Hierarchical Bayesian models in ecology: reconstructing species interaction networks from non-homogeneous species abundance data

Andrej Aderhold[1], Dirk Husmeier[2], Jack J. Lennon[3], Colin M. Beale[4], V. Anne Smith[1]

[1] *School of Biology, University of St Andrews, St Andrews, UK*
[2] *School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QW, UK*
[3] *The James Hutton Institute Craigiebuckler, Aberdeen AB15 8QH, UK*
[4] *Department of Biology (Area 18), P.O. Box 373, University of York, York YO10 5YW UK*

Abstract

The relationships among organisms and their surroundings can be of immense complexity. To describe and understand an ecosystem as a tangled bank, multiple ways of interaction and their effects have to be considered, such as predation, competition, mutualism and facilitation. Understanding the resulting interaction networks is a challenge in changing environments, e.g. to predict knock-on effects of invasive species and to understand how climate change impacts biodiversity. The elucidation of complex ecological systems with their interactions will benefit enormously from the development of new machine learning tools that aim to infer the structure of interaction networks from field data. In the present study, we propose a novel Bayesian regression and multiple changepoint model (BRAM) for reconstructing species interaction networks from observed species distributions. The model has been devised to allow robust inference in the presence of spatial autocorrelation and distributional heterogeneity. We have evaluated the model on simulated data that combines a trophic niche model with a stochastic population model on a 2-dimensional lattice, and we have compared the performance of our model with L1-penalized sparse regression (LASSO) and non-linear Bayesian networks with the BDe scoring scheme. In addition, we have applied our method to plant ground coverage data from the western shore of the Outer Hebrides with the objective to infer the ecological interactions.

Keywords: Species interactions, Bayesian hierarchical model, multiple changepoint process, reversible jump Markov chain Monte Carlo, niche model,

1. Introduction

Understanding the response of ecosystems to perturbation is of paramount importance in a world with diminishing arable and natural land, where global climate change, invasive species, and changing agricultural practices impact world food supplies and biodiversity (Foley et al., 2005). But such understanding is not simple: ecosystems are a complex network of interactions. Modifying populations of one species can produce unexpected effects in others (Henneman and Memmott, 2001); entire ecosystems can respond to changing pressures by shifting to alternative states (Beisner et al., 2003). In order to understand and predict such phenomena, it is necessary to unravel the ecological networks underlying ecosystem stability and fragility (O'Gorman and Emmerson, 2009; Dunne et al., 2002).

Revealing such networks, however, might seem prohibitively difficult when even tracing interactions in simple food webs requires extraordinarily detailed fieldwork (e.g. Memmott et al. (2000)). Direct observation of trophic interactions ignores other relationships, such as inter-specific competition and mutualism, when such interactions may play significant roles in network dynamics (Werner and Peacor, 2003; Cheney and Côté, 2005; Valiente-Banuet and Verdú, 2008; Maestre et al., 2005). Recognising this, ecologists have attempted to measure existence of such indirect interactions (e.g. van Veen et al. (2009); Schmitz et al. (2004)), but quantifying all the effects and identifying all the unexpected interactions within complex real ecosystems may be beyond the scope of traditional fieldwork.

Computational inference of ecological networks presents an alternate route to unraveling ecosystem interactions. Traces of the interactions among species, both trophic and other types, should be present in the resulting distribution of individuals in space. Such species counts are available for a range of ecosystems (e.g. Hagemeijer and Blair (1997)). Computational network inference from such observational datasets has recently been developed in molecular systems biology, e.g. discovering transcriptional regulatory networks from datasets on gene expression (Friedman et al., 2000) and neural information flow from brain activity (Smith et al., 2006). These methods present an avenue for revealing ecological interactions from, rather than observation of interaction, more easily obtainable data on species incidence (Milns et al., 2010; van Oijen et al., 2010; Amstrup et al., 2008; Faisal et al., 2010). Also, by inferring interactions based upon their influence on

species distribution, there is no *a priori* restriction to specific relationship types, allowing competition and other relationships to be revealed alongside trophic interactions.

The objective of the present paper is to adapt a method recently proposed in computational systems biology (Lèbre et al., 2010) for inferring gene interactions from time series of gene expression profiles to the task of inferring species interaction networks from spatial species abundance data, as typically obtained from ecological surveys of fieldwork. The model by Lèbre et al. (2010) is a non-homogeneous dynamic Bayesian network, which combines the Bayesian hierarchical regression model of Andrieu and Doucet (1999) with a multiple changepoint process, as proposed by Punskaya et al. (2002), and pursues Bayesian inference with reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). We adapt this model to the inference of ecological networks in three ways. First, we allow for the fact that we have spatial rather than temporal data. Second, we expand the 1-dimensional changepoint process to two dimensions, by introducing two *a priori* independent changepoint processes in perpendicular directions. Third, we correct for spatial auto-correlation by introducing a parent node (in Bayesian network terminology) explicitly representing the spatial neighborhood of a node. To evaluate the performance of the model, we generate data from an ecological simulation study, which combines a trophic niche model of Lotka-Volterra type predator-prey interactions with a stochastic population model on a 2-dimensional lattice. We have compared the performance of our model with L1-penalized sparse regression (LASSO) and non-linear Bayesian networks (BDe score).

2. Model

Our model is a network in which nodes represent species, and edges (i.e. connections between nodes) represent potential species interactions. We aim to reconstruct the network from spatial species abundance profiles based on the rationale that if species interact, a variation in the abundance of one species should lead to a variation in the abundance of the interacting species. We model this mathematically with an approach based on Bayesian regression, which intrinsically incorporates a regularization effect that discourages the prediction of spurious interactions. We further improve this by explicitly correcting for spatial autocorrelation of the abundance profiles as well as by allowing for unobserved latent variables via a spatial changepoint process. Inference is carried out by sampling the interaction network structure as well as the number and location of spatial changepoints

from the posterior distribution, which is effected with state-of-the-art Monte Carlo algorithms (RJMCMC: reversible jump Markov chain Monte Carlo).

2.1. Interaction Network

The interaction network is represented by a directed graph $\mathcal{G} = \{\pi_1, \ldots, \pi_N\}$ with N species as nodes $n \in \{1, \ldots, N\}$, where π_n denotes the so-called parents of node n, that is the set of nodes with a directed edge pointing to n. $\mathcal{G}_n$ is the subnetwork associated with target species n, which is determined by its parent set π_n. A node cannot be contained in its own parent set, $n \notin \pi_n$, i.e. we rule out self-interactions related to e.g. cannibalism. The species are observed or surveyed at $T_1 \times T_2$ locations defined by their (orthogonal) coordinates (x_1, x_2), at which their abundance levels $y = \{y_n(x_1, x_2)\}_{1 \leq n \leq N, 1 \leq x_1 \leq T_1, 1 \leq x_2 \leq T_2}\}$ are determined.

2.2. Multiple changepoints

The regulatory relationships among the species may be influenced by latent variables, which are represented by spatial changepoints. We assume that latent effects in close spatial proximity are likely to be similar, but locations where spatially close areas are not similar are distinguished by changepoints. They are modelled with two *a priori* independent multiple changepoint processes along the two orthogonal spatial directions: $\xi_i = (\xi_i^1, \ldots, \xi_i^{k_i})$, $\xi_i^0 := 1, \xi_i^{k_i+1} := T_i$, and $i \in \{1, 2\}$. The vector ξ_i thus contains an (*a priori* unknown) number of k_i changepoints, and the changepoint vectors ξ_1 and ξ_2 partition the space into $Z = \prod_{i=1}^{2}(k_i + 1)$ non-overlapping segments, demarcated by the changepoints. We denote the latent variable associated with a segment by $h \in \{1, \ldots, Z\}$. If two locations (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$ are in the same segment, $\xi_1^a \leq x_1, \tilde{x}_1 < \xi_1^{a+1}, \xi_2^b \leq x_2, \tilde{x}_2 < \xi_2^{b+1}$, then they are assigned the same latent variable: $h(x_1, x_2) = h(\tilde{x}_1, \tilde{x}_2)$. We define an isomorphism between segments and changepoints such that segment h is demarcated by changepoints $\{\xi_1^{[f_1(h)-1]}, \xi_1^{f_1(h)}, \xi_2^{[f_2(h)-1]}, \xi_2^{f_2(h)}\}$.

2.3. Regression model

For all species n, the random variable $Y_n(x_1, x_2)$ refers to the abundance of species n at location (x_1, x_2). Within any segment h, this abundance depends on the abundance levels of the species in the parent set of species n, π_n, which we model with a segment specific linear regression model. Define the set of parameters $\{(a_{nm}^h)_{m \in 0..N}, \sigma_n^h\}$, $a_{nm}^h \in \mathbb{R}, \sigma_n^h > 0$. For all $m \neq 0$, $a_{nm}^h = 0$ if $m \notin \pi_n$. For all species n, for all locations (x_1, x_2) in segment h, $Y_n(x_1, x_2)$ depends on the N

variables $\{Y_m(x_1, x_2)\}_{1 \leq m \leq N, m \neq n}$ according to

$$Y_n(x_1, x_2) \;=\; a^h_{n0} + \sum\nolimits_{m \in \pi_n} a^h_{nm}\, Y_m(x_1, x_2) + \varepsilon_n(x_1, x_2) \tag{1}$$

where the latent variable h depends on the location (x_1, x_2) and the change-point vectors ξ_1 and ξ_2 defined in the previous subsection. The noise $\varepsilon_n(x_1, x_2)$ is assumed to be Gaussian with mean 0 and variance $(\sigma^h_n)^2$, $\varepsilon_n(x_1, x_2) \sim N(0, (\sigma^h_n)^2)$. We define $a^h_n = (a^h_{nm})_{n \in 0..N}$ to denote the vector of all regression parameters of species n. This includes the parameters defining the strength of interactions with other species m, a^h_{nm}, as well as a species-specific offset term, a^h_{n0}.

2.4. Spatial autocorrelation

Spatial autocorrelation, the phenomenon that observations at nearby locations are more similar than observations at more distant locations, is nearly ubiquitous in ecology and can have a strong impact on statistical inference (Lennon, 2000; Dale and Fortin, 2002). In our case, spatial autocorrelation could lead to the identification of spurious interactions as a mere consequence of two species co-occurring in similar geographical regions. To incorporate potential spatial autocorrelation into the model, we follow an approach proposed by Faisal et al. (2010) and illustrated in Figure 1b. The idea is to connect each node in the network to an enforced parent node that represents the average population at neighboring cells, weighted inversely proportional to the distance of the neighbors:

$$A_n(x_1, x_2) = \frac{\sum_{(\tilde{x}_1, \tilde{x}_2) \in \mathcal{N}(x_1, x_2)} d^{-1}[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)] Y_n(\tilde{x}_1, \tilde{x}_2)}{\sum_{(\tilde{x}_1, \tilde{x}_2) \in \mathcal{N}(x_1, x_2)} d^{-1}[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)]} \tag{2}$$

where $\mathcal{N}(x_1, x_2)$ is the spatial neighborhood of location (x_1, x_2) (e.g. the four nearest neighbors), and $d[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)]$ is the Euclidean distance between (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$. The value of $A_n(x_1, x_2)$, weighted by an additional weight a^h_{nA}, will be included in (1):

$$\begin{aligned} Y_n(x_1, x_2) \;=\; & a^h_{n0} + \sum\nolimits_{m \in \pi_n} a^h_{nm} Y_m(x_1, x_2) \\ & + a^h_{nA} A_n(x_1, x_2) \;+\; \varepsilon_n(x_1, x_2) \end{aligned} \tag{3}$$

In this way the abundance of species n at location (x_1, x_2) is, in the first instance, determined by the spatial neighborhood. Only if the explanatory power of the latter is not sufficient will there be an incentive for the inference scheme to include further edges related to species interactions.

2.5. Prior

To encourage sparse network structures, we impose a truncated Poisson prior with mean Λ and maximum $\overline{m} = 5$ on the number m_n of parents for node n: $P(m_n|\Lambda) \propto \frac{\Lambda^{m_n}}{m_n!}\mathbb{1}_{\{m_n \leq \overline{m}\}}$. There was no noticeable difference in performance compared to higher settings of $\overline{m}$. Conditional on m_n, the prior for the parent set π_n is a uniform distribution over all parent sets with cardinality m_n: $P(\pi_n \,||\pi_n| = m_n) = 1/\binom{N-1}{m_n}$. The overall prior on the network structure $\mathcal{G}$ is given by factorization and marginalization:

$$\begin{aligned} P(\mathcal{G}|\Lambda) &= \prod\nolimits_{n=1}^{N} P(\pi_n|\Lambda); \\ P(\pi_n|\Lambda) &= \sum\nolimits_{m_n=1}^{\overline{m}} P(\pi_n|m_n)P(m_n|\Lambda) \end{aligned} \tag{4}$$

For both spatial directions $i \in \{1, 2\}$, the $k_i + 1$ segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k}_i = T_i - 1$: $P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i \leq \overline{k}_i\}}$. Conditional on k_i changepoints, the changepoint position vector $\xi_i = (\xi_i^1, ..., \xi_i^{k_i})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. There are $(T_i - 1)$ possible positions for the k_i changepoints, thus vector ξ_i has prior density $P(\xi_i|k_i) = 1/\binom{T_i-1}{k_i}$. Conditional on the parent set π_n of size m_n, the $m_n + 2$ regression coefficients, denoted by $a_n^h = (a_{n0}^h, a_{nA}^h, (a_{nm}^h)_{m\in\pi_n})$, are assumed zero-mean multivariate Gaussian distributed with covariance matrix $(\sigma_n^h)^2\Sigma_n$,

$$P(a_n^h|\pi_n, \sigma_n^h) \;=\; |2\pi(\sigma_n^h)^2\Sigma_{n,h}|^{-\frac{1}{2}}\exp\left(-\frac{[a_n^h]^\dagger\Sigma_{n,h}^{-1}a_n^h}{2(\sigma_n^h)^2}\right) \tag{5}$$

where the symbol † denotes matrix transposition, $\Sigma_{n,h} = \delta^{-2}D_{n,h}^\dagger(y)D_{n,h}(y)$ and $D_{n,h}(y)$ is the $s_{n,h} = \prod_{i=1}^{2}(\xi_i^{f_i(h)} - \xi_i^{f_i(h)-1}) \times (m_n + 2)$ matrix whose first column is a vector of 1s, for the constant in (1), the second column is a vector of autocorrelation variables, defined in (2), and the remaining columns contain the observed abundance values $y_n(x_1, x_2)$ for all species $n \in \pi_n$ and all locations (x_1, x_2) in segment h: $\xi_i^{f_i(h)-1} \leq x_i < \xi_i^{f_i(h)}, i \in \{1, 2\}$. This so-called g-prior is widely used in Bayesian statistics; see e.g. Andrieu and Doucet (1999). Finally, the conjugate prior for the variance $(\sigma_n^h)^2$ is the inverse gamma distribution, $P((\sigma_n^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre et al. (2010), we set the hyper-hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that

give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. Following Lèbre et al. (2010), these hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family: $P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1)$ and $P(\delta^2) = \mathcal{IG}(2, 0.2)$.

2.6. Posterior

Equation (3) implies that the Likelihood is

$$P(y_n^h|\xi_1^{f_1(h)-1}, \xi_1^{f_1(h)}, \xi_2^{f_2(h)-1}, \xi_2^{f_2(h)}, \mathcal{G}, a_n^h, \sigma_n^h) = \left(\sqrt{2\pi}\sigma_n^h\right)^{-s_{n,h}} \exp\left(-\frac{(y_n^h - D_{n,h}(y)a_n^h)^\dagger\,(y_n^h - D_{n,h}(y)a_n^h)}{2(\sigma_n^h)^2}\right)$$

From Bayes theorem, the posterior distribution is given by the following equation, where all prior distributions have been defined above:

$$\begin{aligned} P(k_1, k_2, \xi_1, \xi_2, \mathcal{G}, a, \sigma^2, \lambda, \Lambda, \delta^2|y) \propto\ & P(\delta^2)P(\lambda)P(\Lambda)P(\mathcal{G}|\Lambda) \\ & \prod_{i=1}^{2} P(k_i|\lambda)P(\xi_i|k_i) \prod_{h=1}^{Z}\prod_{n=1}^{N} P([\sigma_n^h]^2)P(a_n^h|\pi_n, [\sigma_n^h]^2, \delta^2) \\ & P(y_n^h|\xi_1^{f_1(h)-1}, \xi_1^{f_1(h)}, \xi_2^{f_2(h)-1}, \xi_2^{f_2(h)}, \mathcal{G}, a_n^h, \sigma_n^h) \end{aligned} \tag{6}$$

2.7. Inference

An attractive feature of the chosen model is that the marginalization over the parameters $a = \{a_n^h, 1 \le n \le N, 1 \le h \le Z\}$ and $\sigma^2 = \{(\sigma_n^h)^2, 1 \le n \le N, 1 \le h \le Z\}$ in the posterior distribution of (6) is analytically tractable (Lèbre et al., 2010; Andrieu and Doucet, 1999):

$$\begin{aligned} & P(k_1, k_2, \xi_1, \xi_2, \mathcal{G}, \lambda, \Lambda, \delta^2|y) \ = \\ & \int P(k_1, k_2, \xi_1, \xi_2, \mathcal{G}, a, \sigma^2, \lambda, \Lambda, \delta^2|y)\, da\, d\sigma^2 \end{aligned} \tag{7}$$

The number of changepoints and their location, k_1, k_2, ξ_1, ξ_2, the network structure $\mathcal{G}$ and the hyperparameters $\lambda, \Lambda, \delta^2$ can be sampled from the posterior distribution $P(k_1, k_2, \xi_1, \xi_2, \mathcal{G}, \lambda, \Lambda, \delta^2|y)$ with RJMCMC (Green, 1995), following the

scheme described in Lèbre et al. (2010); Andrieu and Doucet (1999) and (Punskaya et al., 2002). By marginalization and under the assumption of convergence, this gives us a sample of networks from the posterior distribution $P(\mathcal{G}|y)$. By further marginalization, we get the posterior probabilities of all species interactions $P(n \rightarrow \tilde{n}|y)$, which defines a ranking of the interactions in terms of posterior confidence. If the true network structure is known, this ranking allows the computation of the areas under the ROC (AUROC) and precision-recall (AUPRC) curves (Davis and Goadrich, 2006), which are two measures widely used in the systems biology literature to quantify the overall network reconstruction accuracy (Prill et al., 2010), with larger values indicating a better prediction performance overall.

3. Data

3.1. Synthetic data

For an objective measure of network recovery, we tested the model's ability to recover the true network structure from test data generated from a piecewise linear regression model following equation (1). The data was partitioned by 2-dimensional fixed changepoints and the number of grid cells was selected to be 15 in each direction. The changepoints were inserted globally at location 5 and 10 along each dimension. The number of nodes n was set to 10 and the number of parents for each node was sampled from a Poisson distribution. The regression coefficients a_n^h together with the bias a_0^h of each segment h were sampled from a uniform distribution in the interval of $[-1; -0.5]$ and $[0.5, 1.0]$. The noise ε_n was sampled from a normal distribution. Nodes without incoming edge were initialized to a Gaussian random number. The values of the remaining nodes were calculated at each grid cell following equation (1).

3.2. Ecological simulation of trophic interactions

For a more realistic evaluation, we followed Faisal et al. (2010) and generated data from an ecological simulation that combines a niche model (Williams and Martinez, 2000) with a stochastic population model (Lande et al., 2003) in a 2-dimensional lattice.

Niche model. The niche model defines the structure of the trophic network and has two parameters: the number of species N and the connectance (or network density) defined as L/N^2 where L is the number of interactions (edges) in the network. Each species n is assigned a niche value x_n, drawn uniformly from $[0, 1]$. This gives an ordering of the species, where higher values mean that species are higher up in the food chain. For each species a niche range R_n is drawn from

a beta distribution with expected value $2C$ (where C is the desired connectance), and species n consumes all species falling in a range R_n that is placed by uniformly drawing the centre of the range from $[R_n/2, x_n]$. An illustration is given in Figure 1 by Williams and Martinez (2000). Despite its simplicity, it was shown by the same authors that the resulting networks share many characteristics with real food webs.

Stochastic population dynamics. The population model is defined by a stochastic differential equation where the dynamics of the log abundance $X_n(t)$ of species n at time t can be expressed as:

$$\frac{dX_n(t)}{dt} = r_n + \frac{\sigma_d}{\sqrt{e^{X_n(t)}}}\frac{dA_n(t)}{dt} + \sigma_e\frac{dB_n(t)}{dt} - \gamma X_n(t) - \Omega(X) + \sigma_E\frac{dE(t)}{dt} \quad (8)$$

where X is the set of all $X_N(t)$, r_n is the growth rate of species n, σ_d is the standard deviation of the demographic effect, $A_n(t)$ is the species-specific demographic effect, σ_e is the standard deviation of the species-specific environmental effect, $B_n(t)$ is the species-specific environmental effect, γ is the intra-specific density dependence, Ω is the effect of competition for common resources, σ_E is the standard deviation of the general environmental effect and $E(t)$ is the general community environment. The growth rates r_n are location dependent (depending on the cell of a rectangular grid), with a spatial pattern that is generated by noise with spectral density f^β (with $\beta < 0$, and f denoting the spatial frequency at which the noise is measured). An illustration is given in Figure 2. To model species dispersal, we included an exponential dispersal model, where the probability of a species moving from one location to another is determined by the Euclidean distance between the locations.

Interactions. To incorporate the niche model, we modified the term Ω in (8) to include predator-prey interactions in the Lotka-Volterra form. We explored two versions: one where predatory interactions had a relatively strong negative effect on prey (strong predation) and one where the impact of predation was less severe (weak predation). Strong predation is more akin to traditional predator-eat-prey interactions, whereas weak predation is more akin to partially destructive predation (e.g., grazing) or aggression.

Simulation. We applied this model to 10 species living in a 25-by-25 rectangular grid. We simulated the dynamics of this model for 3000 steps and then recorded species abundance levels in all grid cells at the final step; this corresponds to an ecological survey carried out at a fixed moment in time. For each

grid cell we counted the number of species that went extinct. These counts were added up over all cells, yielding a total number of extinctions. A simulation was rejected if these extinctions exceeded the value 50. This threshold was introduced in order to compensate for the unrealistic artifact that is produced by prey being not able to escape from predators beyond grid borders. For each of the spatial β parameters displayed in Figure 4, 30 surveys were collected by running the simulation repeately with different networks and parameter initializations.

3.3. Real world plant data

We have applied the method to real-world data from Lennon et al. (2011), including 106 vascular plants and 12 physical variables collected from a 200m x 2162m land strip at the western shore of the Outer Hebrides representing a Machair vegetation. Samples were taken at 217 locations, each 1m x 1m in size, equally distributed with a 50m spacing. Plant samples were measured as ground coverage in percentage and physical samples as absolute values (such as moisture, pH value, organic matter and slope). The data was log-normal transformed after observing substantial skeweness in the distributions. Each sample point was mapped into a 2D grid (locations lacking data due to geographic limitations (lochs and bare rocks) were left empty). The spatial autocorrelation value for each plant and location was calculated from neighbors inside a radius of 70m. Since we are interested in plant interactions not mediated by different preferences for soil characteristics, we defined that each plant has all 12 physical soil variables as fixed input, i.e., permanent predictor variables. We apply our 2D change-points model along the longitudinal and latitudinal directions.

4. Comparative Evaluation

To evaluate the network reconstruction accuracy for the simulated data, where the true network structure is known, we proceed as follows. Networks $\mathcal{G}$ are sampled from the posterior distribution $P(\mathcal{G}|y)$, and we compute $P(e_{ik}|y)$, the posterior probability of an edge e_{ik} between nodes i and k, which is given by the proportion of networks in the MCMC sample that contain this edge. Let $\mathcal{E}(\theta) = \left\{e_{ik}|P(e_{ik}|y) > \theta\right\}$ denote the set of all edges whose posterior probability exceeds a given threshold $\theta \in [0, 1]$, from which we determine the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) edges. We then compute the sensitivity = $TP/(TP + FN)$, the specificity = $TN/(TN + FP)$, and the complementary specificity = 1-specificity = $FP/(TN + FP)$. Rather than selecting an arbitrary value for the threshold θ,

we repeat this scoring procedure for all possible values of $\theta \in [0, 1]$, and plot the resulting sensitivity scores against the corresponding complementary specificity scores. This gives the so-called *receiver operating characteristics* (ROC) curve shown in Figure 7. The diagonal line indicates the ROC curve under random expectation. The line marked with "perfect predictor" indicates a perfect retrieval of all true edges without a single spurious edge. In general, ROC curves are between these two extremes, with a larger *area under the ROC curve* (AUROC) indicating a better performance. In particular, random expectation corresponds to a value of AUROC=0.5, and a perfect predictor has an AUROC score of 1.0. An alternative approach, which is preferred in many practical applications, is to plot the precision against the recall, where recall is just another name for sensitivity, and precision is defined as the proportion of recovered interactions that are correct, precision = $TP/(TP + FP)$. The area under the precision-recall curve, AUPRC, is again a measure of the global network reconstruction accuracy, with a larger value indicating a better performance. Both measures are widely applied in systems biology (Prill et al., 2010). They have certain pros and cons, as e.g. discussed in Davis and Goadrich (2006), and we therefore use them jointly in our evaluation.

We compared the performance of BRAM, which corresponds to the model in Figure 2c, with two alternative Bayesian regression models: Bayesian regression without changepoints (BR, Figure 2b) and Bayesian regression without changepoints and without allowing for spatial autocorrelation (BR-0, Figure 2a). We included a comparison with L1-regularized linear regression (LASSO: Tibshirani (1996, 2011)), using the optimization algorithm proposed by Grandvalet (1998). This method is widely applied in molecular systems biology (van Someren et al., 2006), has been recommended to be used more widely in ecology (Dahlgren, 2010), and was found to outperform all competing methods by Faisal et al. (2010). The regularization parameter λ that controls the network sparsity was inferred with 10-fold cross-validation, which led to better results than optimizing the BIC score. The method produces edge weights indicating the strength and sign of interactions among species. For obtaining the ROC and precision-recall curves, we ranked the potential interactions based on the absolute values of the non-zero interaction parameters. We further included a comparison with a non-linear Bayesian network, as implemented in the software package BANJO. We discretized the data with Hartemink's pairwise mutual information method described by Hartemink (2001) (implemented in *R* package *bnlearn*)[1]. Search was done using simulated

[1]This method yielded a better performance than quantile discretization. The number of dis-

annealing with random walk proposals. Simulated annealing was run on each dataset until convergence (typically 7 hours of CPU time). Using the top 100 high-scoring (BDe score) networks we computed edge probabilities for ranking. Application of both LASSO and BANJO included taking spatial autocorrelation into account. Finally, we applied BRAM to real world data, revealing putative plant interactions.

5. Results and Discussion

In the following, we show how BRAM outperforms the other tested methods on synthetic data and on trophic simulations having spatial heterogeneity. On simulations lacking clear spatial heterogeneity, where there is intrinsically no room for improvement with a changepoint model, BRAM performs similarly to LASSO. Finally, we explore how BRAM can be applied to real data for analysing ecological systems.

On the synthetic data of Section 3.1, BRAM outperforms all competing schemes (Figure 3). This is not surprising, in that the data have been generated from a process that is consistent with the modeling assumptions of BRAM. However, it is reassuring both that the MCMC inference scheme can successfully deal with the increased model complexity, and that it leads to an improvement over the competing models in terms of actual network reconstruction accuracy. For the data simulated from the niche model, described in Section 3.2, we found that BRAM consistently outperforms BR-0 and BANJO (Figures 4-5). The improvement over BR-0 confirms the importance of allowing for spatial autocorrelation in ecological modeling. The improvement over BANJO underlines the detrimental effect of the information loss inherent in data discretization. The comparison with BR and LASSO leads to results that, on the face of it, appear less conclusive. On the weak predation data BRAM tends to outperform both BR and LASSO (Figure 5), while the latter methods are on a par with BRAM on the strong predation data (Figure 4). The difference between the two datasets rest in the parameter choice for the trophic interaction model described in Section 3.2. For weak predation, the abundance profiles showed much stronger spatial oscillations than for strong predation, or conversely: for strong predation, these abundance profiles were much flatter than for weak predation. This suggests that weak predation leads to much stronger spatial heterogeneity than strong predation. LASSO showed, on average,

cretization levels was chosen to be 3 based on empirical tests carried out by Yu et al. (2004).

the same performance as our simplified model without changepoints. If there is no spatial heterogeneity, then there is not much benefit in using a changepoint model. Hence, for strong predation with little spatial heterogeneity, our proposed model with changepoints does not outperform our simpler model without changepoints, and consequently it also does not outperform LASSO.

This raises the question of why strong predation leads to less spatial heterogeneity in the first place. Spatial heterogeneity implies that in some regions prey are more affected by predators than in others. For strong predation these fluctuations are stronger than for weak predation, in fact so strong that some prey are driven to extinction. However, the way we set up the simulations is such that populations with an extinction rate above a threshold are rejected. This is motivated by the limited size of the spatial area in our simulated ecological landscape. This limited size 'traps' prey in an unnatural way; high extinction rates are rejected as being ecologically unrealistic. Populations with the highest spatial heterogeneity are the ones most affected by extinction, thus our rejection mechanism favours more homogeneous populations when predation is strong, which we confirmed empirically by inspection of the spatial abundance profiles.

Our simulation studies thus suggest that in the absence of spatial heterogeneity, when there is no room for improvement, BRAM shows the same performance as LASSO (Figure 5). This is reassuring, given that LASSO was found to outperform all competing models by Faisal et al. (2010). When there is genuine spatial heterogeneity, BRAM outperforms LASSO and all homogeneous models without changepoints (Figure 4).

We have applied BRAM to the plant abundance data from the ecological survey described in Section 3.3. We sampled interaction network structures from the posterior distribution with MCMC and computed the marginal posterior probabilities of the individual potential species interactions, as described in Section 2.7. We kept all species interactions with a marginal posterior probability above 0.2, resulting in 39 out of 106 species with relevant interactions in the reconstructed network shown in Figure 6. The right panel in this figure shows the recovered network for a higher threshold of 0.5. Negative interactions were displayed as dashed lines and positive interactions as full lines. They were derived as mean edge weights over all segments and multiple samples from the MCMC chain.

Since we had defined the 12 soil attributes as fixed predictors to each plant, the interactions in this network represent plant-plant interactions not mediated by similar soil preferences. This network can lead to the formation of new ecological hypotheses. For instance, *Ranunculus bulbosus* (species 14) is densely connected with five interspecific links above the threshold. Can that be related to its tolerance

for nutrient-poor soil and its preferred occurrence in species-rich patches? There is a noticeable imbalance between positive and negative interactions. The dominance of positive interactions in the Machair vegetation is surprising given that much research in ecology has emphasised the role of competition within communities, though this is now changing as the potentially important role of facilitation is recognised (e.g. Bruno et al. (2003)). It is worth remembering however that the interactions observed in these data occur between species at the same trophic level and as such are but one horizontal slice of a much more complex hierarchical food web involving plant pathogens, insect and mammalian herbivores and their predators. Nonetheless, the relative lack of negative interactions is intriguing in that it suggests that interspecific competition does not dominate this grassland system.

Figure 8 shows, for a selected plant species, the marginal posterior probability of a changepoint along the longitudinal direction as well as the posterior coocurrence matrix, as introduced by Grzegorczyk and Husmeier (2011). We clustered plant species on the basis of these coocurrence matrices, using a simple clustering algorithm (K-means with restarts) combined with the gap statistic for deciding on the number of clusters (Tibshirani et al., 2001; Hastie et al., 2001). The results are shown in Figure 9. Ecologists could make use of clusters like these to, e.g., identify species which share similar ecological sensitivities. These results demonstrate that the proposed method provides a useful tool for explorative data analysis in ecology with respect to both species interactions and spatial heterogeneity.

6. Conclusions

We have addressed the problem of reconstructing species interaction networks from species abundance data. To this end, we have proposed a Bayesian model combining Bayesian piecewise linear regression with multiple changepoint processes. The work is motivated by a model recently proposed in the molecular systems biology literature (Lèbre et al., 2010), but has been adapted from the temporal domain (gene expression time series) to the spatial one (snapshot of species distributions in space, typical of ecological surveys). We have introduced and tested two essential modifications, illustrated and motivated in Figure 1. First, we extended the 1-dimensional changepoint process from Lèbre et al. (2010) by a 2-dimensional one, which corresponds to a richer latent variable structure that allows modeling unobserved effects with smooth geographical variation. Second, we explicitly introduced an additional enforced parent node for each species, which represents the average species abundance from the spatial neighborhood of the current location and thereby allows a correction for spatial autocorrelation.

We tested our model on data from a trophic simulation, which combines spatial species dispersal with demographic and environmental effects and predator-prey interactions of the Lotka-Volterra form defined by a trophic network obtained from a niche model. Our results show that the proposed model consistently outperforms a Bayesian regression model that does not allow for spatial autocorrelation, as well as a non-linear Bayesian network with the BDe score. Comparison with L1-regularized sparse regression (LASSO) and Bayesian regression without changepoints reveals the following. In the absence of pronounced spatial heterogeneity (strong predation), when there is no room for improvement over the homogeneous models, the performance of BRAM is on a par with LASSO and Bayesian regression (Figure 4). In the presence of spatial heterogeneity (weak predation), BRAM clear outperforms all competing models (Figure 5).

An application to plant species abundance data from a recent ecological survey has demonstrated how the proposed method can be used as a tool for hypothesis generation with respect to species interactions and spatial distribution patterns. The main problem with real data analysis is the 'objective' evaluation. In ecology, we currently lack any gold standard, and the situation is more difficult than in molecular systems biology, where several databases about molecular functions and interactions exist. A more thorough evaluation of our model on real data, which is the objective of ongoing work, needs to be done in close collaboration with ecologists and will ultimately be based on somewhat circumstantial evidence. For the purpose of method assessment we will therefore pursue, in parallel, more extensive studies based on simulated data, with the objective to make the underlying models increasingly ecologically realistic.

Future Work. There are two lines along which the current work can be extended. First, the present changepoint model is overly restrictive in the sort of partitions that it produces. For situations in which the properties of the ecosystem change rapidly in some areas, but slowly in others, the model will require a fine partition everywhere as the edges of small squares in rapidly changing areas will extend and bisect the large rectangles in slowly changing areas. This will yield small squares everywhere and as a result more parameters are required leading to less efficient inference. Furthermore, even if the rate of change of parameters is uniform, if the geographic extent of the ecosystem is large, then rectangles will be unnecessarily bisected by edges extending from distant parts of the geography. Instead of a changepoint model in which the x and y axis partitions are independent, an interesting research project would be to use a Mondrian process, as proposed by Teh and Roy (2009). This would allow the level of fineness of the partition to vary, so

that details about the partition in one area do not unnecessarily extend to others. Alternatively, a Pitman-Yor processes (Sudderth and Jordan, 2009) (i.e. a distant dependent Dirichlet process), in analogy with image segmentation, could be attempted. Or, as the locations of the points from which samples are collected are discrete, a Dirichlet process mixture of Gaussians could be tried; this latter option would have the advantage of not increasing the complexity of the implementation.

The second potential improvement concerns the parameter prior. For the current prior on the regression model (3) the coefficients are assumed to be distributed according to a zero-mean multivariate Gaussian with a covariance drawn from an inverse gamma distribution. This prior is symmetric around 0 and hence does not discourage sign changes. A justification can, in fact, be given based on various recent ecology publications, which discuss how the nature of interactions can change with varying environmental conditions (e.g. Callaway and Walker (1997); Valiente-Banuet and Verdú (2008); Maestre et al. (2009); Choler et al. (2001)). Mutualistic interactions may become neutral or antagonistic (i.e. involve a sign change), either temporarily or over parts of the range of the interacting species, and this is not ruled out by the prior we employ. However, the scenarios described above are, overall, quite rare, and they are in particular unlikely to apply to trophic interactions. In fact, if we know that, for two interacting species A and B, A eats B in rectangle 1, we would assume that it is more likely that A also eats B in rectangle 2 than the other way round. This prior notion can be incorporated into the model by putting a species dependent prior on the mean, and drawing the mean independently from this prior for each rectangle. The implementation of this idea effectively adds an extra layer to the Bayesian hierarchy, and has recently been investigated by (Grzegorczyk and Husmeier, 2012) in the context of molecular systems biology.

7. Acknowledgement

AA was supported by the Biotechnology and Biological Sciences Research Council studentship (BBSRC). Part of the work was carried out while DH was employed at Biomathematics and Statistics Scotland (BioSS), and supported by the Scottish Governments Rural and Environment Science and Analytical Services Division (RESAS).

References

Amstrup, S., Marcot, B., Douglas, D., 2008. A bayesian network modeling approach to forecasting the 21st century worldwide status of polar bears. Arctic sea ice decline: observations, projections, mechanisms, and implications. Geophysics monograph series 180, 213–268.

Andrieu, C., Doucet, A., 1999. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. IEEE Transactions on Signal Processing 47 (10), 2667–2676.

Beisner, B. E., Haydon, D. T., Cuddington, K., 2003. Alternative stable states in ecology. Front. Ecol. Environ. 1 (7), 376–382.

Bruno, J. F., Stachowicz, J. J., Bertness, M. D., 2003. Inclusion of facilitation into ecological theory. Evolution 18 (3), 119–125.

Callaway, R., Walker, L., 1997. Competition and facilitation: a synthetic approach to interactions in plant communities. Ecology 78 (7), 1958–1965.

Cheney, K., Côté, I., 2005. Mutualism or parasitism? the variable outcome of cleaning symbioses. Biology letters 1 (2), 162–165.

Choler, P., Michalet, R., Callaway, R., 2001. Facilitation and competition on gradients in alpine plant communities. Ecology 82 (12), 3295–3308.

Dahlgren, J. P., May 2010. Alternative regression methods are not considered in murtaugh (2009) or by ecologists in general. Ecology letters 13 (5), E7–9.

Dale, M. R. T., Fortin, M. J., 2002. Spatial autocorrelation and statistical tests in ecology. Ecoscience 9 (2), 162–167.

Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: Proc. of Int. Conf. on Machine Learning. ACM, pp. 233–240.

Dunne, J. A., Williams, R. J., Martinez, N. D., 2002. Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecology Letters 5, 558567.

Faisal, A., Dondelinger, F., Husmeier, D., Beale, C., 2010. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. Ecological Informatics 5 (6), 451–464.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., 2005. Global consequences of land use. Science 309, 570–574.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comp. Biol. 7, 601–620.

Grandvalet, Y., 1998. Least absolute shrinkage is equivalent to quadratic penalization. In: Niklasson, L., Bodén, M., Ziemske, T. (Eds.), ICANN 98. Vol. 1 of Perspectives in Neural Computing. Springer, pp. 201–206.

Green, P., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Grzegorczyk, M., Husmeier, D., 2011. Non-homogeneous dynamic Bayesian networks for continuous data. Machine Learning 83 (3), 355–419.

Grzegorczyk, M., Husmeier, D., 2012. Bayesian regularization of non-homogeneous dynamic bayesian networks by coupling interaction parameters. In: Fifth Int. Conf. on Artificial Intelligence and Statistics, to appear.

Hagemeijer, W. J. M., Blair, M. J., 1997. The EBCC atlas of European breeding birds: their distribution and abundance. Poyser London.

Hartemink, A. J., 2001. Principled computational methods for the validation and discovery of genetic regulatory networks. Ph.D. thesis, MIT.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer-Verlag.

Henneman, M. L., Memmott, J., 2001. Infiltration of a Hawaiian Community by Introduced Biological Control Agents. Science 293 (5533), 1314–1316.

Lande, R., Engen, S., Saether, B. E., 2003. Stoch. Pop. Dyn. in Ecol. and Cons. Oxford University Press.

Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P. H., Lelandais, G., 2010. Statistical inference of the time-varying structure of gene-regulation networks. BMC systems biology 4, 130.

Lennon, J., Beale, C., Reid, C., Kent, M., Pakeman, R., 2011. Are richness patterns of common and rare species equally well explained by environmental variables. Ecography 34, 529–539.

Lennon, J. J., 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101–113.

Maestre, F., Callaway, R., Valladares, F., Lortie, C., 2009. Refining the stress-gradient hypothesis for competition and facilitation in plant communities. Journal of Ecology 97 (2), 199–205.

Maestre, F., Valladares, F., Reynolds, J., 2005. Is the change of plant–plant interactions with abiotic stress predictable? a meta-analysis of field results in arid environments. Journal of Ecology 93 (4), 748–757.

Memmott, J., Fowler, S., Paynter, Q., Sheppard, A., Syrett, P., 2000. The invertebrate fauna on broom, *Cytisus scoparius*, in two native and two exotic habitats. Acta Oecol. 21 (3), 213–222.

Milns, I., Beale, C. M., Smith, V. A., 2010. Revealing ecological networks using Bayesian network inference algorithms. Ecology 91, 1892–1899.

O'Gorman, E. J., Emmerson, M. C., Aug. 2009. Perturbations to trophic interactions and the stability of complex food webs. Proceedings of the National Academy of Sciences of the United States of America 106 (32), 13393–8.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., Stolovitzky, G., 02 2010. Towards a rigorous assessment of systems biology models: The dream3 challenges. PLoS ONE 5 (2), e9202.

Punskaya, E., Andrieu, C., Doucet, A., Fitzgerald, W., 2002. Bayesian curve fitting using MCMC with applications to signal segmentation. Signal Processing, IEEE Transactions on 50 (3), 747–758.

Schmitz, O. J., Krivan, V., Ovadia, O., 2004. Trophic cascades: the primacy of trait-mediated indirect interactions. Ecol. Lett. 7 (2), 153–163.

Smith, V., Yu, J., Smulders, T., Hartemink, A., Jarvis, E., 2006. Computational inference of neural information flow networks. PLoS Comput. Biol. 2 (11), 1436–1449.

Sudderth, E., Jordan, M., 2009. Shared segmentation of natural scenes using dependent pitman-yor processes. Advances in Neural Information Processing Systems 21, 1585–1592.

Teh, Y., Roy, D., 2009. The mondrian process. Technology 21 (1979), 1–8.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58 (1), 267–288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective (with comments). Journal of the Royal Statistical Society, Series B 73 (3), 273–282.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2), 411–423.

Valiente-Banuet, A., Verdú, M., 2008. Temporal shifts from facilitation to competition occur between closely related taxa. Journal of Ecology 96 (3), 489–494.

van Oijen, M., Cameron, D., Reinds, G., Thomson, A., 2010. Bayesian methods for spatial upscaling of process-based forest ecosystem models. In: AGU Fall Meeting Abstracts. Vol. 1. p. 04.

van Someren et al., E. P., 2006. Least absolute regression network analysis of the murine osterblast differentiation network. Bioinformatics 22 (4), 477–484.

van Veen, F. J., Brandon, C. E., Godfray, H. C., 2009. A positive trait-mediated indirect effect involving the natural enemies of competing herbivores. Oecologia 160 (1), 195–205.

Werner, E. E., Peacor, S. D., 2003. A review of trait-mediated indirect interactions in ecological communities. Ecology 84 (5), 1083–1100.

Williams, R. J., Martinez, N. D., 2000. Simple rules yield complex food webs. Nature 404 (6774), 180–183.

Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E., Dec 2004. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20 (18), 3594–603.

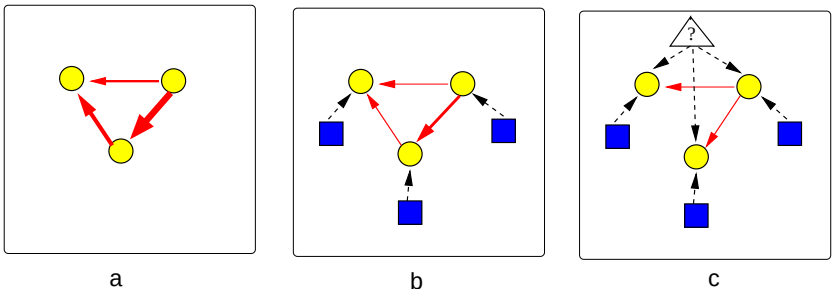

Figure 1: **Illustration of the improved method for ecological network reconstruction.** *Panel (a)* illustrates the naive approach to modeling species interaction networks. Circles represent species (nodes), and arrows present species interactions (edges). Networks inferred from species abundance or population density data alone tend to contain many spurious interactions. *Panel (b): Allowing for spatial autocorrelation.* Each node is hard-wired to an indicator node (square) that represents, via equation (2), the average population density in the spatial neighborhood. *Panel (c): Allowing for missing data.* The model can be further improved by connecting all nodes to a latent node that represents unobserved effects. The observation status at a node is, in the first instance, predicted by the spatial neighborhood and/or the latent variable. Only if the explanatory power of these correction schemes is not sufficient will there be an incentive for the inference scheme to include further edges related to species interactions. Hence the effect of these corrections is to reduce the network connectivity and filter out spurious interactions.

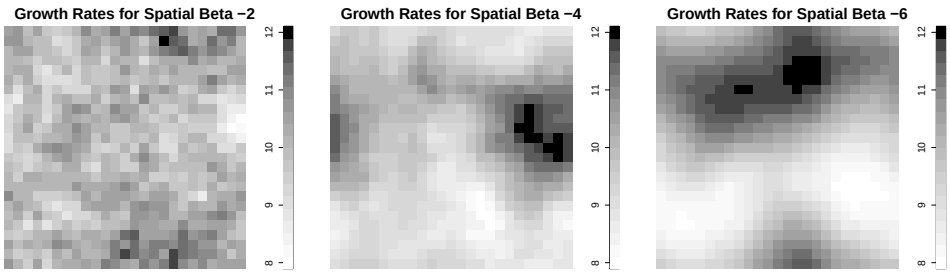

Figure 2: **Spatial autocorrelation.** The figure shows the spatial distribution of growth rates r_n entering equation (8) as the spatial β parameter, defined in Section 3.2, decreases from -2 to -8. A value of 0 would correspond to uniformly random noise, and -2 is Brownian noise.

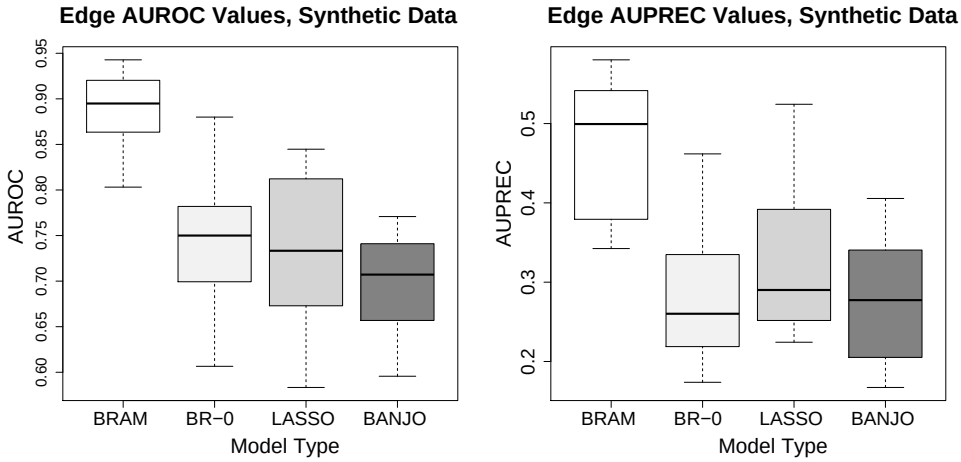

Figure 3: **Comparison on synthetic data.** Boxplots of AUROC (left panel) and AUPRC (right panel) scores obtained with three methods on the synthetic data described in Section 3.1: the proposed model (BRAM), a Bayesian linear regression model without changepoints and correction for spatial autocorrelation (BR-0), sparse L1-regularized linear regression (LASSO), and a homogeneous Bayesian network with the BDe score (BANJO). No correction for spatial autocorrelation is required. The boxplots show the distributions of the scores for 30 independent data sets, where the horizontal bar shows the median, the box margins show the 25th and 75th percentiles and the whiskers indicate data within 2 times the inter-quartile range.

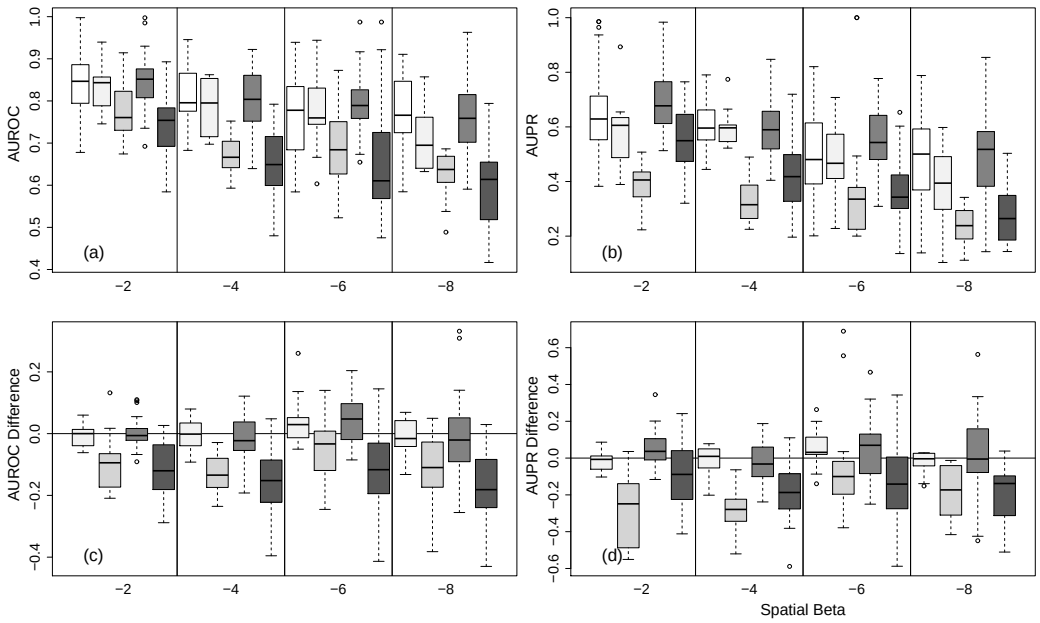

Figure 4: **Comparative evaluation of five network reconstruction methods, strong predation.** AUROC (left column) and AUPRC (right column) scores obtained on the trophic simulated data described in Section 3.2. Top row: absolute scores. Bottom row: difference scores, with the proposed model (BRAM) taken as a reference, i.e. positive (negative) values indicate a better (worse) performance of BRAM. The abscissa represents different values of the spatial β parameter, whose influence is illustrated in Figure 2. Panels: **(a)** Absolute AUROC values for BRAM (white), BR (light gray), BR-0 (gray), LASSO (dark gray), Banjo (darkest gray); **(b)** Absolute AUPR values; **(c)** Pairwise difference of AUROC and **(d)** AUPR. For an interpretation of the boxplots, which show a distribution of the scores over 30 independent datasets, see Figure 3.

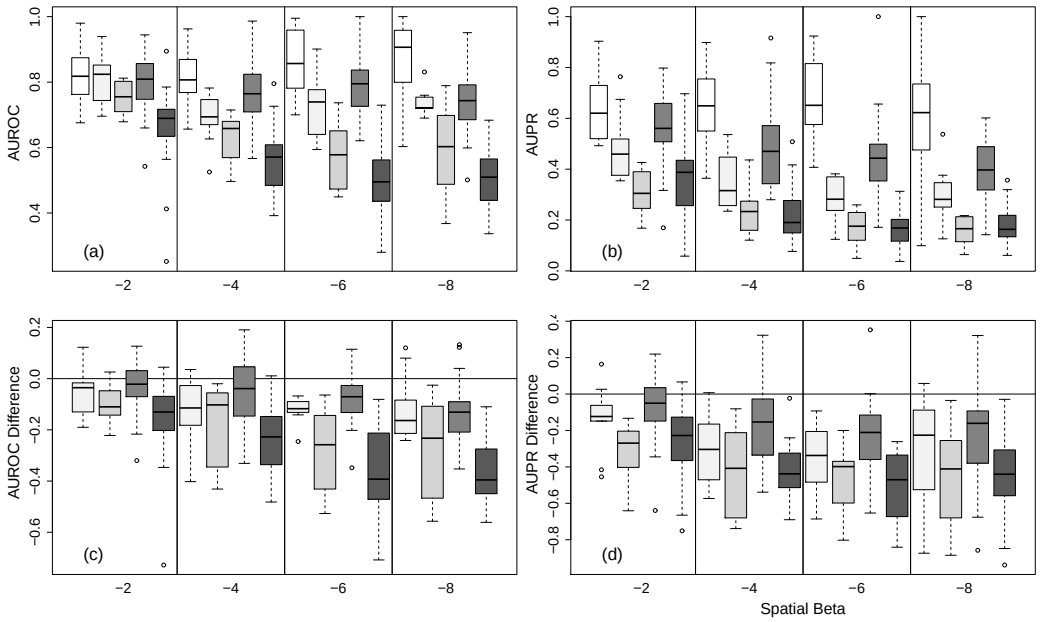

Figure 5: **Comparative evaluation of five network reconstruction methods, weak predation.** AUROC (left column) and AUPRC (right column) scores obtained on the trophic simulated data described in Section 3.2. The simulations were carried out as for Figure 4, but with a weakened influence of the predators on the prey. See the caption of Figure 4 for details. Panels: **(a)** Absolute AUROC values for BRAM (white), BR (light gray), BR-0 (gray), LASSO (dark gray), Banjo (darkest gray); **(b)** Absolute AUPR values; **(c)** Pairwise difference of AUROC and **(d)** AUPR.

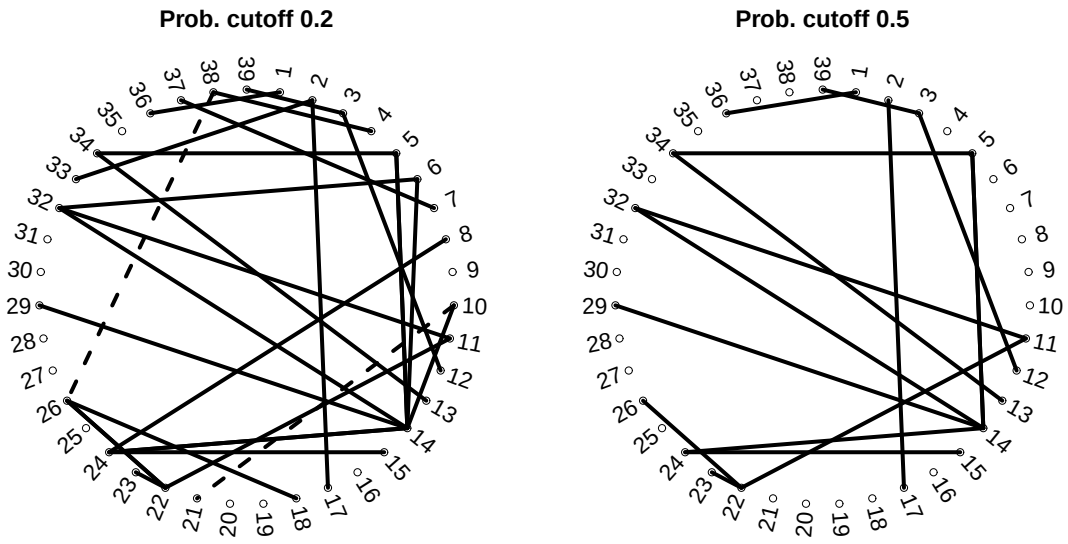

Figure 6: **Species interaction network inferred with BRAM from the ecological survey data described in Section 3.3.** The graph displays species interactions with an inferred marginal posterior probability of 0.2 (left panel) and 0.5 (right panel). Several soil attributes were defined to be fixed inputs to each plant. Solid lines correspond to positive interactions (e.g. mutualism, facilitation) and dashed to negative (e.g. resource competition). The species, represented by numbers, have been ordered phylogenetically, with the four groups of forbs (1-19), grasses (20-29), rushes (30-33) and sedges (34-39). Full species names of the indices are listed in the Supplementary Material, Table Appendix A.

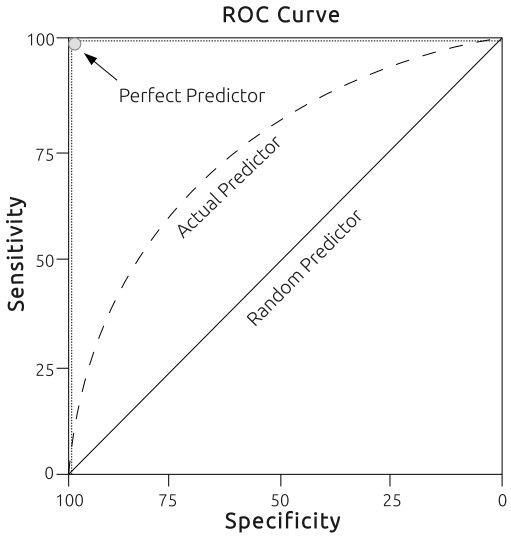

Figure 7: **Receiver operating characteristic (ROC) curve.** The figure shows the ROC curve for a perfect predictor, random expectation, and a typical predictor between these two extremes.

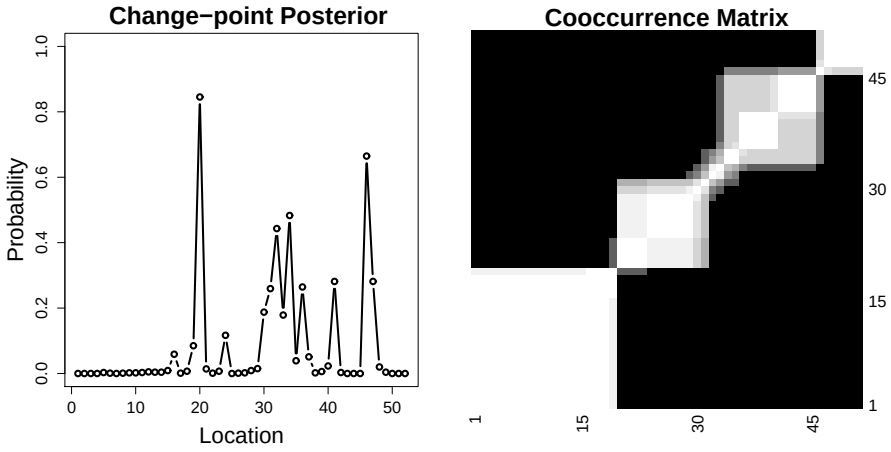

Figure 8: **Inferred spatial segmentation for a selected plant species, *Carex pulicaris*. Left panel:** Marginal posterior probability of a change-point occuring along the longitudinal direction in arbitrary units (corresponding to the plot location ID number in the ecological survey). **Right panel:** Cooccurrence matrix for the selected plant species. The axes represent the position along the longitudinal direction, as before. The grey shading indicates the posterior probability of two longitudinal positions being assigned to the same spatial segment, i.e. of not being separated by a changepoint, ranging from 0 (black) to 1 (white).

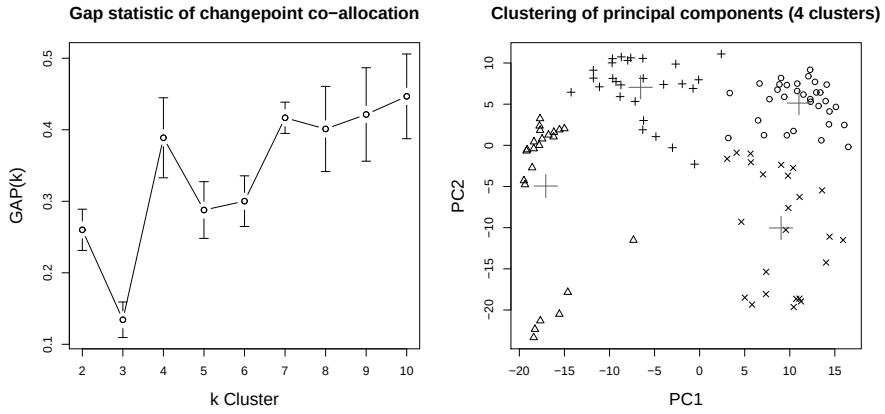

Figure 9: **Clustering of plant species based on their inferred spatial segmentation.** The plant species included in the ecological survey described in Section 3.3 were clustered on the basis of the inferred coocurrence matrices, shown in Figure 8. **Left panel:** The gap statistic, as proposed by Tibshirani et al. (2001) and Hastie et al. (2001), suggests that $k = 2$ and $k = 4$ are reliable cluster numbers because the gap difference to the subsequent cluster, $GAP(k) - GAP(k+1)$, is greater than the standard error at $GAP(k)$. This indicates that the increase of the sum of pairwise distances from k to $k + 1$ is significant and, hence, that k is a reasonable cluster number. **Right panel:** A plot of the plant species in the space spanned by the first principal components that were computed from the inferred coocurrence matrices. The symbols indicate cluster membership and the large crosses the center of each cluster.

Appendix A. Table

Table A.1: Indices with full scientific names as appearing in Figure 6. These plants can be assigned to four taxonomies of forbs (1-19), grasses (20-29), rushes (30-33) and sedges (34-39).

ID	Name
1	*Anagallis tenella*
2	*Calluna vulgaris*
3	*Drosera rotundifolia*
4	*Epilobium palustre*
5	*Galium verum*
6	*Hypochaeris radicata*
7	*Leontodon autumnalis*
8	*Lychnis flos-cuculi*
9	*Odontites verna*
10	*Plantago lanceolata*
11	*Potentilla erecta*
12	*Potentilla palustris*
13	*Prunella vulgaris*
14	*Ranunculus bulbosus*
15	*Ranunculus repens*
16	*Sagina procumbens*
17	*Succia pratensis*
18	*Trifolum repens*
19	*Viola riviniana*
20	*Agrostis capillaris*
21	*Aira praecox*
22	*Anthoxanthum odoratum*
23	*Cynosurus cristatus*
24	*Festuca rubra*
25	*Festuca vivipara*
26	*Holcus lanatus*
27	*Koeleria macrantha*
28	*Molinia caerulea*
29	*Poa pratensis*
30	*Juncus effusus*
31	*Juncus kochii*
32	*Luzula campestris*
33	*Luzula pilosa*
34	*Carex arenaria*
35	*Carex demissa*
36	*Carex dioica*
37	*Carex flacca*
38	*Carex nigra*
39	*Eriophorum angustifolum*