*Ecology Letters*, (2009) **12**: 144–154 doi: 10.1111/j.1461-0248.2008.01270.x

LETTER

# Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States

**A. M. Latimer,[1]* S. Banerjee,[2] H. Sang,[3] E. S. Mosher[4] and J. A. Silander Jr[5]**

[1]*Department of Plant Sciences, University of California, Davis, CA, 95616, USA*

[2]*Department of Biostatistics, University of Minnesota, Minneapolis, MN, 56093, USA*

[3]*Department of Statistics, Texas A&M University, College Station, TX, 77843, USA*

[4]*CT River Coastal Conservation District, Middletown, CT, 06457, USA*

[5]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, 06269, USA*

**Correspondence: E-mail: amlatimer@ucdavis.edu*

## Abstract

Many critical ecological issues require the analysis of large spatial point data sets – for example, modelling species distributions, abundance and spread from survey data. But modelling spatial relationships, especially in large point data sets, presents major computational challenges. We use a novel Bayesian hierarchical statistical approach, 'spatial predictive process' modelling, to predict the distribution of a major invasive plant species, *Celastrus orbiculatus*, in the northeastern USA. The model runs orders of magnitude faster than traditional geostatistical models on a large data set of *c.* 4000 points, and performs better than generalized linear models, generalized additive models and geographically weighted regression in cross-validation. We also use this approach to model simultaneously the distributions of a set of four major invasive species in a spatially explicit multivariate model. This multispecies analysis demonstrates that some pairs of species exhibit negative residual spatial covariation, suggesting potential competitive interaction or divergent responses to unmeasured factors.

## Keywords

Bayesian, computational limitation, invasive species, multivariate spatial models, point-referenced, spatial modelling, spatial predictive process, species distributions.

*Ecology Letters* (2009) 12: 144–154

## INTRODUCTION

Many of the most pressing issues in ecology, theoretically and practically, involve the analysis of spatial data (Guisan & Zimmerman 2000; Guisan & Thuiller 2005). Assessing species abundances and suitable habitat relies on spatial environmental and species occurrence data (Parmesan *et al.* 2005; Helmuth *et al.* 2006; Loarie *et al.* 2008), as does evaluating explanations for diversity patterns and gradients (Palmer 1996; Whittaker *et al.* 2001; Hawkins *et al.* 2003). The ecological mechanisms involved in generating these patterns (Chesson 2000; Chave *et al.* 2002) strongly suggest that the spatial relationships among the data points often contain information important to making good predictions, inferences and interpretations (Ver Hoef *et al.* 2001; Banerjee *et al.* 2004; Beale *et al.* 2007; Van Teeffelen & Ovaskainen 2007). Nonetheless, ecological studies often neglect to model spatial relationships (Beale *et al.* 2007), often for the simple reason that it has remained computationally prohibitive to run spatially explicit models for large, point-based data sets (Banerjee *et al.* 2004). This constraint is becoming more acute as large (>> 10 000 points) spatial data sets continue to be produced through large-scale inventories and remote sensing. In this paper, we present a novel and straightforward approach, the 'spatial predictive process' model, which dramatically lowers this computational barrier (Banerjee *et al.* 2008). This approach makes it feasible to use spatially explicit models for very large spatial data sets, as well as to simultaneously model multiple ecological processes to explore spatial relationships among them.

A rich spatial statistics literature has developed on models for georeferenced point data (Ripley 1988; Cressie 1993; Møller & Waagepetersen 2003; Banerjee *et al.* 2004; Wackernagel 2006). This work has focused primarily on 'point-referenced' (or 'geostatistical') models, which use functions of the spatial relationships among fixed sample points to draw inferences about spatial autocorrelation and to make predictions (Dormann *et al.* 2007). The simplest point-referenced models assume that the process that generated

the observed data has a spatial component, and represent this spatial component through functions of distance between observation points. The best-known example of this approach is kriging, which was developed to predict locations of ore deposits from point-level drill samples (Banerjee *et al.* 2004). Ecological point-referenced modelling has been limited to relatively modest-sized data sets until recently (Latimer *et al.* 2006; Bellier *et al.* 2007; Illian *et al.* 2007). General-purpose spatial statistical software such as GeoBUGS (Thomas *et al.* 2004) becomes overwhelmed by spatial process models using many more than 100 locations – a small number compared to the size of many ecological data sets, and vastly smaller than most climatological and remotely sensed data sets. To take full advantage of the growing flood of spatial and spatiotemporal data, we need models that can incorporate arbitrarily large numbers of points. It would also be useful to be able to fit multiple spatial processes simultaneously – for example, when modelling spatial dependence for multiple time steps in the spread of disease (LaDeau *et al.* 2007) or an invading organism (Wikle 2003), each species in an assemblage (Illian *et al.* 2007), or each genetic marker in an assay (Vounatsou *et al.* 2000; Yeang & Haussler 2007).

We present a solution to this 'many-sites' problem that rests on the insight that modelled representations of a spatial process do not have to be tied to the sample point locations in the data set. Instead, we can anchor the spatial process at a smaller number of points – i.e. a lower dimensional space – and use spatial prediction to link this unobserved or 'latent' process back to the sample locations. Models using this approach are called 'predictive process' models, in reference to the predictive link between the lower dimensional spatial process and the full set of sampling locations (Banerjee *et al.* 2008). Using conditional modelling in a Bayesian hierarchical framework, this method can be implemented as a fully specified statistical model for even large data sets, so that modelling assumptions are explicit and the model quantifies the uncertainty associated with its predictions.

Ecologists often collect and analyse measurements from point locations ('point-referenced data'), e.g. plot or sample locations, specimen collections, weather stations, etc. But there is currently no standard modelling package that can perform spatially explicit point-level predictive modelling of large (much more than *c.* 1000 points) data sets. The packages that do perform species distribution modelling are growing ever-more sophisticated, but most cannot perform spatially explicit analysis of point data, a potentially serious shortcoming, given the importance of spatial autocorrelation in much ecological data and the increasing prevalence of point data (Ver Hoef *et al.* 2001; Beale *et al.* 2007). Methods for spatially explicit analysis of point data, on the other hand, remain limited to relatively small data sets for computational reasons. To the extent that existing species distribution modelling packages include spatially explicit modelling, they do so via random effects at an areal level, as in conditional autoregressive models (Besag 1974; Banerjee *et al.* 2004), or through empirical, plug-in estimates of spatial parameters such as the overall variance and scale of spatial autocorrelation. Moving from point to areal level modelling sacrifices explicit inference about correlation structure and scale, as well as the ability to predict geospatially to new locations. Using plug-in estimates for spatial parameters, on the other hand, means that uncertainty in these parameters is not propagated through the models to the model predictions and inferences about regression parameters.

We demonstrate the spatial predictive process method to address a common and important ecological problem, prediction of species distributions using field survey records of presence/absence. We use data for invasive plant species at geo-referenced locations in the New England region the northeastern USA, and model at two nested scales: regional and local landscape. We focus on the invasive woody vine *Celastrus orbiculatus* ('Asiatic bittersweet'), which is one of the most prevalent and rapidly spreading invasive plant species in eastern North America (Mehrhoff *et al.* 2003). We use cross-validation to compare the predictive performance of this spatially explicit model to three widely used alternative methods for handling point data: generalized linear models, generalized additive models (GAMs) and geographically weighted regression (GWR).

Moving from this larger, regional scale to a nested, local scale (Fig. 1), we demonstrate how this approach can also be used to analyse the fine-scale spatial association among species. Using a multivariate spatial regression to analyse the local distributions of four invasive plant species, including *C. orbiculatus*, within a smaller area (< 100 $km^2$), we quantify the degree of residual spatial association among these species. We conducted this study to assess the effects of present environmental conditions and land use history on invasive woody plant species that have become widespread in this heterogeneous landscape.

## MATERIALS AND METHODS

### Data

At the regional scale, we used one of the best available data sets on the presence/absence of species at point locations – the IPANE data set, assembled by a combination of volunteers and professional botanists, which records the presence and absence of invasive plant species throughout the New England region of North America (Mehrhoff *et al.* 2003). This data set currently includes 5000+ data points

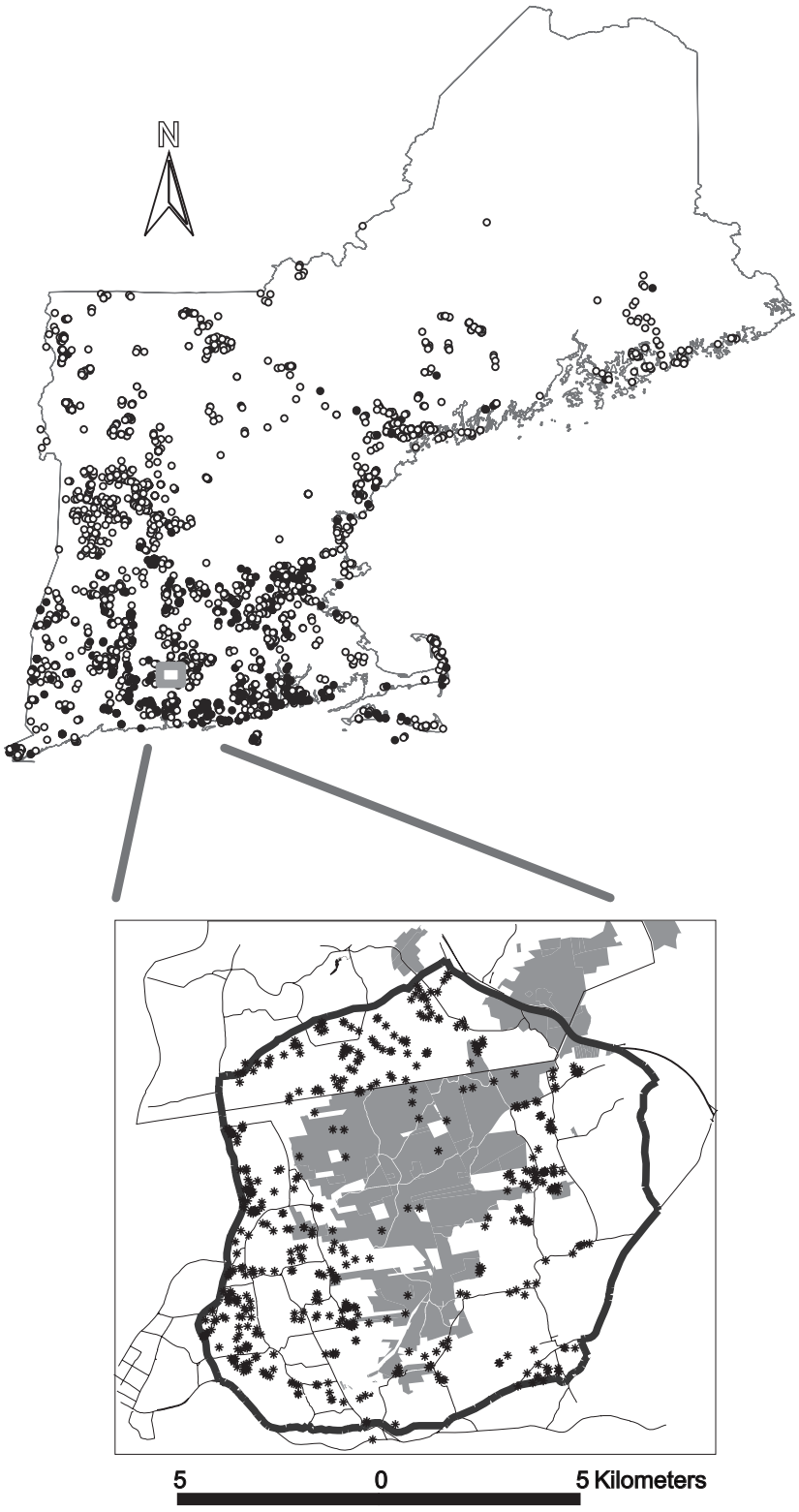


**Figure 1** Map of the New England region of North America, with circles denoting sample locations from the IPANE data set and presence (filled circles) or absence (unfilled circles) of *Celastrus orbiculatus*. The pullout box shows the Meshomasic Forest landscape, with all sample locations shown as asterisks. For context, major roads are shown as thin, black lines, and state forest land is shaded grey.

and covers the entire region (*c.* 180 000 km$^2$; Fig. 1). To assess the predictive performance of the predictive process model at this large scale, while eliminating potential problems of overfitting the modelled data, we held out a randomly selected subset of 10% of the data, and made predictions to these points for cross-validation.

In the local-scale analysis, we used a smaller but still spatially rich and complex data set from an intensive, stratified random sampling of invasive species in central Connecticut, USA (Fig. 1). At each of 603 geocoded points in this area, the presence and absence of several invasive woody plant species were surveyed by direct field observation. In this local-scale analysis, we focused on the four-most abundant woody invasive species: the vine *C. orbiculatus* and the shrubs *Berberis thunbergii* ('Japanese barberry'), *Rosa multiflora* ('multiflora rose') and *Euonymus alatus* ('winged burning bush').

### 'Predictive process' – what and why?

In a standard geostatistical or point-process model, the response variable is related at every sample point to explanatory variables and to spatially correlated errors. The errors may also include a component of pure, uncorrelated error (a 'nugget'), or all error may be assumed to be spatial. A simple spatially explicit regression model with a continuous response variable, normally distributed errors and one explanatory variable can be written as:

$$y_i = x_i\beta + w_i,$$

where $i \in \{1, \ldots, n\}$ indexes sample location, and where the vector of errors ($\boldsymbol{W}$) is given a multivariate normal distribution: $\boldsymbol{W} \sim \mathrm{MVN}(0, \boldsymbol{\Sigma})$ (Congdon 2003; Banerjee *et al.* 2004). The covariance matrix $\boldsymbol{\Sigma}$ incorporates the spatial association. If spatial association is assumed to vary only with distance (i.e. isotropic or independent of direction), then $\boldsymbol{\Sigma}$ can be represented as a function ($H$) representing the decay in correlation between pairs of points with distance multiplied by an error parameter ($\sigma$). $H$ can take several forms, the most common being exponential and Gaussian, and incorporating at least one parameter describing how rapidly correlation declines with distance between points $i$ and $j$ ($\delta_{ij}$). For exponential correlation between points $i$ and $j$, $H(\delta_{ij}) = \mathrm{e}^{-\phi\delta_{ij}}$, and so the elements of the covariance matrix are $\boldsymbol{\Sigma}_{ij} = \sigma\mathrm{e}^{-\phi\delta_{ij}}$, where $i,j \in \{1, \ldots, n\}$.

The problem with fitting even this simple model is that fitting the model to the data requires obtaining the inverse of the dense covariance matrix. As the number of points $n$ increases, $\boldsymbol{\Sigma}$ grows with $n^2$, and the number of computations to invert $\boldsymbol{\Sigma}$ scales as $n^3$. The predictive process approach reduces the dimension of the matrix that has to be inverted, greatly speeding the computation. Instead of directly modelling each error $w_{ik}$, we introduce a second stage to the model, which consists of a spatial process tied to a smaller number of points $m < n$, which we call 'knots'.

The basic idea behind the predictive process is that a representative set of locations ('knots') in the spatial domain should contain enough information to estimate the underlying spatial process while using all the sampling locations is likely to be computationally wasteful. This idea has also been used to develop low-rank smoothing splines by Kamman & Wand (2003). When data locations are fairly evenly distributed across the domain, it is sensible to select knots on a uniform grid overlaid on the domain. A set of knots can be selected from such a grid using a formal design-based approach to minimize some spatially averaged predictive variance criterion (see, e.g. Diggle & Lophaven 2006). However, in our regional invasive species data set, spacing of the locations is highly irregular, generating substantial areas of sparse observations where we wish to avoid placing many knots, as they would be 'wasted' there. So a better approach is to use a space-covering design, such as that developed in Royle & Nychka (1998) implemented in the R library fields (Fields Development Team 2006). For our regional model, we used this algorithm to select different numbers of knots ranging from 50 to 500, fitted the model for each number and selected a number of knots (377) where the model comparison scores and cross-validation performance had approximately levelled out. The precise number used will depend on the individual problem, but our experience is that typically 100–400 knots are sufficient (Banerjee *et al.* 2008).

In the single-species, regional model, we specify a spatial process $\boldsymbol{W}^*$ that is anchored at these $m$ knots, rather than at the $n$ sample locations:

$$\boldsymbol{W}^* \sim \mathrm{MVN}(0, \boldsymbol{\Sigma}^*),$$

where the elements of the covariance matrix $\boldsymbol{\Sigma}^*$ are, for an exponential correlation function:

$$\boldsymbol{\Sigma}^*_{ij} = \sigma H(\delta_{ij}) = \sigma\mathrm{e}^{-\phi\delta_{ij}}, \quad \text{where } i,j \in \{1, \ldots, n\}.$$

The key step relates this lower-dimensional spatial process back to the $n$ sample locations. This is done through prediction (analogous to kriging), where the values of the $n$ spatial random effects ($\boldsymbol{W}$) are predicted from the $m$ values of the predictive process ($\boldsymbol{W}^*$):

$$\boldsymbol{W} = \boldsymbol{\Sigma}'_{(\boldsymbol{W},\boldsymbol{W}^*)}\boldsymbol{\Sigma}^{*-1}\boldsymbol{W}^*,$$

where $\boldsymbol{W}$ is the $1 \times n$ vector of spatial random effects for the sample locations, and $\boldsymbol{W}^*$ is the $1 \times m$ vector of realizations of the latent spatial process at the predictive process points. The matrix $\boldsymbol{\Sigma}_{(\boldsymbol{W},\ \boldsymbol{W}^*)}$ is the cross-covariance matrix between $\boldsymbol{W}$ and $\boldsymbol{W}^*$, which describes the spatial relationships between the $m$ knots and the $n$ sample locations. In this matrix, the $i$, $j$th entry is $\sigma H(\delta_{ij})$, where $\delta_{ij}$ is the distance between the $i$th sample location and the $j$th knot. This predictive step is fully integrated into the model: the spatial parameters $\sigma$ and $\boldsymbol{\Phi}$ are fitted in a lower-dimensional subspace along with the spatial surface $\boldsymbol{W}^*$, then these fitted spatial parameters are used to interpolate $\boldsymbol{W}^*$ to the higher-dimensional space defined by the sampling locations, where the likelihood is evaluated. Therefore, the scale and underlying noise

parameters and their associated uncertainty propagate through the first (regression) stage of the model, preserving the advantages of statistically modelling spatial association (Banerjee *et al.* 2008; Finley *et al.* 2008).

The predictive process model is thus a two-stage hierarchical model. The first stage is the ecological process model that relates the observed data to the explanatory variables and the spatial component $\boldsymbol{W}$. The second stage is the spatial process model for $\boldsymbol{W}^*$. The levels are linked by predicting $\boldsymbol{W}$ from $\boldsymbol{W}^*$. In our example of invasive plant species, presence/absence (the response variable) is binary, so we use a probit link to relate it to the environmental explanatory data. The probit specification introduces a latent 'intensity surface' $\boldsymbol{Z}$ that represents the relative probability that the species is present. Alternatively we could use a logistic link, but the probit has computational advantages in fitting the model, because it allows the regression coefficients to be sampled using the more efficient Gibbs sampler rather than Metropolis-Hastings updates (Congdon 2003; Robert & Casella 2004). So, in sum, the ecological process part of the model is a standard probit model with spatially correlated errors:

$$y_i = \begin{vmatrix} 0 \ldots z_i \leq 0 \\ 1 \ldots z_i > 0 \end{vmatrix}, \quad i \in \{1, \ldots n\}, \text{and}$$

$$\boldsymbol{Z} = \mathbf{X}'\beta + \boldsymbol{W}$$

where $\boldsymbol{Z}$ is the vector of the values of the latent probit-scale intensity surface, $\mathbf{X}$ is the matrix of explanatory variables (including an initial column of ones for an intercept), $\beta$ is the vector of regression coefficients and $\boldsymbol{W}$ is the vector of spatial random effects. Note that this approach can easily be generalized to predict abundances or number of species through a cumulative probit link to ordinal abundance classifications (Congdon 2003).

We can then complete the specification of a Bayesian hierarchical model by assigning prior distributions to the parameters. We assigned vague normal distributions to the regression coefficients ($\beta$). We constrained the spatial decay parameter ($\varphi$) to a biologically plausible range of values, because typically it is not possible to identify both the spatial decay and spatial variance parameters well in spatial models (Banerjee *et al.* 2004). We fitted the model using Markov Chain Monte Carlo (MCMC) methods, which uses repeated stochastic sampling to characterize the posterior distributions of the model parameters (Gelman *et al.* 2004). The models presented here were implemented as stand-alone programs in R 2.7 (R Development Core Team 2008), and code is available on request, but we note that there is now an R library available for running spatial predictive process models for Gaussian data (Finley *et al.* 2007), and extensions for this library are planned that will enable the analysis of binary data as here.

## Model comparison

To assess whether the spatial predictive process model provides a predictive advantage over alternative models, we also fitted three alternative models to the same data and assessed their performance in cross-validation with held out data. The simplest model was a logistic regression, in which spatial relationships were represented as terms for 'northing' and 'easting' (i.e. spatial trend), which is an approach still commonly used (Guisan & Zimmerman 2000). We also fitted what is currently probably the most widely used kind of model for species distributions, the GAM, again with trend terms. Finally, we fitted a GWR to the data. The GWR uses a least-squares fitting approach to allow the coefficients for the environmental covariates to vary spatially by conducting weighted local regressions at each point (Fotheringham *et al.* 2002). We compared the three models using cross-validation; we summarize the performance using the area under the receiver operating characteristics curve as an integrated measure of the discriminative power of the models without assigning an arbitrary classification threshold (Fielding & Bell 1997; Brotons *et al.* 2004).

## Multivariate spatial models

Ecologists often face questions of whether multiple ecological phenomena are spatially correlated and whether these correlations remain after taking into account environmental variation. For example, we could be interested in the distributions of two potentially interacting species such as two plant species or two disease organisms (Plotkin *et al.* 2002; Congdon 2003; Illian *et al.* 2007). We can use standard regression and multivariate statistical methods to check for association among the species occurrences, and between them and a set of explanatory variables. But we can rarely capture all spatial pattern through explanatory variables, and frequently we want to know whether the species distributions exhibit residual autocorrelation or other spatial patterns (Bellier *et al.* 2007; Illian *et al.* 2007). More fundamentally, species distributions are generated by ecological mechanisms, including competition, facilitation and dispersal, which are known to generate spatial clustering and other non-random occurrence patterns, so we would generally not expect the 'background' environment to eliminate spatial pattern from ecological data (Ver Hoef *et al.* 2001).

In our local-scale, multispecies case study, we know that all four species are dispersed by many of the same bird species (especially starlings, themselves a spreading invasive species), which suggests we might expect some residual spatial association even after accounting for environmental variation (Lafleur *et al.* 2007). Another motivation for modelling association among the spatial random effects

surfaces for these species is to detect one of two possible patterns. If the spatial association were primarily derived from a common response to unmeasured environmental variables – say fine distinctions among different edge habitats – then we would expect that the more ecologically similar species would exhibit positive cross-correlation. On the other hand, if competition among similar species were an important influence in these species' fine-scaled distribution patterns, we would expect the more ecologically similar species to have negative cross-correlations.

The basic requirement for these kinds of multivariate models is to fit multiple spatial processes simultaneously, while linking them with cross-covariance parameters that represent association among the processes for different species. The simplest way of structuring this kind of model is to include a single cross-covariance parameter for each pair of species; these parameters then summarize the overall positive or negative association among them. More complex models could allow these cross-covariance parameters to vary over time or space (Banerjee *et al.* 2008). A class of model that builds this cross-covariance structure is the 'linear model of coregionalization' (Wackernagel 2006). Here, we implement a linear coregionalization model for the four invasive plant species. As above in the single-species model, this multispecies model assumes that spatial association is isotropic and can be adequately described by a simple, smooth decay function. As we have a binary response variable, we again use a probit link and introduce latent intensity surfaces for the $K = 4$ species. The ecological process model is then the multivariate regression:

$$z_{ik} = \mu_k + x_i\beta_k + w_{ik}, \quad \text{where } i \in \{1, \ldots, n\} \text{ and } k \in \{1, \ldots, K\}.$$

The spatial random effects $w_{ik}$ are predicted from a latent process $\boldsymbol{W}^*$, which now consists of five spatial surfaces related by cross-covariance parameters. In the linear model of coregionalization, the spatial process is separated into two components, a spatial correlation matrix ($\mathbf{H}$) and a matrix of cross-correlations between species ($\boldsymbol{\Gamma}$).

As in the single species model, the spatial component of $\boldsymbol{\Sigma}$ is implemented at a smaller number of knots ($m = 100$) in a second hierarchical stage, then the spatial random effects in the first stage (the $w_{ik}$) are predicted from this latent process. For this analysis, as the sampling scheme was more regular, we used a regular grid of 100 points to anchor the knots. Because of this reduction in dimension, the cross-covariance matrix $\boldsymbol{\Sigma}$ has dimensions $mk \times mk$. Fitting this model to our data set is roughly computationally equivalent to fitting a traditional point-referenced model for a single species with about 500 data points; not trivial, but clearly feasible. The critical economy of this approach is that we could greatly increase the number of observation points $n$ without affecting the dimensions of $\boldsymbol{\Sigma}$. We update the cross-covariance parameters using a Hastings step with corrections for change-in-variables (Green 2003), and update the other parameters as in the single species model.

## RESULTS

### Regional-scale prediction of invasive species prevalence

The regression analysis shows that even when spatial autocorrelation is taken into account, the coefficients for many of the environmental and land-use explanatory variables still figure prominently in explaining pattern (Table 1). *Celastrus orbiculatus* is negatively associated with canopy closure, and positively associated with warmer temperatures and the proximity of roads (Table 1). Other studies have shown that land use is also significantly associated with the presence/absence of this species (Ibanez I., unpublished data). The predicted probability of occurrence shows high probabilities in the southern part of the region and along the coast (Fig. 2a), while the spatial random effect surface has a trough in the northwestern corner (Vermont near Lake Champlain), which reflects the absence of the species from the area, despite high predicted environmental suitability (Fig. 2b).

Cross-validation performance of the spatial predictive process model was higher than for the logistic regression, GAM, and the GWR (Table 2).

### Local-scale, multispecies model of invasive species distribution and correlation

The multispecies spatial model provides simultaneous inference about the individual environmental responses of

**Table 1** Regression coefficient estimates for the regional-scale model of presence/absence of *Celastrus orbiculatus* across the New England Region

| Variable | Posterior mean | 0.025 Quantile | 0.975 Quantile |
|---|---|---|---|
| Maximal temperature of warmest month | 0.222* | 0.092 | 0.361 |
| Minimal temperature of coldest month | 0.282† | −0.010 | 0.631 |
| Annual precipitation | −0.007 | −0.251 | 0.265 |
| Precip seasonality | −0.091 | −0.180 | 0.351 |
| Precip of warmest quarter | 0.008 | −0.395 | 0.377 |
| Roads within 10 km | 0.097† | −0.007 | 0.208 |
| Canopy closure | −0.102* | −0.155 | −0.047 |

*Significant coefficients (95% credible interval excludes 0).
†Suggestive coefficients (90% credible interval excludes 0).

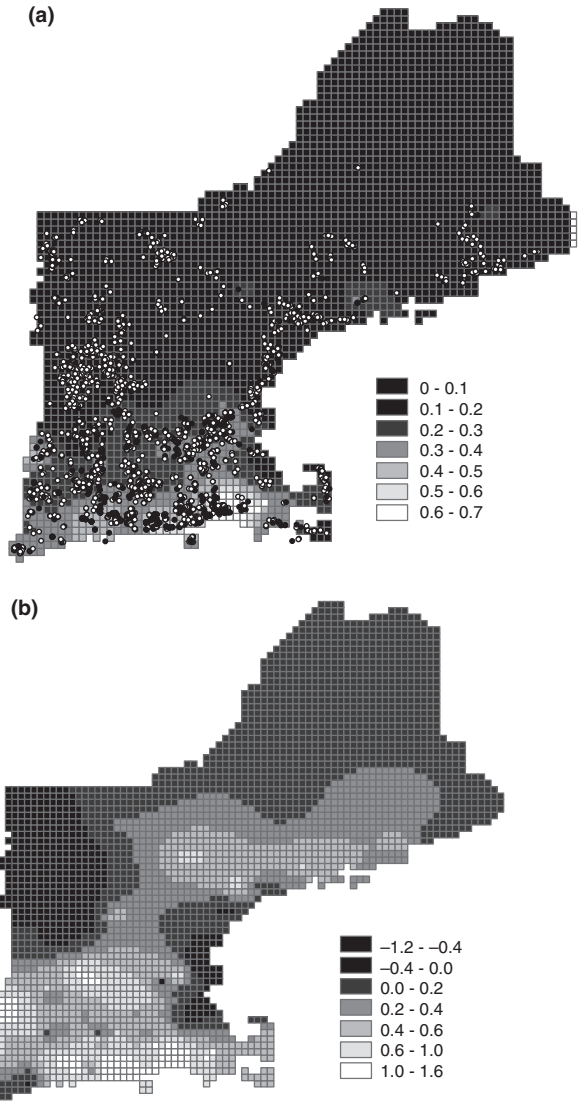


**Figure 2** Spatial predictive process results for the invasive liana *Celastrus orbiculatus* across the New England region: (a) predicted probability of occurrence, with lighter colours indicating increased probability of occurrence; (b) spatial random effects, with lighter colour representing increasingly positive values for the random effect.

the species (Fig. 3), and shows that there are significant differences in their responses to environmental characteristics. The species are grouped into two similar pairs with respect to light levels (canopy closure), with *Berberis* and *Euonymus* exhibiting shade tolerance by responding positively to some level of canopy closure, while *Celastrus* and *Rosa* show no effect. With respect to land use, all species are positively associated with abandoned agricultural areas that have reverted to the forest, and all but *Berberis* are positively associated with residential use and recently abandoned fields. All species respond negatively to intensive management (i.e. mowing or plowing) and positively to association with edge habitat (i.e. decreasing prevalence with increasing distance to edge habitat) (Fig 3).

**Table 2** Predictive performance in cross-validation of the three models: logistic regression with northing and easting terms, generalized additive model with northing and easting terms, geographically weighted regression and spatial predictive process model

| Model | AUC score |
|---|---|
| Logistic regression | 0.665 |
| GAM | 0.696 |
| GWR | 0.671 |
| Predictive process | 0.709 |

The area under the receiver operating characteristics curve is provided as an integrated measure of the power of the model to correctly predict presences and absences.
GAM, generalized additive model; GWR, geographically weighted regression.

The cross-covariance parameters among the species' random effects surfaces, converted to correlations for ease of interpretation and presented in Table 3, show significant negative residual correlations among species. In the case of *Celastrus* vs. both *Berberis* and *Rosa*, the credible intervals exclude 0, indicating a significant negative residual correlation between the spatial surfaces of *Celastrus* and these two shrubs (random effects surfaces shown in Fig. 4).

## DISCUSSION

The performance improvement provided by the predictive process approach will allow spatially explicit point models to be easily implemented on standard desktop computers for many spatial data sets, even those of very large size, so that computation will no longer pose as high a barrier for statistical inference for spatially spatial autocorrelated point data. In the model for *C. orbiculatus* in the northeastern USA, the spatial predictive process approach reduces computation time for a single model-fitting step from *c.* 30 min to 4 s. This improvement of more than 2 orders of magnitude makes it possible to use MCMC methods that would be impractical for the full covariance matrix (in this case, time for a typical 10 000-iteration run would increase from *c.* 11 h to *c.* 200 days without the computational savings from the spatial process model). Improvements would be even more dramatic for larger data sets. This increase in speed also makes many extensions possible, including modelling anisotropy (Banerjee *et al.* 2008) and modelling spatial surfaces that evolve through time (cf. Wikle 2003).

At the regional scale, the model provides robust inference about the environmental relations of the species; for

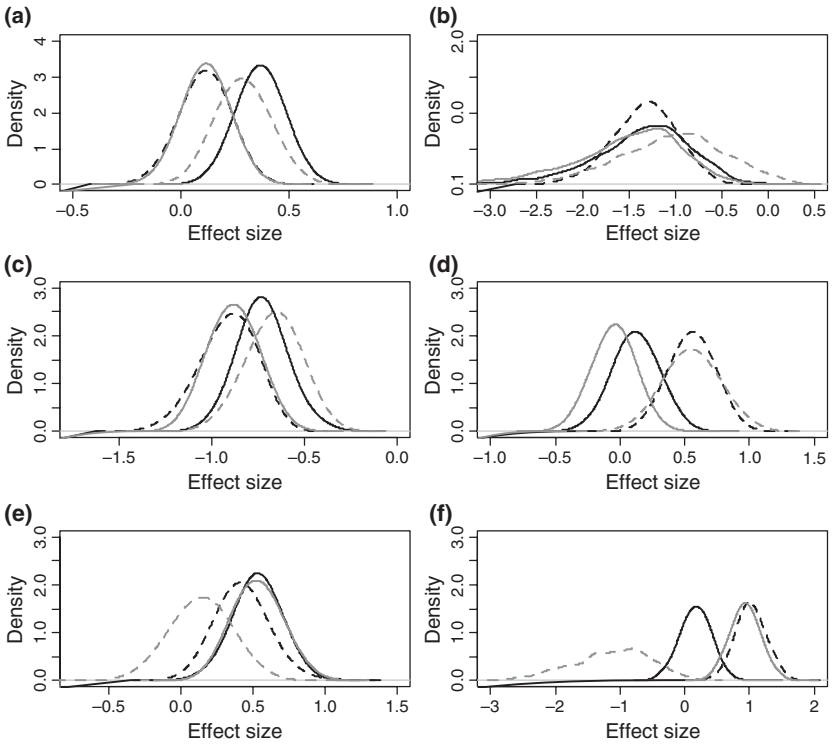


**Figure 3** Posterior probability densities of the regression coefficients for the local-scale model for four invasive species: *Berberis thunbergii* (solid black lines); *Celastrus orbiculatus* (dashed black lines); *Rosa multiflora* (solid grey lines) and *Euonymus alatus* (dashed grey lines). Each panel represents one explanatory variable: (a) canopy closure; (b) heavily managed (1/0); (c) distance from a vegetation edge; (d) residential land use; (e) reforested agricultural land; (f) currently abandoned agricultural land.

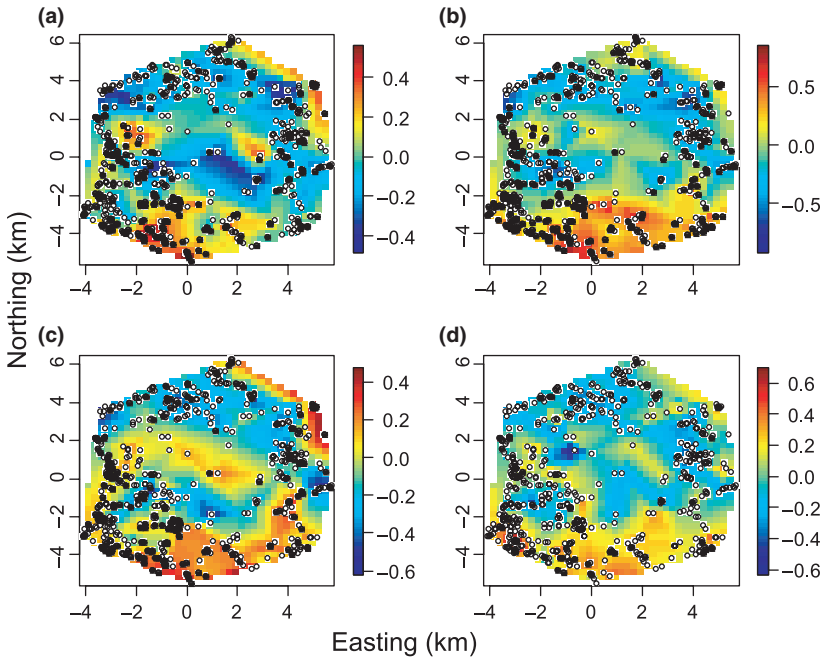


**Figure 4** Results from the multivariate spatial model for four invasive plant species: spatial random effects surfaces for (a) *Berberis thunbergii*; (b) *Celastrus orbiculatus*; (c) *Rosa multiflora* and (d) *Euonymus alatus*. Black squares mark presences; white circles mark absences.

*C. orbiculatus* we see that the species is associated with warmer temperatures and inhibited by high forest canopy cover, as expected for this edge-adapted species (Herron *et al.* 2007; Leicht-Young *et al.* 2007). The predictive power of the model is also superior, with cross-validation performance superior to the logistic regression with trend variables and to GAMs and GWR. Beyond better prediction, we also obtain a full distribution of the predicted probabilities, as well as of the regression coefficients. Unlike the other methods, we obtain inference about the scale of

**Table 3** Cross-correlations among the latent spatial processes for the individual species in the multispecies model

| | *Celastrus orbiculatus* | *Berberis thunbergii* | *Rosa multiflora* | *Euonymus alatus* |
|---|---|---|---|---|
| *C. orbiculatus* | 1 | – | – | – |
| *B. thunbergii* | −0.37* | 1 | – | – |
| *R. multiflora* | −0.25* | 0.01 | 1 | – |
| *E. alatus* | 0.02 | −0.26* | −0.17* | 1 |

Positive values indicate that species are more likely to occur in the same area than otherwise predicted by the environment and the spatial structure of each species' own random effects surface, while negative values indicate the reverse.
*Significant cross-correlation (95% credible interval excludes 0).

residual spatial autocorrelation: the posterior mean of $\Phi$ for the New England regional model (0.11), corresponds to a decline in the correlation of 0.5 in *c.* 6.5 km. Uncertainty about this parameter is relatively large (CI = 0.04 – 0.22) meaning we cannot estimate it precisely; nevertheless it is important to propagate this uncertainty through the model to obtain more robust predictions.

At the local level, the multispecies model provides information about the ecological contrasts among the species. In particular, the regression coefficients for canopy closure rank the species in terms of their tolerance to shade (Fig. 3a). *Celastrus orbiculatus* and *Rosa multiflora* exhibit the greatest response to light availability with an open canopy, while *Euonymus alatus* and *Berberis thunbergii* are more shade-tolerant (Silander & Klepeis 1999), and indeed the latter two are understory shrubs in their native range in Japan (J.A. Silander, Jr., unpublished data). The stronger association of *Celastrus* and *Rosa* with edge habitat (more negative coefficients for distance to a vegetation edge) also reflects their preference for a more open canopy. The model also confirmed that land use influences the local-scale distribution of these species. As none of the land use coefficients were negative for any species, and most were positive, our data support the widespread finding that large stands of mature forest are relatively resistant to invasion by new plant species (DeGasperis & Motzkin 2007). Looking more closely at differences among species reveals some important differences in the effects of land use on the species. As most of the landscape has already reverted to forest, *Berberis*' best window of opportunity for colonization may have passed, while the species better adapted to residential development, *Celastrus* and *Euonymus*, are likely to continue to increase in prevalence (DeGasperis & Motzkin 2007).

The model also reveals substantial spatial variation and covariation that is not readily explained by observed environmental factors. When the spatial patterns for contrasting mechanisms are sufficiently different, we can use such spatially explicit multivariate models to help assess the plausibility of alternative mechanisms or unmeasured environmental variables (e.g. whether spatial clustering and cross-correlation are more consistent with competition or with localized recruitment; Bellier *et al.* 2007; Illian *et al.* 2007; Van Teeffelen & Ovaskainen 2007). This is challenging because often the scale parameters in geospatial models are poorly constrained by the data (Banerjee *et al.* 2004), and it can be difficult in a purely statistical model to link covariance scale with a particular process like the behaviour of a disperser.

Despite these challenges, here we do gain some insight by modelling the species' spatial covariation. Where there are correlations among the species' spatial processes, they tend to be negative (Table 3). For the climbing and edge-adapted species *C. orbiculatus* and *R. multiflora,* this may indicate competitive displacement, although we cannot rule out a role for fine-scale, unmeasured environmental differences. For species that we would consider *a priori* ecologically contrasting, for example *C. orbiculatus* and the shade-tolerant understory shrub *B. thunbergii*, negative residual spatial correlation suggests unmeasured environmental variation to which the species are responding in divergent ways, or, alternatively, pattern retained from contrasting introduction histories.

## CONCLUSION

The spatial predictive process model described here can greatly speed computation in ecological models for point data. This approach offers a statistical method for analysing large point-referenced data sets to learn about environmental relationships in the presence of spatially correlated errors. In addition to making regressions robust to spatial autocorrelation, this approach can help us learn whether two processes are significantly associated in space, and what the scale of their spatial autocorrelation is. This is broadly applicable, as we often want to answer questions such as whether the prevalence of a particular plant trait or animal behaviour is spatially associated with environmental factors, or whether trophically or competitively interacting species show residual spatial association. The simplicity, power and many important potential applications make the spatial predictive process approach a useful addition to ecologists' toolbox.

## ACKNOWLEDGEMENTS

This research was funded by USDA grant NRI 2005–02217 and NSF grant DEB 008901 and DEB 056320 to JAS, and by NSF grant DMS 0706870 to SB. We acknowledge generous assistance and intellectual input from Alan Gelfand of Duke University, and helpful comments from Jérôme Chave and three anonymous referees.

## REFERENCES


Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2004). *Hierarchical Analysis and Modeling for Spatial Data*. Chapman and Hall/CRC, Boca Raton.

Banerjee, S., Gelfand, A.E., Finley, A.O. & Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *J. R. Stat. Soc. B*, 70, 825–848.

Beale, C.M., Lennon, J.J., Elston, D.A., Brewer, M.J. & Yearsley, J.M. (2007). Red herrings remain in geographical ecology: a reply to Hawkins *et al.*. *Ecography*, 30, 845–847.

Bellier, E., Monestiez, P., Durbec, J.-P. & Candau, J.-N. (2007). Identifying spatial relationships at multiple scales: principal coordinates of neighbour matrices (PCNM) and geostatistical approaches. *Ecography*, 30, 385–399.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. B*, 36, 192–236.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27, 437–448.

Chave, J., Muller-Landau, H.C. & Levin, S.A. (2002). Comparing classical community models: theoretical consequences for patterns of diversity. *Am. Nat.*, 159, 1–23.

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.*, 31, 343–366.

Congdon, P. (2003). *Applied Bayesian Models*. John Wiley, Chichester.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc., Chichester.

DeGasperis, B.G. & Motzkin, G. (2007). Windows of opportunity: historical and ecological controls on *Berberis thunbergii* invasions. *Ecology*, 88, 3115–3125.

Diggle, P.J. & Lophaven, S. (2006). Bayesian geostatistical design. *Scand. J. Stat.*, 33, 53–64.

Dormann, C.F., McPherson, J.M., Araújo, M., Bivand, R., Bollinger, J., Carl, G. *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30, 609–628.

Fielding, A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.*, 24, 38–49.

Fields Development Team (2006). *Fields: Tools for Spatial Data*. National Center for Atmospheric Research, Boulder, CO. Available at: http://www.R-project.org, last accessed 24 November 2008.

Finley, A.O., Banerjee, S. & Carlin, B.P. (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *J. Stat. Softw.*, 19, 4.

Finley, A.O., Sang, H., Banerjee, S. & Gelfand, A.E. (2008). Improving the performance of predictive process modeling for large datasets. *Comput. Stat. Data Anal.*, doi:10.1016/j.csda.2008.09.008.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd edn. Chapman and Hall/CRC, Boca Raton.

Green, P.J. (2003). Trans-dimensional Markov chain Monte Carlo. In: *Highly Structured Stochastic Systems* (eds Green, P.J., Hjort, N.L. & Richardson, S.), 1–20. Oxford University Press, Oxford.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, 8, 993–1009.

Guisan, A. & Zimmerman, N.E. (2000). Predictive habitat distribution models in ecology. *Ecol. Modell.*, 135, 147–186.

Hawkins, B.A., Porter, E.E. & Diniz, J.A.F. (2003). Productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds. *Ecology*, 84, 1608–1623.

Helmuth, B., Briotman, B.R., Blanchette, C.A., Gilman, S., Halpin, P., Harley, C.D.G. *et al.* (2006). Mosaic patterns of thermal stress in the rocky intertidal zone: implications for climate change. *Ecol. Monogr.*, 76, 461–479.

Herron, P.M., Martine, C.T., Latimer, A.M. & Leicht-Young, S.A. (2007). Invasive plants and their ecological strategies: prediction and explanation of woody plant invasion in New England. *Divers. Distrib.*, 13, 633–644.

Illian, J.B., Møller, J. & Waagepetersen, R.P. (2007) Spatial point process analysis for a plant community with high biodiversity. *Environ. Ecol. Stat.*, DOI: 10.1007/S10651-007-0070-8.

Kamman, E.E. & Wand, M.P. (2003). Geoadditive models. *Appl. Stat.*, 52, 1–18.

LaDeau, S.L., Kilpatrick, A.M. & Marra, P.P. (2007). West Nile virus emergence and large-scale declines of North American bird populations. *Nature*, 447, 710–713.

Lafleur, N., Rubega, M. & Elphick, C. (2007). Invasive fruits, novel foods, and choice: an investigation of frugivory in European starlings and American robins. *Wilson J. Ornithol.*, 119, 429–438.

Latimer, A.M., Wu, S., Gelfand, A.E. & Silander, J.A. (2006). Building statistical models to analyze species distributions. *Ecol. Appl.*, 16, 33–50.

Leicht-Young, S.A., Silander, J.A. & Latimer, A.M. (2007). Comparative performance of invasive and native *Celastrus* species across environmental gradients. *Oecologia*, 154, 273–282.

Loarie, S.R., Carter, B.E., Hayhoe, K., McMahon, S., Moe, R., Knight, C.A. *et al.* (2008). Climate change and the future of California's endemic flora. *PLoS ONE*, 3, e2502.

Mehrhoff, L.J., Silander, J.A., Leicht, S.A., Mosher, E.S. & Tabak, N.M. (2003). IPANE: Invasive Plant Atlas of New England. Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT, USA. http://www.ipane.org. Last accessed 24 November 2008.

Møller, J. & Waagepetersen, R.P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton.

Palmer, M.W. (1996). Variation in species richness: towards a unification of hypotheses. *Folia Geobot et Phytotaxon (Praha)*, 29, 511–530.

Parmesan, C., Gaines, S., Gonzales, L., Kaufman, D.M., Kingsolver, J., Peterson, A.T. *et al.* (2005). Empirical perspectives on species borders: from traditional biogeography to global change. *Oikos*, 108, 58–75.

Plotkin, J.B., Chave, J.M. & Ashton, P.S. (2002). Cluster analysis of spatial patterns in Malaysian tree species. *Am. Nat.*, 160, 629–644.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. Last accessed 24 Nov. 2008.

Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

Robert, C.P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York.

Royle, J.A. & Nychka, D. (1999). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Comp. & Geosci.*, 24, 479–488.

Silander, J.A. Jr & Klepeis, D.M. (1999). The invasion ecology of Japanese barberry (*Berberis thunbergii*) in the New England landscape. *Biol. Invasions*, 1, 189–201.

Thomas, A., Best, N., Lunn, D., Arnold, R. & Spiegelhalter, D. (2004) *GeoBUGS User Manual*. Available at: http://www.mrc-bsu.cam.ac.uk/bugs, last accessed 24 November 2008.

Van Teeffelen, A.J.A. & Ovaskainen, O. (2007). Can the cause of aggregation be inferred from species distributions? *Oikos*, 116, 4–16.

Ver Hoef, J.M., Cressie, N., Fisher, R.N. & Case, T.J. (2001). Uncertainty and spatial linear models for ecological data. In: *Spatial Uncertainty in Ecology* (eds. Hunsaker, C.T., Goodchild, M.F., Friedl, M.A. & Case, T.J.), 214–237. Springer Verlag, New York.

Vounatsou, P., Smith, T. & Gelfand, A.E. (2000). Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics*, 1, 177–189.

Wackernagel, H. (2006). *Multivariate Geostatistics*, 2nd edn. Springer-Verlag, Heidelberg.

Whittaker, R.J., Willis, K.J. & Field, R. (2001). Scale and species richness: towards a general, hierachical theory of species diversity. *J. Biogeogr.*, 28, 453–470.

Wikle, C.K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84, 1382–1394.

Yeang, C.-H. & Haussler, D.. (2007) Detecting the coevolution in and among protein domains. *PLOS Comput. Biol.*, 3, e211.

Editor, Jerome Chave
Manuscript received 12 July 2008
First decision made 15 August 2008
Second decision made 20 October 2008
Manuscript accepted 4 November 2008