Quantifying SpatioTemporal Variation of Invasion Spread

Murali Haran[1]
joint with Joshua Goldstein[2], Jaewoo Park[1], Ottar Bjørnstad[3], Andrew Liebhold[4]

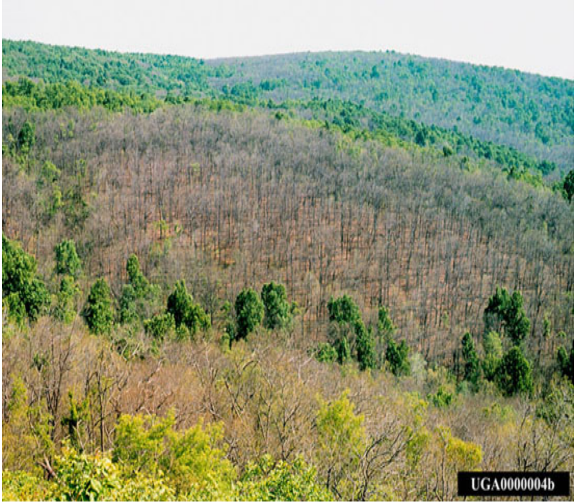
[1] Department of Statistics, Penn State University
[2] Social Data Analytics Lab, University of Virginia
[3] Center for Infectious Disease Dynamics, Penn State University
[4] USDA Forest Services/Entomology, Penn State University

MIDAS Meeting, Atlanta. April 2019

Outline

- We consider data that consist of waiting times to first appearance of an infection or invasive species.
- I will describe a Gaussian process-based approach for fitting surface of this “waiting times to invasion” data.
- Our method is based on estimating gradients of this surface that allows us to *approximately* characterize
  - local speed of spread along the invasion front
  - dominant direction of spread
  - associated uncertainties/significance
- We demonstrate the application of our methods to
  - Historical data on gypsy moths and hemlock woolly adelgid.
  - (Just begun) Malawi measles data

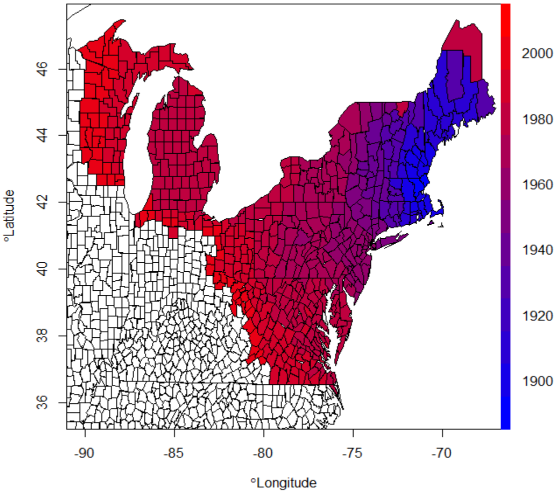
Gypsy Moth Damage



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

Example Data: Year of First Appearance by County

Gypsy moth invasion: Blue (first), purple, red (last)



Of Interest

Want descriptive summaries of the spread:

- 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.
- May link local rates of spread to environmental covariates.

Important: We are just doing a careful exploration of the data. Hopefully this will give some guidance for a future full-fledged analysis.

Illustration: Gypsy Moth Invasion in the U.S.

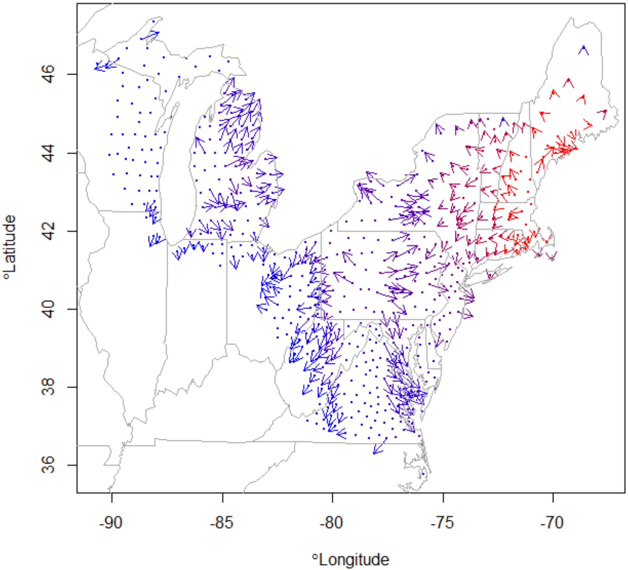
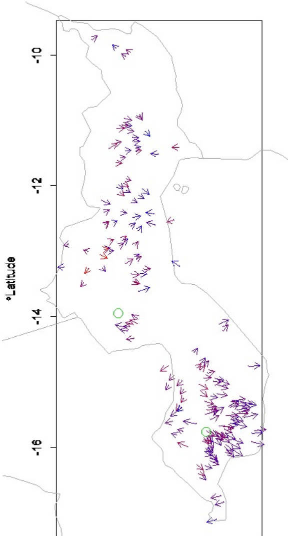


Illustration: Measles in Malawi

Red: early cases, blue: later cases. Green circles: cities

Based on data from 2009 – 2012 (Source: MSF)



Remainder of Talk

- Sketch of how we go from the time of first appearance data to constructing map of direction and speed of spread

Gaussian Process Model for the Invasion

- We model a continuous surface of "time of first appearance" using a Gaussian process
- Motivation
  - Large differences of time of first appearance between neighboring regions indicate slow spread (small differences indicate fast spread)
  - Such differences are modeled via the gradient length of a time surface
  - The reciprocal of the gradient length of this surface is a measure of the invasion speed
- Notation
  - $Y(s)$: time of first appearance at location $s \in R^2$
  - $\nabla Y(s)$: gradient of $Y(s)$ by taking the derivative with respect to spatial direction
  - $\Gamma(r)$: total gradient of $Y(s)$ measures the change in the

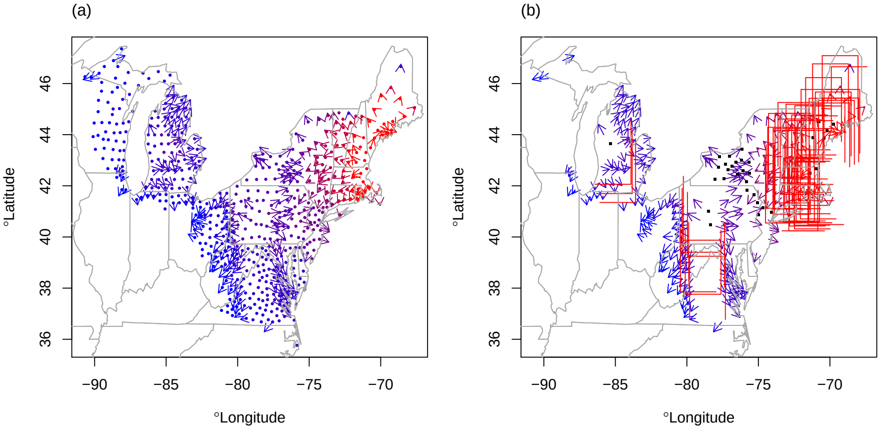
Inferential Procedure

- The Gaussian process model is fit to $Y(s)$ via an MCMC
  - Obtain posterior samples of the mean and covariance parameters $\Theta = (\beta, \theta)$
- Detecting diffusive expansion (local spread)
  - $\nabla Y(s_i) | Y(s_i), \Theta$ is a normal distribution (Banerjee et al., 2003)
  - The mean speed of spread is estimated as $\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\|\nabla Y(s_i)\|}$
  - By plotting all statistically significant gradients we can visualize the vector field of spread
- Detecting sources and long-range jumps
  - $\Gamma(r)_i | Y(s_i), \Theta$ is a normal distribution (Banerjee et al., 2003)
  - We flag a location as a potential site of a long-range introduction if the spread is significant for at least two out of the four cardinal directions

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.
- For gypsy moth: only *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.

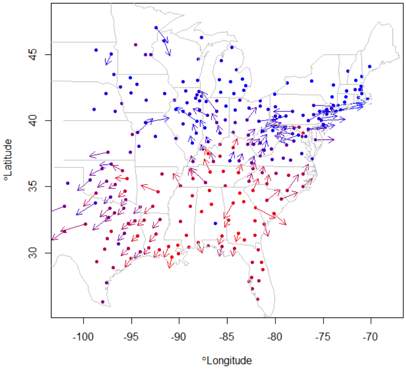
Gypsy Moth Example



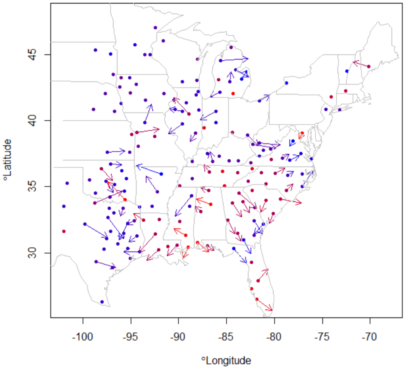
Trying Other Examples: Influenza Data

- Weekly data in over 300 US cities between 2002–2010.
- Compare/characterize diffusion in pandemic (2003/4, 2009) vs epidemic years?

Flu spread fall 2009



Flu spread 2006-2007



Conclusions

- Our methods allow us to identify spatial features: sources, sites of rapid spread, 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.

Reference and Software

Goldstein, J., Park, J., Haran, M., Liebhold, A., and Bjornstad, O.N. (2019) Quantifying Spatio-Temporal Variation of Invasion Spread. *Proceedings of the Royal Society B*, 286, 1894

R package: `InvasionSpeed`

`http://www.personal.psu.edu/muh10/invasionSpeed.html`

Sundries

Gypsy Moths

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

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.

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 **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).

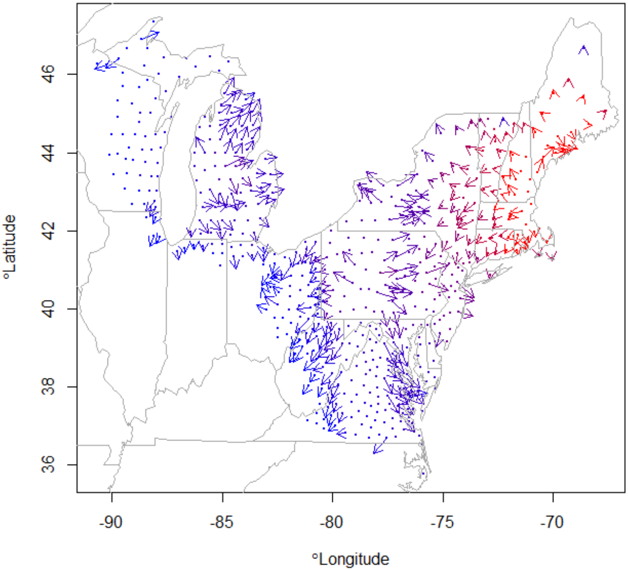
Inference and Prediction

- $\mu(s) = \beta_0 + \beta_1 s_x + \beta_2 s_y$, 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.

Significant Speed and Direction of Spread

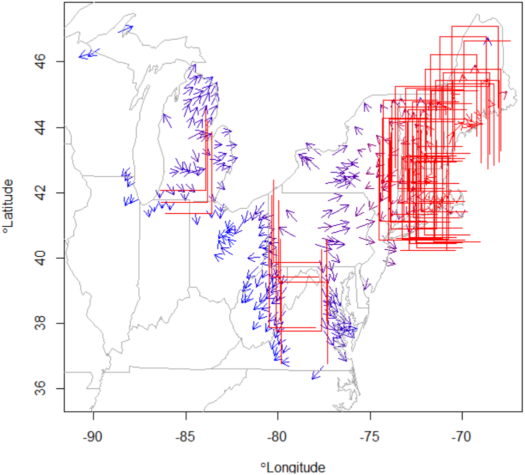


Direct Test of Distribution of Gradient

- 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.
- We can obtain the total gradient over a curve, which is also a Gaussian process (Banerjee and Gelfand, 2006). 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.

Regions of Long Range Jumps

Searching over a grid reveals three regions of potential long-range jumps.



Details: Direct Test of Distribution of Gradient

- 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.

Details: 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

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

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.)