PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Goldstein J, Park J, Haran M, Liebhold A, Bjørnstad ON. 2019 Quantifying spatio-temporal variation of invasion spread. *Proc. R. Soc. B* **286**: 20182294.
http://dx.doi.org/10.1098/rspb.2018.2294

Received: 10 October 2018
Accepted: 3 December 2018

Subject Category:
Ecology

Subject Areas:
ecology, environmental science, health and disease and epidemiology

Keywords:
invasive species, gypsy moth, hemlock woolly adelgid, Gaussian process, spatial gradients

Author for correspondence:
Murali Haran
e-mail: mharan@stat.psu.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4330631.

THE ROYAL SOCIETY PUBLISHING

Quantifying spatio-temporal variation of invasion spread

Joshua Goldstein[1], Jaewoo Park[2], Murali Haran[2], Andrew Liebhold[3] and Ottar N. Bjørnstad[4]

[1]Social and Data Analytics Laboratory, Virginia Tech, 900 N Glebe Rd, Arlington, VA 22203, USA
[2]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
[3]US Forest Service Northern Research Station, Morgantown, WV 26505, USA
[4]Departments of Entomology and Biology, Pennsylvania State University, University Park, PA 16802, USA

MH, 0000-0003-4440-8625

— 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 combine a set of statistical methods in a novel way to characterize local spread properties and demonstrate their application using simulated and historical data on invasive insects. Our method uses a Gaussian process fit to the surface of waiting times to invasion in order to characterize the vector field of spread.
— Using this method, we estimate with statistical uncertainties the speed and direction of spread at each location. Simulations from a stratified diffusion model verify the accuracy of our method.
— We show how we may link local rates of spread to environmental covariates for two case studies: the spread of the gypsy moth (*Lymantria dispar*), and hemlock woolly adelgid (*Adelges tsugae*) in North America. We provide an R-package that automates the calculations for any spatially referenced waiting time data.

1. Introduction

When a non-native species successfully establishes in an exotic environment it enters the spread phase of biological invasions during which the species expands its range into suitable habitat [1]. Ecological theory has shown that the speed of invasion spread is a joint function of the dispersal rate and the population growth rate of the invading species [2,3]; any habitat characteristic that influences population growth or dispersal can thus influence the rate of spread. Rates of spread may vary considerably among species and for a given species, spread rates may vary across heterogeneous landscapes [4,5]. Understanding the mechanisms causing heterogeneity in the rate of invasion spread is key to predicting future rates of spread and identifying important locations for management.

In this work, we propose automated statistical methods for estimating local speed and dominant direction of spread along invasion fronts. Our approach can be applied to identify statistically significant environmental and geographical determinants of local invasion rates and likely epicentra of invasion resulting from long-range introductions.

In addition to environmentally driven heterogeneity in rates of spread, there is considerable variation among species in the extent to which invasion spread is discontinuous (jumps). Spread of some species occurs via continuous expansion of the range into contiguous areas. For example, the North American muskrat, *Ondatra zibethica*, invaded central Europe from 1905 to 1927 via gradual expansion of its range in concentric circles [2]. The spread of other species is highly discontinuous, characterized by a pattern referred to as stratified diffusion [6]; following initial establishment, expansion may happen with long-range jumps

© 2019 The Author(s) Published by the Royal Society. All rights reserved.

into isolated uninvaded areas, founding new colonies that expand and eventually coalesce to form a contiguously invaded zone. This pattern is observed in many species of invading organisms, such as invasion of North America by the Argentine ant, *Linepithema humile* [7] and the gypsy moth, *Lymantria dispar* [8].

Quantifying the spread of non-native species and relating invasion speed to habitat heterogeneity is important for predicting and managing biological invasions. Several methods have been developed for studying processes that control spread rates of species. Species distribution models [9–13] are widely used to predict distributions of invasive species, for example, by using generalized linear models or generalized additive models. A variety of methods [14–20] combine dynamic equations within the framework of a hierarchical Bayesian model. These novel approaches embed dynamic equations within statistical models, allowing for a scientific interpretation of their fitted models. The above work has largely used spatial counts or presence–absence disease data; by contrast, the data we use is the time of first appearance of an invasive species.

We note that there are numerous other ways to model data on the spread of invasive species, including data in the form of point-level spatial data (cf. [21,22]).

Several methods have been developed for measuring spread based upon fitting range size to time since establishment or estimating spread by directly quantifying displacement of range boundaries over time [23–25]. These methods are generally well-suited for quantifying average spread range and temporal variation therein, but they are limited in their ability to quantify local spread rates and their relation to local habitat characteristics. Also, these methods are generally designed to quantify spread as a continuous process; identification of long-range jumps in stratified dispersal is usually done visually in a non-automated fashion. These gaps in existing methodology provide our motivation for combining recent developments in spatial statistics methodology in order to provide an automated approach to estimate local speed and direction of spread. Here, our focus is on constructing a spatial surface that describes the direction and speed of spread of an invasive species. Our method can help researchers learn about characteristics of the spread of the invasive species, including both local speed and direction as well as long range. We take advantage of recent statistical theory on the estimation of spatial gradients. We test our methods on simulated data generated from a stratified diffusion model and apply them to two detailed case studies of biological invasions, the historical spread of the gypsy moth and the hemlock woolly adelgid, *Adelges tsugae*, in North America.

2. Data

(a) Gypsy moth

Native to Europe and Asia, the gypsy moth was accidentally introduced from France to Massachusetts in the late 1860s [26], it has since spread throughout much of the northeastern USA. The gypsy moth is now established in a large area composed of the North Atlantic states and bordering Canadian provinces, as well as a second focus resulting from a long-range jump event to Michigan around 1980 [5,27,28].

The invasion of the gypsy moth across North America has been slow compared to the rate of spread of many other alien species [29]. Mean spread was estimated at 21 km per year from 1960 to 1990 [27]. The relatively slow rate of spread can be attributed, in part, to the fact that females of North America populations are flightless. Gypsy moth populations spread by short-range windborne dispersal of 1st instar larvae through a process known as 'ballooning' [30]. Egg masses are also accidentally transported across longer distances on wood or human-made objects, forming new colonies ahead of the invasion front and resulting in a pattern of stratified diffusion [8].

The full invasion history of the gypsy moth in the USA is reflected in the year of government designation of gypsy moth quarantine by county. County-level quarantine records for the gypsy moth are maintained by the United States Department of Agriculture (US Code of Federal Regulations, Title 7, ch. III, §301.45). Historically, an entire county was usually designated part of the quarantined area when established gypsy moth populations were first detected anywhere within the county. These records are updated annually and exist from 1934 to the present. From 1900 to 1934, the year when counties were first infested has been described in various other published sources (e.g. [27,31,32]). As additional covariates, we used county-level data derived from a national forest inventory system on the per cent of the forest basal area comprised of oaks, which is a favoured food plant of the gypsy moth, and the size (square kilometre) of each county [33].

(b) Hemlock woolly adelgid

Hemlock woolly adelgid (HWA) is an insect species responsible for defoliation of its host trees, eastern hemlock and Carolina hemlock [34,35]. Native to East Asia, it was first discovered in the eastern USA in Virginia in the 1950s [36]. HWA life stages can be transported by wind, wildlife, especially birds, and humans. Since its discovery, it has gradually expanded its range into much of the northeastern USA [35,37]. By 1969, it was found in southern Pennsylvania and it invaded southern New England by 1985, spreading at an estimated speed of 20–30 km year^{-1} [35].

As with the gypsy moth, historical spread of the HWA was recorded at the county level. Records from the US Forest Service Forest Health Protection are available for 1951, 1971, 1981, 1996, and from 2001 to 2011. We use the basal area of hemlock [38] and plant hardiness zone [39] for each county as additional covariates for our analysis.

3. Methods

Historical spread of the gypsy moth has previously been estimated as averages over space. [27] Estimated spread rates have been determined for five geographical regions by the slope of a least-squares regression of time on distance to a reference point in each region. Spread rates have also been estimated by measuring the average displacement of range boundaries over time [23,24].

Previous research on quantifying spatial gradients from georeferenced biological data has focused on detecting zones or boundaries of rapid change across space using geostatistical *wombling* [40]. Wombling methods involve estimating local vector gradients by fitting bilinear functions over a lattice of points. This method has been applied to genetic [41] as well as ecological [42] data. More recent wombling methods for areal data feature Bayesian hierarchical spatial models in order to identify significant boundaries after accounting for

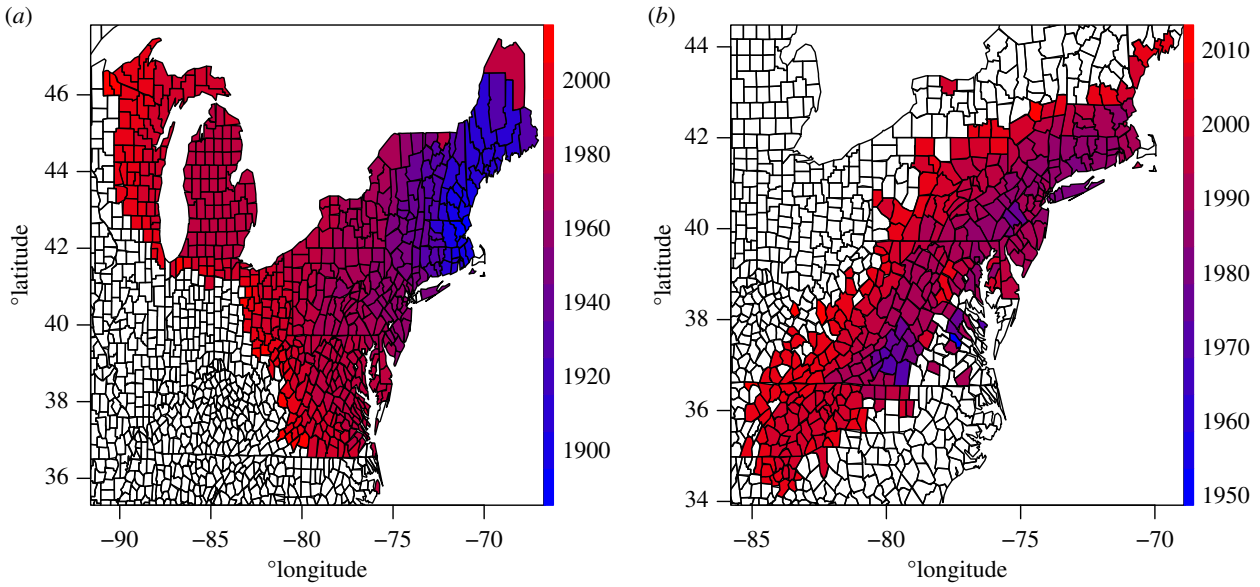

Figure 1. Year of first appearance by county for the gypsy moth (*a*) and hemlock woolly adelgid (*b*).

spatial dependence via Markov random fields [43–45], with applications to ecology and epidemiology.

The use of spatial gradients to estimate biological spread is motivated by the fact that if the surface is the waiting time to first appearance, then the reciprocal of the gradient length is a measure of the invasion speed: fast spread leads to shallow waiting time surfaces, while slow spread results in steep surfaces. Previously [46] estimated spread gradients using a thin plate spline applied to waiting times (as measured by wavelet phase angles) to study outbreak spatial dynamics of the larch budmoth. [47] used a similar spline surface approach to study spread of avian influenza. The thin plate spline approach yielded gradients which reflect the magnitude and direction of the spread, a simple general-purpose approach for visualization, but does not yield measures of statistical uncertainty associated with local spread estimates which prevents rigorous inference regarding whether, for example, any observed spatial variation is significant. In order to facilitate understanding the models and inferential procedure, we summarize our approach in the following sections. The mathematical details for the Gaussian process gradient models are provided in the electronic supplementary material.

(a) Estimating gradient surface using Gaussian processes

Given data on time of first appearance of an invasive species, we are interested in constructing a surface that describes the direction and speed of spread of the invasive species. We use Gaussian process models as a convenient and rigorous approach to estimate such a surface. Gaussian processes are commonly used for spatial interpolation [48]. We use a Gaussian process to spatially interpolate time of first appearance. The gradient of this Gaussian process is known to also follow a Gaussian process [49].

Based on fitting a Gaussian process to our data, we develop methods for estimating speed and direction of the spread of the invasive species, and for detecting sites of long-range dispersal. We also provide, in the electronic supplementary material, computer code for an R [50,51] software package that automates the inference.

We assume we have observations of the year of first appearance $\boldsymbol{Y} = \{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$ at locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $\boldsymbol{s}_i \in \mathbb{R}^2$. For our examples, data are county-level quarantine records and the spatial locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ are taken to be the centroids of counties for the gypsy moth ($n = 571$) counties (figure 1*a*) and for the HWA ($n = 340$) counties (figure 1*b*). The data are discrete (areal) in space as they represent counties. In order to use a Gaussian process gradient model, we treat the data as if they are from the centroid of each county. In order to investigate the potential sensitivity of our conclusions to this approximation, we perturb the locations of the centroids of each county and perform the analysis with this perturbed data. We find that the estimated spread patterns of the perturbed datasets are similar to those of the original dataset (see electronic supplementary material for details). Our methods are ideally suited to data that are point-level, that is, where we can identify individual locations of invasion, or when the data are obtained at an aggregate (areal) level where the areal units are reasonably similar in size and shape. We note that we have not studied the sensitivity of our methods to problems where the size or shape of the areal units are considerably different. Hence the results from applying our method to data with highly variable sized or shaped areal units should be treated with caution. Coordinates are projected using the Albers equal-area conic projection with standard parallels 29°30′ and 45°30′. $Y(\boldsymbol{s}_i)$ is the year county i was added to the quarantine. We assume $Y(\boldsymbol{s})$ can be modelled using an isotropic Gaussian process. For our applications, we assume the original process $Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$, with mean function $\mu(\boldsymbol{s}) = \beta_0 + \beta_1 s_x + \beta_2 s_y$, correlated spatial error $w(\boldsymbol{s}) \sim GP(0, K(\cdot))$ with Matérn covariance smoothness $\nu = \frac{3}{2}$, which takes the explicit form $K(r) = \sigma^2(1 + \phi r)\exp\{-\phi r\}$, and uncorrelated error $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$, where τ^2 is a nugget effect that captures measurement error.

The gradient of waiting time $\nabla Y(\boldsymbol{s})$ can be defined by taking the derivative of $Y(\boldsymbol{s})$ with respect to spatial directions over $\mathbb{R}^2$. The spatial gradient vector $\nabla Y(\boldsymbol{s}) \in \mathbb{R}^2$ indicates the dominant direction of spread. When $Y(\boldsymbol{s})$ is the time of first appearance of the species, the gradient length $\|\nabla Y(\boldsymbol{s})\|$ measures the change in waiting time for spread of the species. Small change in time surfaces means fast spread, while large change indicates slow spread. Therefore, the reciprocal of the gradient length $1/\|\nabla Y(\boldsymbol{s})\|$ represents the speed of spread. Because $Y(\boldsymbol{s})$ is a Gaussian process, well-established results [49] show how we can obtain the distribution of $\nabla Y(\boldsymbol{s})$ by using its direct relationship to the distribution of $Y(\boldsymbol{s})$. This allows us to estimate both the direction and speed of spread of the invasive species based on the observations of time of first appearance.

Our other interest is in detecting long-range jumps. For each spatial location, our goal is to investigate whether they represent

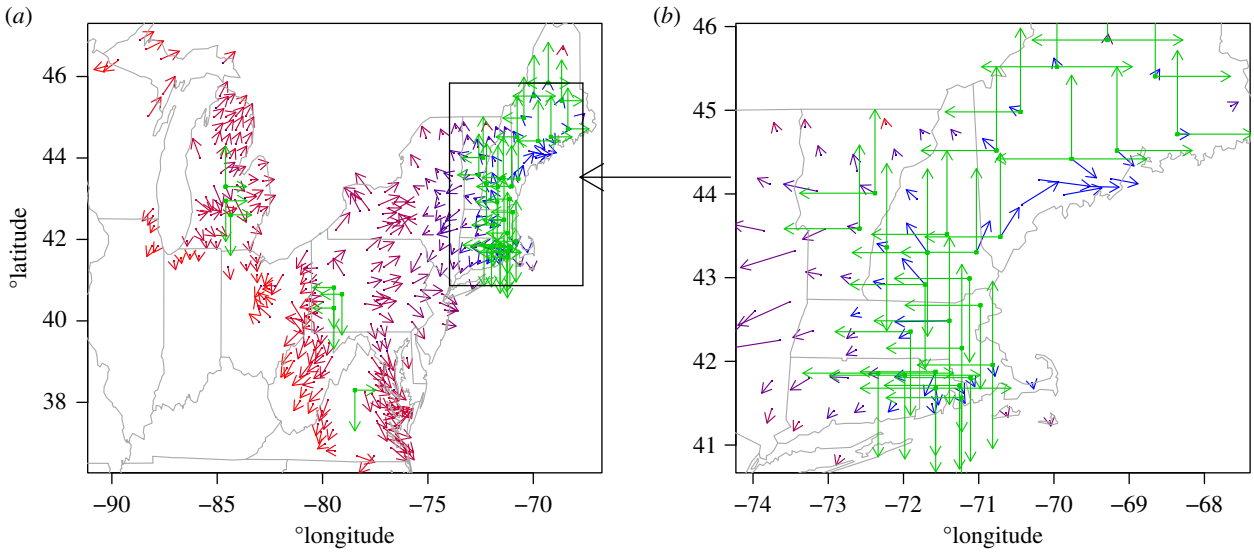

Figure 2. (*a*) Patterns of spread of the gypsy moth. Blue and red arrows indicate local speeds and directions of spread, and are plotted where spread is significant. The length of the arrows indicates the speed of spread—longer arrows indicate faster spread. The colour of each arrow represents the time of first appearance of the process. Blue implies the earliest appearance, and red indicates the latest appearance. Green points indicate potential sites of long-range jumps. Green arrows around a point indicate significant directions of long-range jumps. (*b*) Zoomed in figure of northeastern USA.

plausible introduction well ahead of the general spatial diffusion. For this we use the concept of 'total gradient' function, $\Gamma(r)$. For a particular location and for a given cardinal direction, the total gradient $\Gamma(r)$ measures the change in the waiting time for the spread of the species to a distance r away from the current location. Small $\Gamma(r)$ means shallow time surfaces which comes from fast spread of the species. This implies a potential long-range spread in that direction. Because $Y(\boldsymbol{s})$ is a Gaussian process, we can easily also obtain the distribution of $\Gamma(r)$ [49]. Based on this result, we can learn about the conditional distribution of $\Gamma(r)\,|\,Y(\boldsymbol{s})$ to search for any such long-range jumps.

In addition to total gradient, we also investigate the use of a Rayleigh test from circular statistics [52]. Although we find that this test is not a perfect method, it may still be a useful fast preliminary test for long-range jumps. Details for the Rayleigh test are provided in the electronic supplementary material.

(b) Inferential procedure

Our approach combines well-established spatial statistics tools in a novel way. Our inferential procedure is based on the Gaussian process gradient model and may be summarized as follows.

1. The Gaussian process model is fit to $Y(\boldsymbol{s})$:

   We infer the mean and covariance parameters $\Theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \phi, \tau^2)$ of the Gaussian process $Y(\boldsymbol{s})$ based on a Bayesian approach. Θ is sampled from the posterior distribution using a Markov chain Monte Carlo (MCMC) algorithm. The posterior mean is estimated as $\hat{\Theta} = (1/m)\sum_{i=1}^{m}\Theta_i$.

2. Detecting diffusive expansion:

   We are interested in learning about local speed and direction of spread. For each location $\boldsymbol{s}_i$ and a given posterior sample Θ, the gradient $\nabla Y(\boldsymbol{s}_i)$ has the distribution $\nabla Y(\boldsymbol{s}_i)\,|\,Y(\boldsymbol{s}_i), \Theta$, which is a normal distribution because $Y(\boldsymbol{s}_i)$ is modelled as a Gaussian process.
   - The mean speed of spread is estimated as $(1/n)\sum_{i=1}^{n} 1/\|\nabla Y(\boldsymbol{s}_i)\|$.
   - By plotting all statistically significant gradients (figure 2) we can visualize the vector field of spread.

3. Detecting sources and long-range jumps:

   For each location $\boldsymbol{s}_i$ and a given posterior mean $\hat{\Theta}$, we obtain the total gradient $\Gamma(r)$ from the conditional distribution $\Gamma(r)\,|\,Y(\boldsymbol{s}_i), \hat{\Theta}$ which also follows a normal distribution.
   - We flag a location as a potential site of a long-range introduction (figure 2) if: (i) the spread is significant for at least two out of the four cardinal directions, and (ii) for the remaining directions it is not significantly small.

(c) Driving factors of spread

We can gain insight into drivers of spread by relating the geographical variation in spread to habitat characteristics. To account for spatial dependence we fit a Bayesian spatial regression model to log-speeds using the *spBayes* R package [53]. We apply a log transformation to the response since the speeds have right-skewed distributions. If the mean speed at location $\boldsymbol{s}_0$ is given by $V(\boldsymbol{s}_0)$, then we assume

$$\log V(\boldsymbol{s}_0) = X^T(\boldsymbol{s}_0)\boldsymbol{\beta} + w(\boldsymbol{s}_0) + \epsilon(\boldsymbol{s}_0),$$

where $X(\boldsymbol{s})$ is a vector of the spatially varying environmental and geographical covariates of interest. We assume $w(\boldsymbol{s}) \sim GP(0, G(\cdot))$, $G(\cdot)$ has Matérn covariance smoothness with smoothness ν, range ϕ and partial sill σ^2 and $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$. Priors are selected as before and joint estimation is done via MCMC for $\Theta = \{\boldsymbol{\beta}, \sigma^2, \phi, \tau^2, \nu\}$.

4. Results

(a) Gypsy moth

Significant speeds and directions of historical spread of the gypsy moth are plotted at the locations of each invaded county in figure 2. The mean speed over all counties is 22.6 km year^{-1}, with a median of 15.7 km year^{-1}. Distributions for the magnitude of spread at each location tend to be right-skewed, where the 95% credible interval is (1.7 km year^{-1}, 64.9 km year^{-1}).

In figure 2, we also test whether there are long-range jumps of length $r = 1^\circ$ in the four cardinal directions. Points

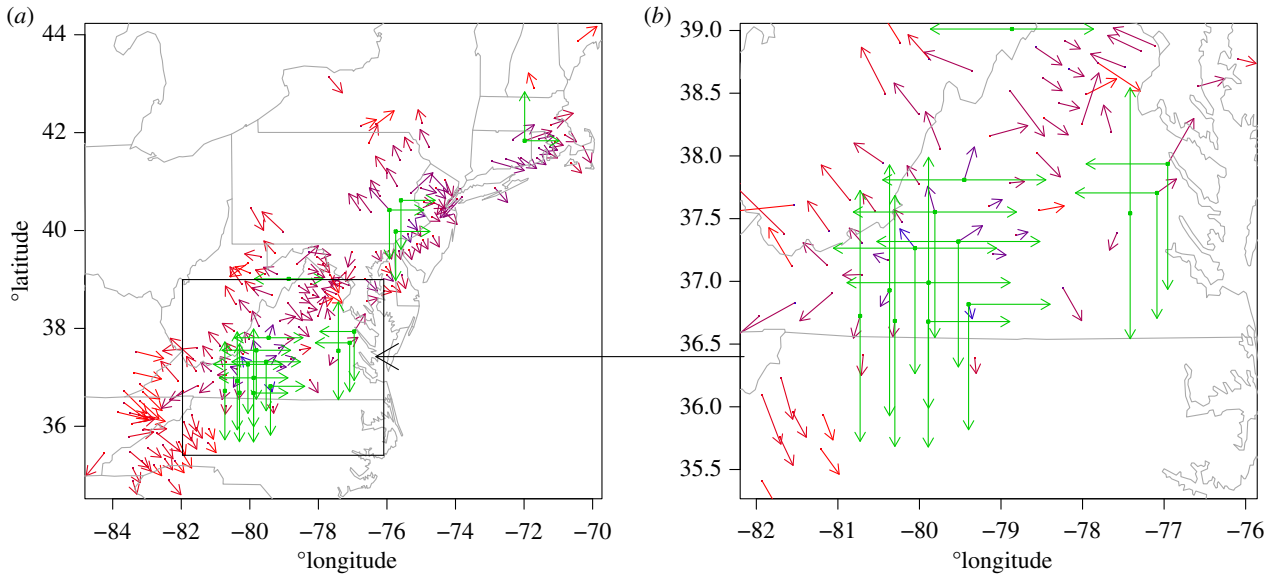

Figure 3. (*a*) Patterns of spread of the hemlock woolly adelgid. Blue and red arrows indicate local speeds and directions of spread, and are plotted where spread is significant. The length of the arrows indicates the speed of spread—longer arrows indicate faster spread. The colour of each arrow represents the time of first appearance of the process. Blue implies the earliest appearance, and red indicates the latest appearance. Green points indicate potential sites of long-range jumps. Green arrows around a point indicate significant directions of long-range jumps. (*b*) Zoomed in figure of Richmond area.

Table 1. Results of a spatial regression of speeds of spread (km year^{-1}) for the gypsy moth (a) and hemlock woolly adelgid (b) including posterior means and 95% credible intervals obtained using the highest posterior density interval algorithm [54].

(a) gypsy moth	β
intercept	−1.6 (−11.0, 9.3)
longitude	−5.1 (−8.1, −2.2)
latitude	−(−6.6, 1.2)
county size	−0.00007 (−0.00020, 0.00002)
quarantine date	0.0006 (−0.0044, 0.0056)
basal% susceptible trees	0.0023 (0.0000, 0.0042)
(b) HWA	β
intercept	19.5 (3.1, 36.6)
longitude	−9.8 (−14.9, −4.8)
latitude	8.5 (1.7, 16.0)
quarantine date	−0.003 (−0.009, 0.003)
$I_{\text{presence of hemlock}}$	0.09 (0.01, 0.07)
plant hardiness zone	0.014 (−0.19, 0.23)

identified as probable long-range jumps are marked in green in figure 2, along with green arrows which indicate significant directions of jumps. Our method identifies three potential sites around the northeastern coast, Michigan, and central-western Pennsylvania. Prior analysis confirms two of these sites, as the population was first introduced in Massachusetts in the 1860s and a discrete population was later established in Michigan [27]. A close examination of figure 1 also highlights a jump to Centre County, PA in the mid-1970s.

We relate speed of spread to latitude and longitude, quarantine date, county size, and finally the per cent basal area comprised of trees preferred as hosts of the gypsy moth. Estimated parameters of the spatial regression model are given in table 1a. We verify that, on average, the gypsy moth spread faster as it moved west. We also found that basal area of susceptible host trees is significantly associated with faster invasion, consistent with the concept that local growth rates will be larger in the face of more favourable habitat, and should consequently enhance invasion spread rates.

(b) Hemlock woolly adelgid

Significant speeds and directions of spread for the HWA are plotted at each county in figure 3. We find a mean speed of spread of 20.5 km year^{-1} across counties, with a median speed of 13.5 km year^{-1}. Distributions for the magnitude of spread at each location tend to be right-skewed, where the 95% credible interval is (3.0 km year^{-1}, 59.2 km year^{-1}).

Probable sites of long-range introductions are also identified in figure 3. We detect areas of apparent long-range dispersal near Richmond, VA, and southern PA, suggesting a pattern of stratified diffusion also for this species. Morin *et al*. [35] previously found that expansion is significantly influenced by availability of host trees. Low winter temperatures can cause extensive mortality in HWA populations and limit expansion to the north [55]. Therefore, we relate speeds of spread to environmental features including the presence or absence of hemlock trees, and the average plant hardiness zone for each county, an index based on the mean annual minimum winter temperature [39]. Estimates from the regression model are given in table 1b. We observed evidence that historically expansion is faster to the west and north. We also find as in [35] that spread is significantly associated with the abundance of host trees. We also tested the interaction between plant hardiness zone and latitude and found that for a given latitude, HWA spread significantly slower through areas with lower (colder) plant hardiness zones [$\beta = 3.4$ (0.4, 6.3)].

(c) Simulation

We tested the ability of our method to recover the effects that spatially varying habitats have on the speed of spread. To accomplish this, data are simulated from a stratified

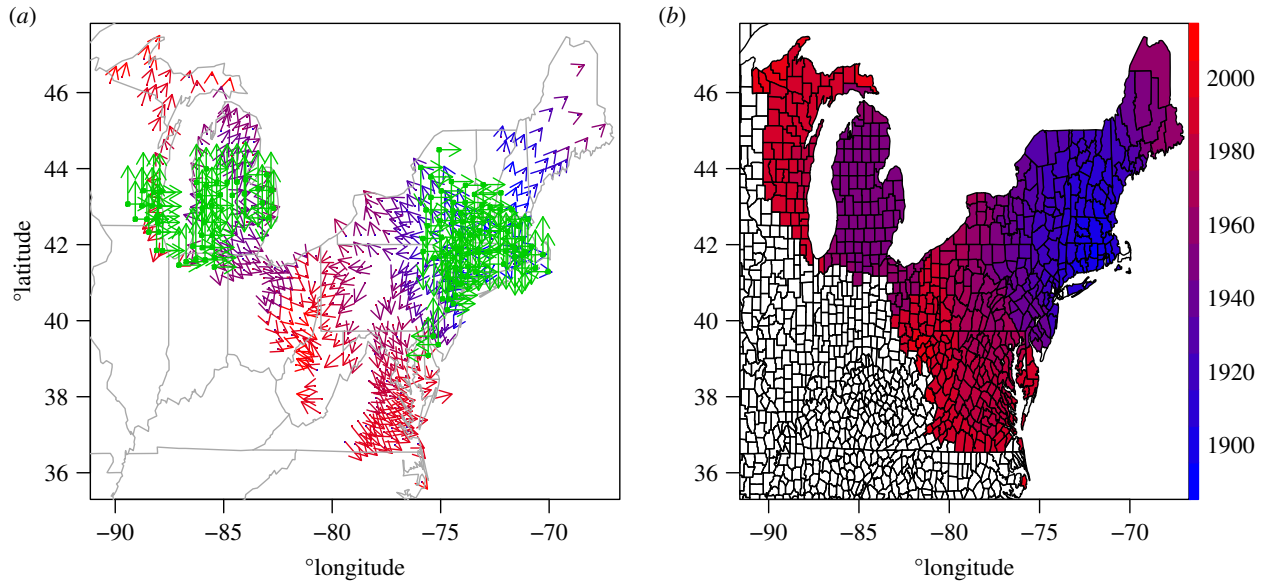

Figure 4. (*a*) Patterns of spread of the simulated invasion. Blue and red arrows indicate local speeds and directions of spread, and are plotted where spread is significant. The length of the arrows indicates the speed of spread—longer arrows indicate faster spread. The colour of each arrow represents the time of first appearance of the process. Blue implies the earliest appearance, and red indicates the latest appearance. Green points indicate potential sites of long-range jumps. Green arrows around a point indicate significant directions of long-range jumps. (*b*) Waiting times of the stratified diffusion simulation [6].

diffusion model following [6]. Stratified diffusion is a combination of neighbourhood diffusion and long-distance dispersal. As the size of the original colony expands, new colonies are more likely to be created by long-distance migrants.

The simulation starts with a single colony, centred at the initial point of invasion. The occupied area grows out in a circle with the radius r growing at constant rate c. This colony can then form offspring colonies from long-distance migrants in a random direction at a distance L from the invasion front. New colonies form at a rate $\lambda(r)$ that is a function of the colony radius. These offspring colonies grow at speed c and form offspring colonies of their own. The stratified diffusion simulation approach may be summarized as follows.

Algorithm 1. The stratified diffusion simulation approach.

Initialize with the first colony with the coordinates $\boldsymbol{s}_0$ and radius r_0.

for $t = 1 : T$ **do**

Given nth colony $\boldsymbol{s}_n$ with radius $r_{n,t}$ at time t.

1. Obtain the radius $r_{n,t+1}$ from nth colony: $r_{n,t+1} = r_{n,t} + cdt$, where dt is a time difference. (e.g. $dt = t + 1 - t = 1$)
2. With probability $\lambda(r_{n,t+1})$, a new colony $\boldsymbol{s}_{n+1}$ is generated in a random direction at a distance L

end for

Return coordinates for N number of simulated colonies $(\boldsymbol{s}_0, \ldots, \boldsymbol{s}_N)$. Note that N may be much smaller than T if λ is small.

We begin with an initial introduction in Massachusetts in 1900. Colony range expansion c varies by longitude to simulate a slow period of initial expansion; $c = 10$ km year^{-1} east of -78° and $c = 20$ km year^{-1} west of -78°. New colonies form at rate $\lambda(r) = 0.1\,r$ a distance $L = 10$ km from the invasion front. Additionally, to mimic the observed gypsy moth data an artificial long-range jump is introduced in Michigan in 1950. The simulation is run for 107 years with an annual timestep.

The time until the invasion front reaches each county is recorded as the simulated quarantine data (figure 4*b*). Figure 4*a* indicates that our automated method successfully identified the two fixed colony introductions as regions of long-range jumps. We recover mean spread rates in the west of 10.7 km year^{-1} and in the east of 21.4 km year^{-1}, close to the true values used in the simulation. We also test our method under two different simulation scenarios—slow spread and fast spread of the invasive species. Our method successfully detects long-range jumps and recovers the true spread rates well under both scenarios (see electronic supplementary material for details).

5. Discussion

To study the establishment and spread of biological invasions, we present a new method to estimate local rates and direction of spread, and identify key spatial features including sources, sites of rapid spread, and long-range jumps. We visualize and make inferences on historical patterns of spread of the gypsy moth and HWA as well as validate the methodology on simulated data. Posterior inference in a Bayesian setting allows us to test the significance of spread patterns and spatial features of these invasions in a statistically rigorous way.

Taking our local estimates of gypsy moth spread and averaging them across time yields results in line with previous estimates [27]. We find an average speed of 11.4 km year^{-1} across counties quarantined from 1900 to 1915, followed by a slow spread (5.0 km year^{-1}) across counties from 1916 to 1965, and then a period of very rapid expansion (25.8 km year^{-1}) from 1966 to 2000. These changes may also be related to the differences in Allee effects among different regions along the invasion front as evidenced in [56]. From 2000 to present, coincident with USDAs 'Slow the Spread' program of control [8] we calculate an average speed of 14.6 km year^{-1}.

Our estimates for the spread of HWA when spatially averaged are also in line with previous estimates (e.g. [36]). There is evidence that HWA range expansion is limited both by a lack of host trees and in the north by winter temperatures. We note the important role of wildlife, especially migratory birds, as a means for HWA movement [57], in addition to windborne dispersal and human transport [35].

Our abilities to identify patterns of spread are constrained by the spatial and temporal resolution of our data. County-level quarantine data are typically coarser than, for example, gypsy moth pheromone trap count data, though [24] showed the two sources of gypsy moth data provided similar spread estimates. Additionally, the original Gaussian process must be sufficiently smooth for a gradient process to exist (we take the Matérn model with smoothness parameter $\nu = \frac{3}{2}$), with the consequence that some information is lost at local scales. We rely for the most part on annual records, but before 2001 the range of HWA was recorded at less frequent intervals. This is a potential source of bias in our early analysis of HWA spread.

For large spatial datasets, fitting a Gaussian process is a computational burden. Once the original Gaussian process is fitted, however, we can draw samples by composition from the gradient process quickly. When the number of spatial locations is in the thousands we have to rely on approximations such as the predictive process model of [58].

Generally, whenever the data are point-referenced waiting times, the speeds of spread can be estimated from the inferred gradient process. Therefore, the methods presented here should be generally applicable to both ecological and epidemiological invasions. These methods are also potentially applicable to non-invasion problems such as the spread of an advantageous allele [59], or recurrent outbreak waves [46]. An R package that automates the inference is available in the electronic supplementary material.

Ethics. This study was conducted in accordance with the laws, guidelines, and ethical standards of the United States, where this statistical methodology was developed and data sets were analysed.

Data accessibility. Data available from the following repository: http://www.personal.psu.edu/muh10/invasionSpeed.html.

Authors' contributions. All authors contributed to project conception. J.G., J.P., M.H., and O.N.B. developed the statistical methods used in the manuscript. J.G. and J.P. wrote the computer code. J.G. and J.P. analysed the data and conducted simulation studies, with input from M.H. and O.N.B., J.G., J.P., and M.H. wrote the manuscript and electronic supplementary material, with input from O.N.B. and A.L.; O.N.B. and A.L. provided the datasets used in the analysis.

Competing interests. We have no competing interests.

Funding. This work has been funded by the Bill and Melinda Gates Foundation and by the National Science Foundation, grant no. DEB-1354819.

References

1. Lockwood JL, Hoopes MF, Marchetti MP. 2013 *Invasion ecology*. New York, NY: John Wiley & Sons.
2. Skellam JG. 1951 Random dispersal in theoretical populations. *Biometrika* **38**, 196–218. (doi:10.1093/biomet/38.1-2.196)
3. Okubo A, Okubo A. 1980 *Diffusion and ecological problems: mathematical models*, vol. 10. Berlin, Germany: Springer.
4. Shigesada N, Kawasaki K, Teramoto E. 1987 The speeds of traveling frontal waves in heterogeneous environments. In *Mathematical Topics in Population Biology, Morphogenesis and Neurosciences*, pp. 88–97. Berlin, Germany: Springer.
5. Tobin PC, Whitmire SL, Johnson DM, Bjørnstad ON, Liebhold AM. 2007 Invasion speed is affected by geographical variation in the strength of Allee effects. *Ecol. Lett.* **10**, 36–43. (doi:10.1111/j.1461-0248.2006.00991.x)
6. Shigesada N, Kawasaki K, Takeda Y. 1995 Modeling stratified diffusion in biological invasions. *Am. Nat.* **146**, 229–251. (doi:10.1086/285796)
7. Suarez AV, Holway DA, Case TJ. 2001 Patterns of spread in biological invasions dominated by long-distance jump dispersal: insights from Argentine ants. *Proc. Natl Acad. Sci. USA* **98**, 1095–1100. (doi:10.1073/pnas.98.3.1095)
8. Sharov AA, Leonard D, Liebhold AM, Roberts EA, Dickerson W. 2002 "Slow The Spread": a national program to contain the gypsy moth. *J. Forestry* **100**, 30–36.
9. Guisan A, Zimmermann NE. 2000 Predictive habitat distribution models in ecology. *Ecol. Model.* **135**, 147–186. (doi:10.1016/S0304-3800(00)00354-9)
10. Stauffer DF. 2002 Linking populations and habitats: Where have we been? Where are we going? In *Predicting Species Occurrences: Issues of Accuracy and Scale*, pp. 53–61. Washington, D.C.: Island Press.
11. Guisan A, Thuiller W. 2005 Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009. (doi:10.1111/ele.2005.8.issue-9)
12. Elith J, Leathwick JR. 2009 Species distribution models: ecological explanation and prediction across space and time. *Ann. Rev. Ecol. Evol. Syst.* **40**, 677–697. (doi:10.1146/annurev.ecolsys.110308.120159)
13. Elith J. 2015 Predicting distributions of invasive species. In *Risk-based decisions for biological threats*. Cambridge, UK: Cambridge University Press.
14. Wikle CK. 2003 Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* **84**, 1382–1394. (doi:10.1890/0012-9658(2003)084[1382:HBMFPT]2.0.CO;2)
15. Hooten MB, Wikle CK, Dorazio RM, Royle JA. 2007 Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics* **63**, 558–567. (doi:10.1111/biom.2007.63.issue-2)
16. Hooten MB, Wikle CK. 2008 A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environ. Ecol. Stat.* **15**, 59–70. (doi:10.1007/s10651-007-0040-1)
17. Wikle CK, Hooten MB. 2010 A general science-based framework for dynamical spatio-temporal models. *Test* **19**, 417–451. (doi:10.1007/s11749-010-0209-z)
18. Hooten MB, Wikle CK. 2010 Statistical agent-based models for discrete spatio-temporal systems. *J. Am. Stat. Assoc.* **105**, 236–248. (doi:10.1198/jasa.2009.tm09036)
19. Bled F, Royle JA, Cam E. 2011 Hierarchical modeling of an invasive spread: the Eurasian Collared-Dove Streptopelia decaocto in the United States. *Ecol. Appl.* **21**, 290–302. (doi:10.1890/09-1877.1)
20. Broms KM, Hooten MB, Johnson DS, Altwegg R, Conquest LL. 2016 Dynamic occupancy models for explicit colonization processes. *Ecology* **97**, 194–204. (doi:10.1890/15-0416.1)
21. Latimer A, Banerjee S, Sang Jr H, Silander Jr J. 2009 Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecol. Lett.* **12**, 144–154. (doi:10.1111/ele.2009.12.issue-2)
22. Hanks EM. 2017 Modeling spatial covariance using the limiting distribution of spatio-temporal random walks. *J. Am. Stat. Assoc.* **112**, 497–507. (doi:10.1080/01621459.2016.1224714)
23. Sharov AA, Liebhold AM, Roberts AE. 1997 Methods for monitoring the spread of gypsy moth (Lepidoptera: Lymantriidae) populations in the Appalachian Mountains. *J. Econ. Entomol.* **90**, 1259–1266. (doi:10.1093/jee/90.5.1259)
24. Tobin PC, Liebhold AM, Anderson Roberts E. 2007 Comparison of methods for estimating the spread of

a non-indigenous species. *J. Biogeogr.* **34**, 305–312. (doi:10.1111/jbi.2007.34.issue-2)

25. Gilbert M, Liebhold A. 2010 Comparing methods for measuring the rate of spread of invading populations. *Ecography* **33**, 809–817. (doi:10.1111/j.1600-0587.2009.06018.x)
26. Liebhold A, Mastro V, Schaefer P. 1989 Learning from the legacy of Leopold Trouvelot. *Bull. ESA* **35**, 20–22. (doi:10.1093/besa/35.2.20)
27. Liebhold AM, Halverson JA, Elmes GA. 1992 Gypsy moth invasion in North America: a quantitative analysis. *J. Biogeogr.* **19**, 513–520. (doi:10.2307/2845770)
28. Johnson DM, Liebhold AM, Tobin PC, Bjørnstad ON. 2006 Allee effects and pulsed invasion by the gypsy moth. *Nature* **444**, 361–363 (doi:10.1038/nature05242)
29. Liebhold AM, Tobin PC. 2008 Population ecology of insect invasions and their management. *Annu. Rev. Entomol.* **53**, 387–408. (doi:10.1146/annurev.ento.52.110405.091401)
30. Mason C, McManus M. 1981 Larval dispersal of the gypsy moth. *The gypsy moth: research toward integrated pest management. US Department of Agriculture Technical Bulletin* **1584**, 161–202.
31. Burgess AF. 1913 *The dispersion of the gypsy moth*, Bulletin 119, p. 62. Washington, D.C.: United States Department of Agriculture Bureau of Entomology.
32. Burgess AF. 1915 *Report on the gipsy moth work in New England*, Bulletin 204, p. 32. Washington, D.C.: United States Deptartment of Agriculture.
33. Liebhold AM, Gottschalk KW, Luzader ER, Mason DA, Bush R, Twardus DB. 1997 Gypsy moth in the United States: an Atlas. General Technical Report-Northern Research Station, USDA Forest Service, (NE-233).
34. Orwig DA, Foster DR, Mausel DL. 2002 Landscape patterns of hemlock decline in New England due to the introduced hemlock woolly adelgid. *J. Biogeogr.* **29**, 1475–1487. (doi:10.1046/j.1365-2699.2002.00765.x)
35. Morin RS, Liebhold AM, Gottschalk KW. 2009 Anisotropic spread of hemlock woolly adelgid in the eastern United States. *Biol. Invasions* **11**, 2341–2350. (doi:10.1007/s10530-008-9420-1)
36. Ward JS, Montgomery ME, Cheah CA-J, Onken BP, Cowles RS. 2004 Eastern hemlock forests: guidelines to minimize the impacts of hemlock woolly adelgid. USDA Forest Service Northeastern Research Station Research Paper.
37. Evans AM, Gregoire TG. 2007 A geographically variable model of hemlock woolly adelgid spread. *Biol. Invasions* **9**, 369–382. (doi:10.1007/s10530-006-9039-z)
38. Morin RS, Liebhold AM, Luzader ER, Lister AJ, Gottschalk KW, Twardus DB. 2004 Mapping host-species abundance of three major exotic forest pests. USDA Forest Service Northeastern Research Station Research Paper, (NE-726).
39. Cathey HM. 1990 *USDA plant hardiness zone map*. Misc. Publication 1475. Washington, D.C.: United States Department of Agricultural Research Service.
40. Womble WH. 1951 Differential systematics. *Science* **114**, 315–322. (doi:10.1126/science.114.2961.315)
41. Barbujani G, Oden N, Sokal R. 1989 Detecting regions of abrupt change in maps of biological variables. *Syst. Biol.* **38**, 376–389. (doi:10.2307/2992403)
42. Fortin M-J. 1994 Edge detection algorithms for two-dimensional ecological data. *Ecology* **75**, 956–965. (doi:10.2307/1939419)
43. Banerjee S, Gelfand AE, Carlin BP. 2004 *Hierarchical modeling and analysis for spatial data*. CRC Press.
44. Fitzpatrick MC, Preisser EL, Porter A, Elkinton J, Waller LA, Carlin BP, Ellison AM. 2010 Ecological boundary detection using Bayesian areal wombling. *Ecology* **91**, 3448–3455. (doi:10.1890/10-0807.1)
45. Lu H, Reilly C, Banerjee S, Carlin B. 2007 Bayesian areal wombling via adjacency modeling. *Environ. Ecol. Stat.* **14**, 433–452. (doi:10.1007/s10651-007-0029-9)
46. Johnson DM, Bjørnstad ON, Liebhold AM. 2004 Landscape geometry and travelling waves in the larch budmoth. *Ecol. Lett.* **7**, 967–974. (doi:10.1111/ele.2004.7.issue-10)
47. Farnsworth ML, Ward MP. 2009 Identifying spatio-temporal patterns of transboundary disease spread: examples using avian influenza H5N1 outbreaks. *Vet. Res.* **40**, 1–14. (doi:10.1051/vetres/2009003)
48. Krige DG. 1951 A statistical approach to some basic mine valuation problems on the witwatersrand. *J. Chem. Metall. Min. Soc. South Africa* **52**, 119–139.
49. Banerjee S, Gelfand A, Sirmans C. 2003 Directional rates of change under spatial process models. *J. Am. Stat. Assoc.* **98**, 946–954. (doi:10.1198/C16214503000000909)
50. Ihaka R, Gentleman R. 1996 R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314. (doi:10.2307/1390807)
51. R Core Team. 2013 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
52. Jammalamadaka SR, Sengupta A. 2001 *Topics in circular statistics*, vol. 5. Singapore: World Scientific.
53. Finley AO, Banerjee S, Carlin BP. 2007 spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *J. Stat. Softw.* **19**, 1–24. (doi:10.18637/jss.v019.i04)
54. Chen M-H, Shao Q-M, Ibrahim JG. 2000 *Monte Carlo methods in Bayesian computation*. New York, NY: Springer.
55. Trotter RT, Shields KS. 2009 Variation in winter survival of the invasive hemlock woolly adelgid (Hemiptera: Adelgidae) across the eastern United States. *Environ. Entomol.* **38**, 577–587. (doi:10.1603/022.038.0309)
56. Tobin PC, Robinet C, Johnson DM, Whitmire SL, Bjørnstad ON, Liebhold AM. 2009 The role of Allee effects in gypsy moth, *Lymantria dispar* (L.), invasions. *Population Ecol.* **51**, 373–384. (doi:10.1007/s10144-009-0144-6)
57. McClure MS. 1990 Role of wind, birds, deer, and humans in the dispersal of hemlock woolly adelgid (Homoptera: Adelgidae). *Environ. Entomol.* **19**, 36–43. (doi:10.1093/ee/19.1.36)
58. Banerjee S, Gelfand AE, Finley AO, Sang H. 2008 Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. B (Stat. Methodol.)* **70**, 825–848. (doi:10.1111/rssb.2008.70.issue-4)
59. Fisher RA. 1937 The wave of advance of advantageous genes. *Ann. Eugenics* **7**, 355–369. (doi:10.1111/j.1469-1809.1937.tb02153.x)