A Dynamical Mechanistic Spatio-Temporal Framework for Modeling the Spread of Infectious Diseases

Ali Arab

Associate Professor
Department of Mathematics & Statistics
Georgetown University

August 6, 2023
JSM 2023, Toronto

Outline

- (Re)Emerging Epidemics
- Modeling Challenges & Approaches
- Modeling the Dynamics: Focus on Early Stages of Disease Spread
- Motivating Problem: Spread of Lyme in Virginia
- Bayesian Hierarchical Modeling Framework
- Choices for Data & Process Models
- Drivers of Disease Spread Dynamics
- Case Studies: Lyme & COVID-19 Cases in Virginia
- Mechanistic Spatio-Temporal Models

(Re)Emerging Epidemics

One of the main challenges of modeling emerging epidemics is data sparsity (low number of cases over space and time) and of course, detectability issues. Thus, the modeling issues are similar to modeling **Rare Events** in many regards:

- Events that are low frequency, low probability, or "unexpected"
- Relative rareness over time and space.
- Spatio-temporal dynamics of rare events may be quite complicated depending on the context of the problem.

Modeling Challenges & Approaches

Some of the challenges of these type of data include:

- Excess zeroes
- Small sample size of non-zero values (and typically spatially clustered)
- Large range of values: many zeroes, few non-zeroes, sometimes very few very large values (heavy tailed)
- complex dynamics over space and/or time (e.g., Lyme disease is expanding, becoming more common in areas that it used to be rare).

In addition to excess zeroes and possibly heavy tails, in order to model the dynamics of emerging epidemics, we need to understand and address spatial and temporal structure in the data.

Model Choice

Common choices for modeling these data include zero-modified models (See Arab 2015 for a discussion) including zero-inflated and hurdle models (i.e., mixture of a zero generating process and a count model), For example, a hurdle Poisson model

$$p_{i,t} I_{(y_i=0)} + (1 - p_{i,t}) \text{Poi}(\lambda_{i,t} > 0),$$

or a negative binomial hurdle model:

$$p_{i,t} I_{(y_i=0)} + (1 - p_{i,t}) \text{NegBin}(n, q_{i,t} > 0).$$

Model Choice

Other (less common) choices include models for heavy-tailed data, and models that address both heavy tails and excess zeroes such as a double hurdle model (Balderama et al. 2016).

These models are in particular preferred when prediction is of interest.

Modeling the Dynamics

Dynamics of emerging epidemics over space and time are complex and are often modeled using autoregressive models which may have limited capacity to mimic the true dynamics.

Alternative approaches include physical-statistical models such as PDE-based models (Wikle 2003), or agent based Models (Hooten and Wikle 2010).

Here, I will look at drivers of dynamics in two different cases (Lyme disease and COVID-19 in Virginia) to motivate modeling early stages of epidemics.

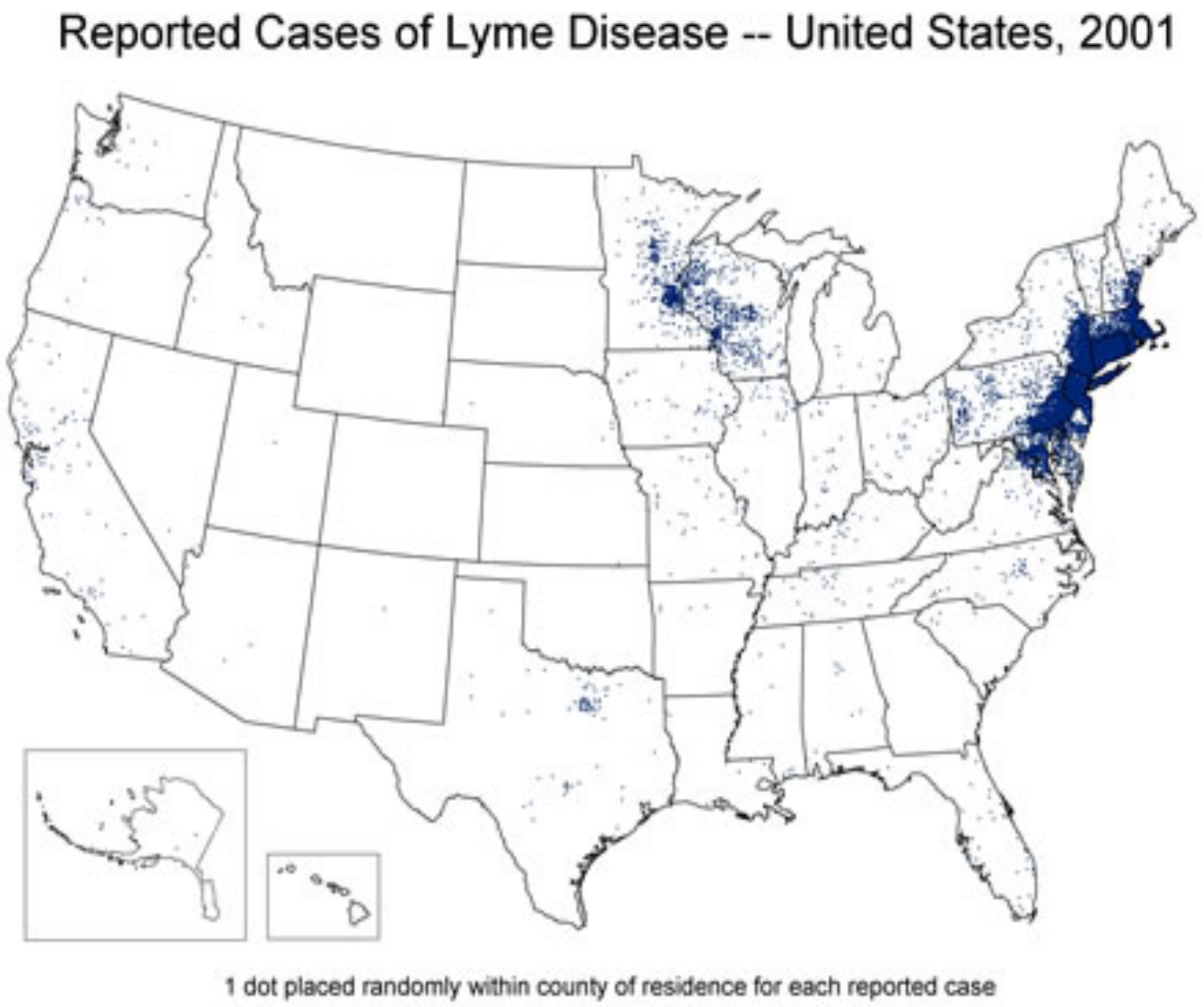
Motivating Problem: Lyme Disease

First identified in Lyme, Connecticut, Lyme disease is a bacterial illness that can cause fever, fatigue, joint pain, and skin rash, as well as more serious joint and nervous system complications. It is the most common vector-borne disease in the United States.

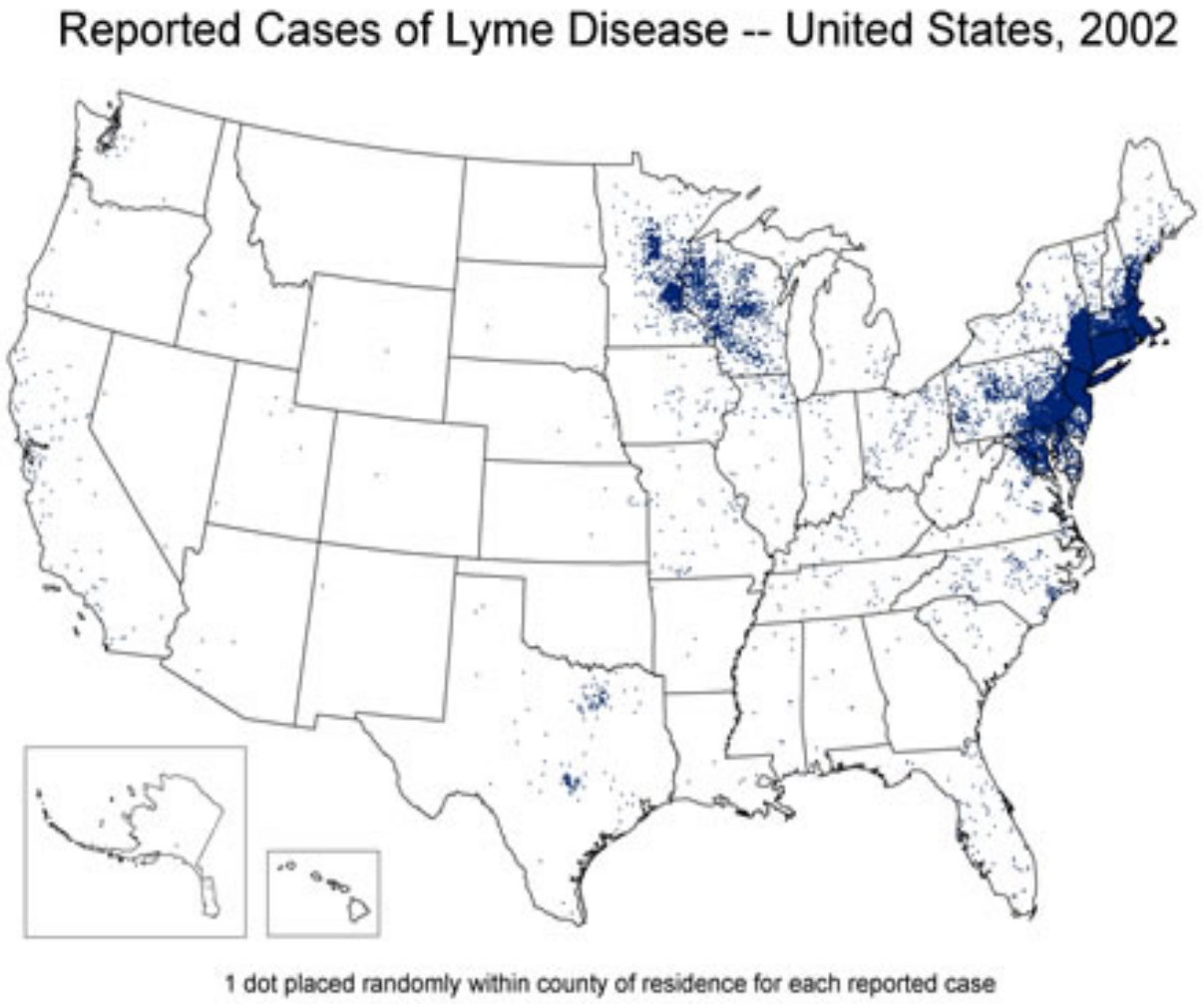
Lyme is common in parts of the upper East Coast with a high incidence rate but uncommon in other areas (e.g., almost non-existent in Arkansas).

“The incidence of Lyme disease in the United States has approximately doubled since 1991, from 3.74 reported cases per 100,000 people to 7.95 reported cases per 100,000 people in 2014.” (EPA site on Climate Change Indicators)

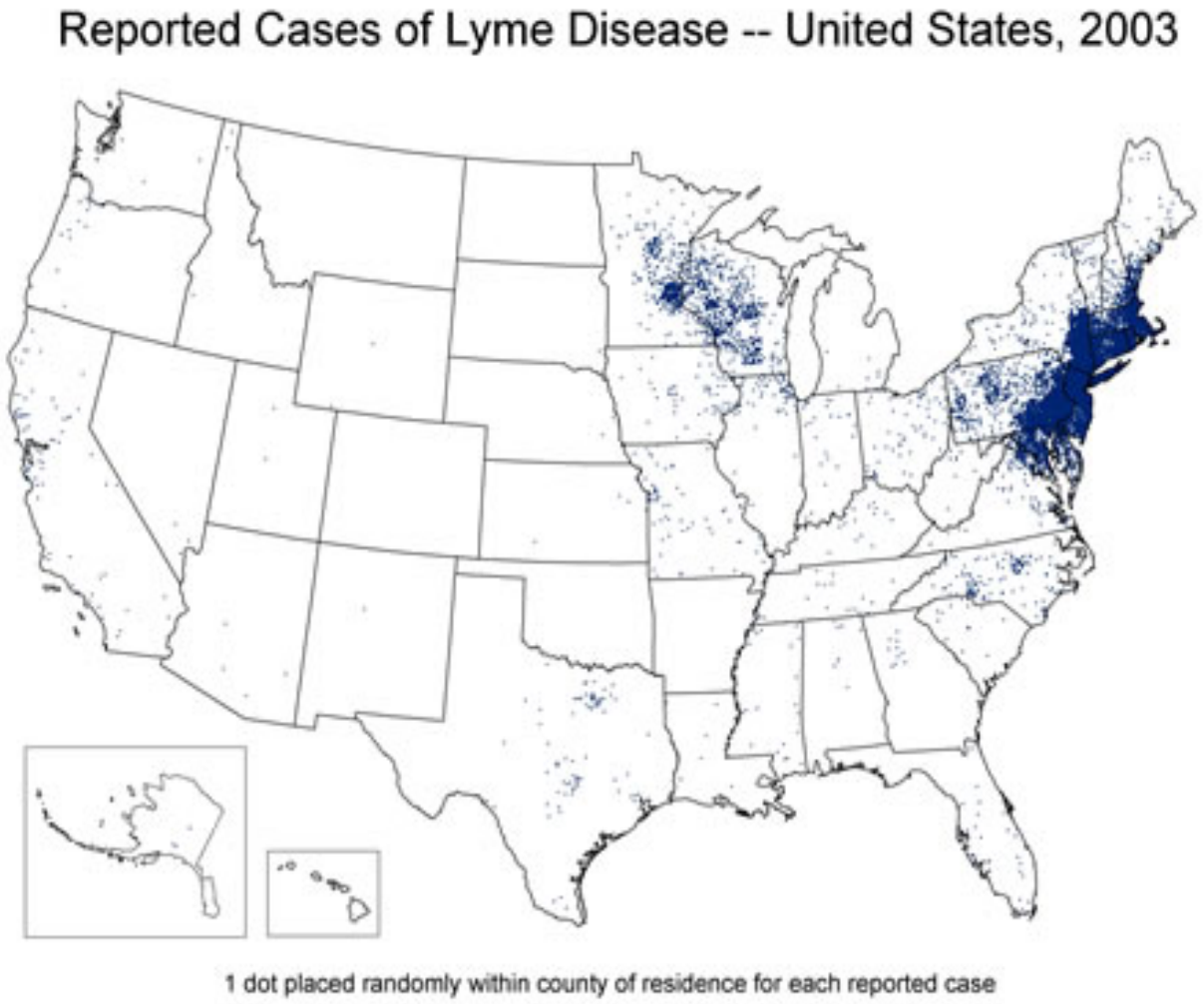
Lyme Disease



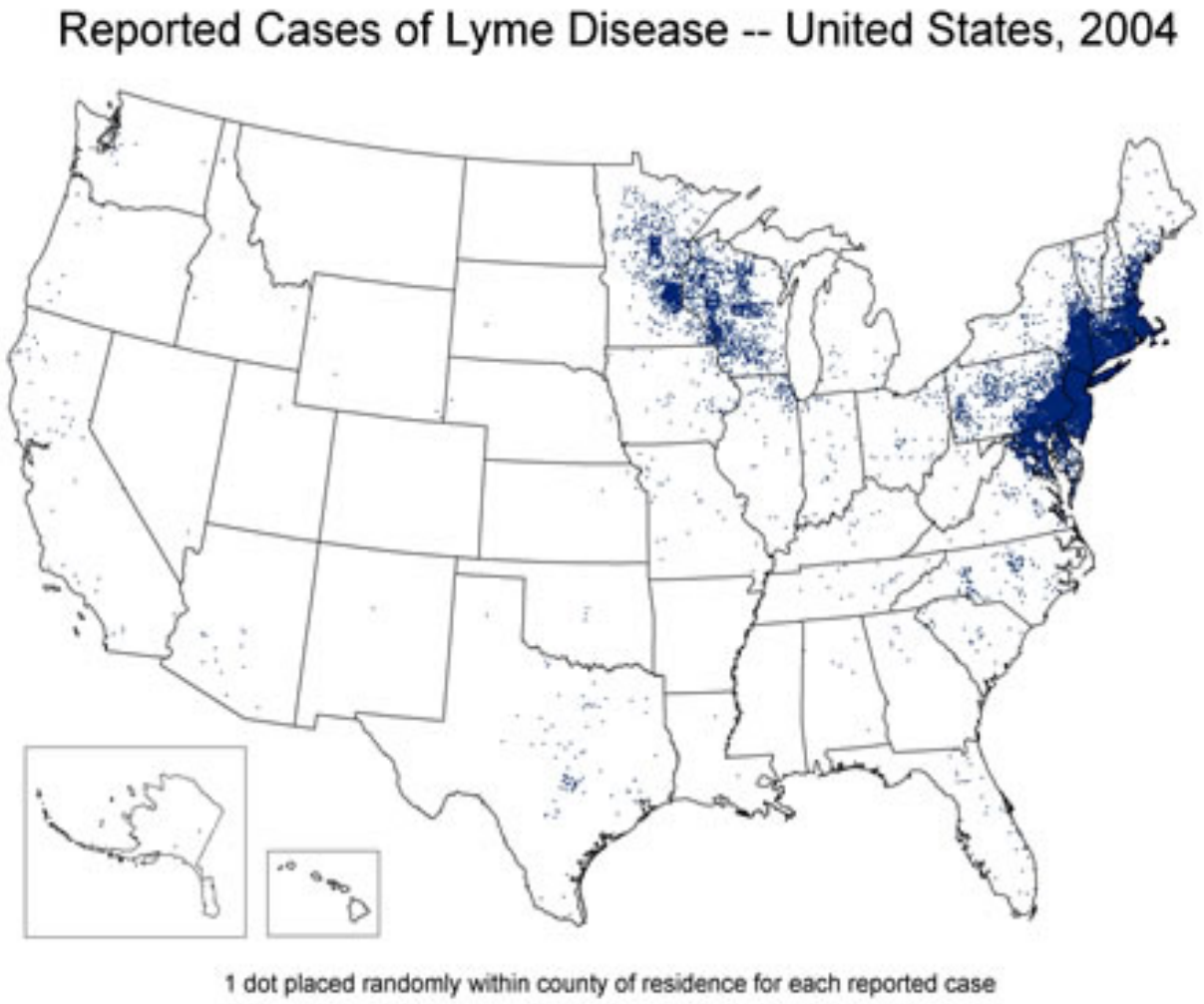
Lyme Disease



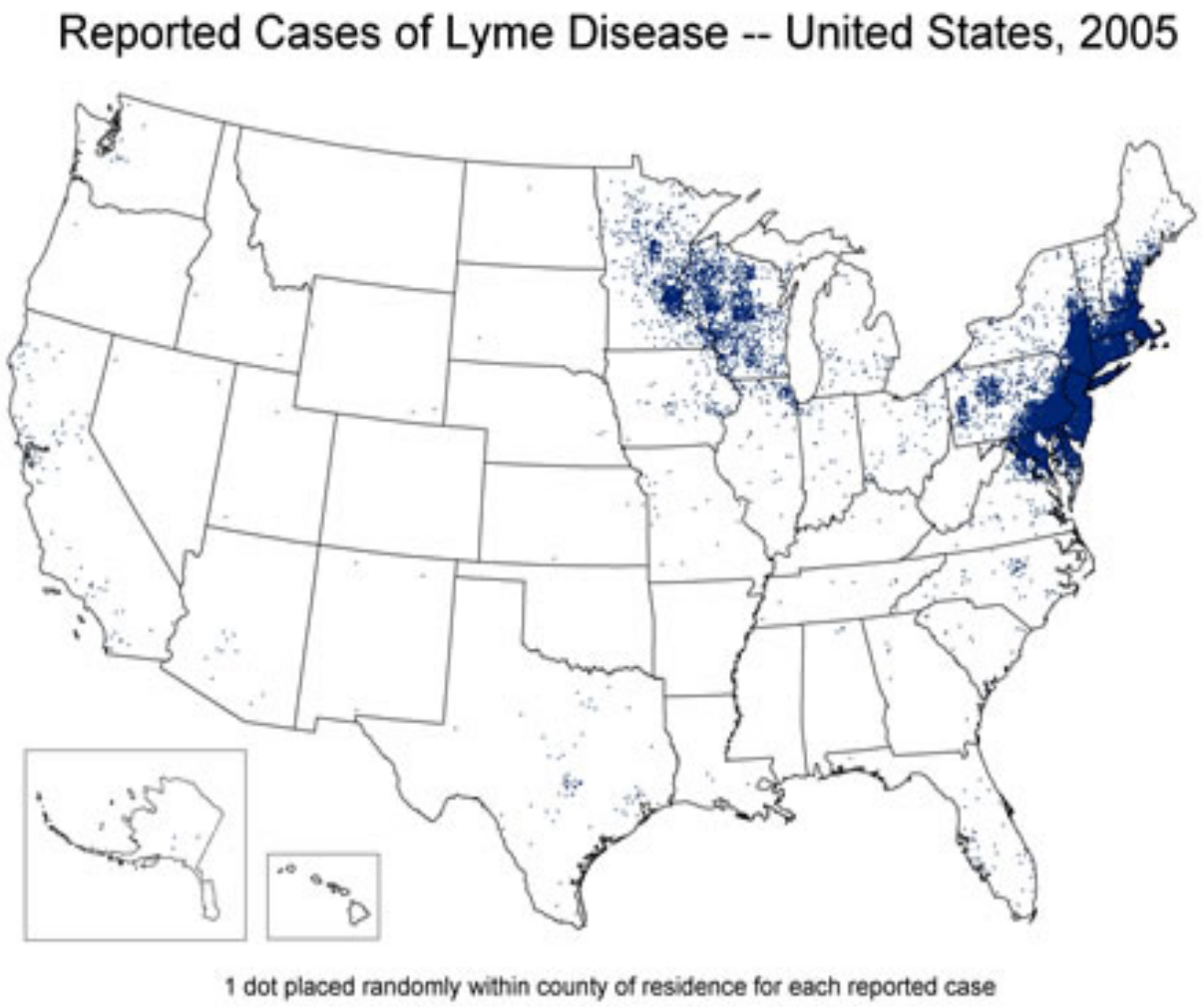
Lyme Disease



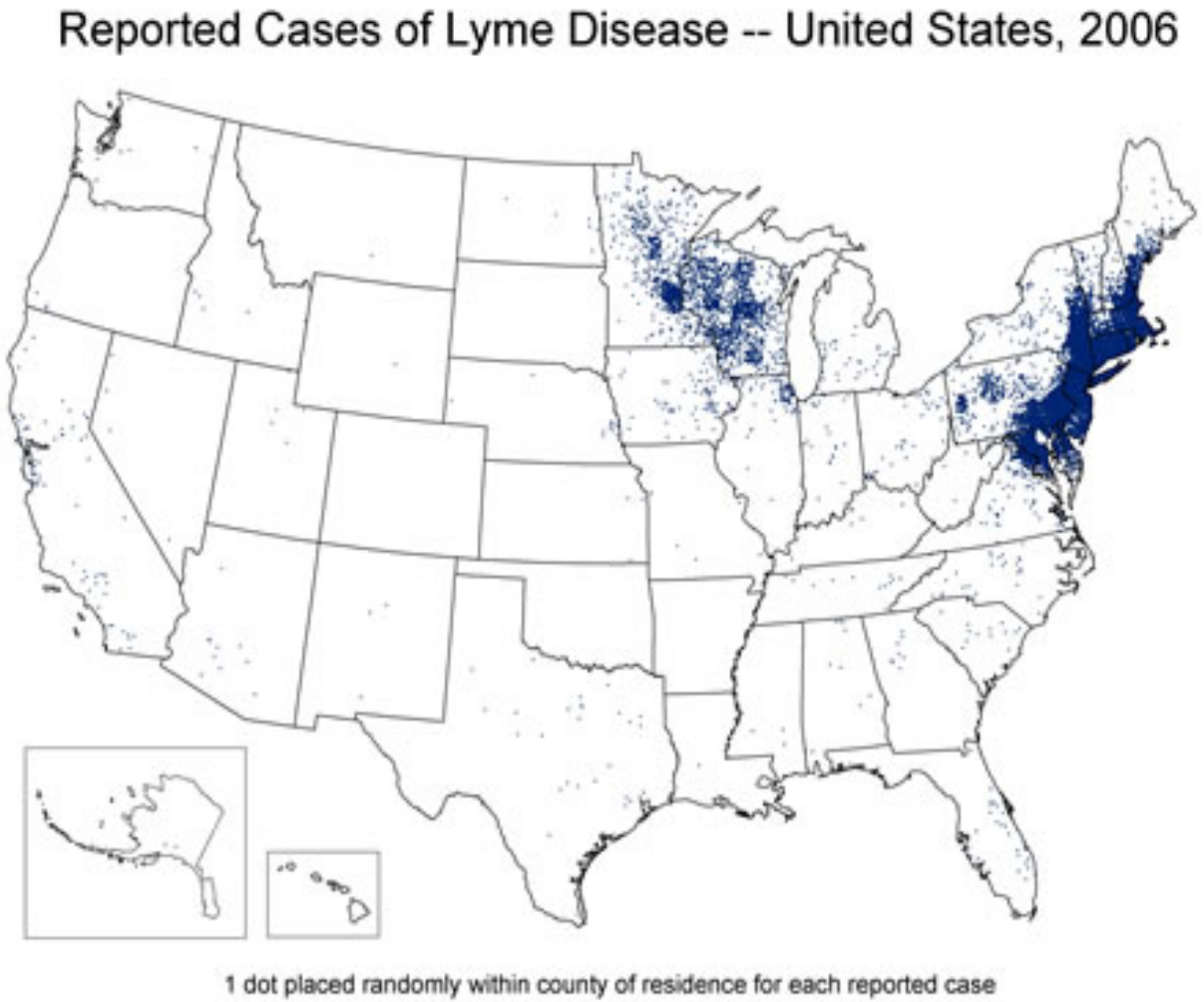
Lyme Disease



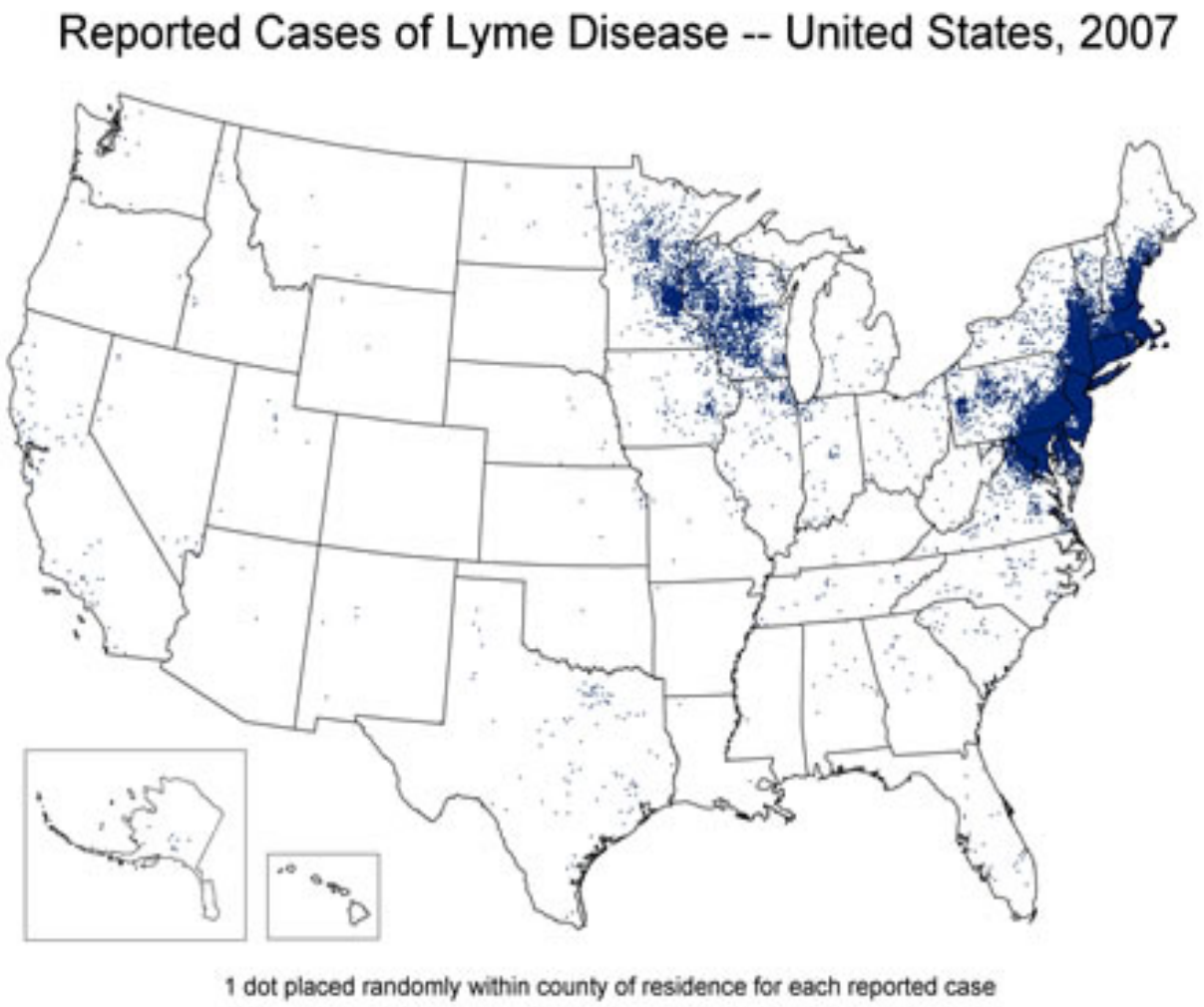
Lyme Disease



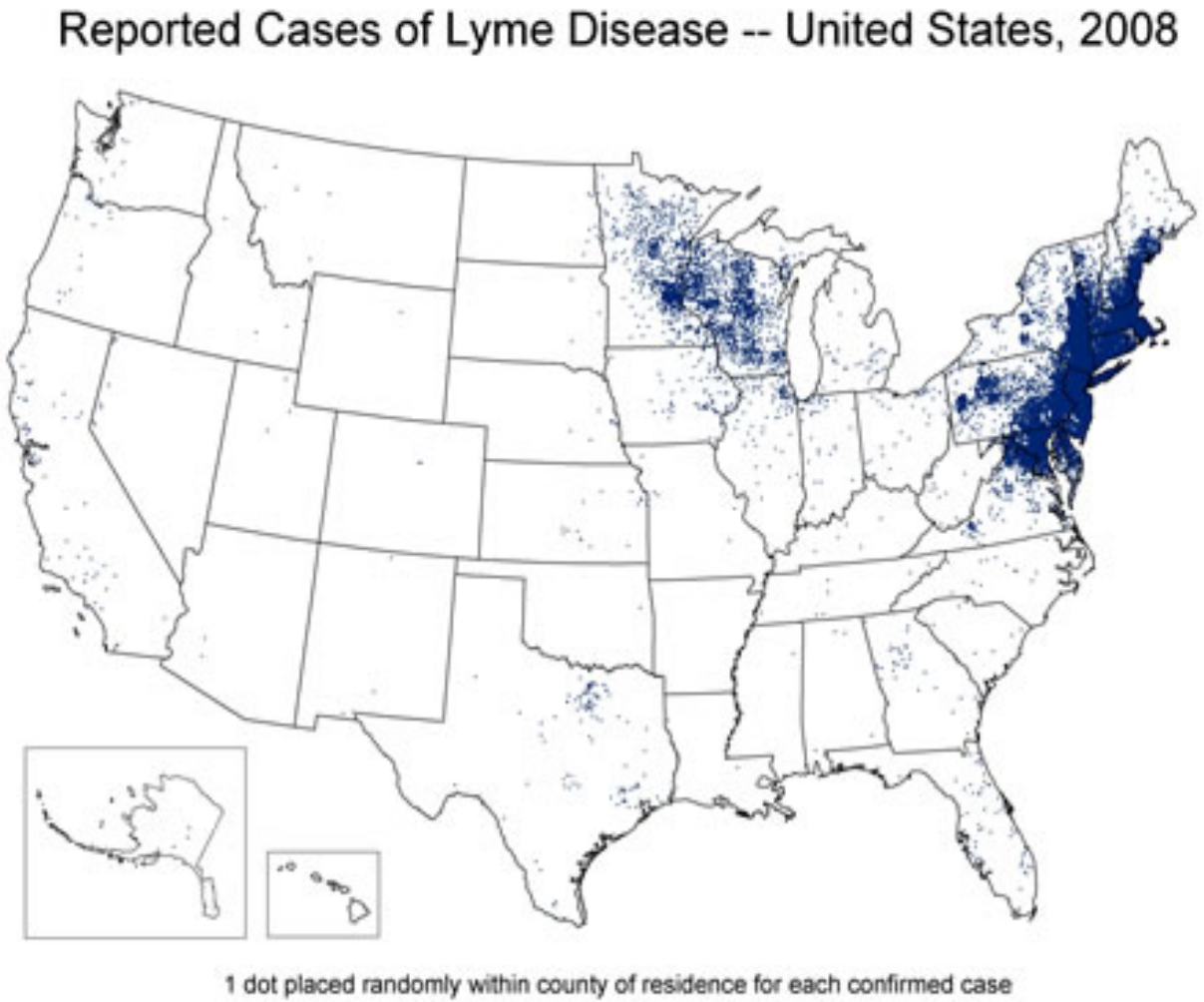
Lyme Disease



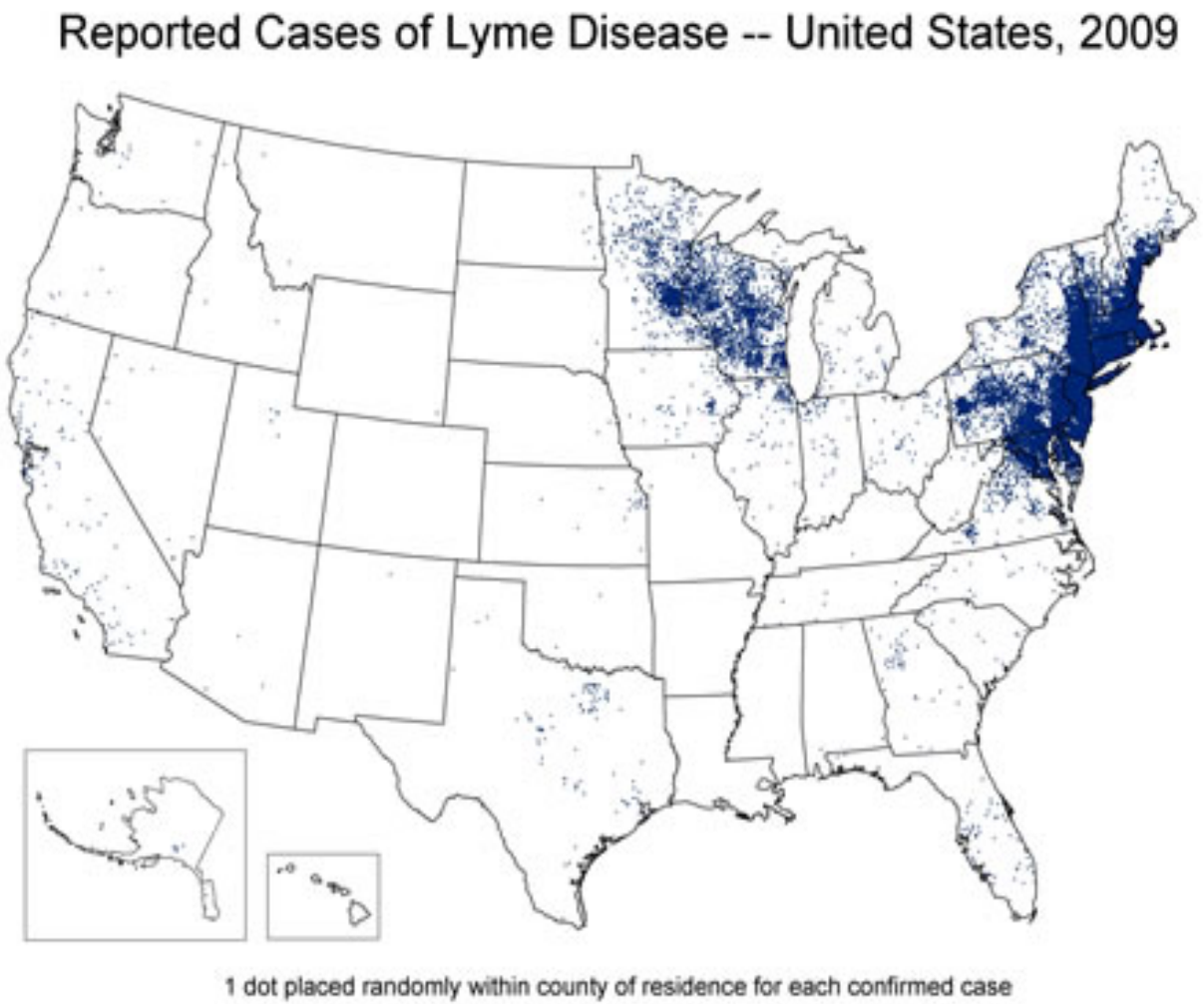
Lyme Disease



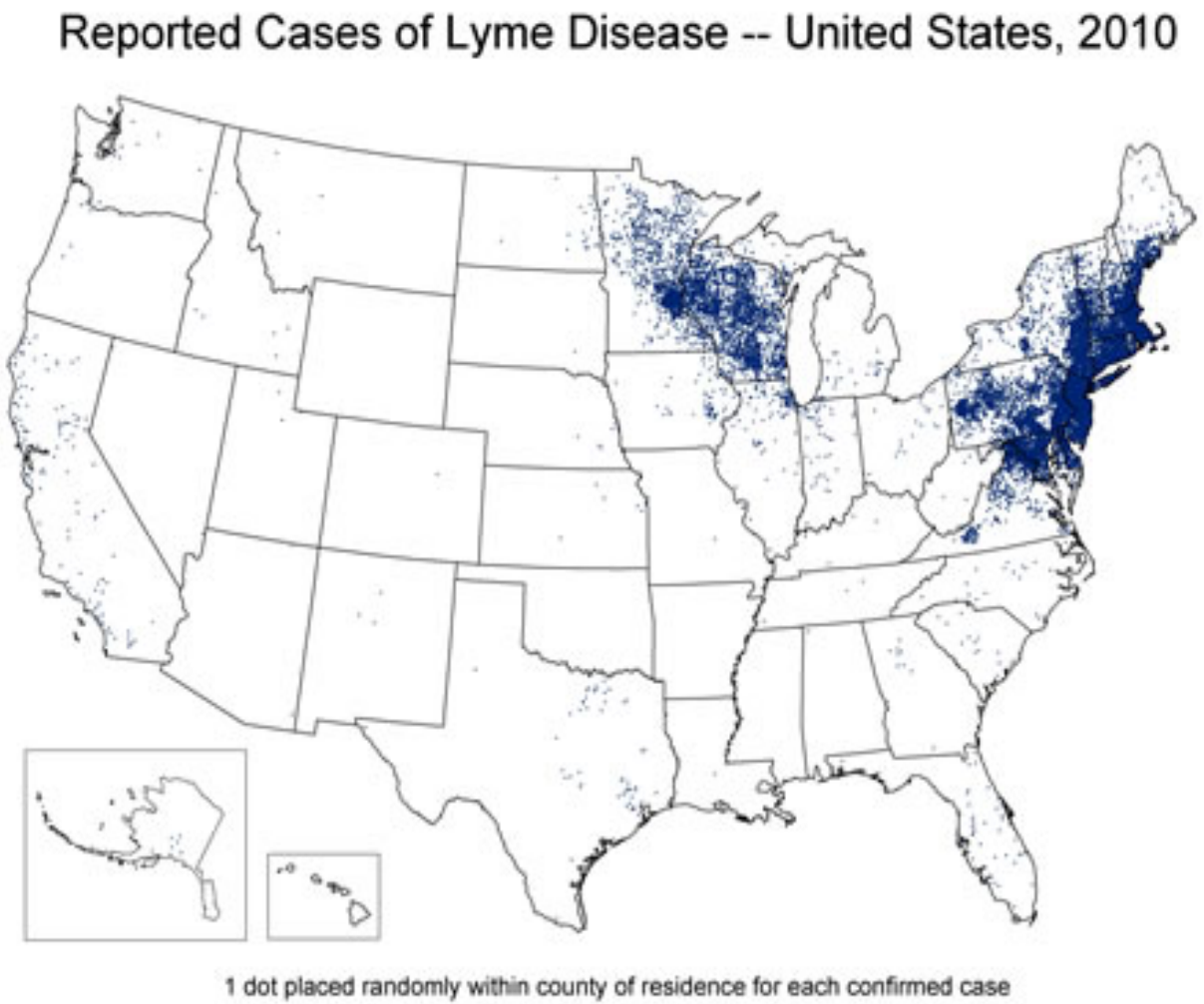
Lyme Disease



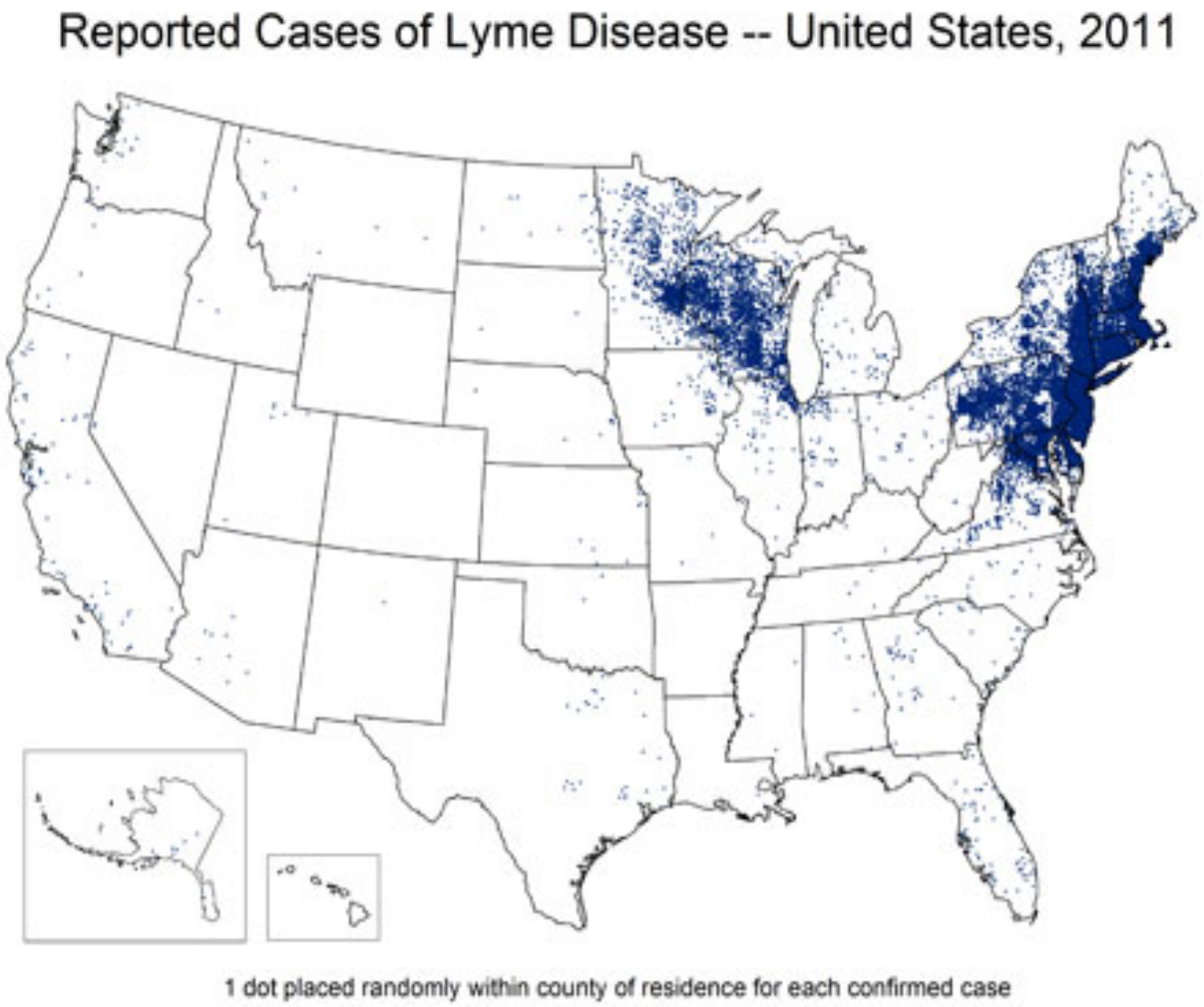
Lyme Disease



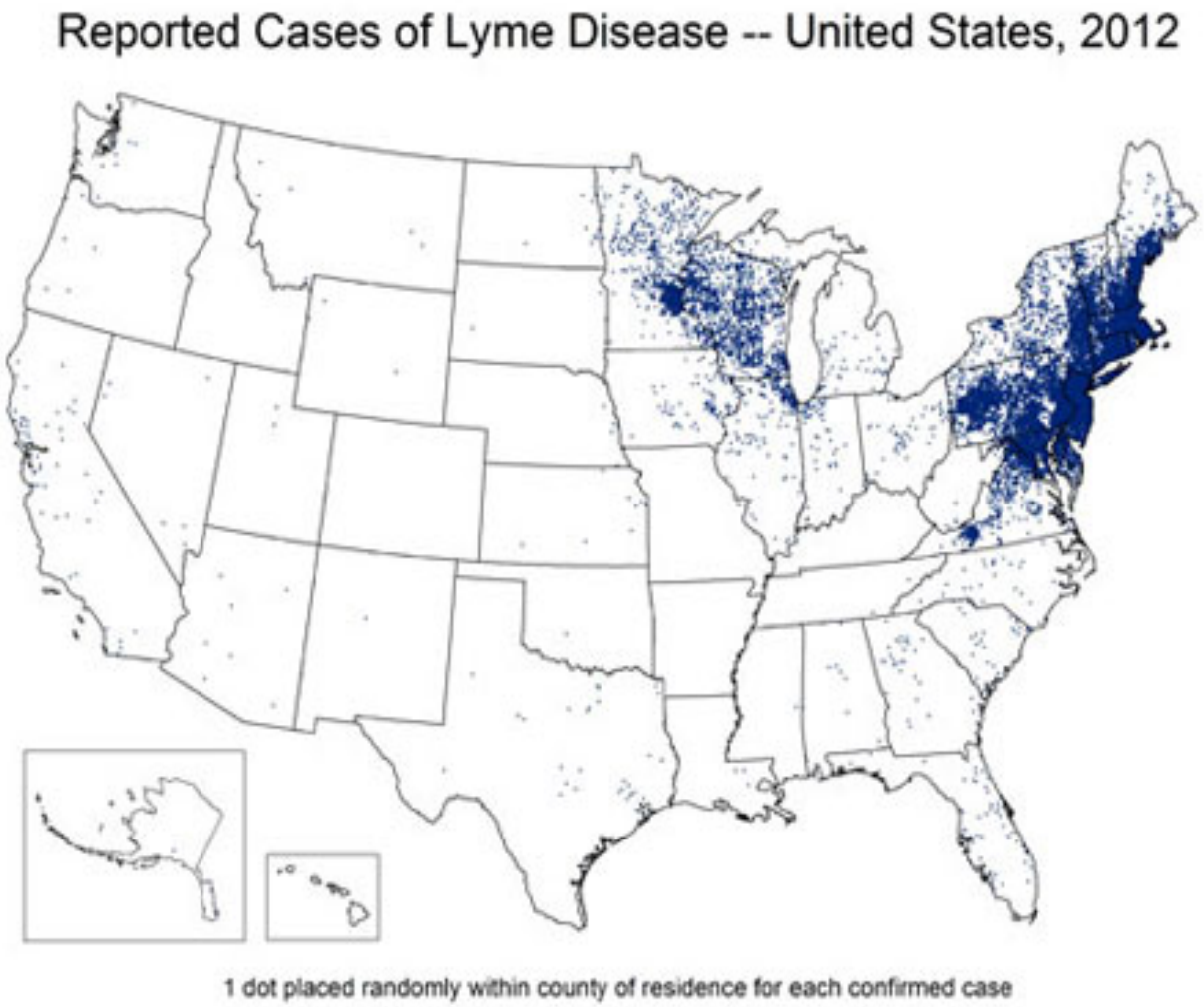
Lyme Disease



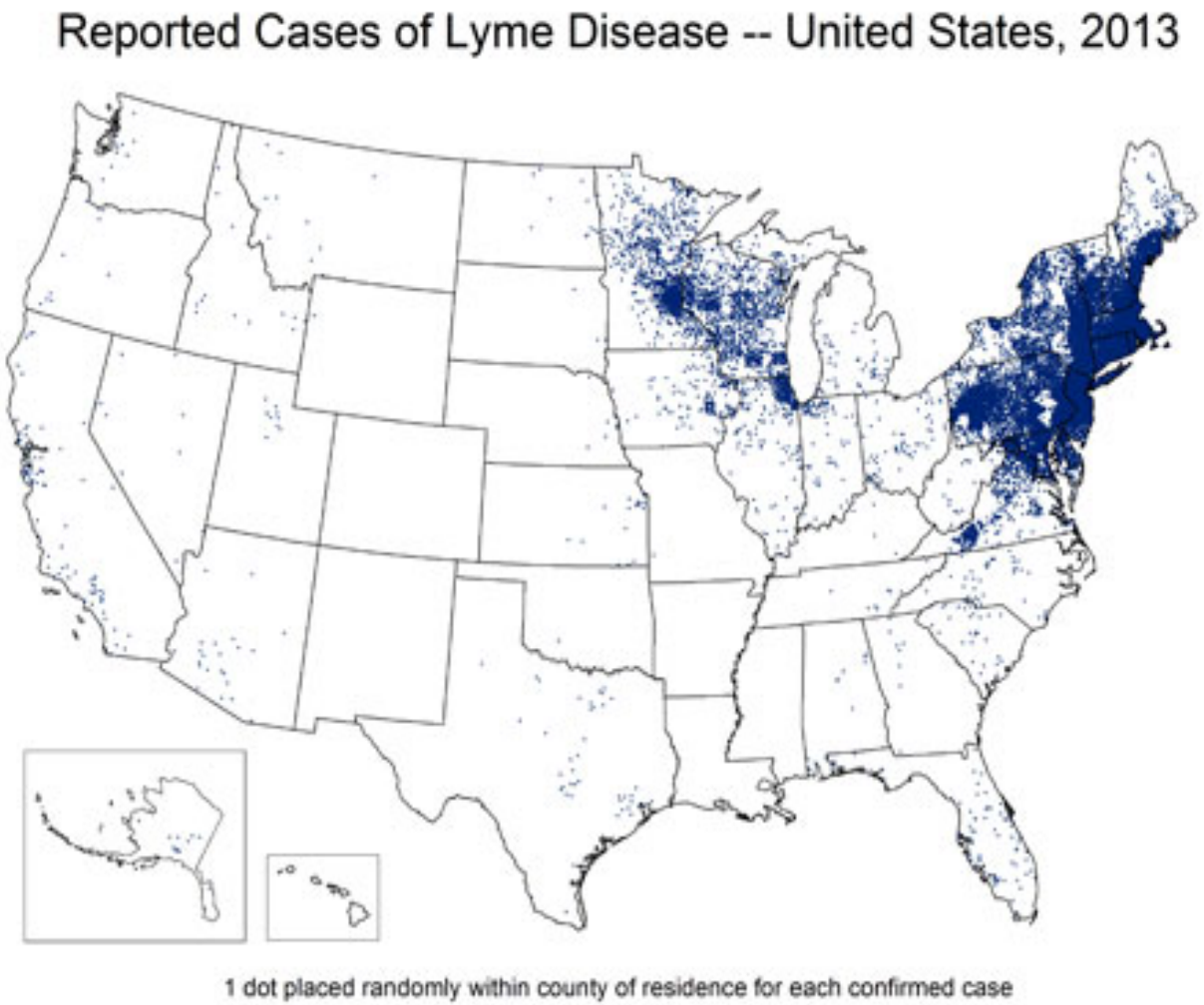
Lyme Disease



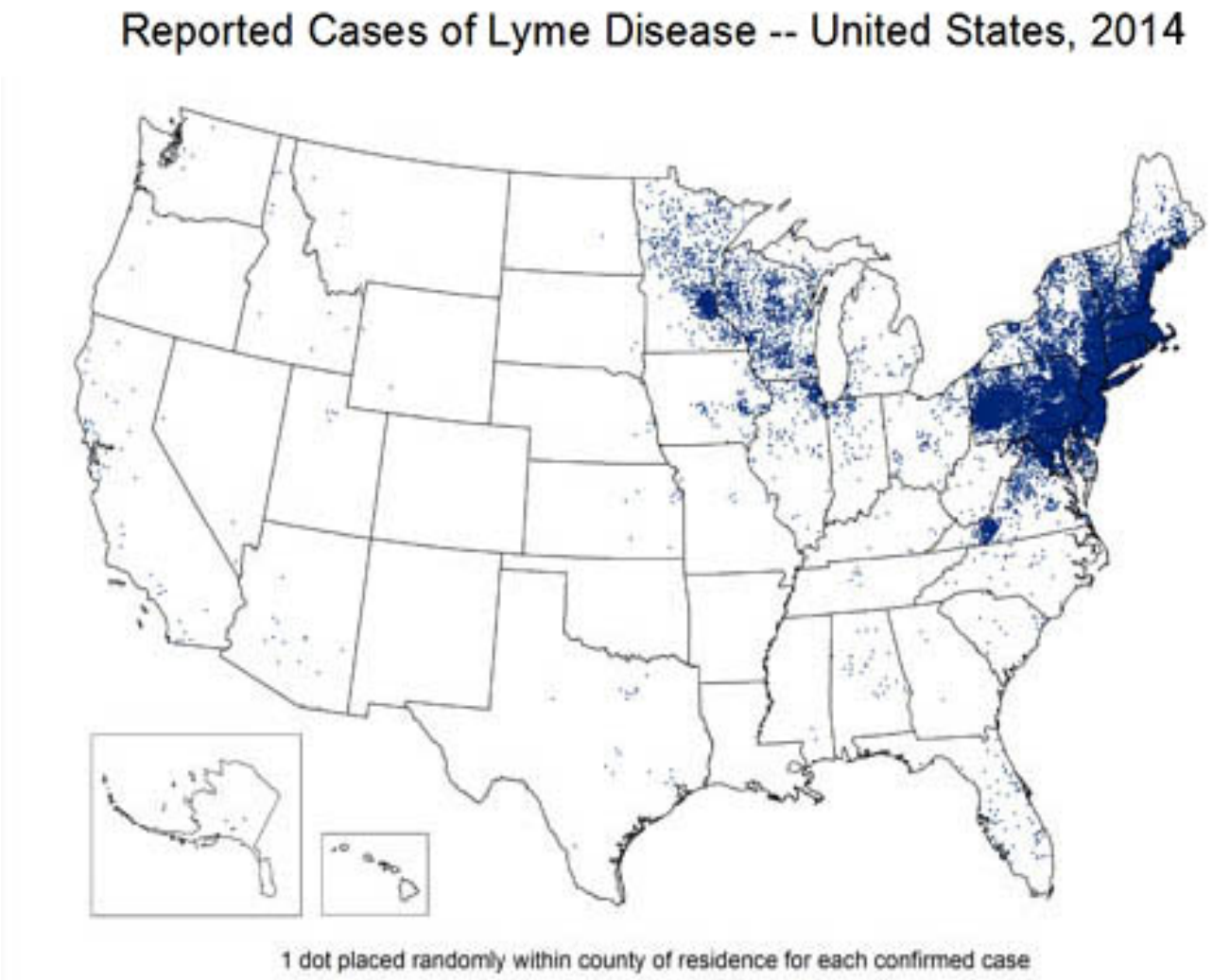
Lyme Disease



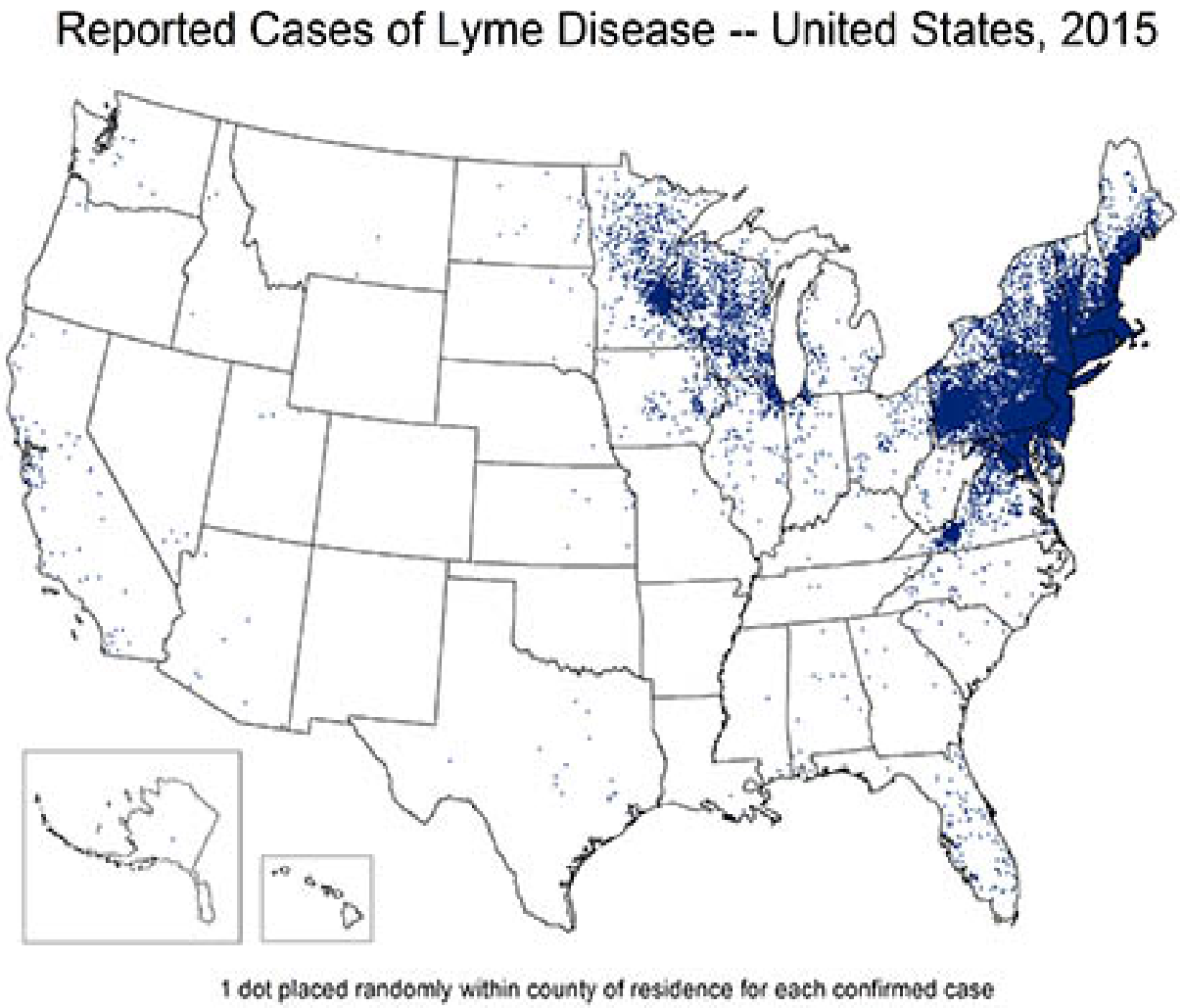
Lyme Disease



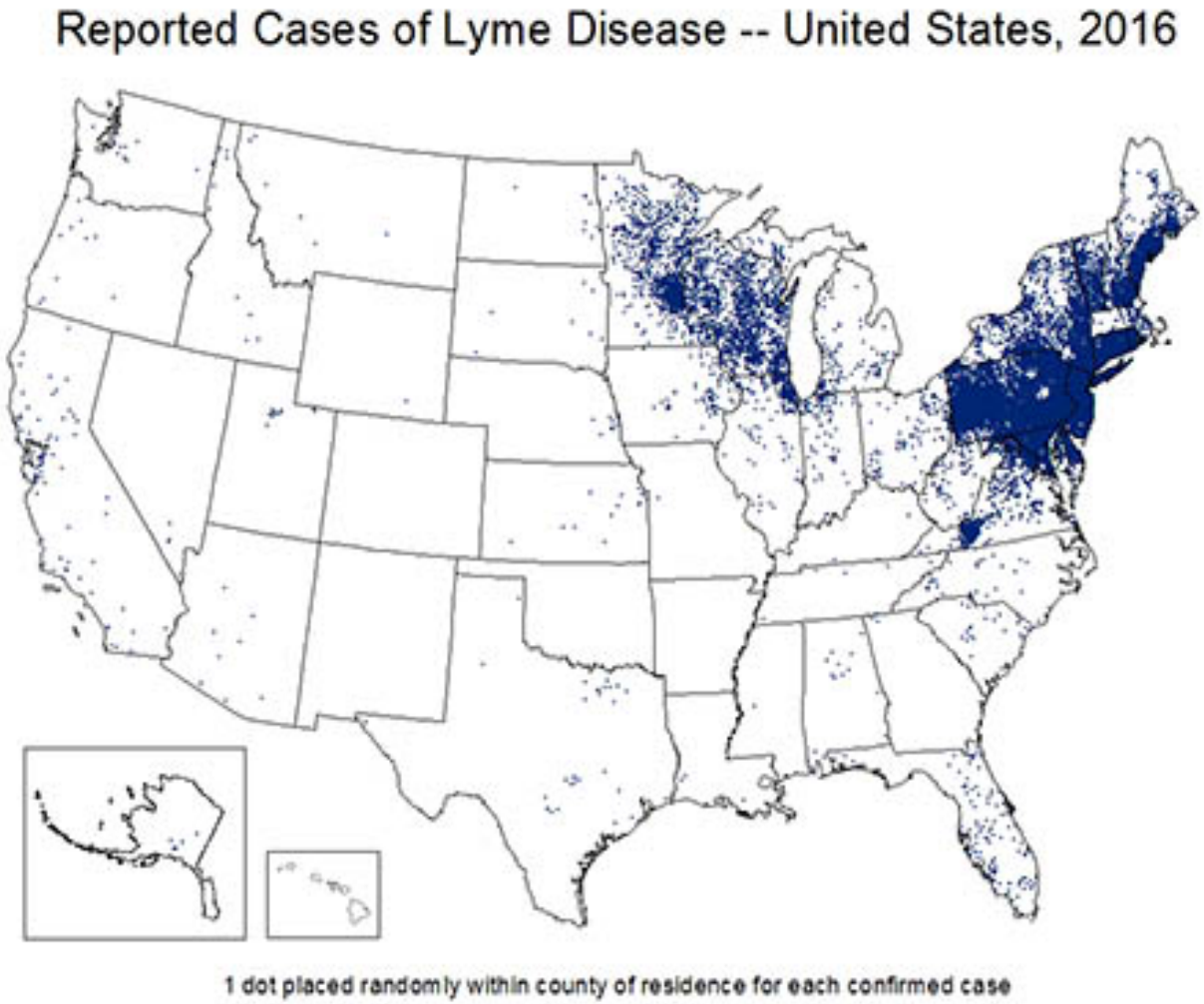
Lyme Disease



Lyme Disease

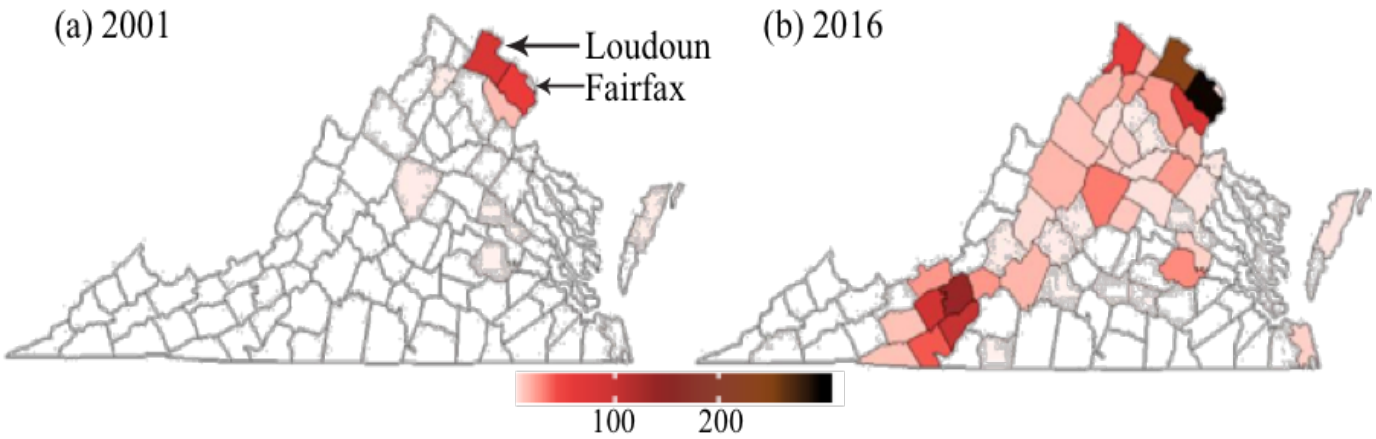


Lyme Disease

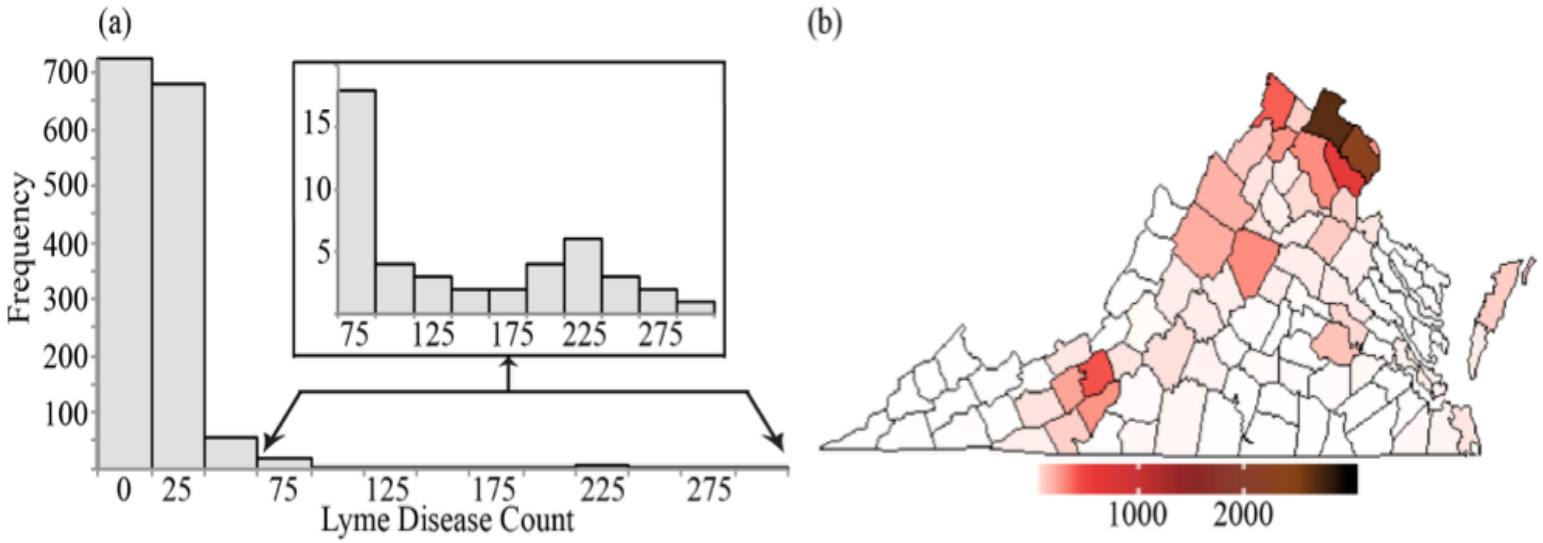


Lyme Disease in Virginia

Annual number of confirmed cases of Lyme at the county level (Source: CDC)



Lyme Disease in Virginia



Bayesian Hierarchical Model

Basic Hierarchical Model (Berliner 1996)

1. [data|process, parameters]
2. [process|parameters]
3. [parameters]

Using the Bayes Theorem, the posterior [process,parameters|data] can be written as proportional to the product of these three distributions!

Data Model

Let the random variable $Y_{i,t}$ represent the count of the disease:

$$[Y_{i,t}|\mu_{i,t}, p_{i,t}] \sim \text{NegBinHurdle}(\mu_{i,t}, p_{i,t}).$$

where the negative binomial hurdle model is defined as

$$p_{i,t} I_{(y_i=0)} + (1 - p_{i,t})\ \text{NegBin}(\mu_{i,t} > 0).$$

Notation:

$$\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})',$$

$$\mathbf{p}_t = (p1, t, \ldots, p_{n,t})'.$$

$$\boldsymbol{\mu}_t = (\mu 1, t, \ldots, \mu_{n,t})'.$$

Process Model

We consider generalized additive models for two process models: a logistic regression model for the presence probabilities

$$\text{logit}(\mathbf{p}_t) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}_s + \boldsymbol{\varepsilon}_t,$$

$$\varepsilon_{i,t} = \phi\varepsilon_{i,t-1} + \xi_{i,t}$$

and a log-linear model for the mean of a zero-truncated negative binomial:

$$\log(\boldsymbol{\lambda}_t) = \mathbf{X}^*\boldsymbol{\alpha} + \mathbf{w}_s + \boldsymbol{\eta}_t.$$

Similarly, an autoregressive error term may be considered for the log-linear model.

Parameter Models

In the Bayesian setting, the parameter models for a hierarchical model are the priors!

We define relatively non-informative prior densities for the unknown parameters (e.g., normal distribution with mean 0 and large variance; inverse-Gamma distributions with small mean and large variance for variance components).

Drivers of Dynamics

An important factor to consider in modeling the diseases dynamics is the mechanism of disease spread and its drivers. The drivers of dynamics of disease spread are often specific to classes of infectious diseases and may even be disease-specific.

For example, the dynamics of disease spread for vector-borne diseases are often a function of the environment (and ecology), and human behavior (and mobility). While the dynamics in human-to-human transmission is largely a function of human behavior, population and mobility.

Drivers of Dynamics

Understanding the large scale drivers of disease dynamics and utilizing data (often proxy data) on these factors can greatly improve modeling the dynamics and allow for near real-time prediction of diseases spread.

In addition to conventional data sources (population, temperature, greenness, etc.), organic data offers a promising source of information for modeling dynamics. For example, social media data may serve as proxy for mobility, identifying early cases of disease, etc.

Other examples include: Search data (disease symptoms, treatments, etc.), traffic cameras (mobility), point of sales data (medication sales), among others.

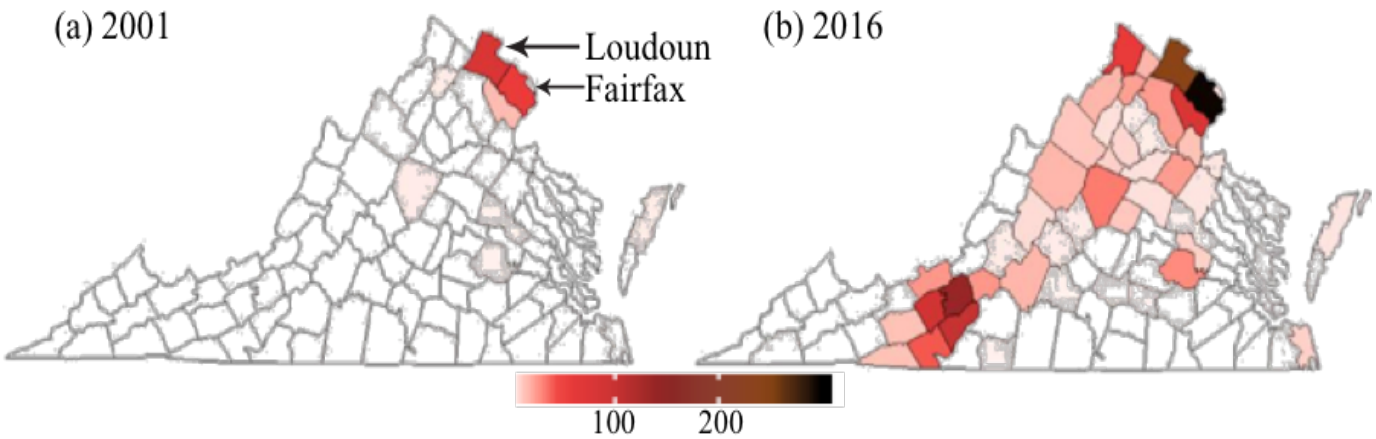
Case Studies

Let's consider two different problems: Lyme disease and COVID-19.

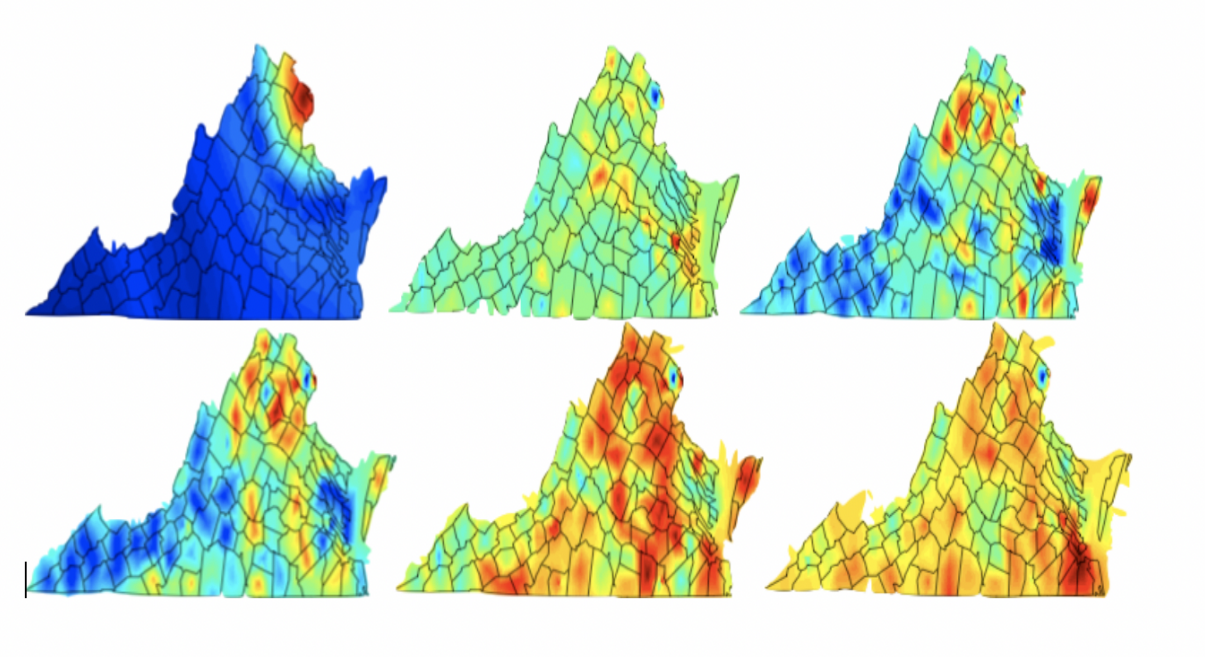
Both emerging epidemics with different rates of spread (Lyme is slow-spreading over several years, COVID-19 is fast-spreading over several weeks).

Both are introduced in the most populous part of the state, Northern Virginia, and then start to spread to other parts.

Spatial Variability of Lyme in Virginia



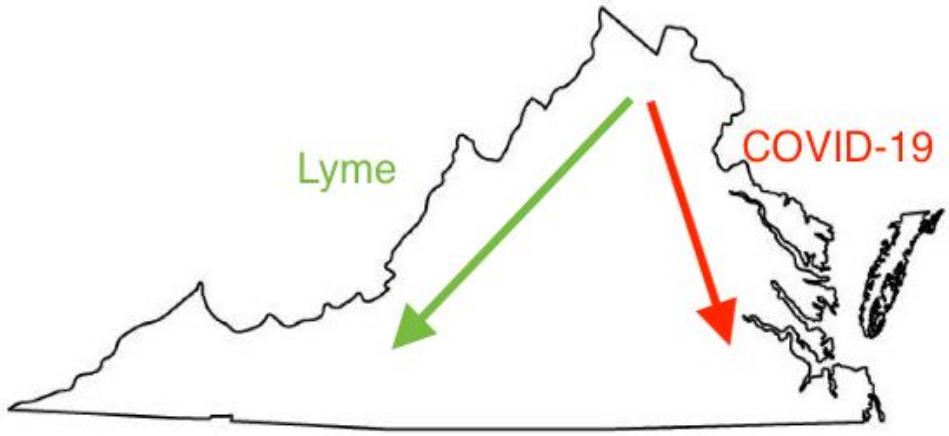
Spatial Variability of COVID-19 in Virginia



(Weeks 1, 4, 8, 12, 16, 20; Weekly values between March 8-July 19.)

Case Studies

Although the spread of both Lyme and COVID-19 start in Northern Virginia, the geographical direction of spread is quite different.



This may be due to drivers of dynamics being different (of course, there is a fundamental difference between the two diseases: Lyme is transmitted from ticks to human, and COVID-19 is transmitted from human to human)

Case Studies

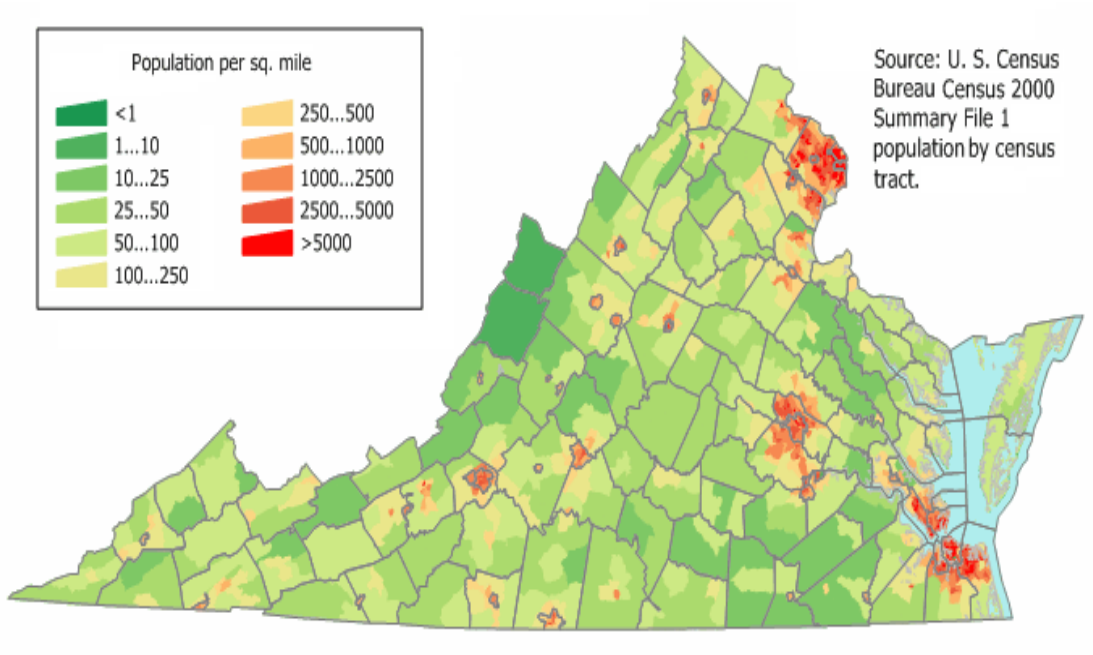
Factors that may be important in spread/prevalence of Lyme include elevation, greenness, wind, temperature, precipitation.

Not much is known about important factors in spread of COVID-19 beyond population (human-to-human transmission and related effects such as travel, etc.).

Also, it is important to recognize the level of intervention as another important difference between the two cases: not much active intervention to control the spread of Lyme (beyond awareness raising campaigns) vs. significant level of intervention to control COVID-19 (e.g., masking, distancing, lock-downs, etc.).

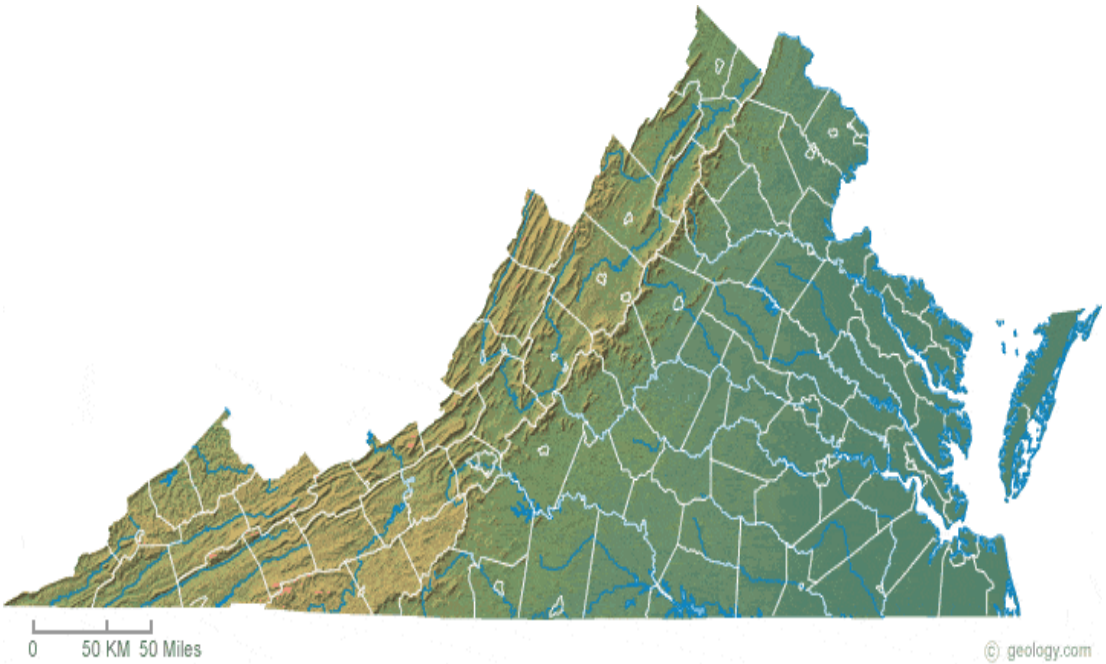
Discussion: COVID-19

Population Density



Discussion: Lyme

Environmental factors such as elevation, greenness, wind, temperature, and precipitation.

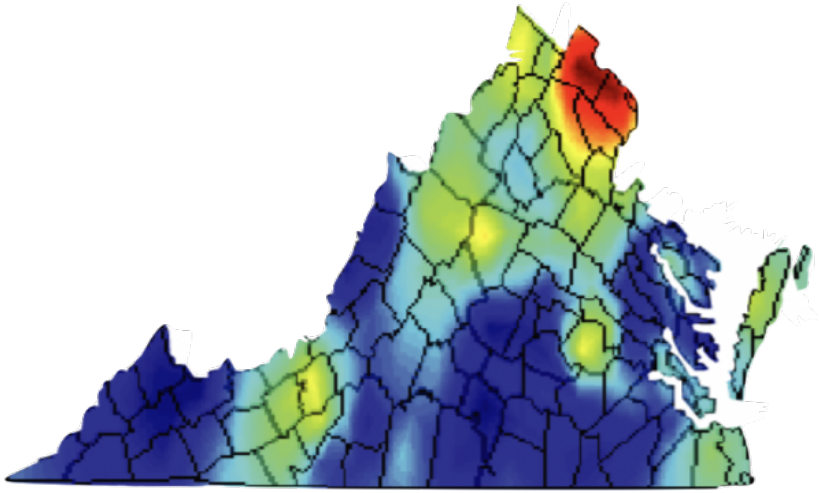


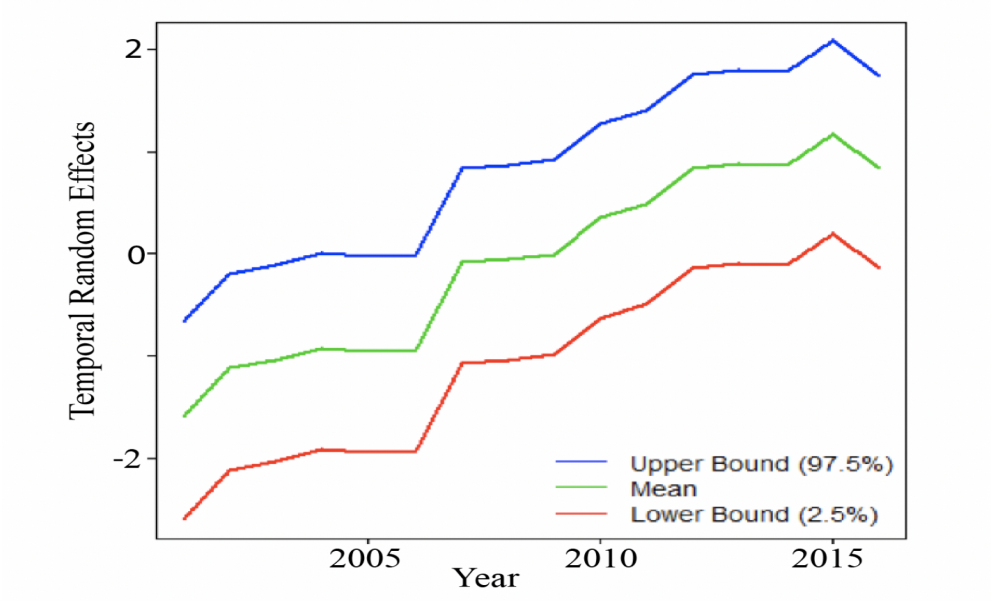
Modeling

To better understand the spatial and temporal variabilities, we fit negative binomial hurdle models with autoregressive error terms to both data sets.

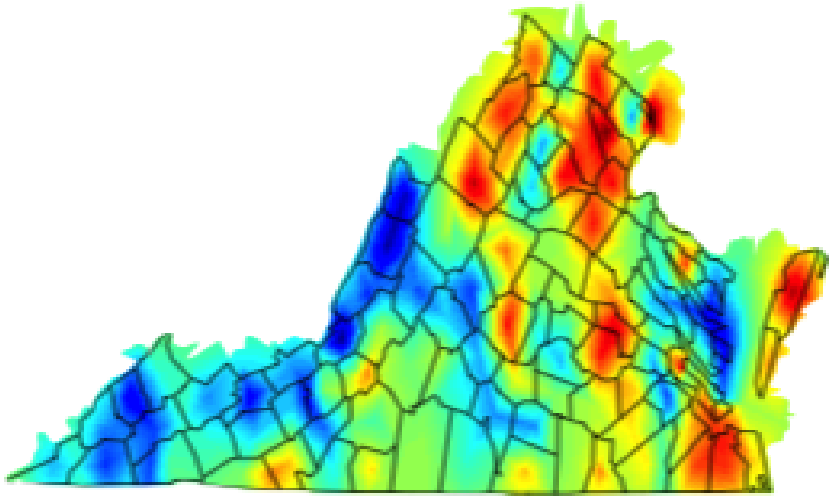
Again, there are fundamental differences between the two diseases, however, our goal is to understand large scale drivers of dynamics, and to illustrate the effect of these drivers on early stages of the spread of the disease.

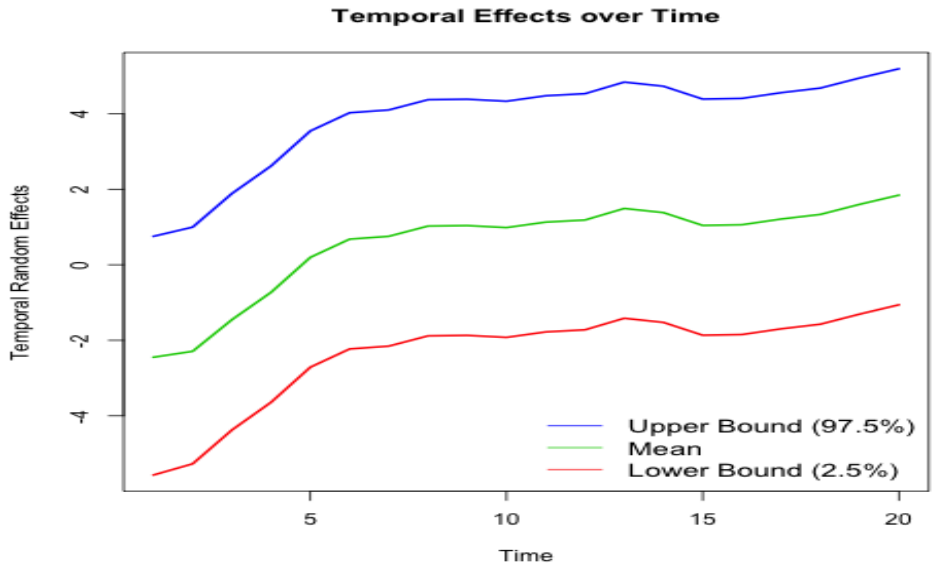
Lyme: Spatial+Temporal Effects Posterior Means





COVID-19: Spatial+Temporal Effects Posterior Means





Mechanistic PDE-Based Process Model

As an alternative to the autoregrssive approach, we consider a mechanistic (physical-statistical) model for modeling the dynamics of Lyme over space and time.

Here, the logistic regression model for the presence/absence probabilities is described as follows:

$$\text{logit}(\mathbf{p}_t) = \mathbf{u}_t + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{u}_t = (u_{1,t}, \ldots, u_{n,t})'$ is a latent process which may be characterized based on the diffusion PDE, and ε_t's are *iid* error terms (again due to conditional independence):

$$\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}).$$

Process Model: PDE

The latent process will be characterized based on the following two-dimensional diffusion equation:

$$\frac{\partial u_t(x,y)}{\partial t} - D\left(\frac{\partial^2 u_t(x,y)}{\partial x^2} + \frac{\partial^2 u_t(x,y)}{\partial y^2}\right) = 0.$$

where D denotes the diffusion parameter.

Process Model

Using a forward differencing time discretization:

$$\frac{\partial u_t}{\partial t} = \frac{u_t - u_{t-\Delta t}}{\Delta t},$$

and centered differences in space:

$$\frac{\partial^2 u_t}{\partial x^2} = \frac{u_t(x+\Delta x, y) - 2u_t(x,y) + u_t(x-\Delta x, y)}{\Delta^2 x},$$

$$\frac{\partial^2 u_t}{\partial y^2} = \frac{u_t(x, y+\Delta y) - 2u_t(x,y) + u_t(x, y-\Delta y)}{\Delta^2 y},$$

the system of equations may be discretized and presented as a dynamical model, $\mathbf{u}_t = \mathbf{H}\mathbf{u}_{t-1}$, where the transition matrix H is **sparse** and only a function of D, Δt, Δx, and Δy.

Process Model

The process model includes the dynamical discretized PDE with additive error:

$$\mathbf{u}_t = \mathbf{H}\mathbf{u}_{t-1} + \boldsymbol{\eta}_t.$$

where:

$$\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

and the variance-covariance matrix $\boldsymbol{\Sigma}$ is parameterized to reflect spatial dependence among the locations:

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{R}(\theta),$$

$$\mathbf{R}(\theta) = \exp(-\theta ||d||)$$

$\mathbf{R}(\theta)$ represents the spatial correlation based on the Euclidean distance between locations ($||d||$). Many other choices!

Model Implementation & Results

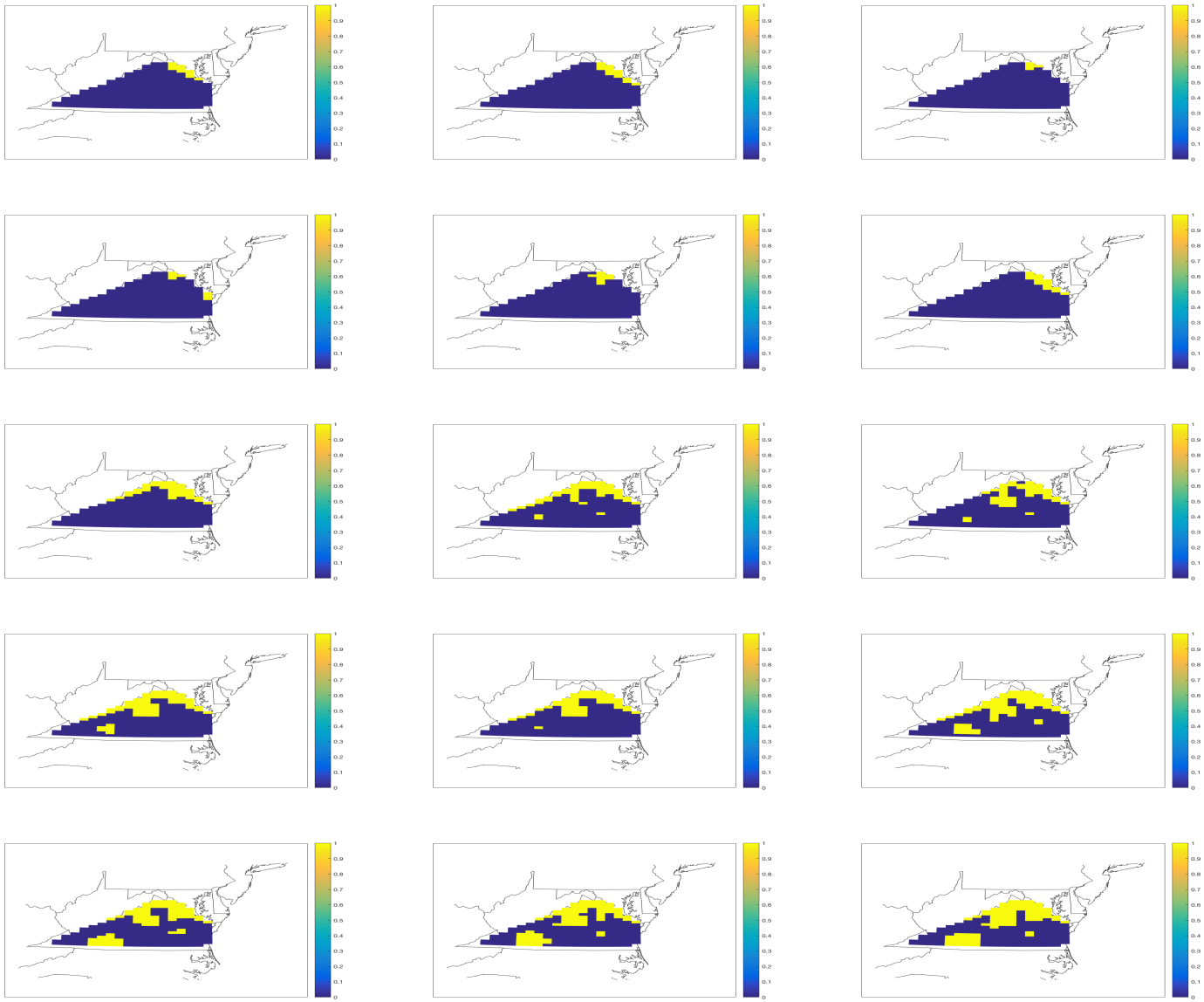
Model Implementation is based on Markov Chain Monte Carlo (MCMC):

- Metropolis-Hastings within Gibbs Sampling,
- 20,000 iterations, and 2000 burn-in period, and quick convergence.

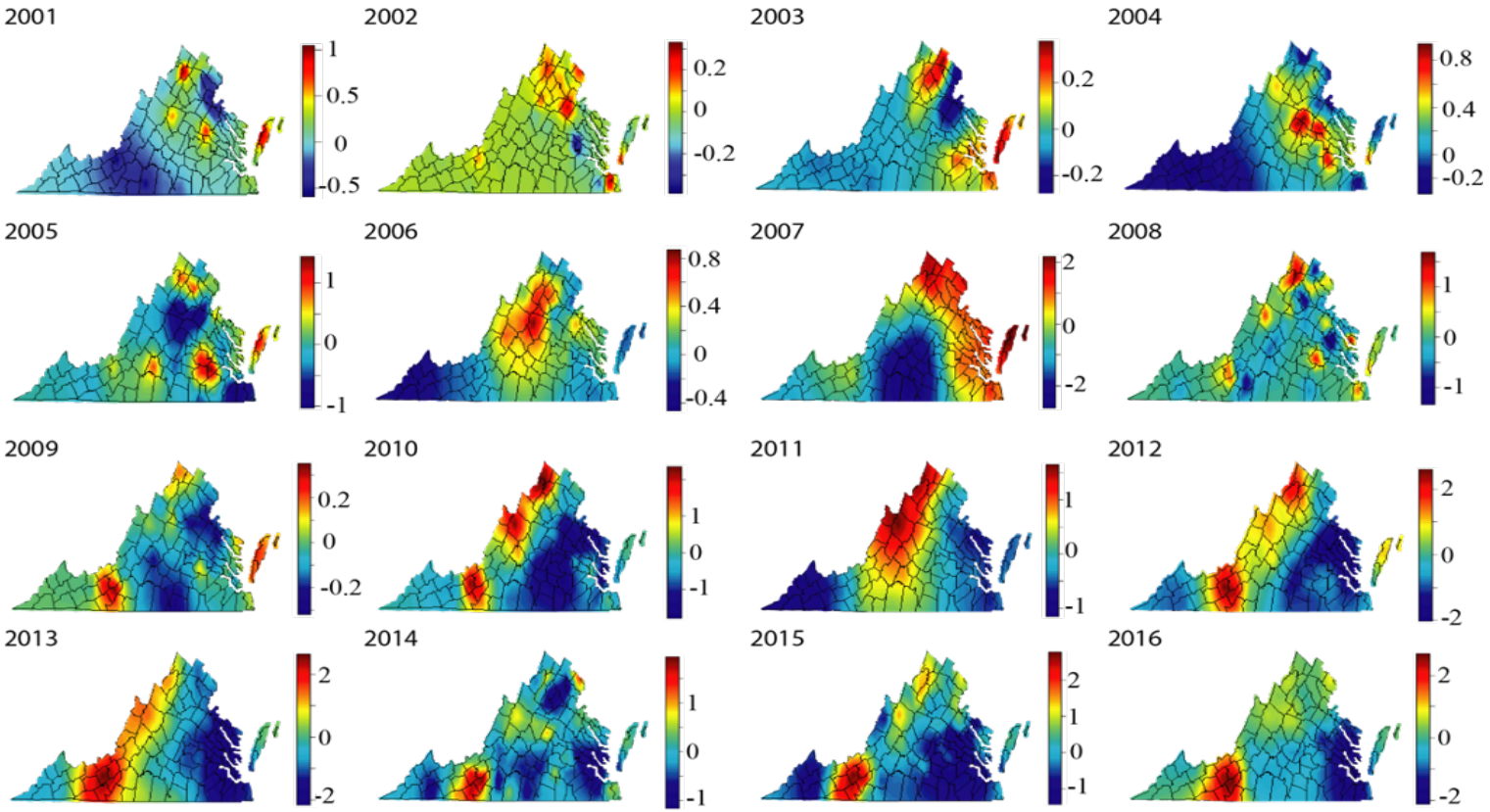
Parameter	**Post.Mean**	**95%CI**
D	0.283	(0.023, 0.598)

Table: Posterior results for the diffusion parameter of the PDE-based model

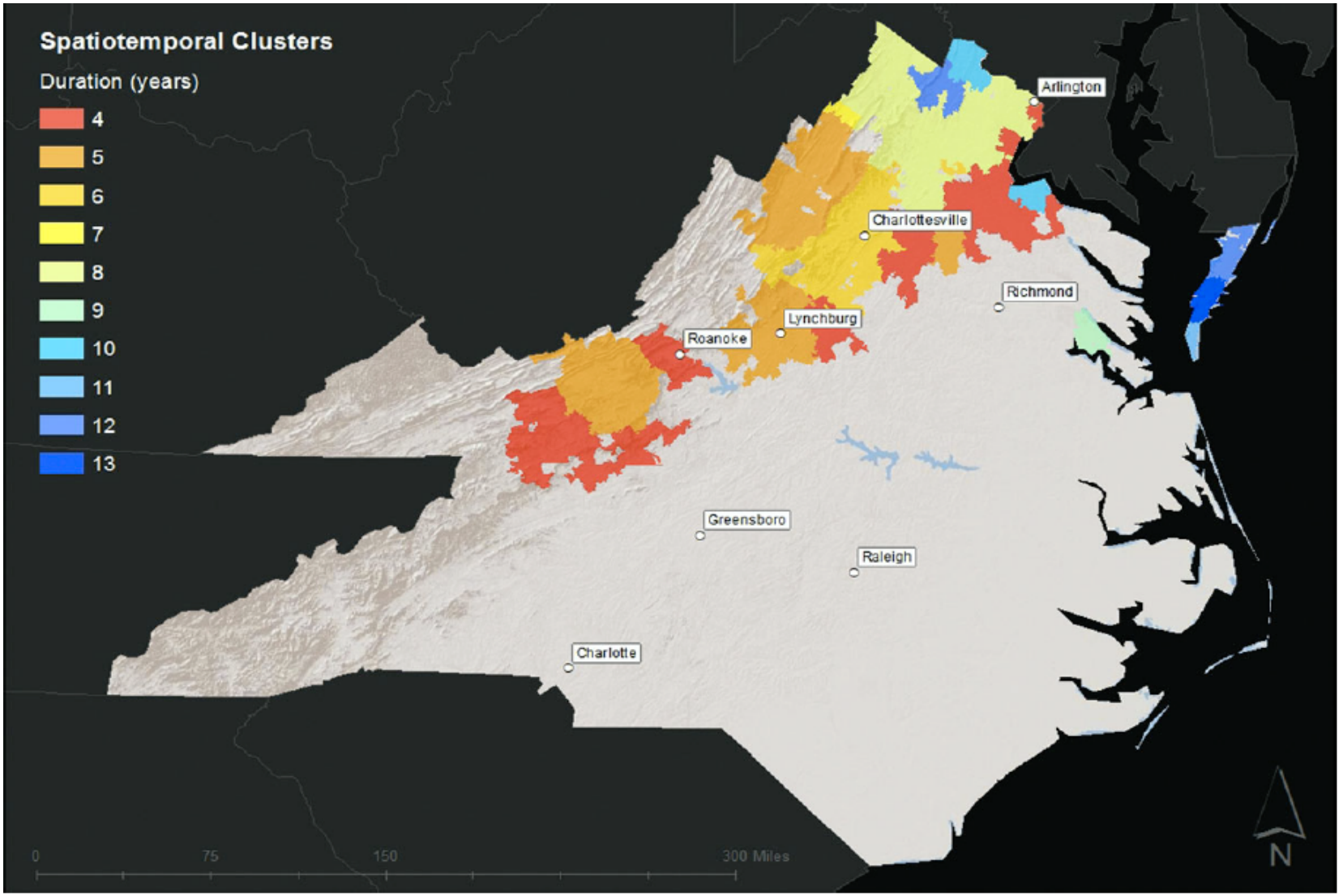
Posterior Means: Presence/Absence



Posterior Means: Counts



Spread of Lyme



From Lantos et al. (2015)

Future Work

- Use drivers of dynamics (using conventional and organic data) to inform the model (in particular, the parameters that govern the dynamics). This may be effectively done using agent based models but other modeling frameworks including PDE-based models may be useful too.
- For COVID-19, accounting for interventions (closings, restrictions, re-opening phases) may be insightful. There are critical variations both in timing and intensity of these measures that should be taken into account.
- Understanding risk categories and using demographic information to better inform the dynamics.
- Assimilate epidemiological modeling information (e.g., SIR models) through the hierarchical framework.

Acknowledgements

Acknowledgements:

Earlier version of this work based on an autoregressive model for Lyme disease in Virginia ia a joint work with Naresh Neupane (Georgetown University) and Ari Goldbloom-Helzner. [Ticks and Tick-borne Diseases, 2021]

Simulation studies related to mechanistic PDE-based model is based on joint collaboration with Zhen Liu (former postdoctoral researcher) and Ryan Ripper (MS student).

References

- Neupane, N., Goldbloom-Helzner, A., & Arab, A. (2021). Spatio-temporal modeling for confirmed cases of lyme disease in Virginia. Ticks and Tick-borne Diseases, 12(6), 101822.
- Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. International journal of environmental research and public health, 12(9), 10536-10548.
- Lantos, P. M., et al. (2015). Geographic expansion of Lyme disease in the southeastern United States, 2000–2014. In Open forum infectious diseases. Oxford University Press.
- Wikle, C. K., & Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. Test, 19(3), 417-451.
- Balderama, E., Gardner, B., & Reich, B. J. (2016). A spatial–temporal double-hurdle model for extremely over-dispersed avian count data. Spatial Statistics, 18, 263-275.
- Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology, 84(6), 1382-1394.

Thank You! Questions?

Predictor Data

We consider monthly values of the following environmental factors:

- Surface temperature,
- Relative humidity,
- Precipitation,
- Wind components (V-component along latitude, and U-component along longitude) from the North American Regional Reanalysis,
- Vegetation index (NDVI).

Weather & Climate Effects?

Log-linear model Results:

Coefficient	Mean	SD	95% CI
January NDVI	-0.16	0.06	(-0.28, -0.03)
March NDVI	0.23	0.09	(0.06,0.39)
September NDVI	-0.27	0.10	(-0.47,-0.07)
April V-wind	0.28	0.06	(0.16,0.39)
May V-wind	-0.20	0.05	(-0.30,-0.11)
September V-wind	-0.19	0.05	(-0.29,-0.09)
August U-wind	-0.36	0.09	(-0.53,-0.19)
October U-wind	0.23	0.06	(0.11,0.35)
March Temperature	-0.39	0.11	(-0.60,-0.18)
January Relative Humidity	-0.17	0.07	(-0.30,-0.04)
March Relative Humidity	-0.33	0.11	(-0.53,-0.12)
July Relative Humidity	-0.29	0.10	(-0.49,-0.09)
September Relative Humidity	0.35	0.08	(0.19,0.51)
January Precipitation	-0.16	0.04	(-0.25,-0.08)
June Precipitation	-0.16	0.04	(-0.23,-0.08)
October Precipitation	-0.16	0.038	(-0.24,-0.09)
November Precipitation	0.09	0.04	(0.01,0.16)

Weather & Climate Effects?

Logistic Regression Model Results:

Coefficient	Mean	SD	95% CI
February NDVI	0.48	0.16	(0.17, 0.79)
August V-wind	0.24	0.11	(0.03, 0.45)
October V-wind	0.60	0.16	(0.28, 0.92)
November V-wind	0.57	0.21	(0.16, 0.98)
December V-wind	-1.18	0.24	(-1.66, -0.72)
August U-wind	0.67	0.22	(0.24, 1.09)
October Precipitation	0.26	0.09	(0.09, 0.43)

spatial & Temporal Random Effects