016 in U.S.. Square root of count is adopted to amplify the differences.


Reported Cases in 2016
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2017 (NB)
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2017 (poisson)
sqrt(count)
40
30
20
10
0
Lyme Disease Expected Count in 2017 (z-poisson)
sqrt(count)
40
30
20
10
0

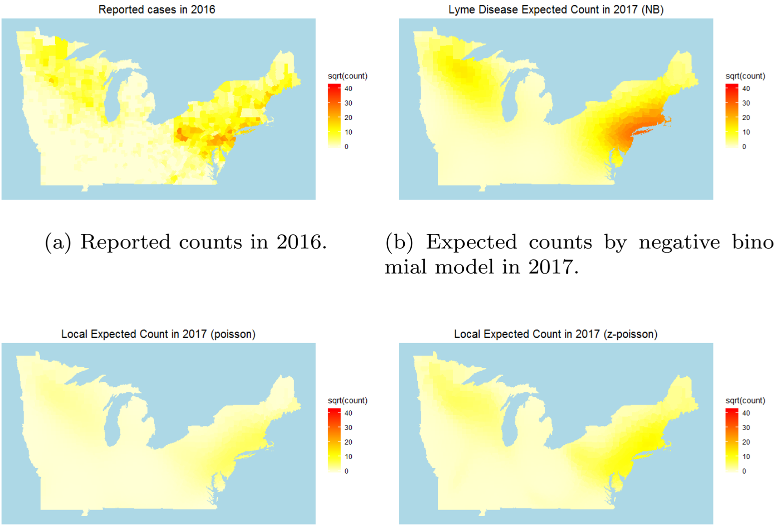
(a) Reported counts in 2016.

(b) Expected counts by negative binomial model in 2017.

(c) Expected counts by Poisson model in 2017.

(d) Expected counts by zero-inflated Poisson model in 2017.

Fig. 5: Reported counts in 2016 and expected counts in 2017 in U.S.. Square root of count is adopted to amplify the differences.


Reported cases in 2016

Lyme Disease Expected Count in 2017 (NB)

Local Expected Count in 2017 (poisson)

Local Expected Count in 2017 (z-poisson)

(a) Reported counts in 2016.

(b) Expected counts by negative binomial model in 2017.

(c) Expected counts by Poisson model in 2017.

(d) Expected counts by zero-inflated Poisson model in 2017.

Fig. 6: Reported counts in 2016 and expected counts in 2017 in high incidence areas. Square root of count is adopted to amplify the differences.

Since we don't have the access to the case report of 2017, the forecasts of 2017 are unable to compare with real data. However, according to the CDC statement that confirmed and probable cases reported in 2017 are about 9% more than in 2016, as well as the Geographic expansion during 2000-2016, we could get some clues about the distribution in 2017. So comparison between reported counts in 2016 and expected counts in 2017 are displayed in aspect of countrywide (Figure 5) and high incidence areas(Figure 6). Forecast generated from negative binomial model seems more likely to mimic the potential distribution of 2017 than the other two models, which is also consistent with the AIC result.

4. Discussion

In this study, we explore three generalized additive models that aim to provide information for Lyme disease surveillance and prevention. The three generalized additive models mainly differ on data models, which are from negative binomial, Poisson, zero-inflated Poisson distribution. In most count data analysis, Poisson data model is commonly used and zero-inflated Poisson data model is appropriate for data with an excess of zero counts. However, when dealing with overdispersed data, indicating much greater variance, the assumption of applying Poisson distribution is not satisfied (Zuur *and others* (2009)). An alternative solution is to introduce a Poisson mixture model, such as mixed with gamma distribution in our case, which turns out to be a negative binomial distribution. This accounts for the better performance of negative binomial data model on Lyme disease data which has overdispersion feature.

Since the generalized additive model belongs to descriptive models, the interpolation is reliable to some degree, but the extrapolation is not ideal with wider credible interval due to the absence of information. A better proposal is to develop a mechanistic spatial and temporal model which could be implied by Markov chain Monte Carlo (MCMC) algorithm. Moreover, combinations with additional predictors would provide more information for model fitting. In case of Lyme disease, grass, brush, and shrubs are the main inhabit for ticks (Estrada-Peña *and others* (2018)), indicating the landscape feature could contribute to spread of pathogens and incorporation with geospatial modeling techniques, such as geographic information systems (GIS), would promote studies of epidemiology (Nicholson and Mather (2014)). Ostfeld *and others* (2006) have reported that risk of exposure to Lyme disease is correlated positively with prior abundance of key hosts and with critical food resources for those hosts by using regression models. Also, quantification of the tick-human interactions, including high-risk outdoor activities and potentially indirect transmission by pets, could also contribute to better model fitting (LoGiudice *and others* (2003)). So the improvement of the statistical model for Lyme disease data is still in need and will benefit the study in fields of epidemiology and public health.

References

Aronowitz, Robert A. (2012). The rise and fall of the lyme disease vaccines: a cautionary tale for risk interventions in american medicine and public health. *The Milbank Quarterly* **90**(2), 250–277.

Burgdorfer, Willy, Barbour, Alan G, Hayes, Stanley F, Benach, Jorge L, Grunwaldt, Edgar and Davis, Jeffrey P. (1982). Lyme disease-a tick-borne spirochetosis? *Science* **216**(4552), 1317–1319.

Estrada-Peña, Agustín, Cutler, Sally, Potkonjak, Aleksandar, Vassier-Tussaut, Muriel, Van Bortel, Wim, Zeller, Hervé, Fernández-Ruiz, Natalia and Mihalca, Andrei Daniel. (2018). An updated meta-analysis of the distribution and prevalence of borrelia burgdorferi sl in ticks in europe. *International Journal of Health Geographics* **17**(1), 41.

Logigian, Eric L, Kaplan, Richard F and Steere, Allen C. (1990). Chronic neurologic manifestations of lyme disease. *New England Journal of Medicine* **323**(21), 1438–1444.

LoGiudice, Kathleen, Ostfeld, Richard S, Schmidt, Kenneth A and Keesing, Felicia. (2003). The ecology of infectious disease: effects of host diversity and community composition on lyme disease risk. *Proceedings of the National Academy of Sciences* **100**(2), 567–571.

Nicholson, Matthew C and Mather, Thomas N. (2014). Methods for evaluating lyme disease risks using geographic information systems and geospatial analysis. *Journal of Medical Entomology* **33**(5), 711–720.

Ostfeld, Richard S, Canham, Charles D, Oggenfuss, Kelly, Winchcombe, Raymond J and Keesing, Felicia. (2006). Climate, deer, rodents, and acorns as determinants of variation in lyme-disease risk. *PLoS biology* **4**(6), e145.

PACHNER, ANDREW R AND STEERE, ALLEN C. (1985). The triad of neurologic manifestations of lyme disease meningitis, cranial neuritis, and radiculoneuritis. *Neurology* **35**(1), 47–47.

POLAND, GREGORY A. (2001). Prevention of lyme disease: a review of the evidence. In: *Mayo Clinic Proceedings*, Volume 76. Elsevier. pp. 713–724.

RAHN, DANIEL W. (1991). Lyme disease: clinical manifestations, diagnosis, and treatment. In: *Seminars in arthritis and rheumatism*, Volume 20. Elsevier. pp. 201–218.

STEERE, ALLEN C, BARTENHAGEN, NICHOLAS H, CRAFT, JOSEPH E, HUTCHINSON, GORDON J, NEWMAN, JAMES H, PACHNER, ANDREW R, RAHN, DANIEL W, SIGAL, LEONARD H, TAYLOR, ELISE AND MALAWISTA, STEPHEN E. (1986). Clinical manifestations of lyme disease. *Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene. Series A: Medical Microbiology, Infectious Diseases, Virology, Parasitology* **263**(1-2), 201–205.

STEERE, ALLEN C, BARTENHAGEN, NICHOLAS H, CRAFT, JOSEPH E, HUTCHINSON, GORDON J, NEWMAN, JAMES H, RAHN, DANIEL W, SIGAL, LEONARD H, SPIELER, PHYLLIS N, STENN, KURT S AND MALAWISTA, STEPHEN E. (1983). The early clinical manifestations of lyme disease. *Annals of internal medicine* **99**(1), 76–82.

ZUUR, ALAIN F, IENO, ELENA N, WALKER, NEIL J, SAVELIEV, ANATOLY A AND SMITH, GRAHAM M. (2009). Glm and gam for count data. In: *Mixed effects models and extensions in ecology with R*. Springer, pp. 209–243.

[*Received December 1, 2018; revised October 1, 2010; accepted for publication December 10, 2018*]