Spatial Local Gradient Models of Biological Invasions

Murali Haran[1]

Joshua Goldstein[1], Ottar Bjørnstad[2], Andrew Liebhold[3]

[1]Department of Statistics, Penn State University

[2]Center for Infectious Disease Dynamics, Penn State University

[3]USDA Forest Services/Entomology, Penn State University

ENAR Conference, Miami. March 2015

Summary

- The spread of invasive species can have far reaching environmental and ecological consequences.
- Understanding invasion spread patterns and the underlying process driving invasions are key to predicting and managing invasions.
- We develop methods based on Gaussian processes to characterize local speed and dominant direction of spread along the invasion front.
- We can identify significant environmental and geographical determinants of local invasion rates.
- We demonstrate the application of our methods to
  - Historical data on gypsy moths and hemlock wooly adelgid.
  - Simulated data from a stratified diffusion model.

Gypsy Moths

Native to Europe and Asia, the gypsy moth was accidentally introduced from France to Massachusetts in the 1860s.

Gypsy Moth Damage

Extensive defoliation due to gypsy moths. (figure from U of Illinois College of Agriculture, Consumer, & Environmental Sc.)

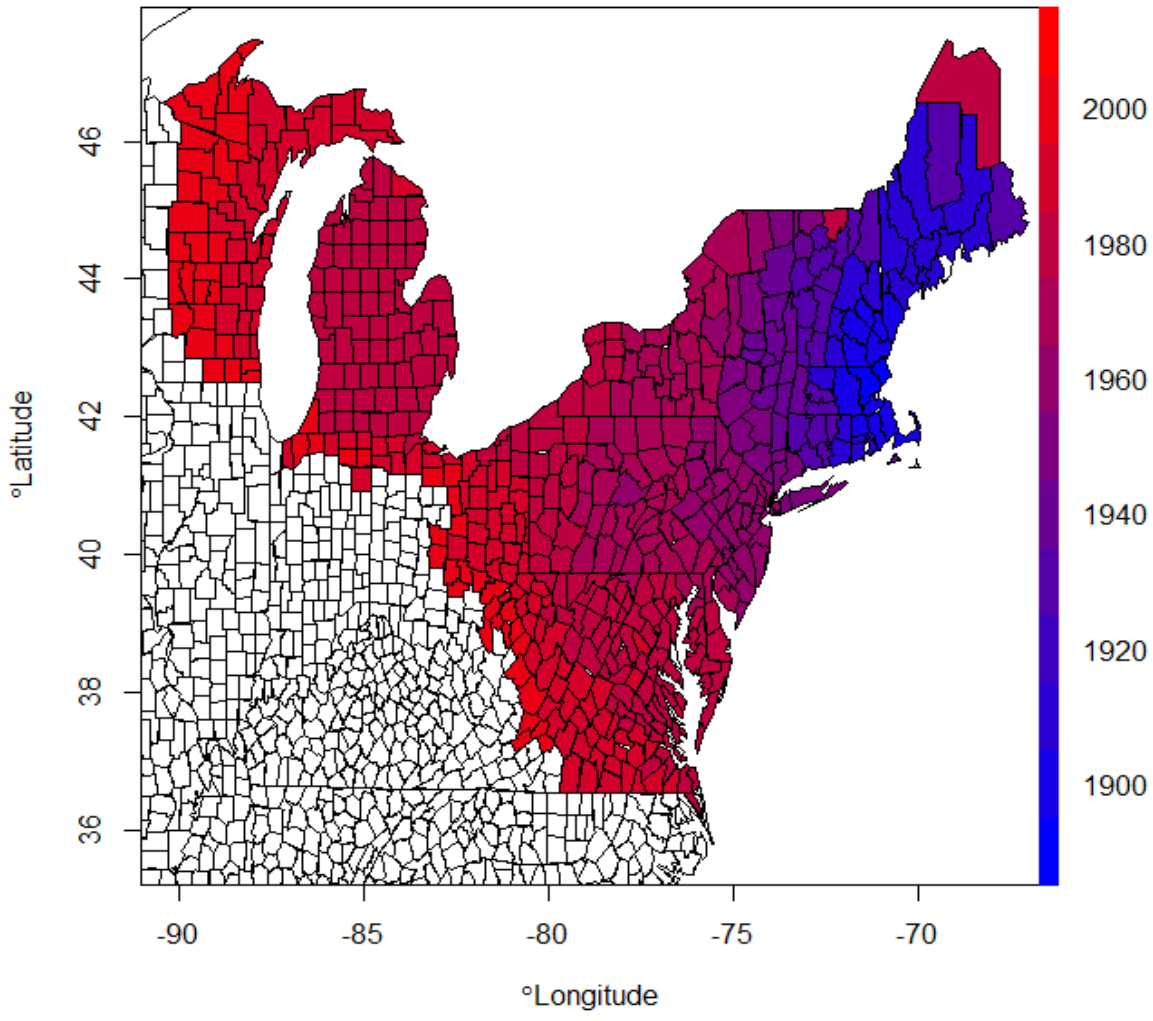
Gypsy Moth Invasions

- Since introduction it has spread through much of the northeastern US. Now: large area including north Atlantic states, bordering Canadian provinces, and second focus resulting from a long-range jump event to Michigan.
- Relatively slow spread partly due to North America females being unable to fly. Mean spread was estimated at 13 miles per year from 1960 to 1990. Spread by:
  - Short-range windborne dispersal of larvae through a process known as ballooning.
  - Egg masses accidentally moved by human transport, forming new colonies ahead of the invasion front, causing a pattern of stratified diffusion.

Historical Data on Spread

- County-level USDA quarantine records. Entire county is part of quarantined area when established gypsy moth populations were first detected anywhere within the county. Annual records from 1934 to present.
- Other published sources for infestations from 1900-1934.
- Percent forest basal area comprised of oaks, favored food plant.
- Estimated size, human population, per capita income for each county.

Year of First Appearance by County



Of Interest

- Direction and speed of spread. Also want uncertainties associated with them since we should only display significant directions.
- Distinguish between wave-like diffusive spread from long-range jumps in spread.

Essentially: want carefully constructed descriptive summaries of the spread.

Gaussian Process Model for the Invasion

- We model a continuous surface of “waiting time to first appearance” using a Gaussian process.
- Motivation: the reciprocal of the gradient length of this surface is a measure of the invasion speed. Fast spread should lead to shallow waiting time surfaces, while slow spread results in steep surfaces.
- Other approaches (e.g. Johnson et al., 2004; Farnsworth and Ward, 2009) provide nice visualization of gradients using thin plate splines.
- In contrast, by using a full statistical model, we obtain uncertainties for local spread estimates. This is important for assessing significance which in turn is important for main questions of interest.

Spatial Process Gradients

- County-level quarantine records, $\mathbf{Y} = \{Y(s_1), \ldots, Y(s_n)\}$ where $\mathbf{s} = \{s_1, \ldots, s_n\}, s_i \in \mathbb{R}^2$. Here, s_i is the centroid of the ith county. $n = 571$.
- $Y(s)$ is modeled as an isotropic Gaussian process with mean $\mu(s)$ and covariance function $K(r)$ at distance r.
- The gradient process $\nabla Y(s)$ and $\mathbf{Y}$ have a joint multivariate normal distribution (Banerjee et al., 2003) given by

$$\begin{pmatrix} \mathbf{Y} \\ \nabla \mathbf{Y} \end{pmatrix} \sim N_{3n}\left[\begin{pmatrix} \boldsymbol{\mu} \\ \nabla \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} K(D) & -\nabla K(D) \\ \nabla K(D)^T & -H_K(D) \end{pmatrix}\right]$$

where D is the $n \times n$ matrix of pairwise distances of $\mathbf{s}$, and $K(D)$ represents the $n \times n$ matrix of $K(\cdot)$ applied element-wise to D. $\nabla \boldsymbol{\mu}$ is a length $2n$ vector, $\nabla K(D)$ is a n x $2n$ matrix and $H_K(D)$ is a $2n$ x $2n$ matrix

Interpolating Spatial Gradients

- The conditional distribution $\nabla \mathbf{Y} | \mathbf{Y}$ (with known parameters Θ) is therefore simply multivariate normal.
- Crucially, this allows us to obtain the distribution of the gradient at any new location s_0,

$$\nabla Y(s_0)|\mathbf{Y}, \Theta \sim N_2\big(\nabla \mu(s_0) - \nabla K(\delta)^T [K(D)]^{-1}(\mathbf{Y} - \boldsymbol{\mu}),$$
$$-H_K(0) - \nabla K(\delta)^T [K(D)]^{-1} \nabla K(\delta)\big),$$

  with $\delta = (s_0 - s_1, ..., s_0 - s_n)$.
- Note: we use a Matern family covariance with $\nu > 1$ so that the gradient process is well-defined (all second order partial derivatives of K must exist).

Inference and Prediction

- $\mu(s) = \beta_0 + \beta_1 s_x + \beta_2 s_y$, Matern covariance parameters σ^2, ϕ, and $\nu = 3/2$. Let $\Theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \phi, \tau^2)$.
- Prior distribution for Θ.
- Infer posterior for Θ via Markov chain Monte Carlo.
- Using posterior samples of Θ, obtain posterior predictive distribution of gradients at new locations.
  - Sample from posterior distribution of Θ via MCMC.
  - For each Θ, obtain draw for gradients at new locations.

Inferring Speed and Direction of Spread

- Data are times of first appearance, the direction of the spread is the opposite direction to the gradient, and steeper gradients correspond to slower speeds. Hence, posterior samples for the gradient $\nabla Y(s_0)$ are transformed.
- Result: posterior samples for the magnitude of speed in the x and y directions at s_0, from which we can infer the speed and direction of spread, with credible intervals.
- Notice: sample-based inference keeps this simple.

Detecting Sources and Long Range Jumps

- Of scientific interest: automated methods to identify candidate locations for foci of long-range jumps ahead of the advancing front. Distinguishing these locations from those with contiguous “wave-like” diffusive spread.
- This is a challenge. We propose two approaches:
  - Rayleigh test
  - Direct test of average gradient on a curve around a point

Using Circular Distributions

- Rayleigh test: test whether a circular distribution is random or non-random.
- When applied to the vectors of spread near a point, a “non-random” distribution of the directional vectors of spread around a point implies a unified directional spread through that point. (“Random distribution” implies no such directional spread.)

Direct Test of Distribution of Gradient

- We can also test directly the distribution of the gradient around a point, which tells us if there is significant radial expansion around that point.
- Using Banerjee and Gelfand (2006) we can obtain the total gradient over a curve, which is also a Gaussian process. Hence, we can find the distribution of the gradient normal to a curve.
- With this method we can test the average flux out of a boundary around an arbitrary point. If this is significant, we can say that there is spread radially outward from this point, marking it as a potential site of a source or long-range dispersal.

Direct Test of Distribution of Gradient: Details

- Define a curve $\mathcal{C}_{t^*} = \{s(t) : t \in [0, t^*]\}$. Let $\eta(s(t))$ be the unit vector normal to the curve at the point $s(t)$. The total gradient normal to $\mathcal{C}_{t^*}$ is

$$\Gamma(t^*) = \int_0^{t^*} \langle \nabla Y(s(t)), \eta(s(t)) \rangle dv,$$

where v is the arc-length of the curve, $v(t^*) = \int_0^{t^*} \|s'(t)\| dt$, and so

$$\Gamma(t^*) = \int_0^{t^*} \langle \nabla Y(s(t)), \eta(s(t)) \rangle \|s'(t)\| dt,$$

- Monte Carlo integration can approximate the conditional distribution $\Gamma(t^*)|\mathbf{Y}, \Theta$ and test average flux out of a boundary. If significant, mark as a potential site or source of long-range dispersal.

Distribution of Total Gradient Over a Curve

- Distribution for total gradient over a curve is a Gaussian process on $[0, T]$, $\Gamma(t^*) \sim GP(\mu_\Gamma(t^*), K_\Gamma(\cdot,\cdot))$ (Banerjee and Gelfand, 2006) with

$$\mu_\Gamma(t^*) = \int_0^{t^*} \langle \mu(s(t)), \eta(s(t)) \rangle \|s'(t)\| dt$$

And $K_\Gamma(t_1^*, t_2^*)$

$$= \int_0^{t_1^*} \int_0^{t_2^*} \eta^T(s(t_1)) H_K\left[s(t_2) - s(t_1)\right] \eta(s(t_2)) \|s'(t_1)\| \|s'(t_2)\| dt_1\, dt_2$$

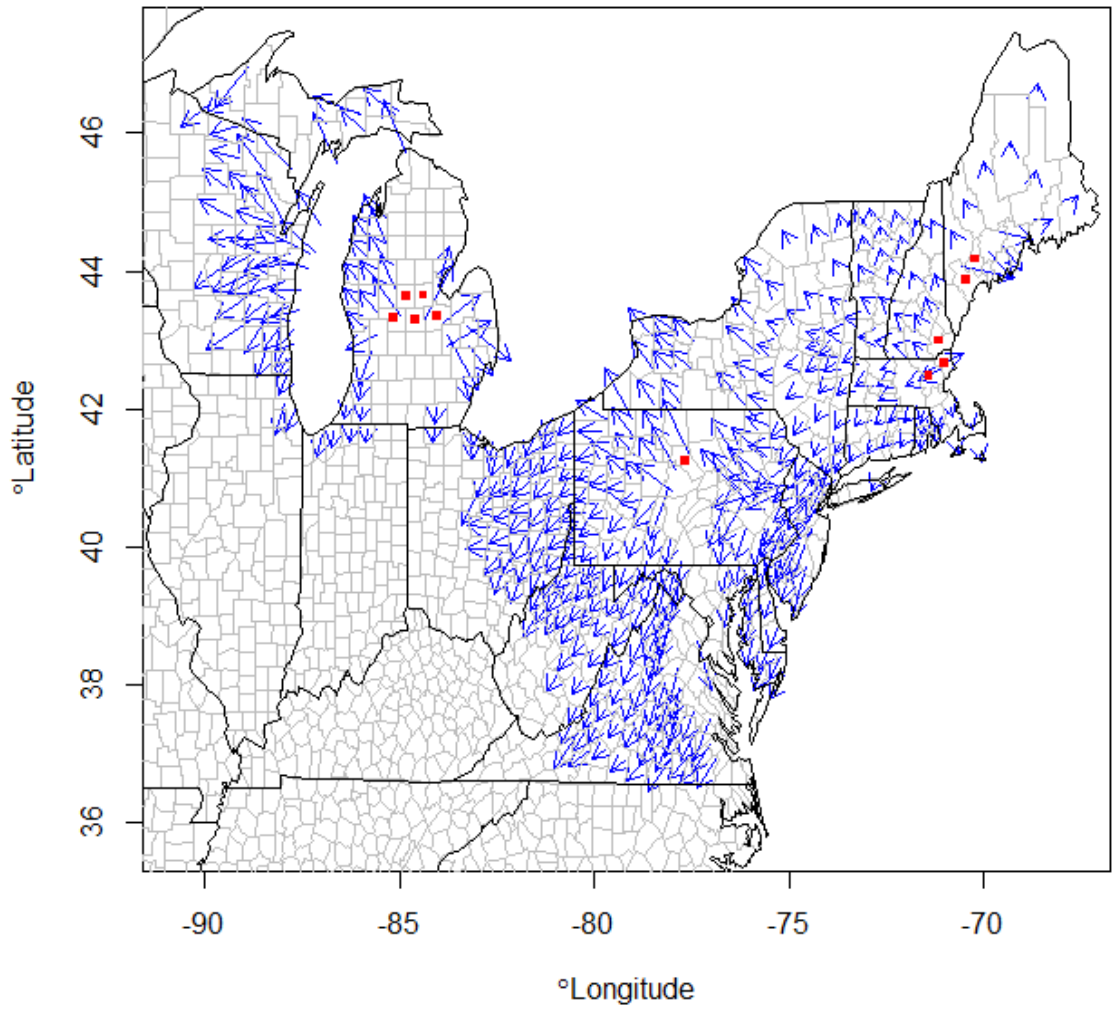
where $\mu(\cdot)$ is the mean of the original process $Y(s)$ and $H_K(\cdot,\cdot)$ is the hessian of the covariance of $Y(s)$.

- The conditional distribution of interest, $\Gamma(t^*)|\mathbf{Y}, \Theta$ is

$$N\big(\mu_\Gamma - \gamma_\Gamma^T(t^*)[K(D)]^{-1}(\mathbf{Y}-\boldsymbol{\mu}), K_\Gamma(t^*,t^*) - \gamma_\Gamma^T(t^*)[K(D)]^{-1}\gamma_\Gamma(t^*)\big).$$

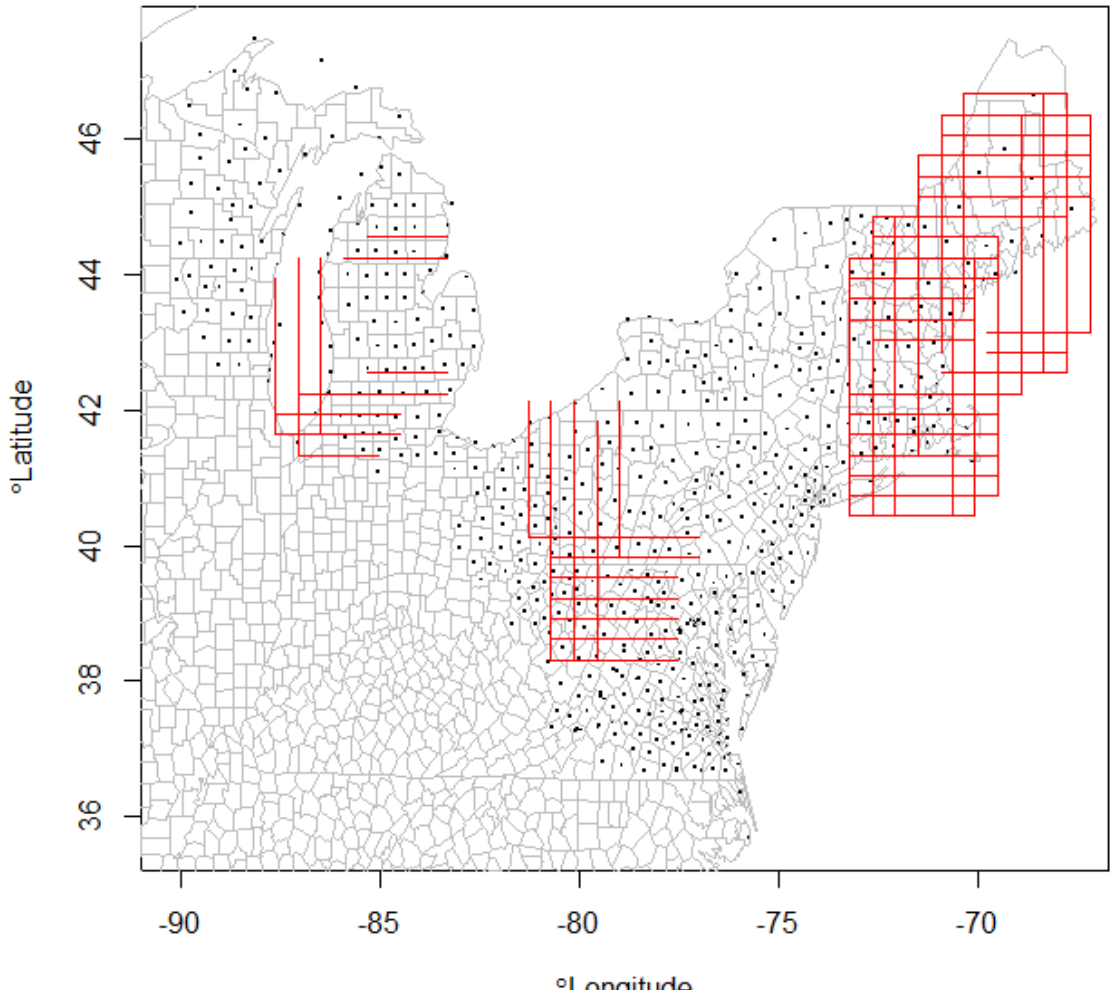
Significant Speeds and Directions of Spread

Red points: sites of long-range jumps (from Rayleigh test).



Regions of Long Range Jumps

Searching over a grid reveals three regions of potential long-range jumps. Red lines indicate sides of a box around a potential source for significant spread (outside the box).



Relationship to Covariates

- We can gain insights into mechanisms of spread by relating the geographic variation in the speed of spreads to characteristics of habitat.
- Results obtained from spatial regression using gradient samples from posterior predictive distribution.
- *Basal area of susceptible host trees* is significantly associated with speed of spread, consistent with the concept that local growth rates will be larger in the face of more favorable habitat, and should consequently enhance invasion spread rates.
- County size, population, income are not significant.

Conclusions

- Our methods allow us to identify key spatial features including sources, sites of rapid spread and quick long-range growth.
- (Skipped) Application to data from stratified diffusion model successfully recover features of this process.
- We are able to test the significance of spread patterns and spatial features of these invasions while accounting for uncertainties. Important for maps of spread rates and directions, as well as long-range jumps.
- Our methods are applicable to a wide range of problems in ecology and epidemiology.
  - (Skipped) Applied to spread of hemlock woolly adelgid.
  - `R` package available. Coming soon to a website...