Modeling Spatio-Temporal Dynamics under Data Sparsity

Ali Arab

Georgetown University
Associate Professor

Mizzou Statistics 60th Anniversary Conference
October 13, 2023

GEORGETOWN UNIVERSITY

Outline

- Challenges of Data sparsity and modeling dynamics
- Motivating examples:
  - (Re)emerging epidemics, Spread of rare conditions (e.g., Lyme disease)
  - Conflict- or climate-driven forced migration (e.g., Forced internal displacement in Iraq)
- Modeling strategy
  - Zero-modified models
  - PDE based dynamics
- Understanding drivers of dynamics
- Exploiting organic data to inform the dynamics
  - Inducing sparsity
  - Bias issues

GEORGETOWN UNIVERSITY

(Re)Emerging Epidemics

Data sparsity is one of the main challenges of modeling emerging epidemics and spread of rare conditions (low number of cases over space and time). Of course, detectability is also an issue in these cases. Generally, 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.

GEORGETOWN UNIVERSITY

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, we need to understand the drivers of the dynamics and address spatial and temporal variability in the data.

GEORGETOWN UNIVERSITY

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 Poisson hurdle model

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

or a negative binomial hurdle model:

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

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

Motivating Problem: Lyme Disease

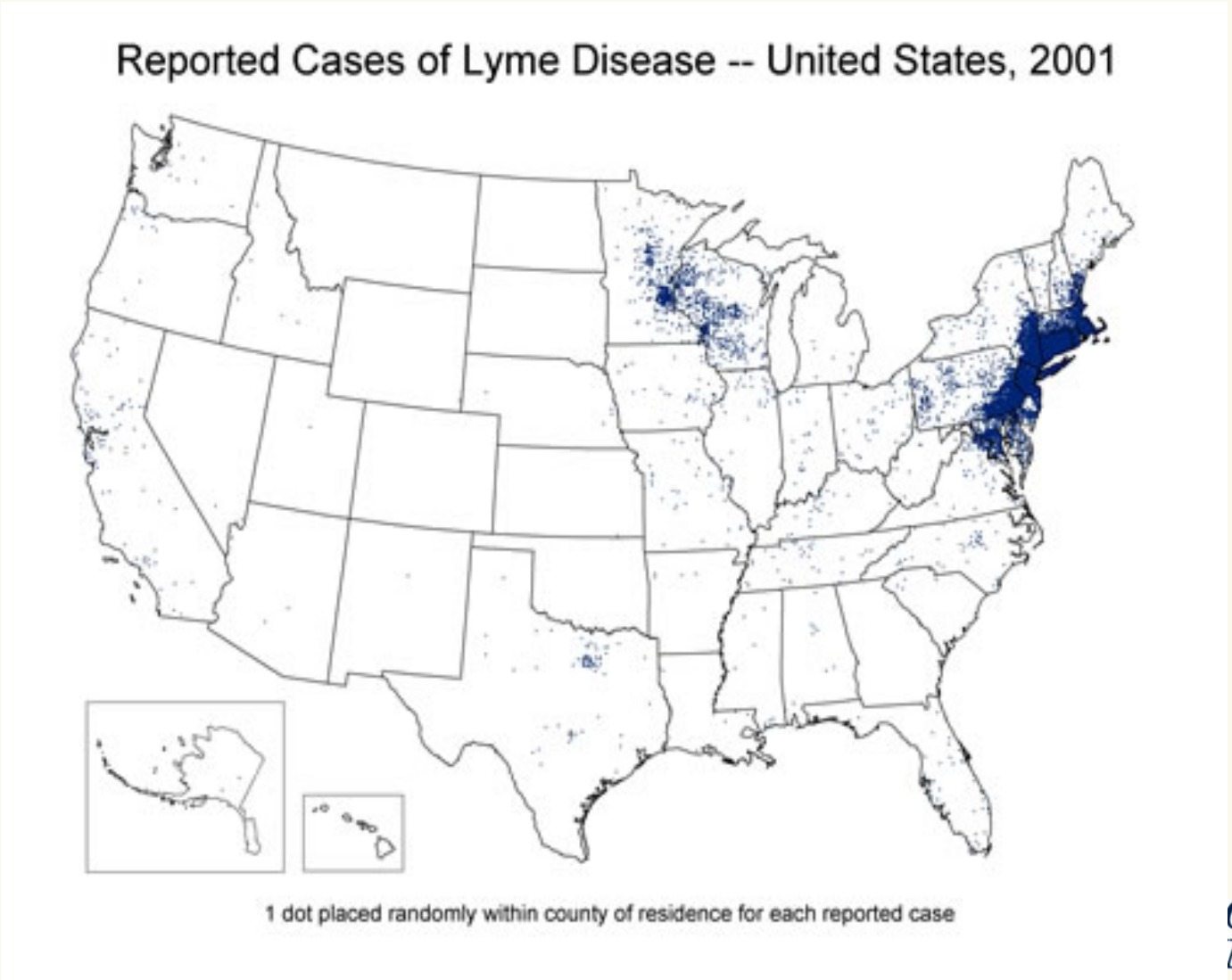
First identified in Lyme, Connecticut, Lyme disease is a bacterial illness that can cause fever, fatigue, joint pain, and skin rash, as well as more serious joint and nervous system complications. It is the most common vector-borne disease in the United States.

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)

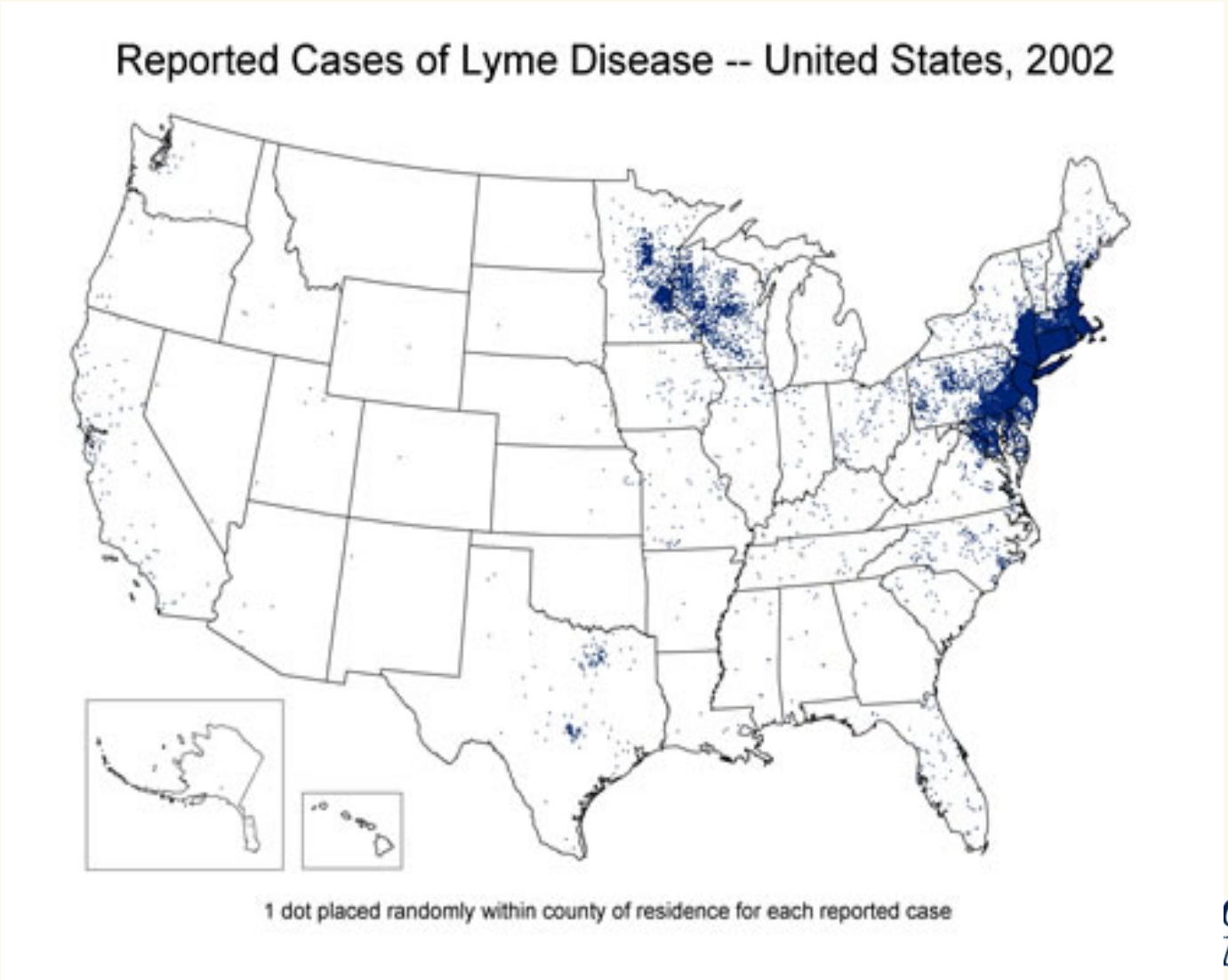
GEORGETOWN UNIVERSITY

Lyme Disease

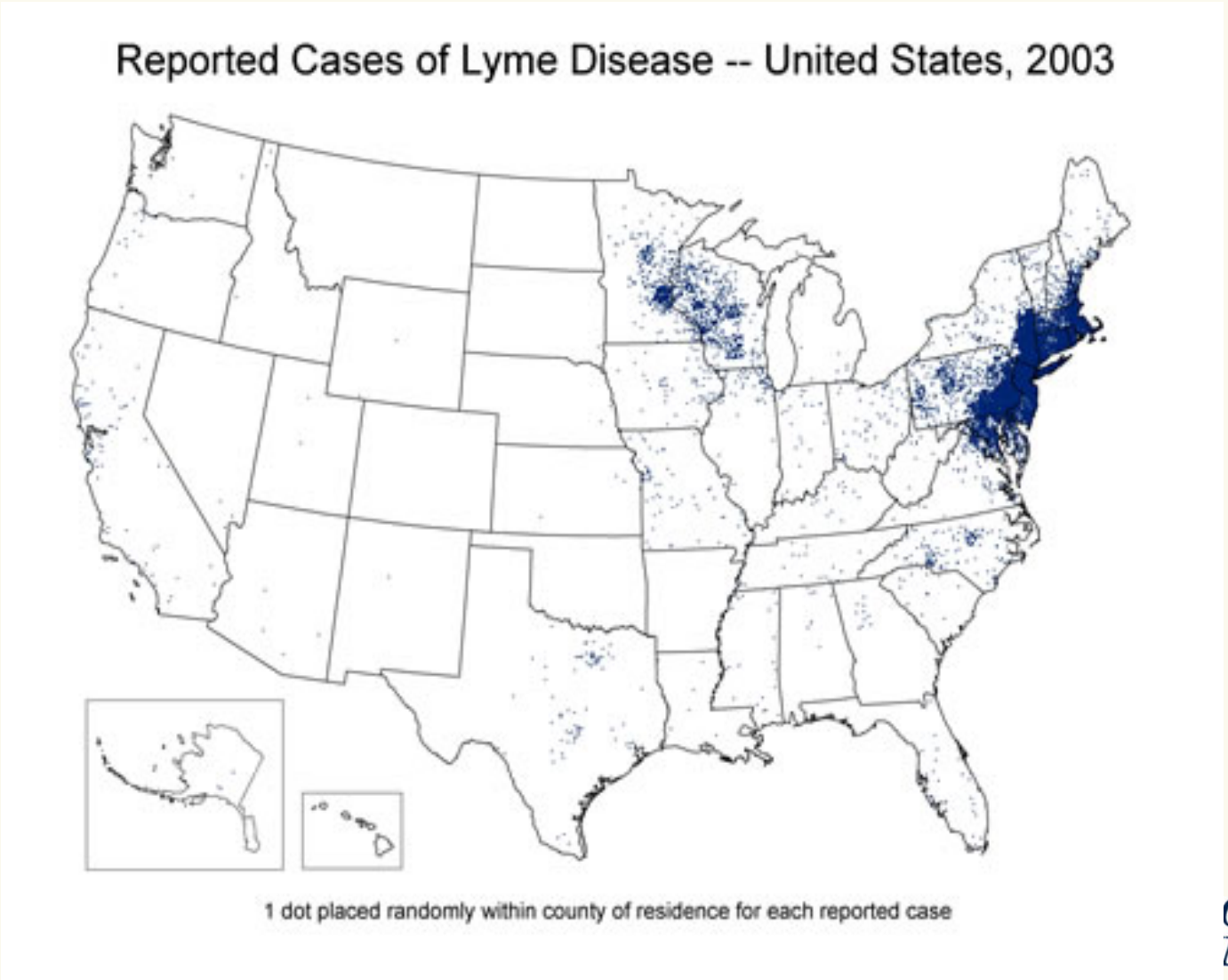


GEORGETOWN UNIVERSITY

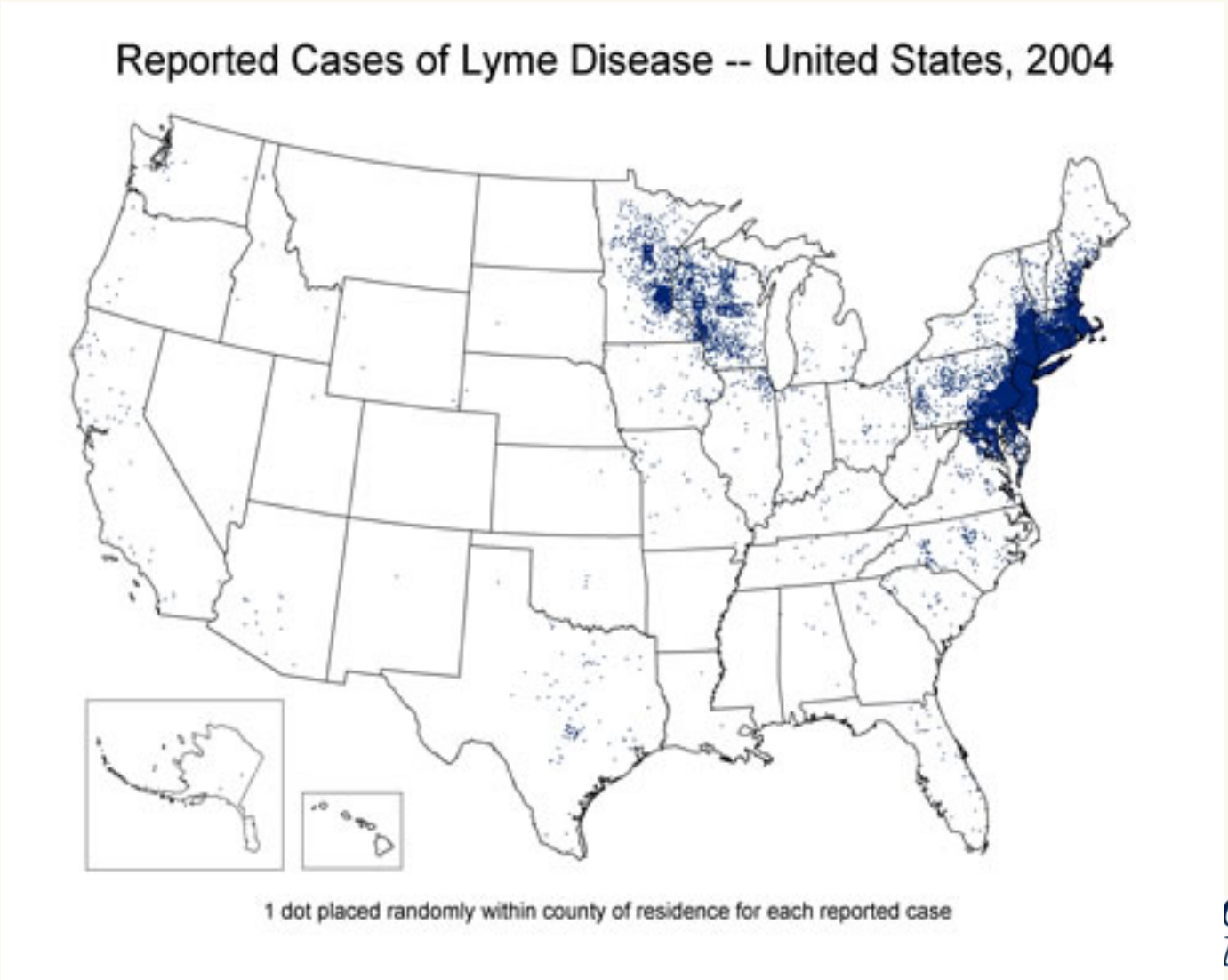
Lyme Disease



Lyme Disease

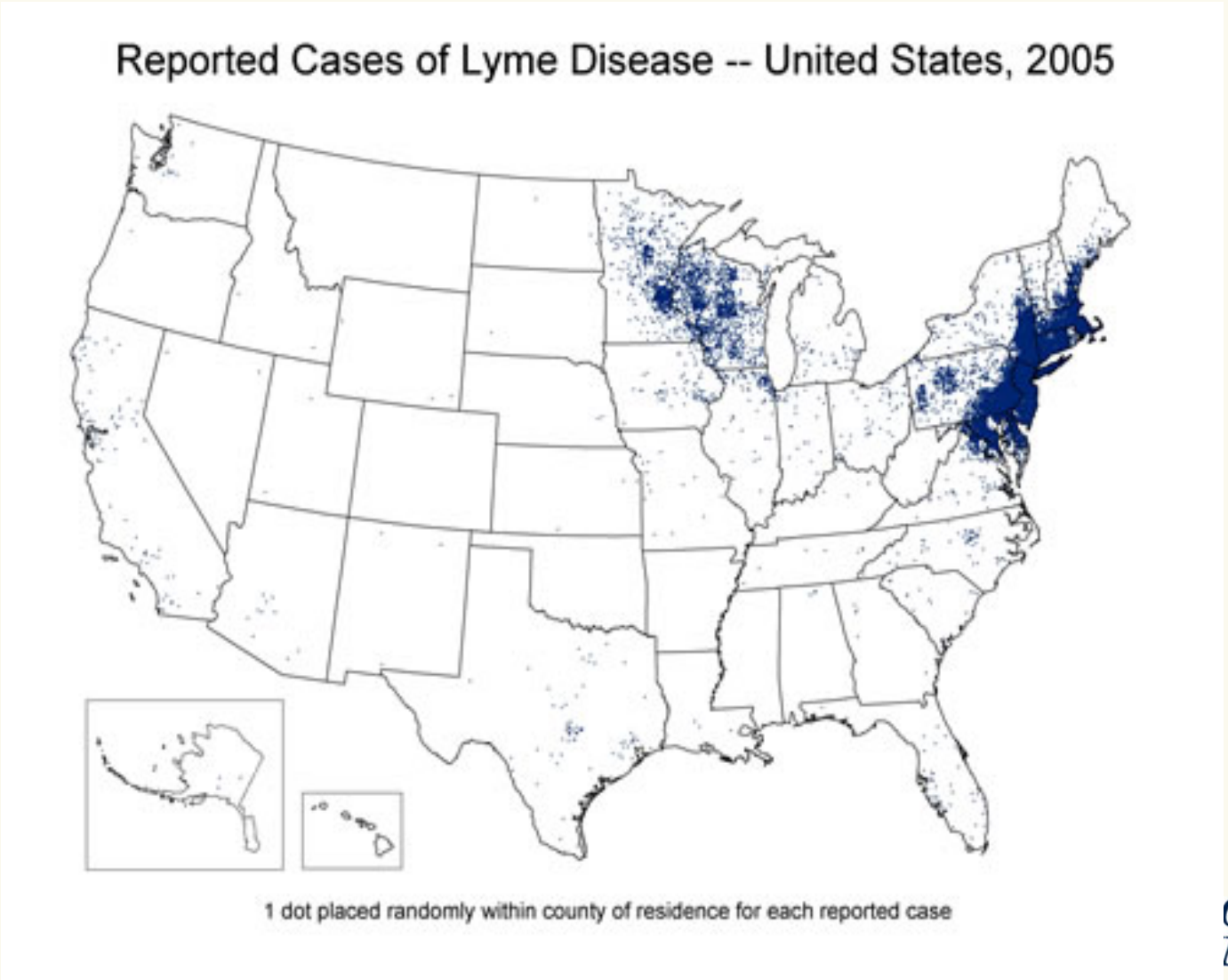


Lyme Disease

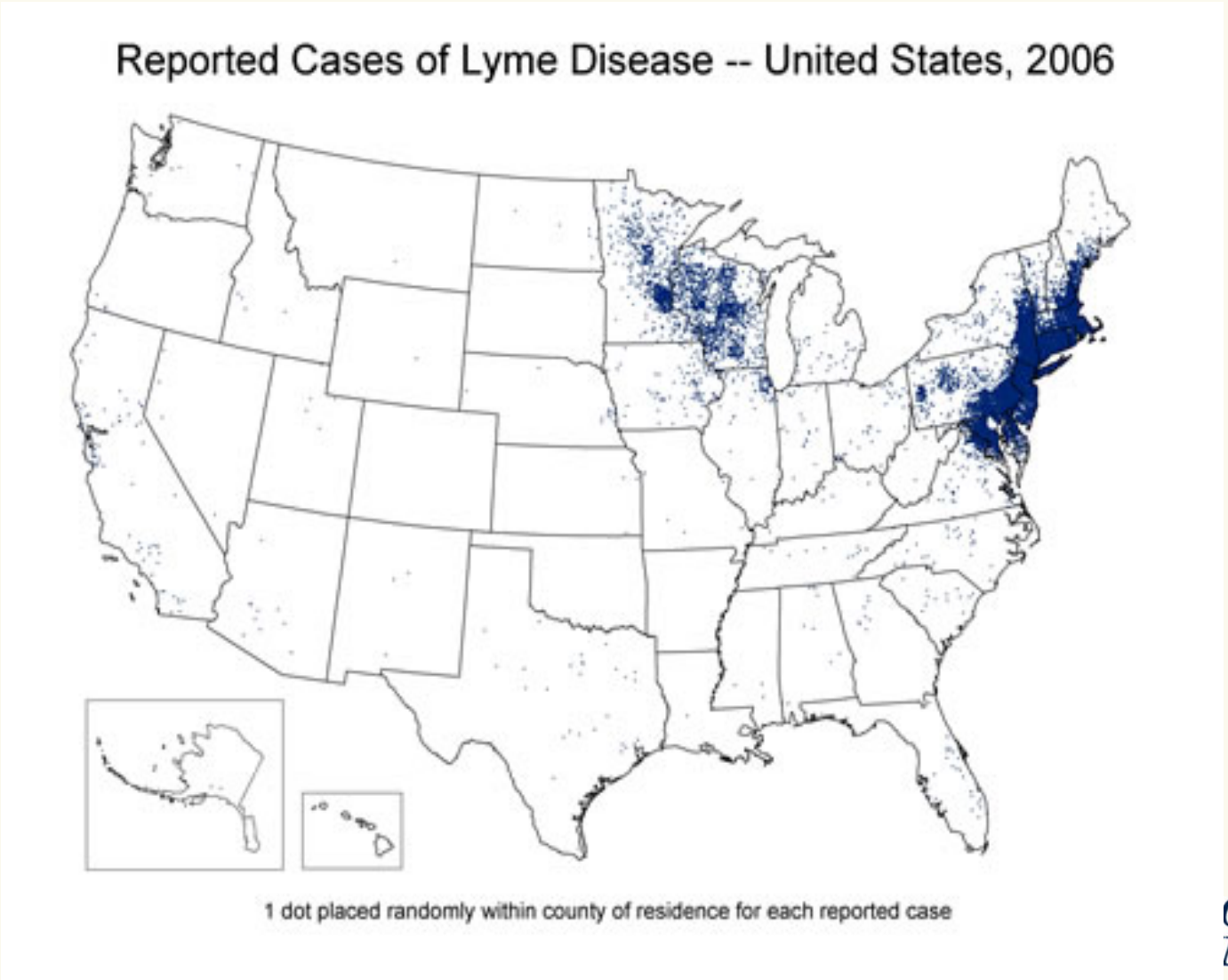


ORGETOWN
NIVERSITY

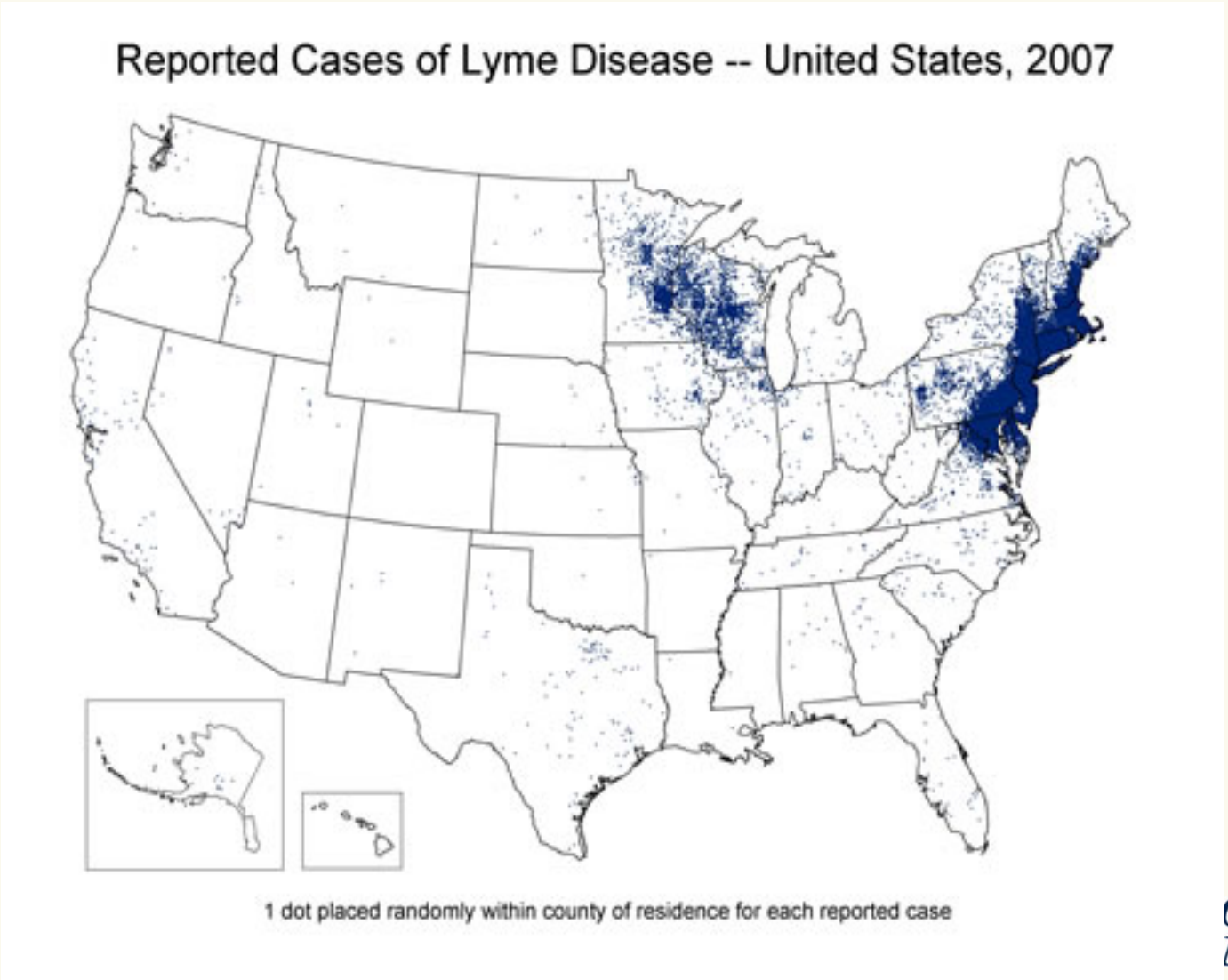
Lyme Disease



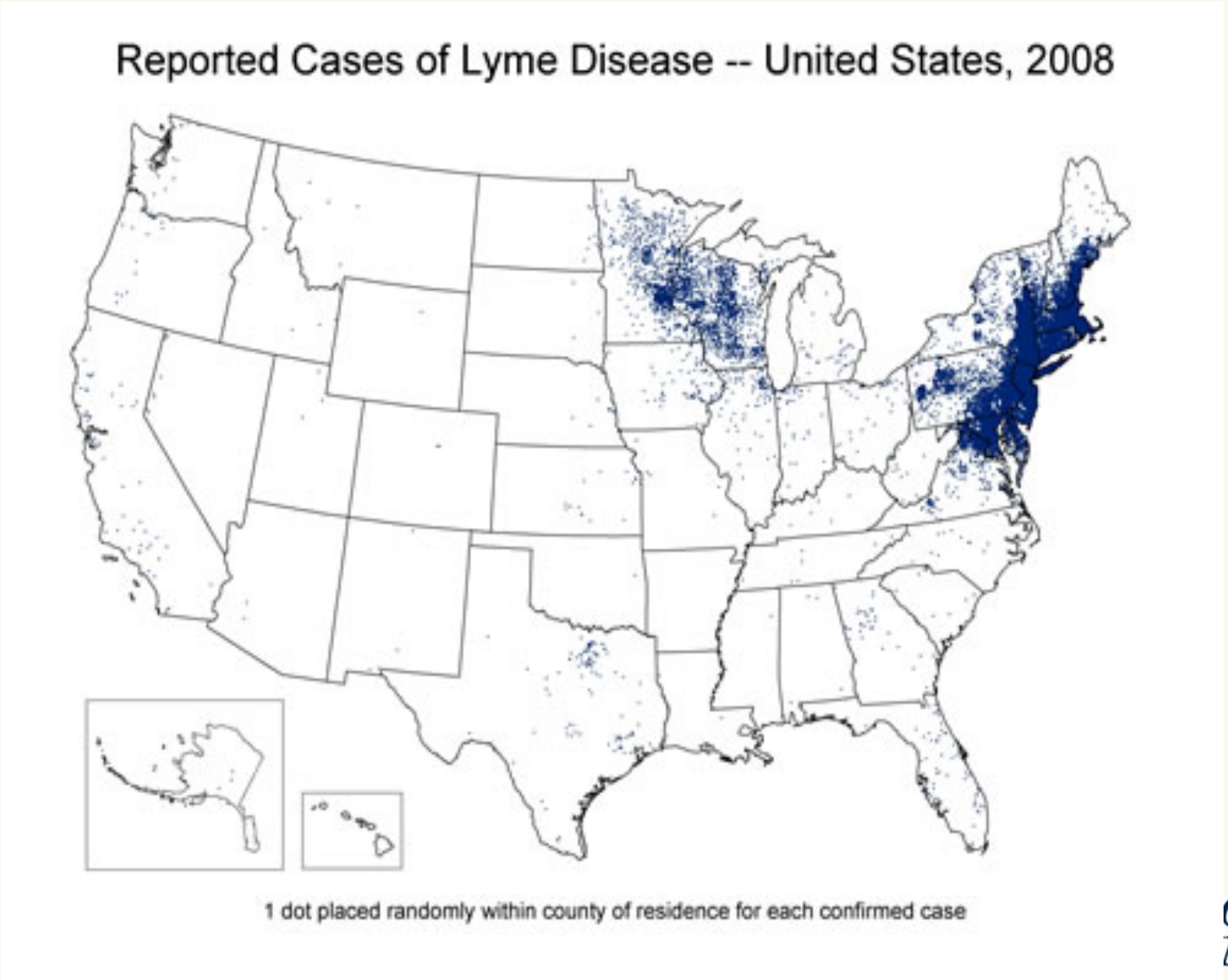
Lyme Disease



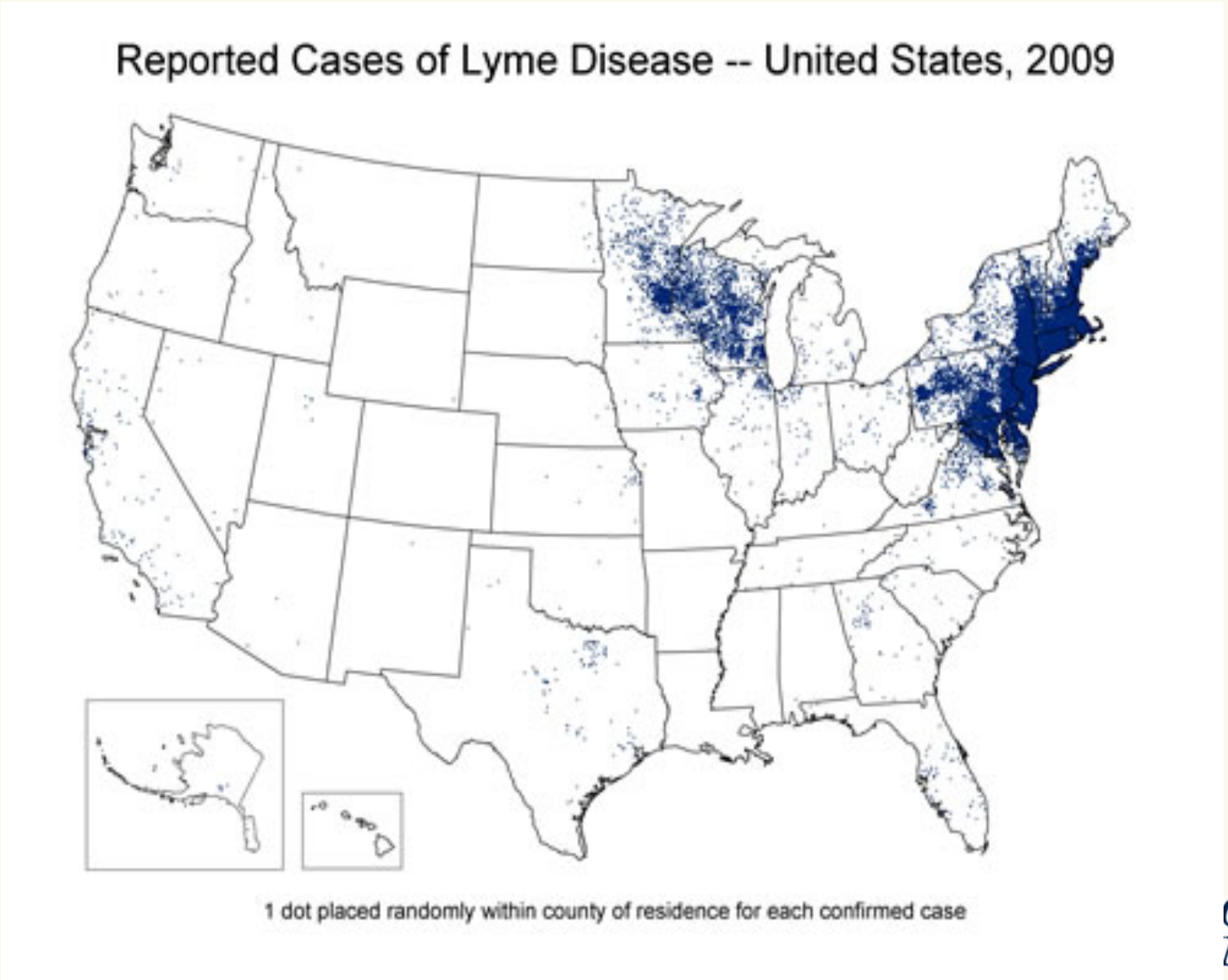
Lyme Disease



Lyme Disease

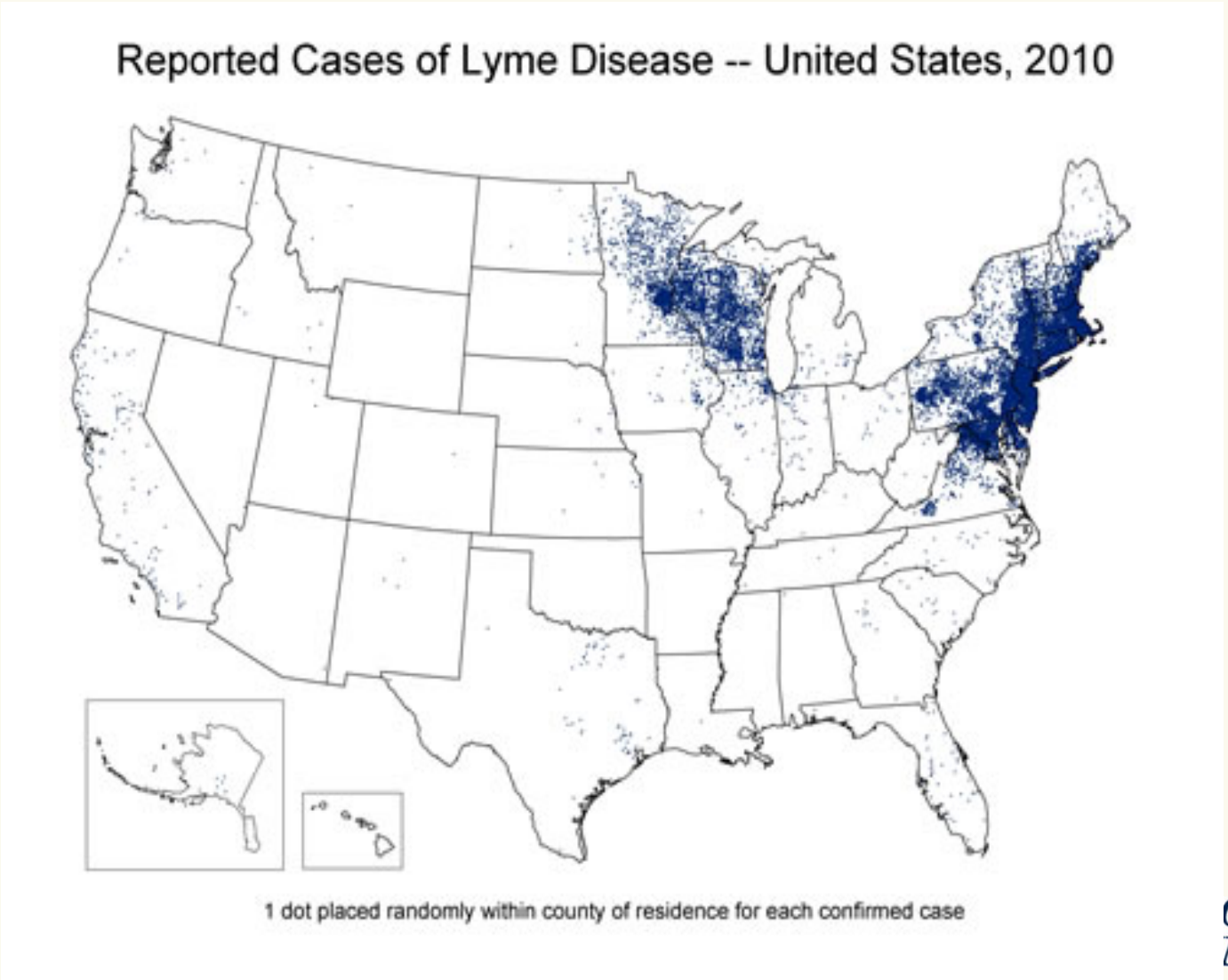


Lyme Disease

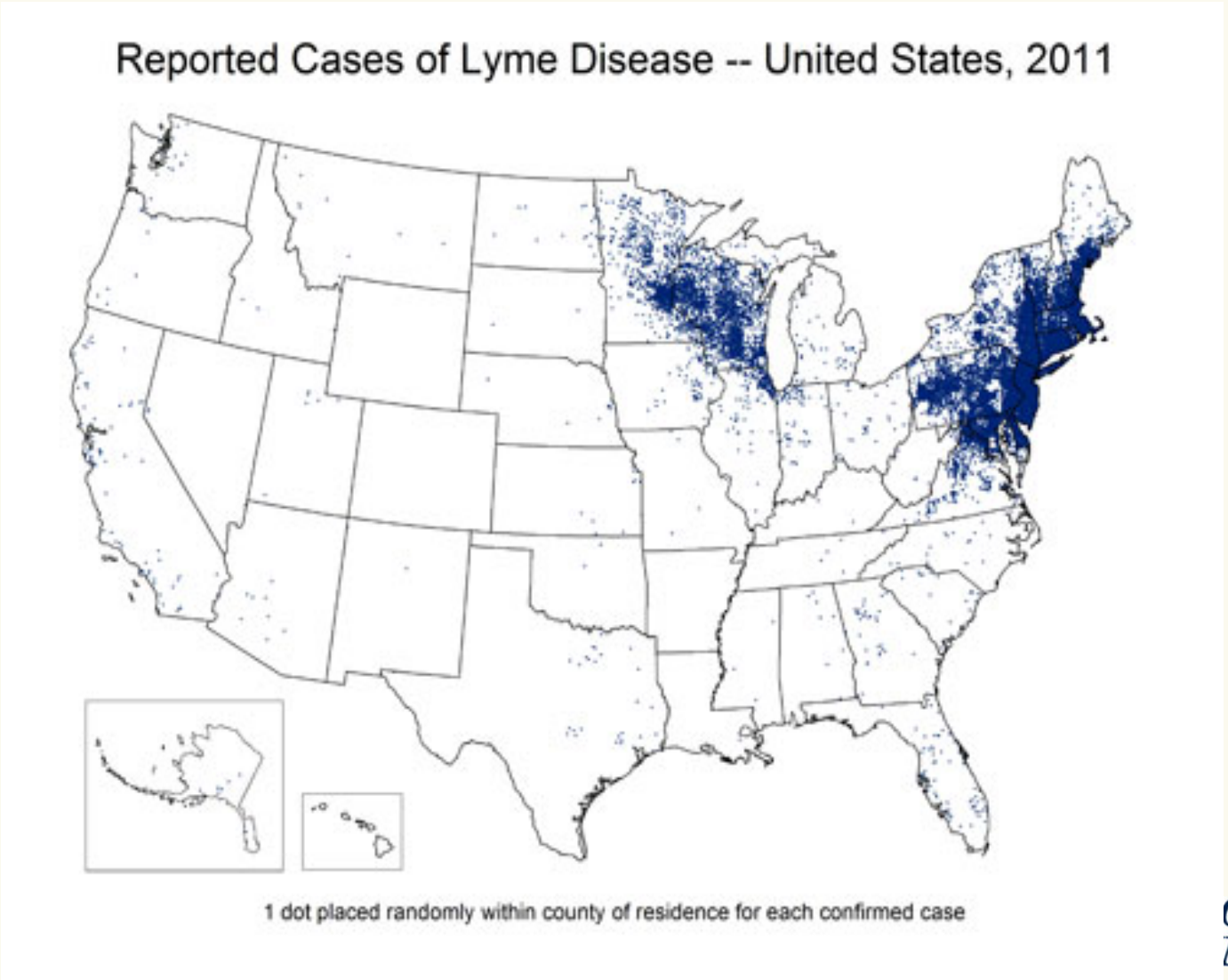


ORGETOWN
NIVERSITY

Lyme Disease

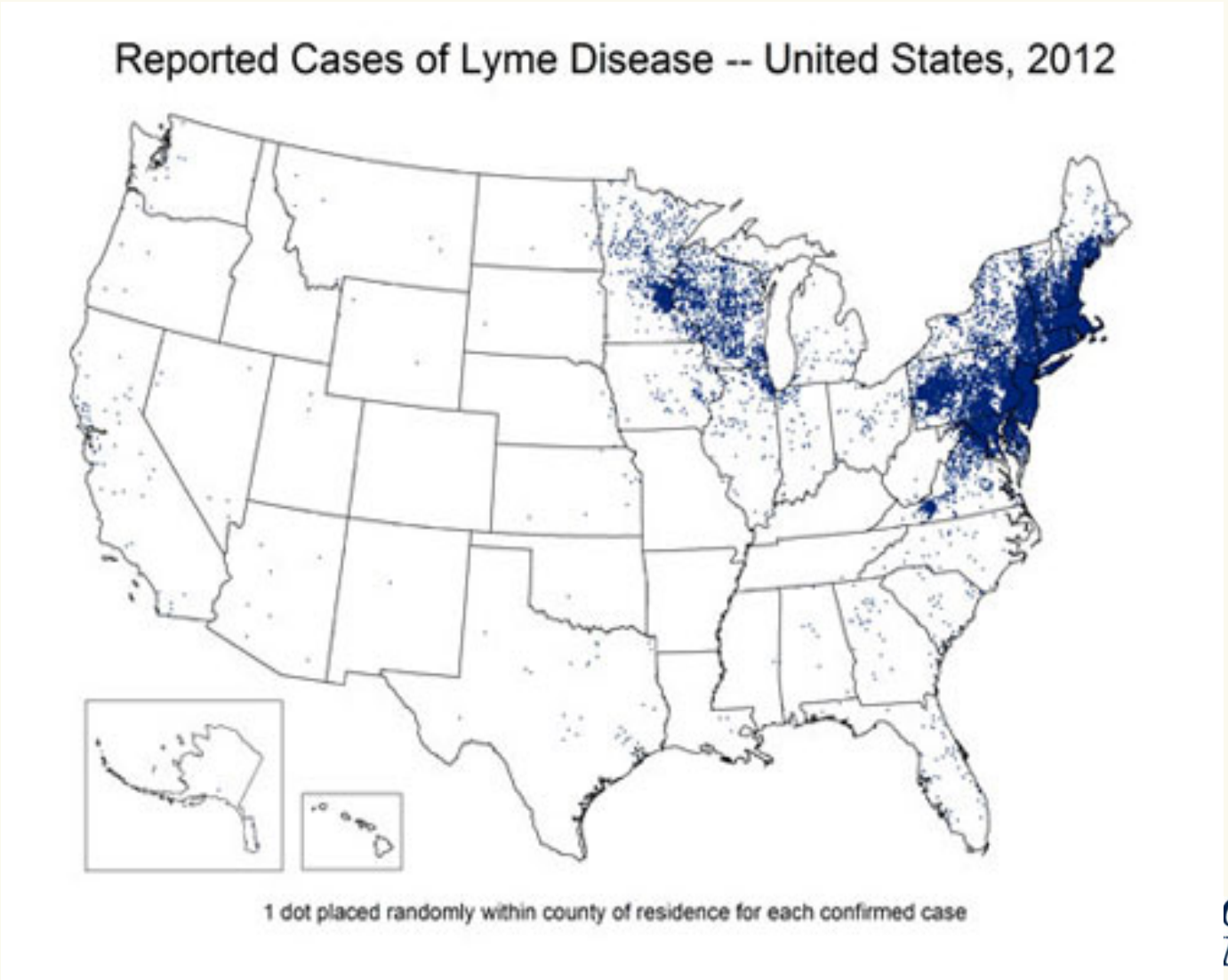


Lyme Disease



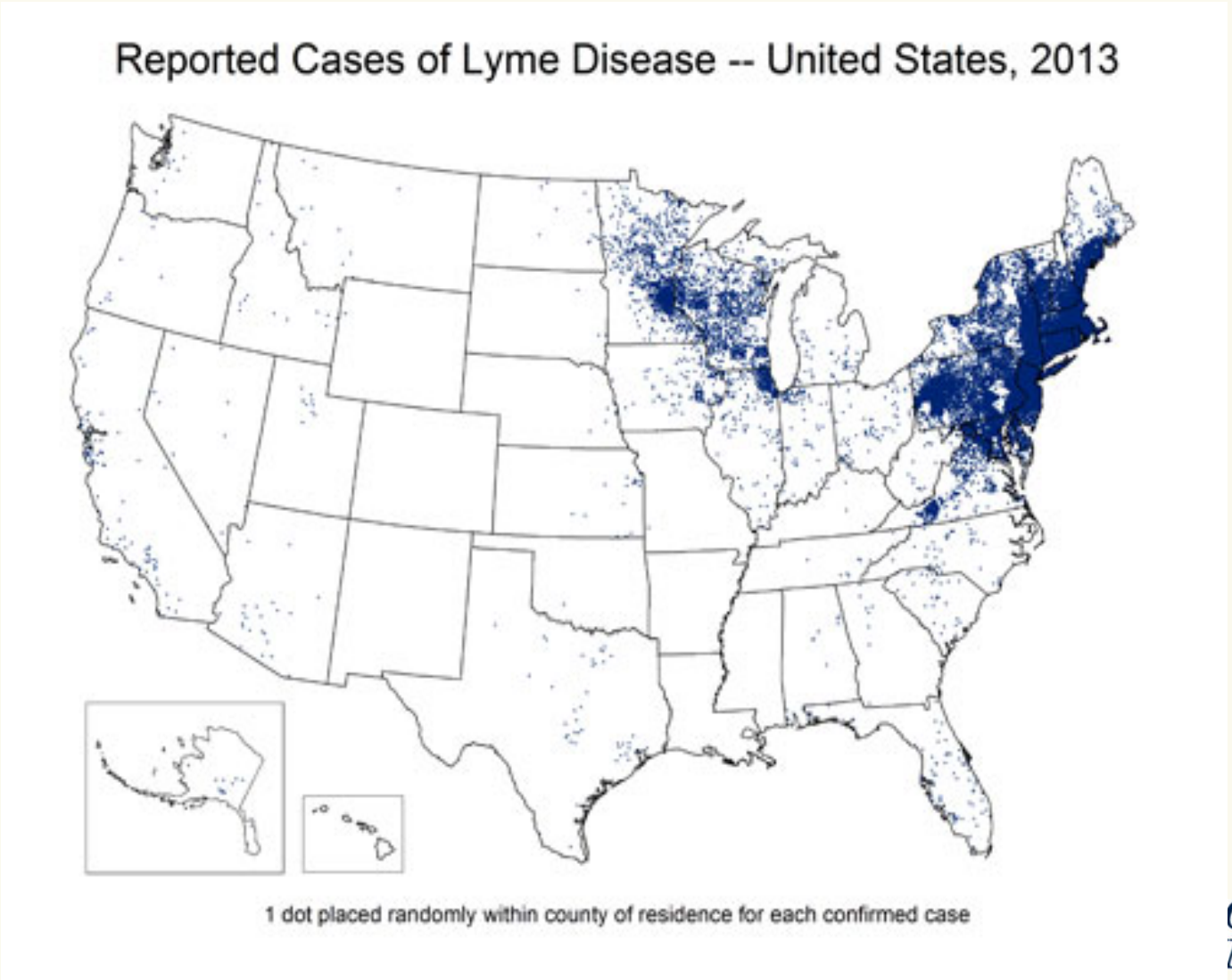
GEORGETOWN UNIVERSITY

Lyme Disease



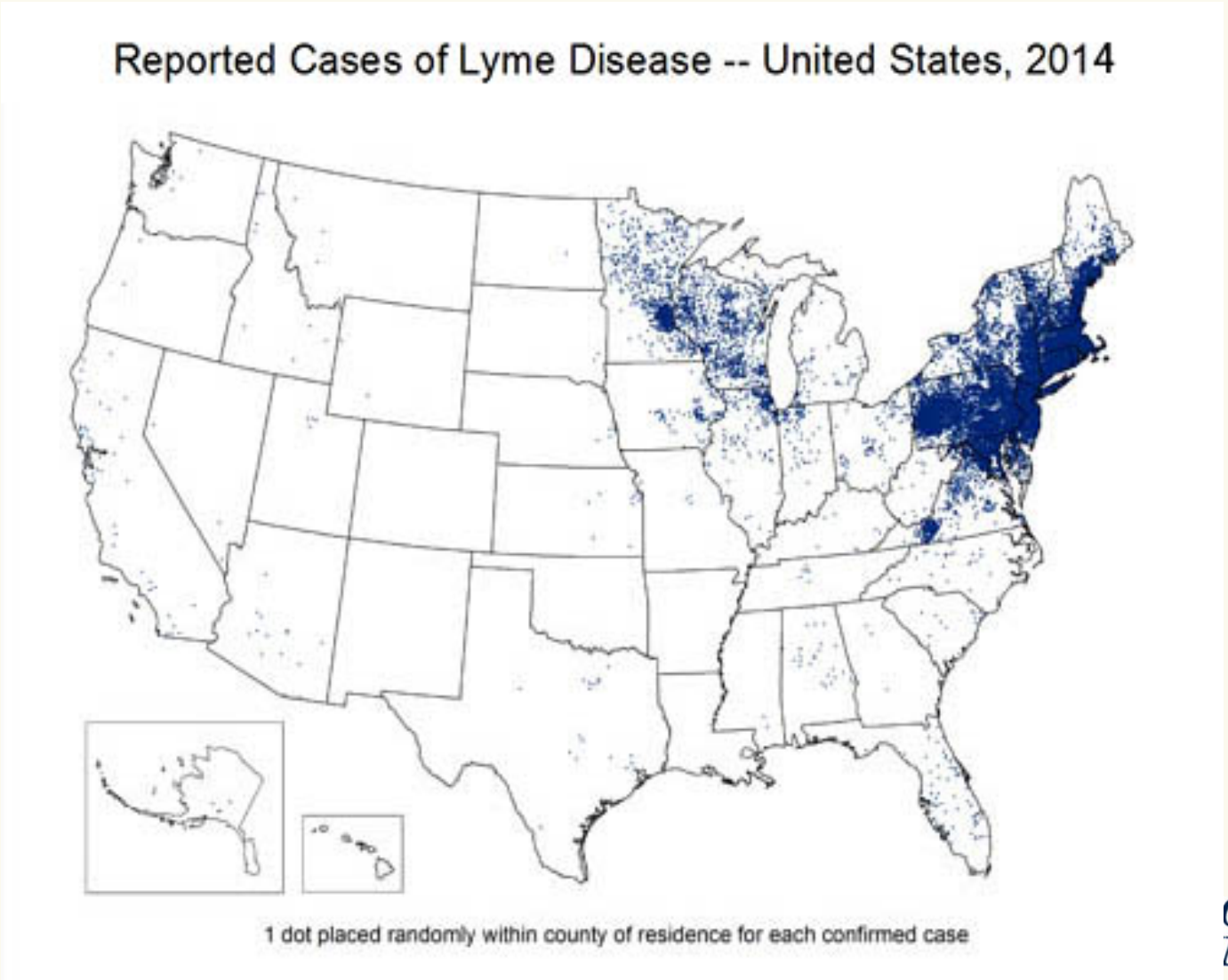
ORGETOWN
NIVERSITY

Lyme Disease

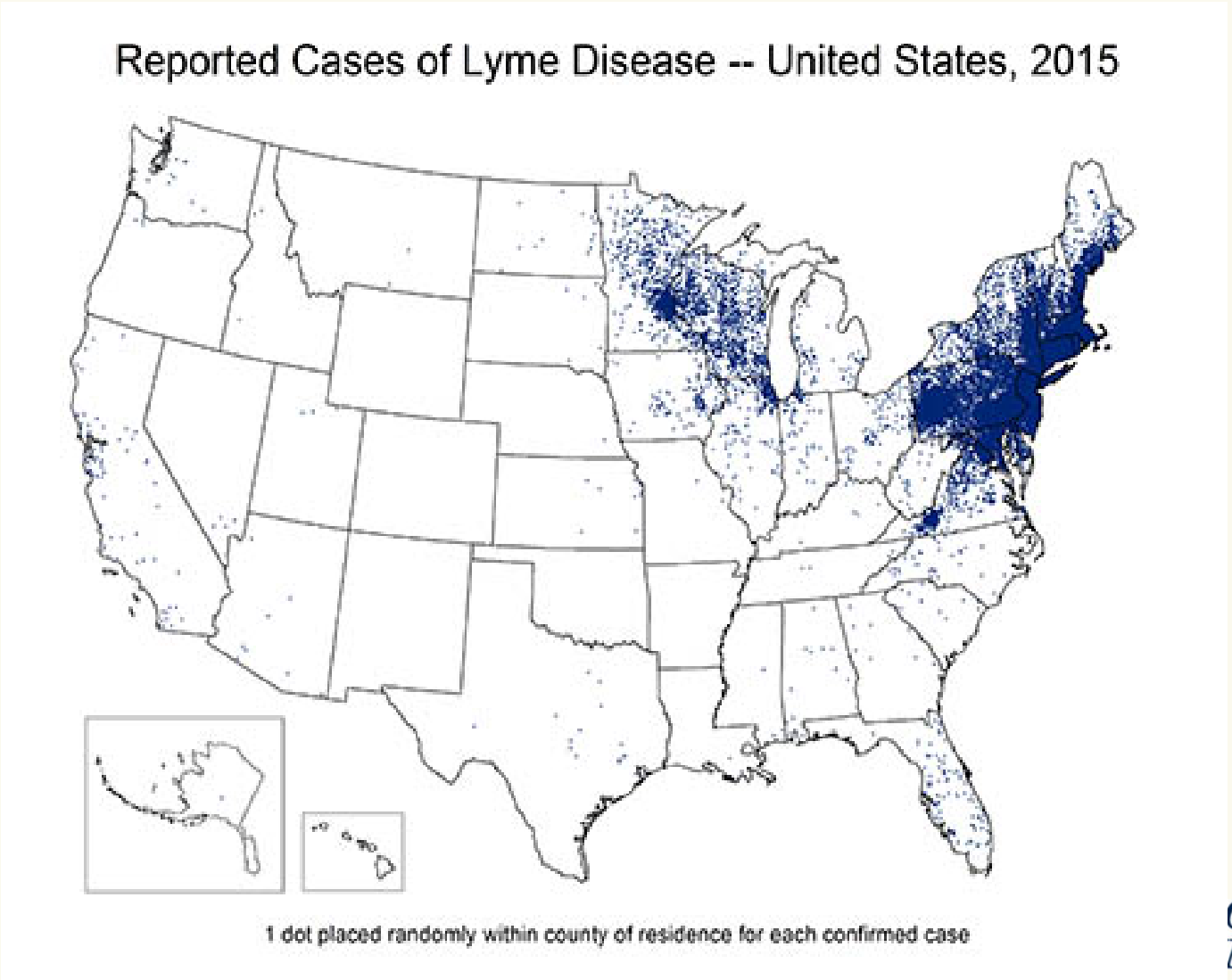


ORGETOWN
NIVERSITY

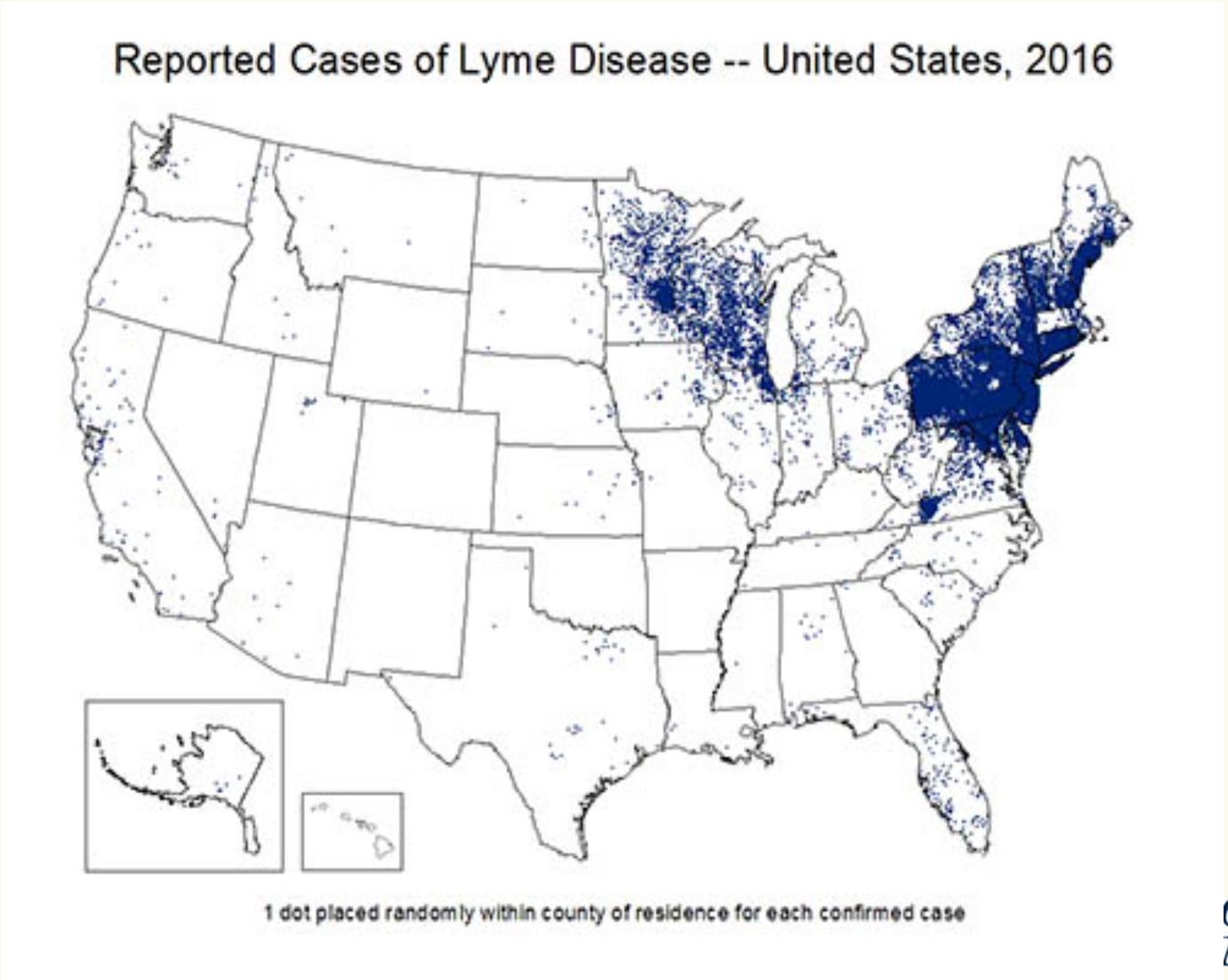
Lyme Disease



Lyme Disease

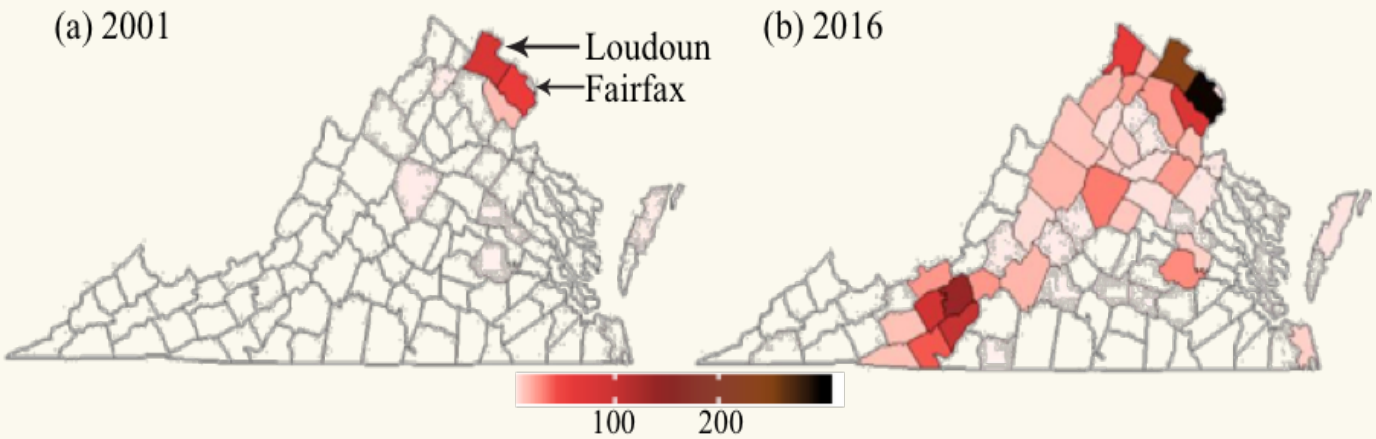


Lyme Disease



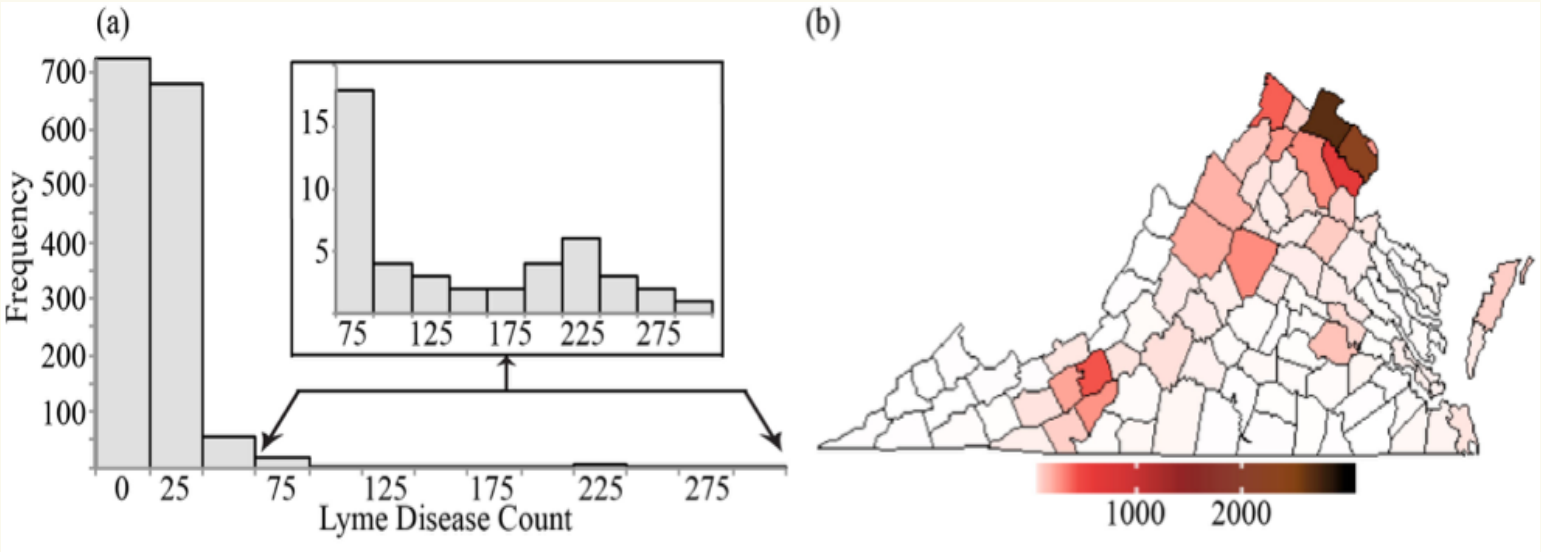
Lyme Disease in Virginia

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



GEORGETOWN UNIVERSITY

Lyme Disease in Virginia



GEORGETOWN UNIVERSITY

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!

GEORGETOWN UNIVERSITY

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:

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

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

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

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

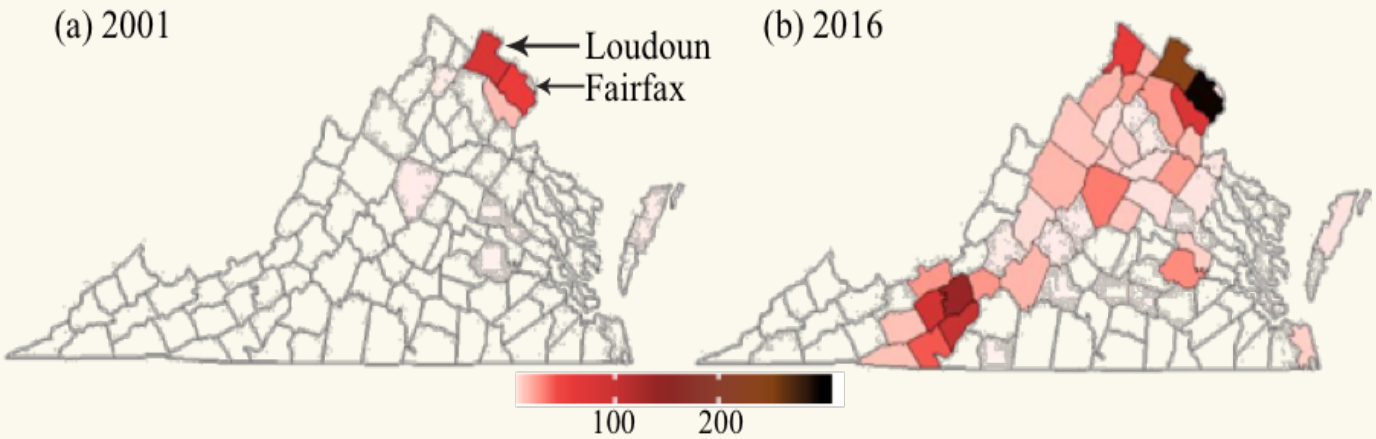
Case Studies

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).

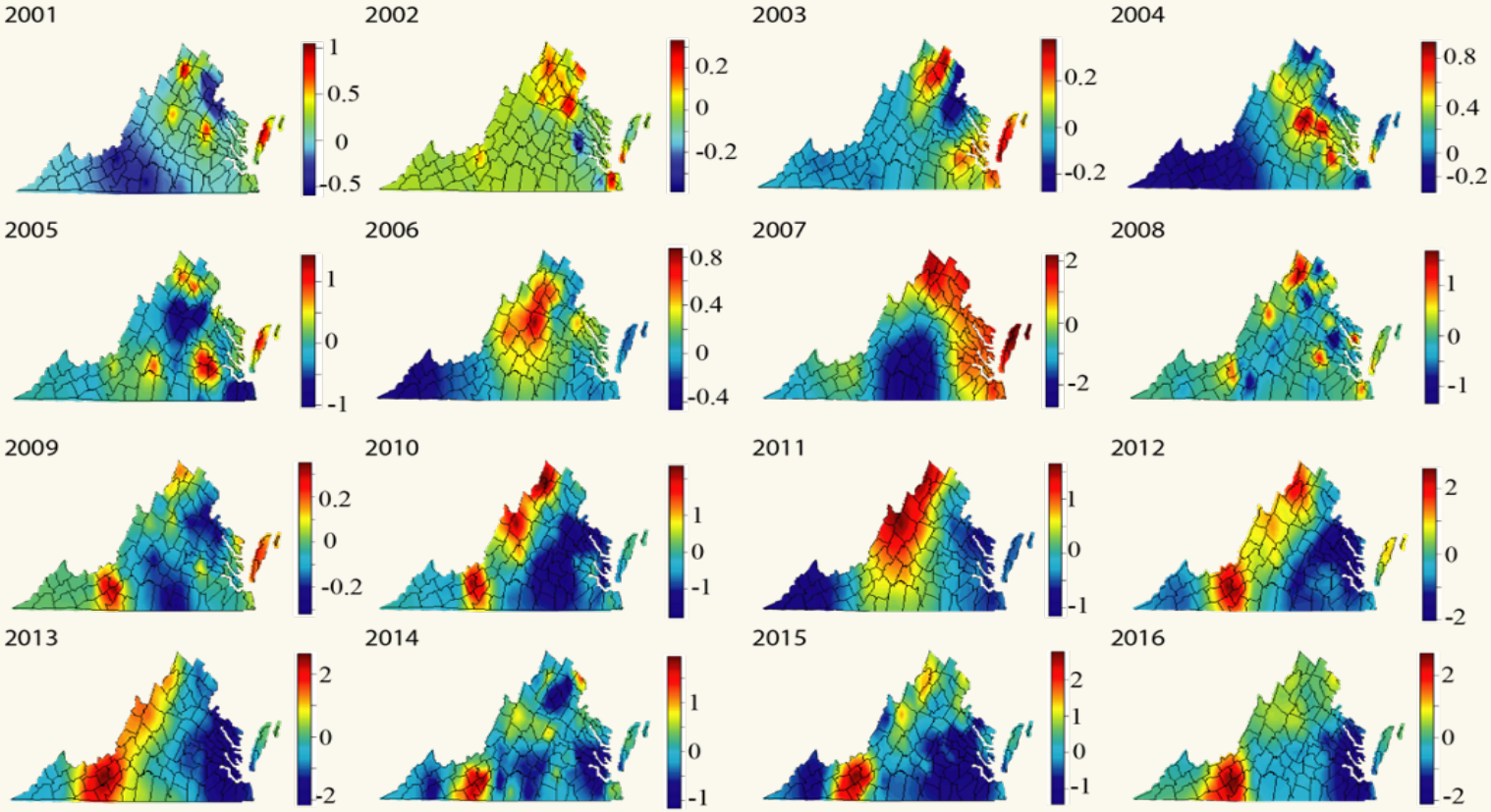
Both are introduced in the most populous part of the state, Northern Virginia, and then start to spread to other parts.

Spatial Variability of Lyme in Virginia



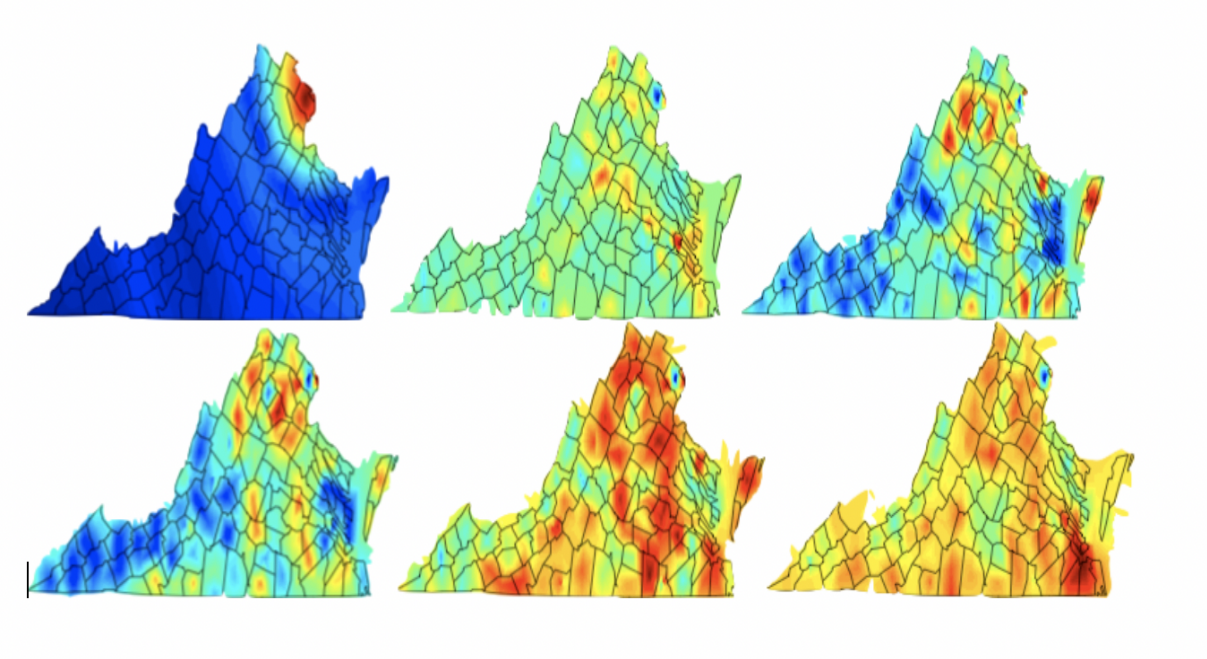
GEORGETOWN UNIVERSITY

Spatio-Temporal Variability of Lyme in Virginia



GEORGETOWN UNIVERSITY

Spatio-Temporal Variability of COVID-19 in Virginia

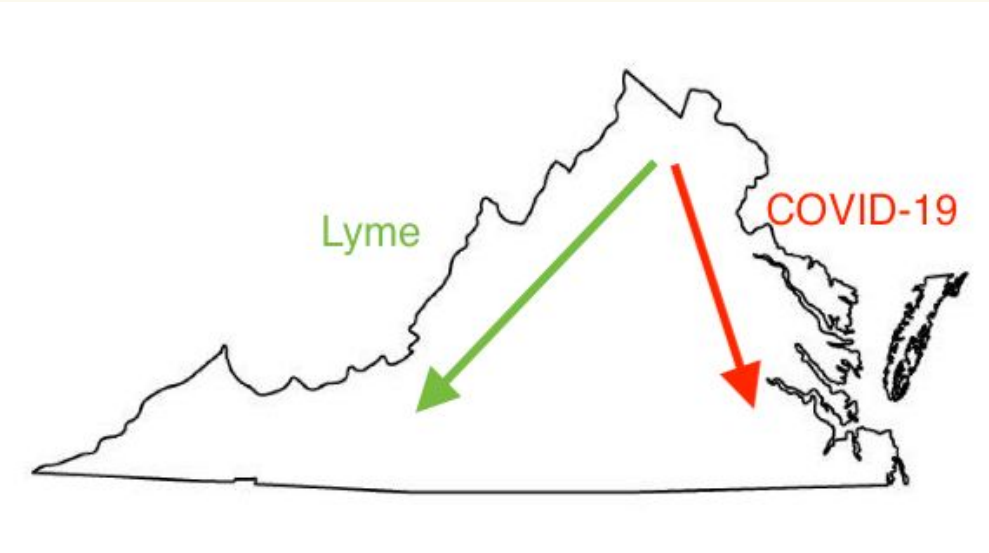


(Weeks 1, 4, 8, 12, 16, 20; Weekly values between March 8-July 19.)

GEORGETOWN UNIVERSITY

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)

GEORGETOWN UNIVERSITY

Case Studies

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

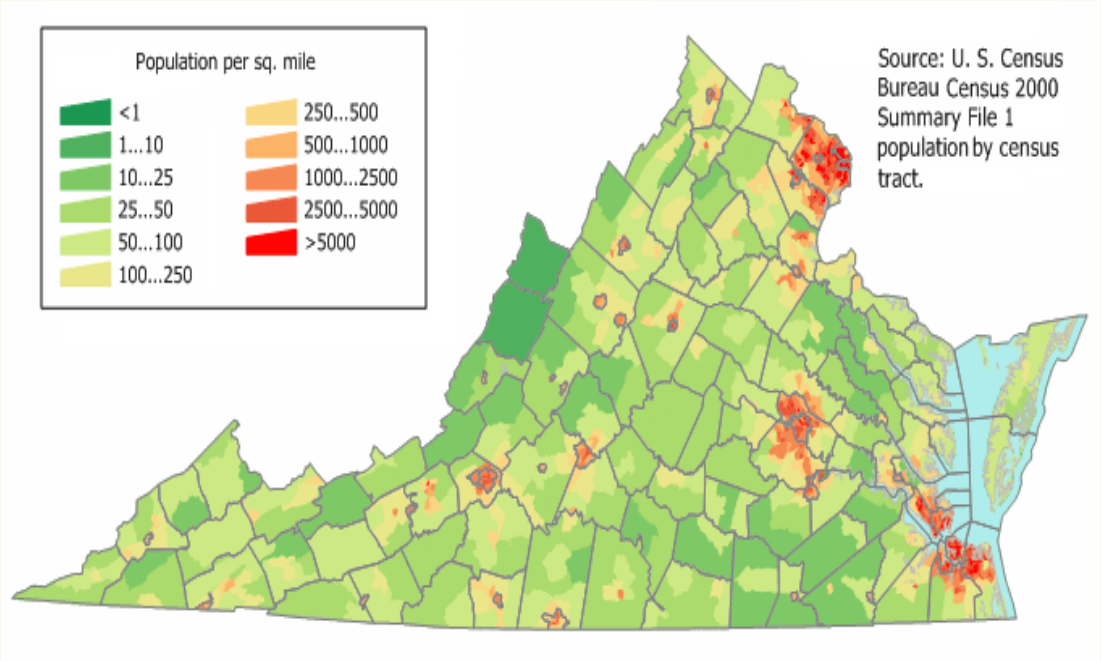
Not much is known about important factors in spread of COVID-19 beyond population (human-to-human transmission and related effects such as travel, etc.).

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.).

GEORGETOWN UNIVERSITY

Discussion: COVID-19

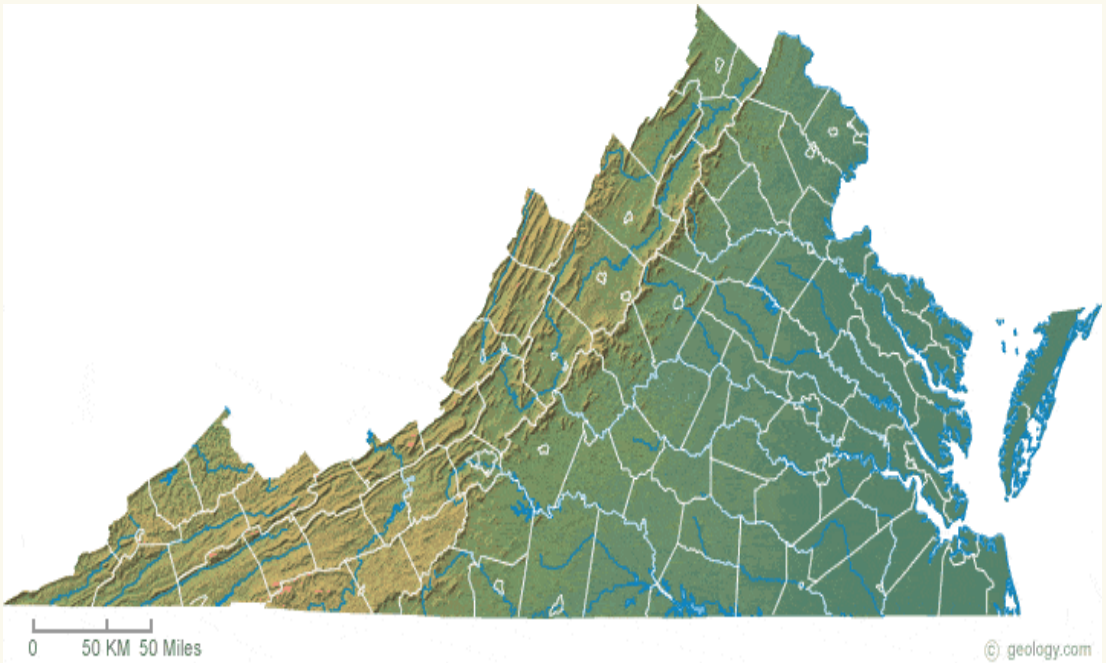
Population Density



GEORGETOWN UNIVERSITY

Discussion: Lyme

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



GEORGETOWN UNIVERSITY

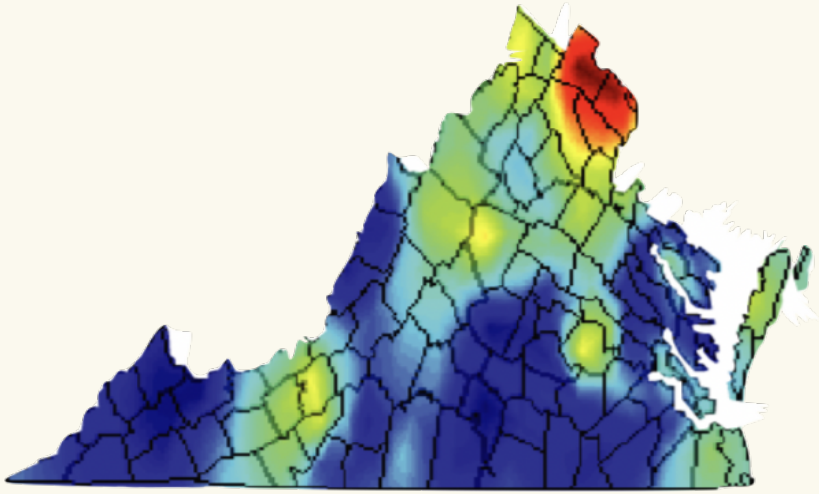
Modeling

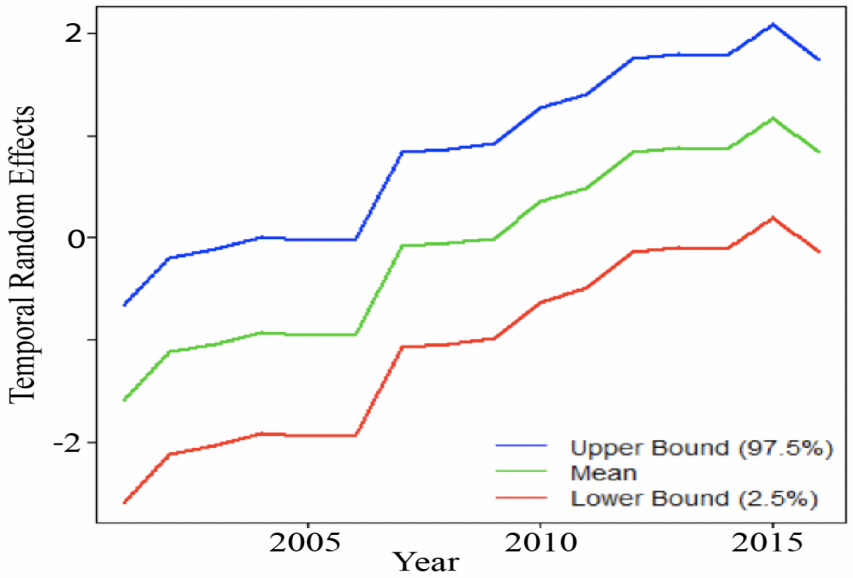
To better understand the spatial and temporal variabilities, we fitted negative binomial hurdle models with autoregressive error terms to both data sets.

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.

GEORGETOWN UNIVERSITY

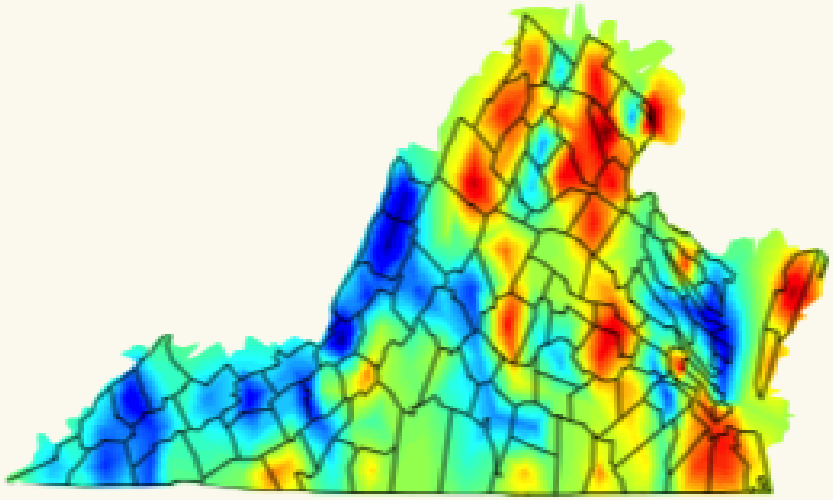
Lyme: Spatial+Temporal Effects Posterior Means

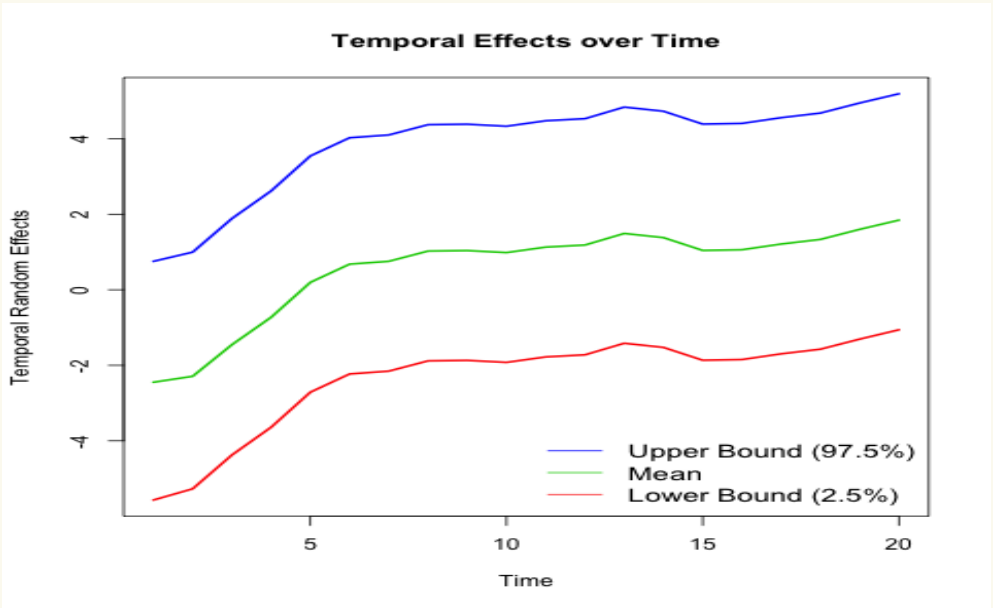




GEORGETOWN UNIVERSITY

COVID-19: Spatial+Temporal Effects Posterior Means





GEORGETOWN UNIVERSITY

Modeling

...But how can we go about modeling and forecasting the dynamics at the early stages of spread of the disease in either of these cases?

GEORGETOWN UNIVERSITY

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):

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

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

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.

GEORGETOWN UNIVERSITY

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!

GEORGETOWN UNIVERSITY

Model Implementation & Results

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

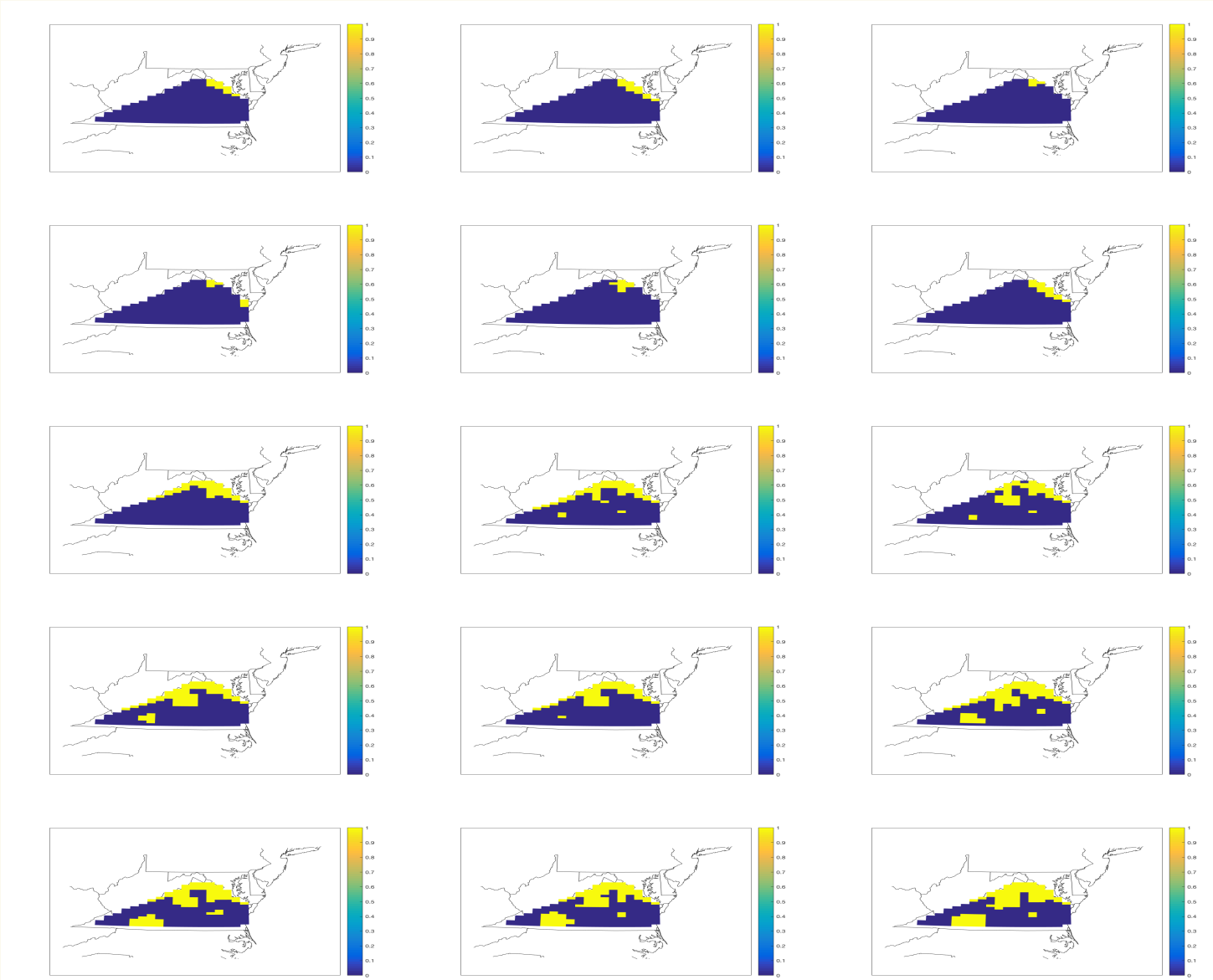
- Metropolis-Hastings within Gibbs Sampling,
- 20,000 iterations, and 2000 burn-in period, and quick convergence.

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

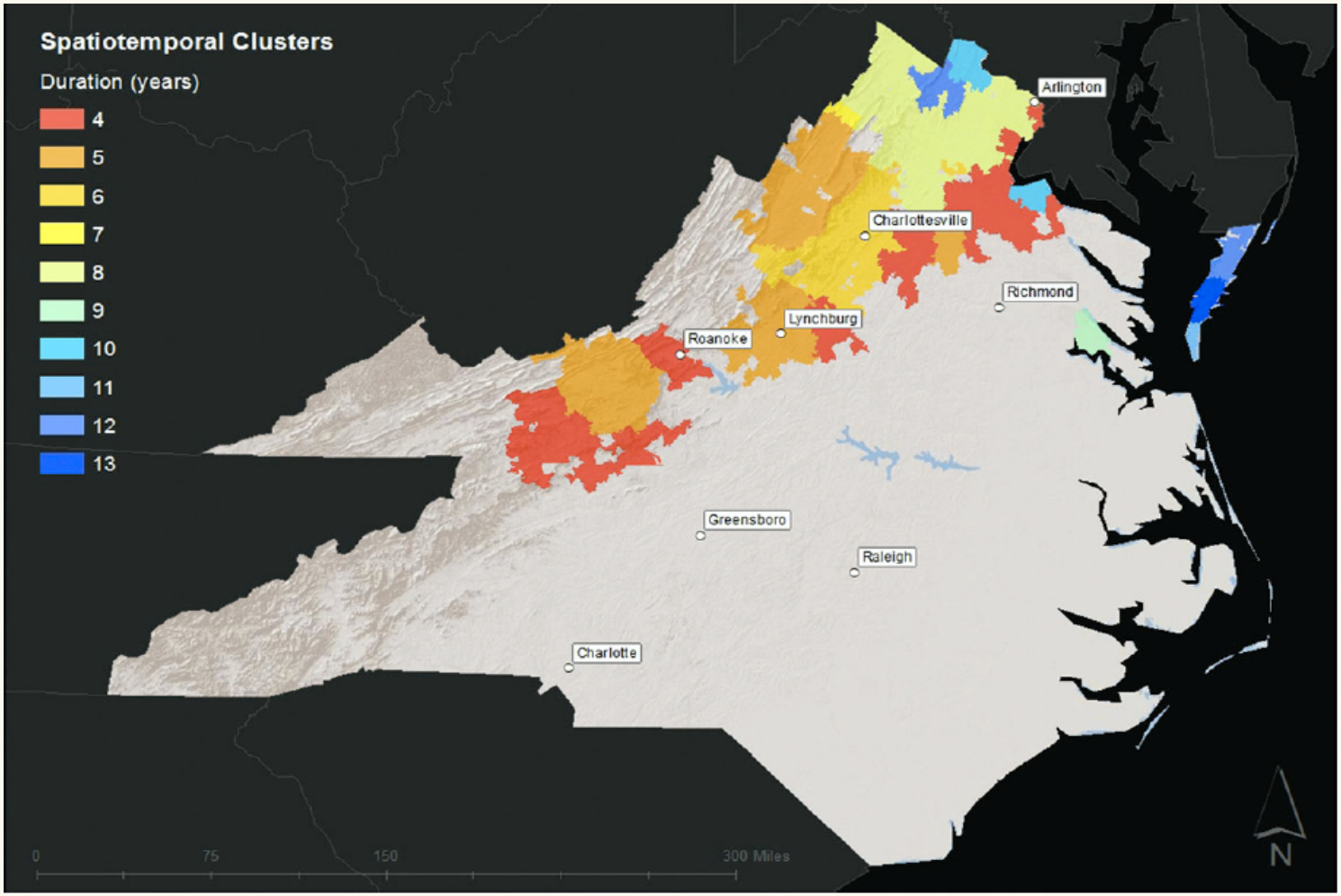
Table: Posterior results for the diffusion parameter of the PDE-based model

GEORGETOWN UNIVERSITY

Posterior Means: Presence/Absence



Spread of Lyme



From Lantos et al. (2015)

GEORGETOWN UNIVERSITY

Case Study 2: Forced Migration

ISIL (ISIS) and the 2013-2017 Iraq Civil War:

In late 2013, tensions between Sunni and Shia Muslims in the Anbar province of Iraq boiled over into violence between the Shia government and its allies against the Islamic State. The violence led to significant internal displacement.

In the absence of traditional migration variables, we consider the use of social media for capturing indirect indicators of migration. Specifically, we focus on novel conversation buzz and insecurity predictors based on Twitter data.

GEORGETOWN UNIVERSITY

Displacement Data

We use the flow data for 2016-2017 provided by the International Organization for Migration's Iraq Displacement Tracking Matrix, which gives stocks of displaced families in one of 97 Districts, as well as which of the 18 Governorates they originated from.

The counts are aggregated over variable length periods, with the shortest period consisting of 11 days and longest of 31 days. There are 48 total time periods in our study, leading to a total of 57,024 observations. 96.4% of response values are zero, with a mean of 6.71 families displaced for all time across origin-destination pairs.

GEORGETOWN UNIVERSITY

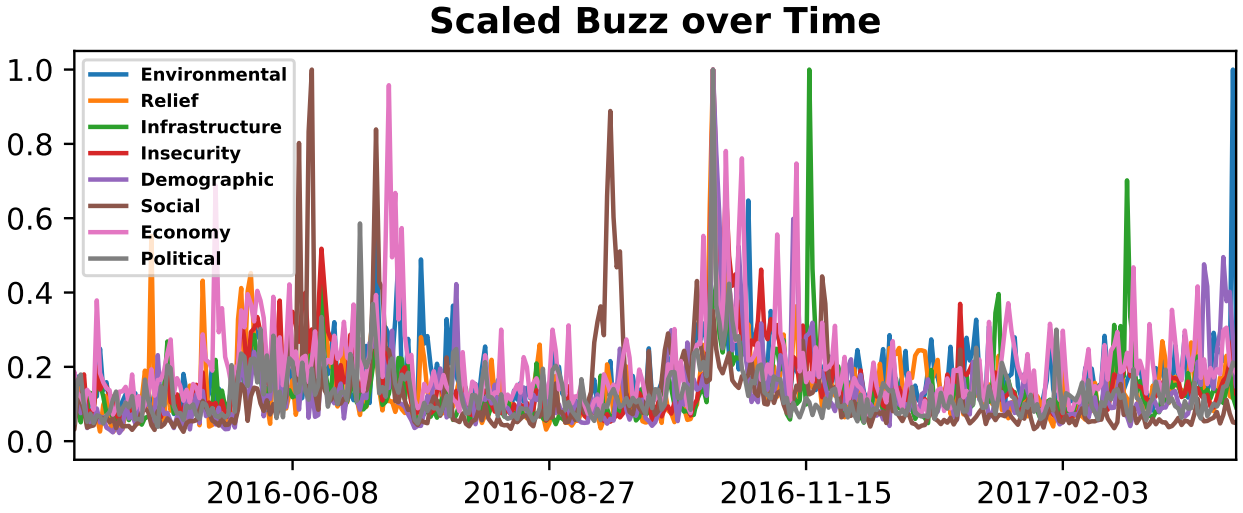
Twitter Data

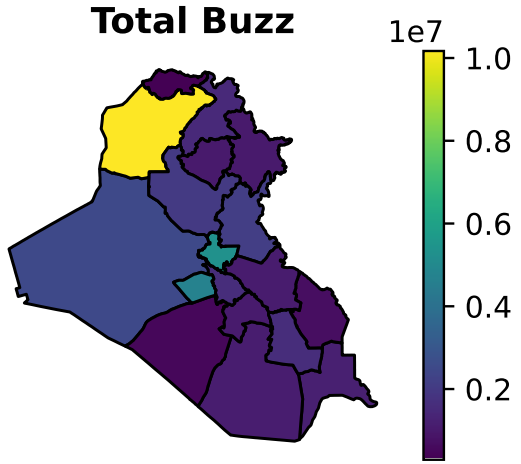
We collected Arabic language tweets between March 2016 and June 2017 from Twitter's Streaming Application Program Interface (API).

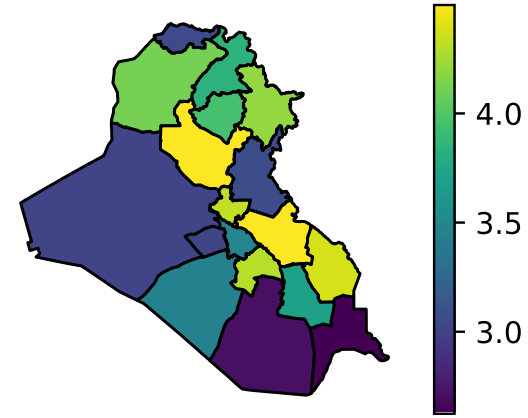
Buzz variables were formed by monitoring counts of keywords selected by Arabic language experts related to various migration-relevant topics, such as ethnic angst and discussion about the economy. We also formed sentiment indicators. Variables such as Social buzz for different demographic groups, (e.g. Social-shia buzz), Insecurity, (e.g. Insecurity-emotions), Environmental (e.g. Environmental-weather) are examples of the 68 variables identified.

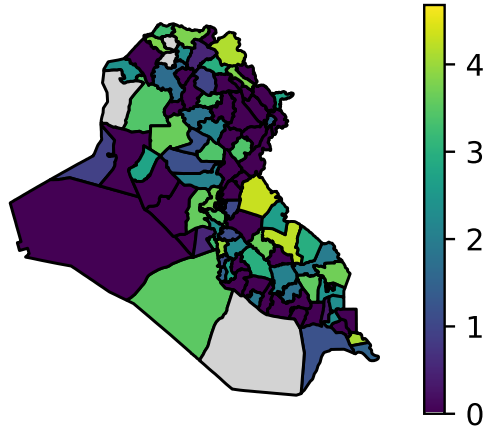
GEORGETOWN UNIVERSITY

Twitter Data









GEORGETOWN UNIVERSITY

Spatio-temporal ZINB Gravity Model

Extends classic Gravity Models to account for complex count data with spatio-temporal variability.

$$y_{t,o,d} \sim \begin{cases} \delta_0 & \text{w.p. } \pi \\ \text{NegBin}(g(\mathbf{z}_{t,o,d}(\mu_l, \sigma_l)^\top \beta + \omega_t + \omega_o + \omega_d), \frac{l_t}{\sigma^2}) & \text{o.w.} \end{cases}$$

$$\boldsymbol{\omega}_j | \phi_j, \rho_j^2 \sim \text{N}\left(\mathbf{0}, \rho_j^2 \Sigma_j(\phi_j)\right) \text{ for } j \in \{\mathscr{T}, \mathscr{O}, \mathscr{D}\}$$

$$\pi \sim \text{logitN}(a_\pi, b_\pi)$$
$$\phi_j \sim \text{logitN}(a_\phi, b_\phi)$$
$$\rho_j^2 \sim \text{logN}(a_\rho, b_\rho)$$
$$\beta_p | \lambda_p \overset{\text{indep}}{\sim} \text{L}(0, \tfrac{1}{\tau\lambda_p})^+$$
$$\lambda_p \overset{\text{iid}}{\sim} \text{C}(0,1)^+$$

GEORGETOWN UNIVERSITY

Refresher on Adaptive Lasso

Given Penalized Parameters $\boldsymbol{\beta}$, Unpenalized Parameters $\boldsymbol{\theta}$, Penalty coefficients $\boldsymbol{\lambda}$, minimize:

$$-\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{p=1}^{P} \lambda_p |\beta_p| \tag{1}$$

- Log Likelihood
- Penalty Function

Adaptive Lasso Iteratively Updates λ_p based on β: bigger β yields smaller λ.

Advantage: Decrease Bias on Important, Nonzero Coefficients.

GEORGETOWN UNIVERSITY

MAP Inference for Adaptive Lasso

MAP-Bayesian Approach: Put prior on λ, optimize.

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{p=1}^{P} \lambda_p |\beta_p| - \log \lambda_p - \log P(\boldsymbol{\lambda}) \tag{2}$$

- Log Likelihood.
- Laplace log-density.
- Penalty Hyperprior.

Black-box case: $\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\log P(\boldsymbol{\lambda})$; can evaluate gradients.

Gradient Descent not applicable because of nonsmooth absolute value.

Subgradient descent methods OK for prediction but bad at determining zero vs nonzero.

GEORGETOWN UNIVERSITY

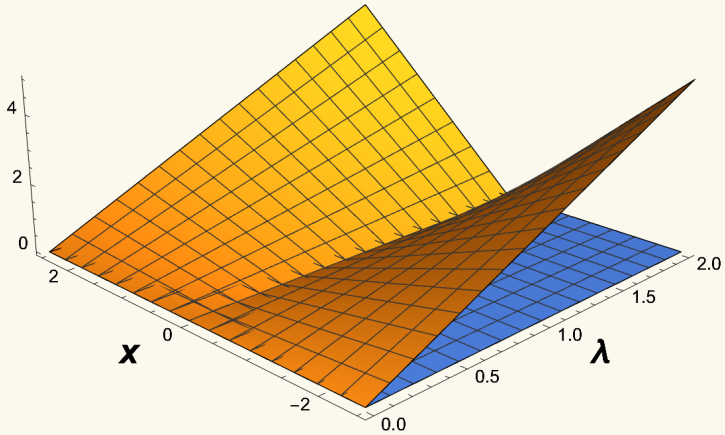
Proximal Gradient Descent

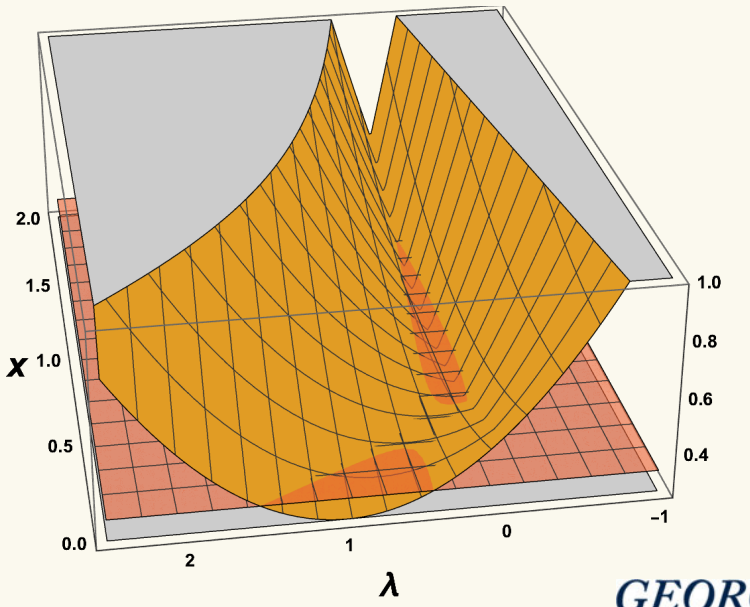
Handle nonsmooth term exactly via its *Proximal Operator*.

Jointly optimizing β, λ requires a new proximal operator.

Allows efficient optimization for arbitrary likelihoods (we've tried GLMs, hurdle models and neural networks).

Subproblem: optimize this nonconvex cost function:





But we show it's available in closed form.

Iraq Civil War & Forced Displacement

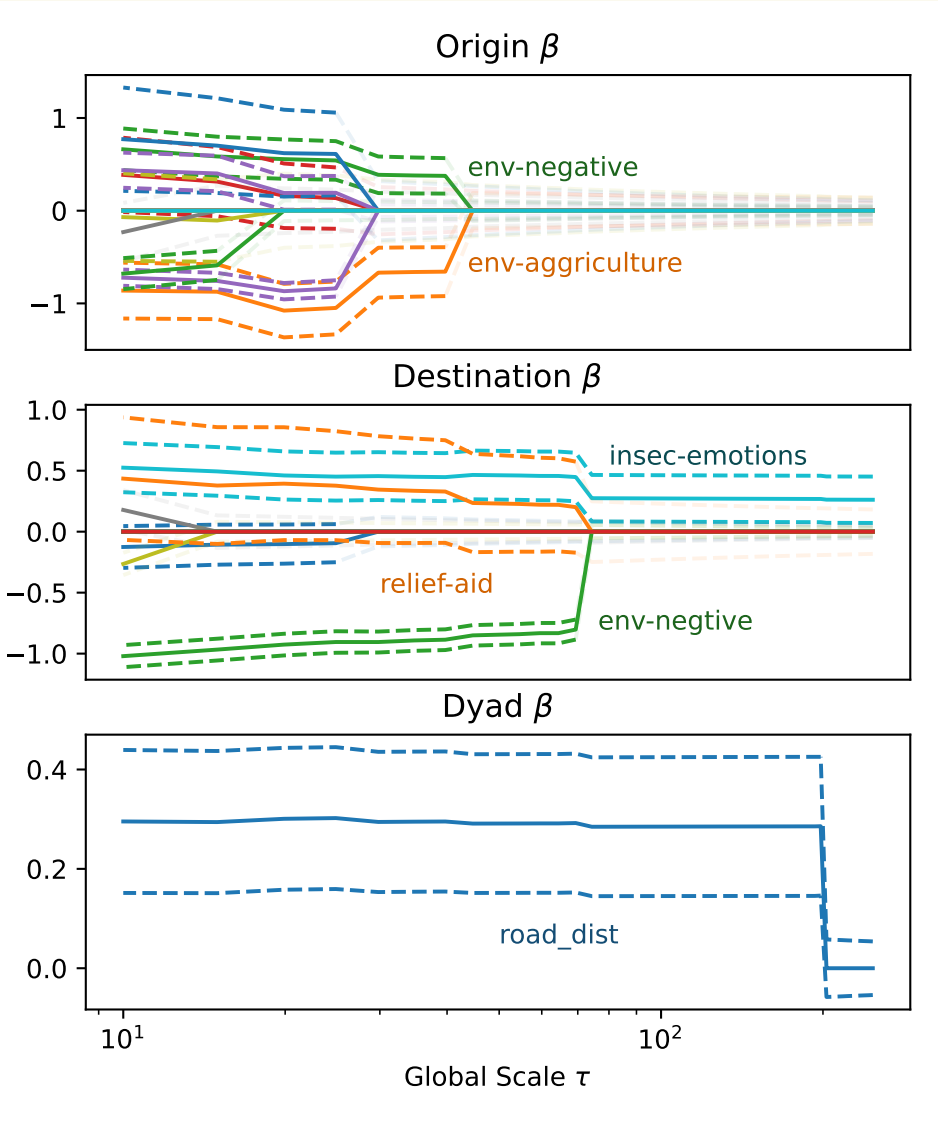
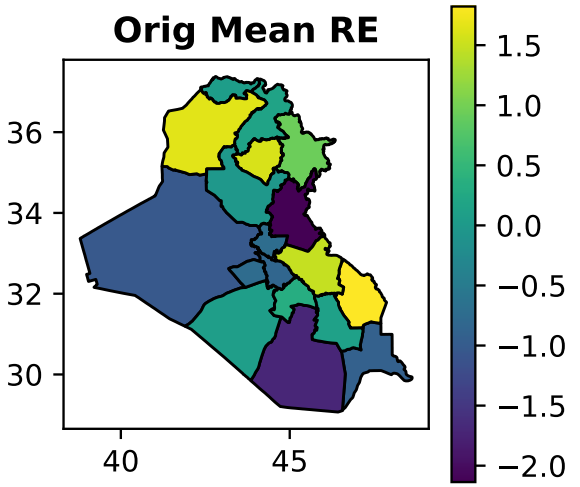
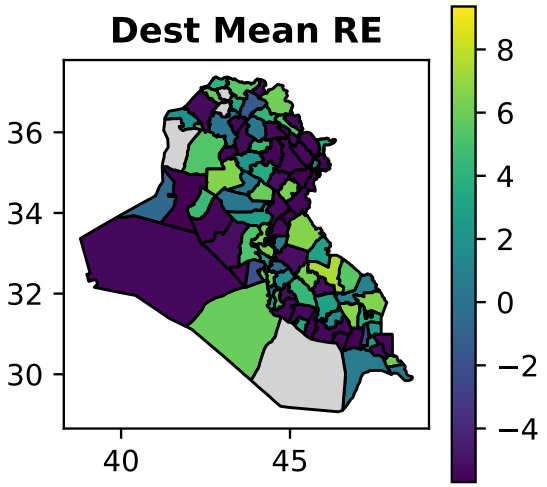


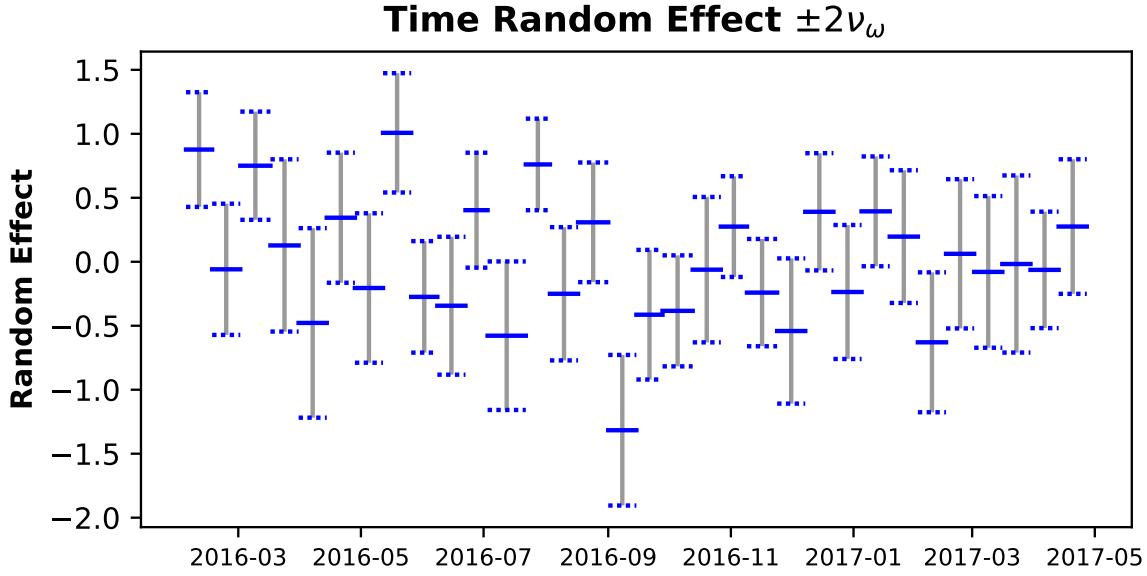
Figure: Variational Bayesian bi-Spatiotemporal Gravity Count Model with Adaptive Penalty.

GEORGETOWN UNIVERSITY

Iraq Civil War & Forced Displacement







GEORGETOWN UNIVERSITY

Iraq Civil War & Forced Displacement

The most positive origin random effects belongs to Ninewa Governorate with capital Mosul, which experienced urban combat against ISIL, followed by Maysan and Kirkuk governorates.

For destinations, the Baghdad area has some of the highest effects, with its constituent Al-Sadr and Kut districts well separated from the rest.

On the lower end we see a different story: many districts which observed no inflow of migration have random effects which are experiencing separation and similar, negative values. The time period with the largest random effect was from 2016-05-12 to 2016-05-26, a period which saw a series of ISIL bomb attacks which left over 100 dead in Baghdad (May 12) as well as combat between Iraqi and ISIL forces, while lowest random effect was 2016-09-01 to 2016-09-15, during which ISIL was more active abroad than in Iraq.

GEORGETOWN UNIVERSITY

Future Work

- Better understand and use the drivers of dynamics. This may be effectively done using agent based models but other modeling frameworks including PDE-based models may be useful too.
- Use blended data (conventional and organic data) to inform the parameters governing the dynamics.
- Address bias issues related to organic data.
- **Lyme:** Understanding risk categories and using demographic information to better inform the dynamics; Utilize organic data.
- **Forced Migration:** implement PDE-based dynamics using finer spatio-temporal resolution for displacement data.

GEORGETOWN UNIVERSITY

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 GU postdoctoral researcher) and Ryan Ripper (former GU graduate student).

The forced migration project is based on collaboration with Lisa Singh (GU Computer Science/Massive Data Institite) and Katharine Donato (GU ISIM), and Nathan Wycoff (GU postdoctoral researcher). This project is supported by funding from the McCourt Institute.

GEORGETOWN UNIVERSITY

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.
- 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.

GEORGETOWN UNIVERSITY