

Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods

Ali Faisal^a, Frank Dondelinger^{b,c,*}, Dirk Husmeier^b, Colin M. Beale^{d,1}

^a Helsinki Institute for Information Technology, Adaptive Informatics Research Centre, Department of Information and Computer Science, Aalto University, P.O. Box 15400, FI-02015 Aalto, Finland

^b Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, United Kingdom

^c Institute for Adaptive Neural Computation, School of Informatics, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom

^d Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen, AB15 8QH, United Kingdom

ARTICLE INFO

Article history:

Received 9 September 2009

Received in revised form 17 June 2010

Accepted 30 June 2010

Keywords:

Network reconstruction

Biotic interaction

Spatial autocorrelation

Bio-climate envelope

ABSTRACT

The complexity of ecosystems is staggering, with hundreds or thousands of species interacting in a number of ways from competition and predation to facilitation and mutualism. Understanding the networks that form the systems is of growing importance, e.g. to understand how species will respond to climate change, or to predict potential knock-on effects of a biological control agent. In recent years, a variety of summary statistics for characterising the global and local properties of such networks have been derived, which provide a measure for gauging the accuracy of a mathematical model for network formation processes. However, the critical underlying assumption is that the true network is known. This is not a straightforward task to accomplish, and typically requires minute observations and detailed field work. More importantly, knowledge about species interactions is restricted to specific kinds of interactions. For instance, while the interactions between pollinators and their host plants are amenable to direct observation, other types of species interactions, like those mentioned above, are not, and might not even be clearly defined from the outset. To discover information about complex ecological systems efficiently, new tools for inferring the structure of networks from field data are needed. In the present study, we investigate the viability of various statistical and machine learning methods recently applied in molecular systems biology: graphical Gaussian models, L1-regularised regression with least absolute shrinkage and selection operator (LASSO), sparse Bayesian regression and Bayesian networks. We have assessed the performance of these methods on data simulated from food webs of known structure, where we combined a niche model with a stochastic population model in a 2-dimensional lattice. We assessed the network reconstruction accuracy in terms of the area under the receiver operating characteristic (ROC) curve, which was typically in the range between 0.75 and 0.9, corresponding to the recovery of about 60% of the true species interactions at a false prediction rate of 5%. We also applied the models to presence/absence data for 39 European warblers, and found that the inferred species interactions showed a weak yet significant correlation with phylogenetic similarity scores, which tended to weakly increase when including bio-climate covariates and allowing for spatial autocorrelation. Our findings demonstrate that relevant patterns in ecological networks can be identified from large-scale spatial data sets with machine learning methods, and that these methods have the potential to contribute novel important tools for gaining deeper insight into the structure and stability of ecosystems.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Darwin's description of a tangled bank describes the everyday complexity of ecology that we overlook at our peril. Tampering with the population of one species can cause surprising and dramatic changes in the populations of others (Cohen et al., 1994; Henneman &

Memmott, 2001). Altering pressures to which ecosystems are exposed can drive them to alternative states (Beisner et al., 2003) or catastrophic failure (Sinclair & Byrom, 2006). Understanding and predicting how ecosystems will respond to change requires untangling the tangled bank and is of enormous importance during a period of rapid global change. Yet such a task can seem impossible given the enormous complexity of ecological systems and the excruciating fieldwork needed to quantify even the simplest of foodwebs (Memmott et al., 2000; Ings et al., 2009).

Currently, most work on ecological networks has focused on quantifying food webs and pollination networks by direct observation of interactions among individuals. This approach has provided important

* Corresponding author. Tel.: +44 131 650 7536; fax: +44 131 650 4901.

E-mail addresses: ali.faisal@tkk.fi (A. Faisal), frankd@bioss.ac.uk (F. Dondelinger), dirk@bioss.ac.uk (D. Husmeier), cb751@york.ac.uk (C.M. Beale).

¹ Present Address: Department of Biology (Area 18), University of York, PO Box 373, York, YO10 5YW, United Kingdom.

insight into the structure and stability of some types of ecological networks, and has also had some limited success in predicting the consequences of anthropogenic changes in managed ecosystems. However, the predictive ability of these types of networks is limited by their assumption that other types of interaction, such as competition or mutualism relationships are unimportant when these have recently been identified as perhaps overwhelming (Werner & Peacor, 2003; Schmitz et al., 2004). Recognising this importance, some recent attempts have been made to include such non-trophic interactions within food web models (e.g. van Veen et al., 2009) but traditional field observations are unable to quantify the strength of these interactions and new methods are required to allow ecological interaction networks to expand beyond the current food web paradigm.

There has recently been a surge of interest in elucidating and modelling the structure of biological networks. A variety of summary statistics for characterising the global properties of networks have been derived, like the degree distribution (Albert & Barabási, 2002), clustering coefficient (Watts & Strogatz, 1998) and average path length (Valiente, 2002). This has been augmented by local characterisations in terms of overrepresented network motifs (Milo et al., 2002), and measures of specialisation based on information theory (Blüthgen et al., 2006). The formation and evolution of a network can then be simulated from a mathematical model, like the simple preferential attachment model of (Barabási et al., 1999), or more realistic models of basic biological processes (de Silva & Stumpf, 2005). The summary statistics obtained from the ensemble of simulated networks can then be compared with those obtained from the real networks, and the discrepancy provides a measure of how accurately the mathematical model captures the true network formation processes.

A critical assumption of the approach delineated above is that the true network is known. In molecular systems biology, the structure of protein interaction networks is commonly obtained from yeast two-hybrid assays. It is well known that these experiments are noisy, that they are susceptible to large proportions of both false positives and false negatives, and that the networks extracted from different assays can differ substantially (e.g. Tong et al., 2002). In ecology, establishing the structure of a species interaction network typically requires minute observations and detailed field work. For instance, the information theoretic summary statistics proposed in Blüthgen et al. (2006) were applied to the plant–pollinator interaction networks obtained in the studies of Memmott (1999) and Vázquez and Simberloff (2002). These studies entailed detailed observations of how often a particular plant was visited by a particular pollinator, for all pollinators and plants in turn. This process is laborious and error-prone. More importantly, it is restricted to specific kinds of interactions. The interactions between pollinators and their host plants are amenable to direct observation. However, other types of species interactions, like competition for resources, are not, and might not even be clearly defined from the outset. Our work therefore aims to adapt a novel type of methodology that has recently been explored in molecular systems biology: to infer the network structure directly from the data. To reword this: rather than taking an “existing” network structure and analysing it in terms of summary statistics, we assume that the interaction network is unknown, and we aim to reconstruct it *in silico* from the species abundance counts.

Information about ecological interactions should be evident in a range of ecological data that are currently available. For example, time-series of the populations of multiple species present in a study site should allow identification of important interactions, and similarly the spatial patterns of coincidence of species should contain information about the interactions among these species, potentially at a range of scales. What is needed is a statistical tool capable of recovering network structures from these types of data sets. Recently, the challenge of identifying regulation networks and signalling pathways from post-genomic data has resulted in the development of a number of statistical and machine learning methods for the

recovery of network structure. Examples are the reconstruction of transcriptional regulatory networks from gene expression data (Friedman et al., 2000), the inference of signal transduction pathways from protein concentrations (Sachs et al., 2005), and the identification of neural information flow operating in the brains of songbirds (Smith et al., 2006). This development has potentially given ecologists a new set of tools for network recovery, if the methods can be applied to typical ecological data sets.

Our aim is to compare different models for recovering ecological interaction networks, similarly to the approach of Tirelli et al. (2009) for modelling presence/absence data of *Salmo marmoratus*. Here, we introduce and seek to test the suitability of four statistical/machine learning methods for the identification of network structure on ecological data: Graphical Gaussian models (GGMs), L1-regularised linear regression with the least absolute shrinkage and selection operator (LASSO), sparse Bayesian regression (SBR), and Bayesian networks. We extend these methods by including explanatory variables to model the effect of spatial autocorrelation and the impact of bio-climate variables. We first test the success of these methods for recovering the structure of simulated food webs, where the true structure is known precisely. We then use the methods to identify the large-scale interactions among 39 species of European warblers (families Phylloscopidae, Cettiidae, Acrocephalidae and Sylviidae), a subset of the European breeding bird data set (Hagemeyer & Blair, 1997) covering Europe west of 30°E and including all probable and confirmed breeding records. These data have been augmented by two bio-climate covariates, related to temperature and water availability. Our work has been motivated by preliminary explorations described in the first two authors' MSc dissertations (Faisal, 2008; Dondelinger, 2008). However, for the present paper, the methodology has been considerably expanded, new methodological concepts have been included, different ways of result and network integration have been explored, and all simulations have been rerun.

2. Materials and methods

2.1. Statistical and machine learning methods for network reconstruction

In the most general case, our aim in describing an ecological network is to model all the interactions between and among species and their environment. It is convenient to think of this network as a ‘graph’ (e.g. Fig. 2), describing species as the ‘nodes’ within the graph, and interactions as the links or ‘edges’ that join the nodes. To identify and infer these graphs we selected four widely used methods for network recovery in postgenomic data analysis: Graphical Gaussian Models (Schäfer & Strimmer, 2005a; Schäfer & Strimmer, 2005b), LASSO regression (Tibshirani, 1996; van Someren et al., 2006), sparse Bayesian regression (Tipping & Faul, 2003; Rogers & Girolami, 2005) and Bayesian Networks (Friedman et al., 2000; Werhli & Husmeier, 2007). All four methods have previously been used to recover gene regulation network structures and there is no *a priori* assumption that any method will perform best on ecological data where other statistical issues such as spatial autocorrelation (Lennon, 2000), small sample sizes, or the influence of other, unmeasured covariates may be important. Each method differs in the mechanism it uses to recover networks from data and as most methods will be unfamiliar to many ecologists we provide, in Section 3, a description of the important features of the methods we tried, along with the full details of the mathematical implementation. All methods were implemented in MATLAB © (The MathWorks, Inc.) or R (<http://www.R-project.org>) (see Table 2).

2.2. Simulation study

In order to have an objective measure of network recovery, we first tested the ability of the models to recover the true network structure

from test data generated by an ecological simulation model. This model combines a niche model (Williams & Martinez, 2000) with a stochastic population model (Lande et al. (2003), chap. 8) in a 2-dimensional lattice. The niche model defines the structure of the network and has two parameters (the number of species and the connectance (or network density) defined as L/S^2 where L is the number of links and S the number of species in the network).

More precisely, to generate a food web consisting of N species, we start off by assigning to each species i a niche value n_i , drawn uniformly from $[0, 1]$. This gives us an ordering of the species by niche value, where higher niche values mean that species are higher up in the food chain. For each species we then draw a niche range r_i from a beta distribution with expected value $2C$ (where C is the desired connectance) to determine the size of the niche that that species preys upon. Then we uniformly draw a centre c_i for the niche from $[r_i/2, n_i]$. This generates networks that share many characteristics with real food webs, such as the fraction of species with no prey, no predators or both prey and predators, and the amount of cannibalism and looping in the network.

The population model is defined by a stochastic differential equation where the dynamics of the log abundance X_i of species i can be expressed as:

$$\frac{dX_i}{dt} = r_i + \frac{\sigma_d}{\sqrt{N_i}} \frac{dA_i(t)}{dt} + \sigma_e \frac{dB_i(t)}{dt} - \gamma X_i - \Omega(X) + \sigma_E \frac{dE(t)}{dt} \quad (1)$$

where r_i is the growth rate of species i , σ_d is the standard deviation of the demographic effect, N_i is the abundance of species i (e^{X_i}), $A_i(t)$ is the species-specific demographic effect, σ_e is the standard deviation of the species-specific environmental effect, $B_i(t)$ 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. In order to incorporate the niche model, the simulation modifies the term Ω to include predator–prey interactions in the Lotka–Volterra form.

In order to extend this model to a 2D arena, the simulation incorporates an exponential dispersal model, where the probability of a species moving from location A to location B is determined by the Euclidean distance between A and B. Locations are arranged on a rectangular grid. Each location has its own growth rates. The spatial pattern of growth rates for a single species is generated by noise with spectral density f^β (with $\beta < 0$, and f the frequency at which the noise is measured), and a normal error distribution.

We simulated the dynamics of this model for 3000 steps (until the system had reached equilibrium), with 10 different network structures to generate 10 independent data sets. The final ‘gold-standard’ network against which the recovered networks were assessed was the structure of the niche model linking among species present in the data set (as some species went extinct during the initial runs to equilibrium). We recovered networks from these data using all methods first without consideration of spatial autocorrelation, then with the inclusion of spatial autocorrelation for methods where this was possible.

2.3. Application to the European bird atlas data

Until relatively recently, climate was considered to be the main factor affecting large-scale (continental) distribution patterns and global climate change is already having measurable effects on the distribution of many species (Gaston, 2003). Lately, however, theoretical models have suggested that biotic interactions may also be important in shaping range limits (Holt & Barfield, 2009), and recent empirical research has suggested that the distribution of many European bird species may not be as strongly related to climate as previously thought (Beale et al., 2008). This weaker than expected

association with abiotic climate variables may be explained if biotic interactions are more important than previously thought. If biotic interactions play an important role in large-scale species distributions, developing a method to identify and predict their influence must be considered a priority. If successful, therefore, application of network recovery methods to mapped data used in ecological analysis would be valuable.

To test the utility of the available methods for network recovery in this large context, we use a subset of the European breeding bird data set (Hagemeyer & Blair, 1997) covering Europe west of 30°E and including all probable and confirmed breeding records. From this data set we extracted the distributions of all 39 old world warbler species breeding in this area (families Phylloscopidae, Cettiidae, Acrocephalidae and Sylviidae). These species are all small insectivores occupying a range of habitat types from boreal forest to Mediterranean reedbeds, several of which are likely to interact at a range of spatial scales (e.g. Murray, 1988). As covariates we include the mean temperature of the coldest month and the water availability for plant growth, two climate variables that had strongest influence on avian distribution (Beale et al., 2008). Climate data were available at 0.5° (data set CRU CL 1.0, New et al., 1999), and because soil types differ in their ability to retain moisture (e.g. sandy soils drain very quickly, whilst clay retains water longer) were combined with soil data (data set WISEAWC, Batjes, 1996) using a bucket model (following Prentice et al., 1992) and interpolated to 50 km resolution. These or similar variables are typically used in distribution modelling exercises (e.g. Thomas et al., 2004; Thuiller et al., 2005; Araujo et al., 2005; Beale et al., 2008; Huntley et al., 2008) not because they are always expected to directly impact bird distributions, but they are perceived to have strong indirect effects on birds and other taxa through effects on food availability or habitat type (Araujo et al., 2005; Beale et al., 2008; Huntley et al., 2008). Other biologically relevant climate variables could also be used but are usually strongly correlated with one of these and have little effect on the strength of associations realised (Beale et al., 2008). As these are real valued variables, we discretise them by maximising the mutual information. For this pre-processing step, we perform a standard quantile discretisation into 20 levels and then use the information bottleneck algorithm, proposed by (Hartemink, 2001), to get a binary variable minimising the expected information loss.

As the simulation studies suggested that sparse Bayesian regression (SBR) consistently underperformed the other methods (see Section 4.1), and as the Gaussian assumption underlying graphical Gaussian models (GGMs) is violated by the binary nature of the data, we only applied L1-regularised regression (LASSO) and Bayesian networks to recover network structures from the real data sets. We used three different data sets that increased in complexity from the simple warbler dataset alone, through inclusion of spatial autocorrelation, to inclusion of the bio-climate covariates. We generated consensus networks for each data set, which should represent successively better models of true network structure. We also attempted to build a latent variable model, but the Markov chain Monte Carlo (MCMC) chains did not converge and we do not consider this further here.

In the absence of complete ecological knowledge of the true network of interactions among these species, success of the modelling methods can only be assessed against known or likely relationships. To validate our methods on these real datasets we therefore determined four tests: firstly, for each pairwise interaction we sought to give an *a priori* interaction score, identifying any published studies and, when these were unavailable, using expert judgement to categorise interactions into likely, unknown or unlikely (see Appendix B). We tested similarity between the recovered network and the *a priori* network using the area under the receiver operator characteristic (ROC) curve and the true positive rate at 5% false positives (TPFP5).

Secondly, as ecological niches are often conserved in evolutionary time (Losos, 2008) we expected there to be a relationship between phylogenetic distance and inferred interaction score. Thirdly, we expected that ecologically similar species were most likely to interact, so for each species we identified the preferred habitat, migrant status (resident, short or long-distance migrant), wing length, body mass and body length and clutch size (all data from <http://www.bto.org/birdfacts/> or Snow & Perrins, 1998) and summarised these variables to generate a measure of ecological distance (see Appendix B for further details). Significance of both these tests with phylogenetic and ecological distance was assessed by correlation. Finally, as well as expecting these measures to be related to the final networks identified, we predicted that the simpler (and less biologically plausible) network models lacking spatial autocorrelation and bio-climate covariates would show weaker associations with the ecological datasets than the full models, and the number of significant interactions among bird species in the network would decline as complexity increases (and spurious interactions are accounted for by the additional complexity of the model).

To characterise the networks recovered using these methods and put them in the context of other ecological networks, we counted the non-zero links with each species in turn, and measured the frequency distribution of these (i.e. we measured the degree distribution: Proulx et al., 2005). We also measured the mean shortest path between all species in the network (Dunne et al., 2002) and a measure of how clustered the network is (related to the proportion of species linked to the neighbours of a focal species that are themselves linked to the focal species. We measured the clustering coefficient: Luce & Perry (1949)). As our recovered networks are not binary but identify continuous probabilities of linkage between two species, we calculated all three values across a range of threshold levels and identified network characteristics that are consistent across all thresholds.

2.4. Units

Table 1 gives an overview of the units for different quantities in our paper, along with the equations where these quantities were used.

3. Theory

3.1. Statistical and machine learning methods for network reconstruction

3.1.1. Graphical Gaussian models (GGMs)

Graphical Gaussian models (GGMs) are undirected probabilistic graphical models that allow the identification of conditional independence relations among the nodes under the assumption of a multivariate Gaussian distribution of the data. The inference of GGMs

Table 1

Units for the different quantities in the paper, along with the equations where they are used (if any).

Symbol/quantity	Equation	Unit
$\mathbf{x}_r, x_i, X_i$	3, 17, 20, 21	Discrete presence/absence value of species over 50 km ² area
$\mathbf{y}_g, y_g$	3, 4, 8, 21	Discrete presence/absence value of species over 50 km ² area
$\mathbf{w}_g, w_g, \mathbf{w}, v$	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 21	Dimensionless weight parameters
a	20, 21	Spatial autocorrelation: discrete presence/absence value of species from averaging over four 50 km ² areas
Temperature covariate	None	Discrete warm/cold value over 50 km ² area
Water covariate	None	Discrete presence/absence value over 50 km ² area

is based on a (stable) estimation of the covariance matrix of this distribution. The element C_{ik} of the covariance matrix $\mathbf{C}$ is proportional to the correlation coefficient between nodes X_i and X_k . A high correlation coefficient between two nodes may indicate a direct interaction, an indirect interaction, or a joint regulation by a common (possibly unknown) factor.

However, only the direct interactions are of interest to the construction of a species interaction network. The strengths of these direct interactions are measured by the partial correlation coefficient ρ_{ik} , which describes the correlation between nodes X_i and X_k conditional on all the other nodes in the network. From the theory of normal distributions it is known that the matrix of partial correlation coefficients ρ_{ik} is related to the inverse of the covariance matrix $\mathbf{C}$, $\mathbf{C}^{-1}$ (with elements C_{ik}^{-1}) (Edwards, 2000):

$$\rho_{ik} = - \frac{C_{ik}^{-1}}{\sqrt{C_{ii}^{-1} C_{kk}^{-1}}} \quad (2)$$

To infer a GGM, one typically employs the following procedure. From the given data, the empirical covariance matrix is computed, inverted, and the partial correlations ρ_{ik} are computed from Eq. (2). The distribution of $|\rho_{ik}|$ is inspected, and edges (i,k) corresponding to significantly small values of $|\rho_{ik}|$ are removed from the graph. The critical step in the application of this procedure is the stable estimation of the covariance matrix and its inverse. Note that the covariance matrix is only non-singular if the number of observations exceeds the number of nodes in the network. This condition might not always be satisfied in a survey study. In order to learn a GGM from a data set in such a scenario, (Schäfer & Strimmer, 2005b) explored various stabilisation methods, based on the Moore-Penrose pseudo inverse and bagging.

In the present work, we apply an alternative regularisation approach based on shrinkage, which Schäfer and Strimmer (2005b) found to be superior to their earlier schemes. The idea is to add a weighted non-singular regularisation matrix, e.g. the unity matrix, to the covariance matrix so as to guarantee its non-singularity. The optimal weight parameter is estimated based on the Ledoit Wolf lemma from statistical decision theory so as to minimise the expected deviation of the regularised covariance matrix from the (unknown) true covariance matrix. The method of GGMs, which are undirected graphs, can be extended to infer putative directions of causal interactions, as proposed in Opgen-Rhein and Strimmer (2007). This scheme is based on the computation of the standardised partial variance, which is the proportion of the variance that remains if the influence of all other variables is taken into account. All significant edges in the GGM network are directed in such a fashion that the direction of the arrow points from the node with the larger standardised partial variance (the more *exogenous* node) to the node with the smaller standardised partial variance (the more *endogenous* node), provided the ratio of the two partial variances is significantly different from 1. For further details, see Opgen-Rhein and Strimmer (2007).

3.1.2. Linear regression and the LASSO

The approach discussed in the previous subsection aims to predict interactions between species based on the partial correlations between their abundance profiles. In the present subsection, we review an alternative paradigm, which pursues a regression approach: given the species abundance profile $\mathbf{y}_g$ of some target species g , we aim to find a set of regulators $\{r\}$ (i.e. other species or exogenous variables related e.g. to the habitat, climate etc.), whose abundance profiles $\{x_r\}$ are good predictors of abundance profile $\mathbf{y}_g$:

$$\mathbf{y}_g = \sum_r w_{gr} \mathbf{x}_r \quad (3)$$

where $\hat{\mathbf{y}}_g$ is a predictor of $\mathbf{y}_g$, and the regression parameters $\mathbf{w}_{gr}$ represent interaction strengths between the target species g and the putative regulators r .

The different concepts are illustrated in Fig. 1. We denote the vector of interaction strengths as $\mathbf{w}_g$, which has w_{gr} as its r th component. The mismatch between the predicted and measured expression profile of target species g is typically measured by the L2 norm:

$$E(\mathbf{w}_g) = \|\mathbf{y}_g - \hat{\mathbf{y}}_g(\mathbf{w}_g)\|^2 \quad (4)$$

Obtaining the optimal interaction parameters $\hat{\mathbf{w}}_g$ by minimising $E(\mathbf{w}_g)$ corresponds to a maximum likelihood estimator under the assumption of isotropic Gaussian noise. In practice, this approach is usually susceptible to over-fitting, which calls for the application of some regularisation scheme. The standard method of ridge regression is given by:

$$\hat{\mathbf{w}}_g = \arg \min_{\mathbf{w}_g} \left(E(\mathbf{w}_g) + \lambda \sum_r \mathbf{w}_{gr}^2 \right) \quad (5)$$

This can be interpreted in three different ways:

1. Maximising the penalised likelihood with an L2-norm penalty term and regularisation parameter λ .
2. Constrained maximisation of the likelihood under the L2-norm constraint $\sum_r \mathbf{w}_{gr}^2 < C$, where λ is a Lagrange parameter.
3. Bayesian *maximum a posteriori* estimate under a zero-mean Gaussian prior on $\mathbf{w}_g$ with diagonal isotropic covariance matrix $\lambda^{-1}\mathbf{I}$:

$$P(\mathbf{w}_g) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}).$$

A disadvantage of ridge regression is that the set of interaction parameters $\{\mathbf{w}_{gr}\}$ does usually not tend to be sparse. This is a consequence of the fact that the derivative of the regularisation term with respect to w_{gr} approaches zero as $w_{gr} \rightarrow 0$. Consequently, there is no “force” pulling the parameters to zero when they are small. According to our current knowledge, species interaction networks are usually sparse, and a stronger regularisation term is therefore desirable. This can be achieved with an L1-norm instead of the L2-norm regularisation term:

$$\hat{\mathbf{w}}_g = \arg \min_{\mathbf{w}_g} \left(E(\mathbf{w}_g) + \lambda \sum_r |\mathbf{w}_{gr}| \right) \quad (6)$$

which can be interpreted as a Bayesian *maximum a posteriori* estimate under a Laplacian prior $\mathbf{w}_g$, as first proposed by (Williams, 1995). The derivative of the regularisation term with respect to the parameters is now constant, which provides a stronger “force” driving small parameters to zero. The discontinuity of the derivative at $w_{gr} \rightarrow 0$

can be exploited to implement an effective pruning scheme for discarding interactions, as discussed in Williams (1995). The L1-norm regularisation term was introduced to the statistics community by Tibshirani (1996), where it was termed the LASSO (least absolute shrinkage and selection operator). One of the first applications to the somewhat related problem of reconstructing gene regulatory networks is reported in van Someren et al. (2006). Grandvalet and Canu (1999) showed that the LASSO estimate of the interaction strengths is equivalent to ridge regression with r -dependent regularisation hyperparameters:

$$\hat{\mathbf{w}}_g = \arg \min_{\mathbf{w}_g} \left(E(\mathbf{w}_g) + \sum_r \lambda_r \mathbf{w}_{gr}^2 \right) \quad (7)$$

subject to the constraint $\sum_{r=1}^R 1/\lambda_r = R/\lambda$, for some predefined constant λ .

The regulatory network between the target species g and the regulators $\{r\}$ is defined by the set of interactions with nonzero interaction strengths w_{gr} . The degree of sparsity is determined by the regularisation hyperparameter λ , with larger values of λ resulting in sparser networks. The question, then, is how to set λ . Williams (1995) suggested integrating λ out; this approach has been subject to some controversy, though (MacKay, 1996). A standard non-Bayesian approach is to estimate λ with k -fold cross-validation. This is the approach that was implemented in the software we applied in the present study, with $k=10$. An alternative Bayesian approach would be to estimate λ by maximising the evidence, as discussed in the next subsection.

Note that the generalisation of the sparse regression approach to more target species g is straightforward: $E(\mathbf{w}_g)$ in Eq. (4) just needs to be replaced by:

$$E(\mathbf{W}) = \sum_g \|\mathbf{y}_g - \hat{\mathbf{y}}_g(\mathbf{w}_g)\|^2 \quad (8)$$

where $\mathbf{W}$ is a matrix with column vectors $\mathbf{w}_g$. If there is no clear separation between the set of target and regulatory species, the effect of species g needs to be excluded when forming the predictor $\hat{\mathbf{y}}_g(\mathbf{w}_g)$. Again, this requirement is straightforward to implement. To avoid notational opacity, we have not described this approach in its full generality, though.

3.1.3. Sparse Bayesian Regression (SBR)

As mentioned in the previous subsection, the minimisation of $E(\mathbf{w}_g)$ in Eq. (4) corresponds to maximising the likelihood $P(\mathbf{D}|\mathbf{w}_g)$ under the assumption of isotropic Gaussian noise, where $\mathbf{D} = \{\mathbf{y}_g, \{\mathbf{x}_r\}\}$ is used to denote the data. The estimates $\hat{\mathbf{w}}_g$ in Eqs. (5) and (7) are equivalent to the *maximum a posteriori* estimates:

$$\hat{\mathbf{w}}_g = \arg \max_{\mathbf{w}_g} P(\mathbf{w}_g|\mathbf{D}, \lambda) = \arg \max_{\mathbf{w}_g} [\log P(\mathbf{D}|\mathbf{w}_g) + \log P(\mathbf{w}_g|\lambda)] \quad (9)$$

under the assumption of an isotropic Gaussian or Laplacian prior $P(\mathbf{w}_g|\lambda)$ on the interaction strengths $\mathbf{w}_g$. If we now want to do this within the Bayesian framework, the hyperparameter λ is optimised by maximising the marginal likelihood or evidence:

$$P(\mathbf{D}|\lambda) = \int P(\mathbf{D}|\mathbf{w}, \lambda) P(\mathbf{w}|\lambda) d\lambda \quad (10)$$

as discussed in MacKay (1992). In the present study, we applied the “sparse Bayesian regression” (SBR) approach of Rogers and Girolami (2005), which is based on the work of Tipping and Faul (2003). Here, the prior on the interaction parameters is chosen to be a product of zero-mean Gaussian distributions:

$$P(\mathbf{w}_g|\lambda) = \prod_r \mathcal{N}(w_{gr}|\mathbf{0}, \lambda_r^{-1}) \quad (11)$$

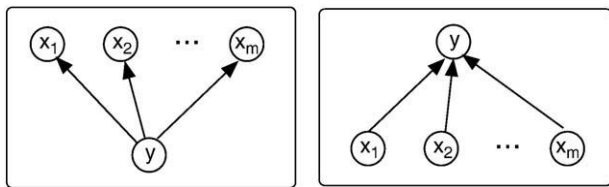


Fig. 1. Schematic of the approach of partial correlation (left) and sparse regression (right). Left: Conditional on y , the species abundance profiles $x_1, x_2, \dots, x_m$ are independent, and the partial correlation coefficients will be small. Right: The approach of sparse regression aims to find a minimal set of predictors $x_1, x_2, \dots, x_m$ to explain species abundance profile y .

with separate hyperparameters for species r . This scheme is similar to Eq. (7), except that the constraint: $\sum_{r=1}^R 1/\lambda_r = R/\lambda$ is missing. We can think of this as ARD (Automatic Relevance determination) in the sense used by MacKay (1992).

The hyperparameters λ_r are optimised with the evidence scheme described above.² Tipping and Faul (2003) showed that the marginal likelihood can be decomposed into separate contributions from the individual regulatory species $\{r\}$. This leads to a fast, iterative maximisation algorithm not only for the hyperparameters λ_r , but also for the network structure: interactions between the target species g and the putative regulatory species $\{r\}$ are progressively added and removed until a local maximum of the marginal likelihood is reached. Specific details of the algorithm can be found in Tipping and Faul (2003).

The reason for the sparsity of sparse Bayesian regression may not be immediately apparent. In fact, as Tipping (2001) points out, it comes from the hierarchical nature of the prior on weights in Eq. (11). Each hyperparameter λ_r has a prior from the Gamma family of distributions. In the algorithm, we assume an uninformative Gamma prior with the shape and inverse scale parameters set to zero, which leads to an improper prior for the weights if we integrate the hyperparameter out:

$$P(\mathbf{w}_g) = \int P(\mathbf{w}_g | \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \quad (12)$$

Which for an individual weight gives:

$$P(w_{gr}) \propto \frac{1}{w_{gr}} \quad (13)$$

This is clearly a sparse prior. In fact, we can make an analogy to LASSO here. If one takes a Bayesian view of the LASSO, as in Park and Casella (2008), then each weight in the LASSO estimate has an independent Laplace prior, so that:

$$P(w_{gr}) \propto \exp(-|w_{gr}|) \quad (14)$$

Both the LASSO and the SBR prior are sparse. However, they differ in the amount of regularisation that they apply, as can be seen by taking the derivative of the negative log likelihood for both priors:

$$\text{SBR} : \frac{d}{dw_{gr}} \{-\log P(w_{gr})\} \propto \frac{1}{w_{gr}} \quad (15)$$

$$\text{LASSO} : \frac{d}{dw_{gr}} \{-\log P(w_{gr})\} \propto \text{const} \quad (16)$$

The regularisation term for LASSO is constant, while the regularisation term for SBR tends to infinity as the weight tends to zero.

3.1.4. Bayesian networks (BNs)

Bayesian networks (BNs) have received substantial attention from the computational biology community as models of regulatory and interaction networks (Friedman et al., 2000; Hartemink et al., 2001; Needham et al., 2007). Formally, a BN is defined by a graphical structure $\mathcal{H}$, a family of (conditional) probability distributions F , and their parameters $\mathbf{q}$, which together specify a joint distribution over a set of random variables of interest. The structure $\mathcal{H}$ of a BN consists of a set of nodes and a set of directed edges. The nodes represent random variables, e.g. species and their abundance values, while the edges indicate conditional dependence relations. The structure $\mathcal{H}$ of a BN is a directed acyclic graph (DAG), which defines a unique rule for

expanding the joint probability in terms of simpler conditional probabilities. Let $X_1, X_2, \dots, X_n$ be a set of random variables represented by the nodes $i \in \{1, \dots, n\}$ in the graph, define $pa[i]$ to be the set of nodes with a directed edge feeding into node i (the “parents”), and let $X_{pa[i]}$ represent the set of random variables associated with $pa[i]$. Then

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{pa[i]}) \quad (17)$$

The objective of learning is to find network structures with high posterior probabilities, i.e. to sample network structures $\mathcal{H}$ from the posterior distribution

$$P(\mathcal{H} | \mathbf{D}) \propto P(\mathbf{D} | \mathcal{H}) P(\mathcal{H}) \quad (18)$$

where $\mathbf{D}$ denotes the training data. This requires a marginalisation over the parameters $\mathbf{q}$:

$$P(\mathbf{D} | \mathcal{H}) = \int P(\mathbf{D} | \mathbf{q}, \mathcal{H}) P(\mathbf{q} | \mathcal{H}) d\mathbf{q} \quad (19)$$

If certain regulatory conditions, discussed in Heckerman (1999), are satisfied and the data are complete, then the integral in Eq. (19) is analytically tractable. Two function families F that satisfy these conditions are the multinomial distribution with a Dirichlet prior (Heckerman et al., 1995) and the linear Gaussian distribution with a normal-Wishart prior (Geiger & Heckerman, 1994). The resulting scores $P(\mathbf{D} | \mathcal{H})$ are usually referred to as the BDe (discretised data, multinomial distribution) or the BGe (continuous data, linear Gaussian distribution) score. Direct sampling from the posterior distribution Eq. (18) is analytically intractable and is therefore approximated with Markov Chain Monte Carlo (MCMC) (Madigan & York, 1995; Friedman & Koller, 2003; Grzegorzczak & Husmeier, 2008). To restrict the size of the configuration space, we restrict the fan-in to a node, i.e. we keep the number of incoming edges from other nodes below a pre-specified threshold (3 in our study). This approach, which is commonly adopted in other studies, e.g. Friedman and Koller (2003), incorporate our prior knowledge that interaction networks are usually sparse.

The ultimate objective is to infer causal relations among the interacting nodes. While such a causal network forms a valid Bayesian network, the inverse relation does not always hold. One reason for this discrepancy is the existence of unobserved nodes. Even under the assumption of complete observation, the inference of causal interaction networks can be impeded by symmetries within so-called equivalence classes, which consist of networks that define the same conditional independence relations. Each Bayesian network corresponds to a whole equivalence class, represented by a complete partially directed acyclic graph (CPDAG); see Chickering (1995). Under the assumption of complete observation, directed edges in a CPDAG can be taken as indications of putative causal interactions (Friedman et al., 2000).

Several tutorials on Bayesian networks have been published; see for instance Heckerman (1999), Husmeier et al. (2005) and Grzegorzczak et al. (2008b) for further details.

3.2. Extension

3.2.1. 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 (Legendre, 1993; Lennon, 2000; Dale & 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. Where possible, we applied an autoregressive approach similar to that of Augustin

² In statistics this is called a type-II maximum likelihood estimation.

et al. (1996) to incorporate potential spatial autocorrelation into the models. To this end, we computed the average population at neighbouring cells, weighted inversely proportional to the distance of the neighbours, which we will call the autocorrelation variable:

$$a = \frac{\sum_{i=1}^N \omega_i x_i}{\sum_{i=1}^N \omega_i} \quad (20)$$

where N is the number of neighbours that we're considering. We chose $N=4$.³ x_i is the population density at neighbour i , and ω_i is the weight given to that neighbour, which is inversely proportional to the Euclidean distance of the neighbour. A slight subtlety when working with real world data that is not distributed in a regular grid is to work out which neighbouring locations to consider. In this work, we have opted for the closest neighbours by Euclidean distance. The extension to the discrete case is straightforward; we simply discretise the autocorrelation variable using a threshold.

The regression then becomes:

$$\hat{y}_g = \sum_r w_{gr} \mathbf{x}_r + va \quad (21)$$

where w_{gr} denotes the weights associated with each species r , and v is the additional weight assigned to the autocorrelation variable. The weight v will catch the effects of the spatial autocorrelation, leaving the other weights to determine the effects of other species on species g .

For Bayesian networks, we connect each node to a parent node whose value is given by Eq. (20), i.e., a representation of the spatial neighbourhood. The incoming edge from the parent node is enforced and excluded from the fan-in count. In this way the observation status at a node is, in the first instance, predicted by the spatial neighbourhood. 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.

Introducing spatial autocorrelations into GGMs is less straightforward. Since we did not apply GGMs to the real data (owing to their binary nature), we did not further pursue this issue in our work.

3.2.2. Bio-climate covariates

We include the bio-climate covariates (discretised temperature and water availability) as extra variables, in the same way as we included the spatial autocorrelation variable. In particular, in the Bayesian networks, we introduce fixed connections between the bio-climate covariates and the other nodes. We modify the fan-in limit so that it does not take these extra variables into account (i.e. if the fan-in limit is three, then that means that a species can have up to six parent nodes: three other species, the covariates, and the spatial autocorrelation node).

3.2.3. Consensus networks

As each of the network reconstruction methods has advantages and disadvantages, it may be useful to combine outputs of different methods into one single recovered network. Such a network would capture the consensus between the various methods, whilst simultaneously allowing the strengths of the different methods to be combined (e.g. interaction size and sign inferred with regression-based methods could inform the marginal posterior probabilities obtained for Bayesian networks). Simulation experiments showed that the expected accuracy of the consensus network is higher than the expected average accuracy of the individual networks.

In the present project, we generated consensus networks by normalising the estimated interaction probabilities and absolute

strengths (where available) from each method to the range [0, 1], then taking the arithmetic mean across all methods included within the consensus graph. This potentially confuses statistical significance (probabilities) with biological significance (strengths). However, for methods where both significance measures were available we found a very strong correlation between the two ($\rho=0.92$).

3.3. Performance evaluation

Each network reconstruction method infers a matrix of interaction strengths among all species (nodes) in the network (graph). The nature of interaction strengths varies among the methods (GGMs: partial correlation coefficients, LASSO and SBR: regularised regression coefficients, Bayesian networks: marginal posterior probabilities). However, all three scores define a ranking of the edges. If the true interaction network is known, this ranking defines a receiver operating characteristic (ROC) curve, where the relative number of real interactions (i.e. the true positive or TP rate) is plotted against the relative number of spurious interactions (the false positive or FP rate) for all possible thresholds on the rank. To assess the network reconstruction accuracy, we follow the procedure outlined in (Werhli et al., 2006) and apply two complementary performance measures. The first measure is the area under the receiver operating characteristic curve (AUC), which is a widely used global measure of reconstruction accuracy. The expectation value for a random predictor is $AUC=0.5$, a perfect predictor gives $AUC=1.0$, and larger values indicate a better reconstruction accuracy overall. As we are particularly interested in the performance of the network recovery methods when setting the threshold to a value that generates few false positives, we also identified the threshold that leads to an FP rate of 5% and counted the proportion of true interactions that were recovered at this threshold. We call this second measure the TP rate at 5% FP rate (the TFP5 score). A good network reconstruction method is characterised by both a high AUC score and a high TFP5 score.

3.4. Implementations

Table 2 shows which software we used for the different network reconstruction methods described in Section 3.1, as well as where to

Table 2
Network reconstruction software used.

Method	Software	Package	Description
GGM	R	GeneNet	The software implementing Graphical Gaussian models is described in (Schäfer & Strimmer, 2005b) and can be found at: http://strimmerlab.org/software/genenet/
LASSO (Linear)	MATLAB	Genelab	For LASSO regression with continuous data, we used software from the Genelab package referenced in (van Someren et al., 2006).
LASSO (Logistic)	C	BBR	For LASSO regression with discrete data, we used the BBR package (for Bayesian Binary Regression), which implements logistic LASSO regression. The package can be found at: http://www.stat.rutgers.edu/~madigan/BBR/
SBR	MATLAB	RegNets	We used the sparse Bayesian regression software referenced in (Rogers & Girolami, 2005) and available here: http://www.dcs.gla.ac.uk/~srogers/reg_nets.htm
Structure MCMC	MATLAB	None	The implementations for Structure MCMC and Structure MCMC with latent variables were developed from code by Marco Grzegorzczuk and can be found at: http://www.bioss.ac.uk/students/frankd.html
Population Simulation	MATLAB	None	The simulation code was developed by Jonathan Yearsley and slightly modified for this project. It can be found at: http://www.bioss.ac.uk/students/frankd.html

³ A comparison with $N=8$ showed no significant differences in the network reconstruction (see Appendix A).

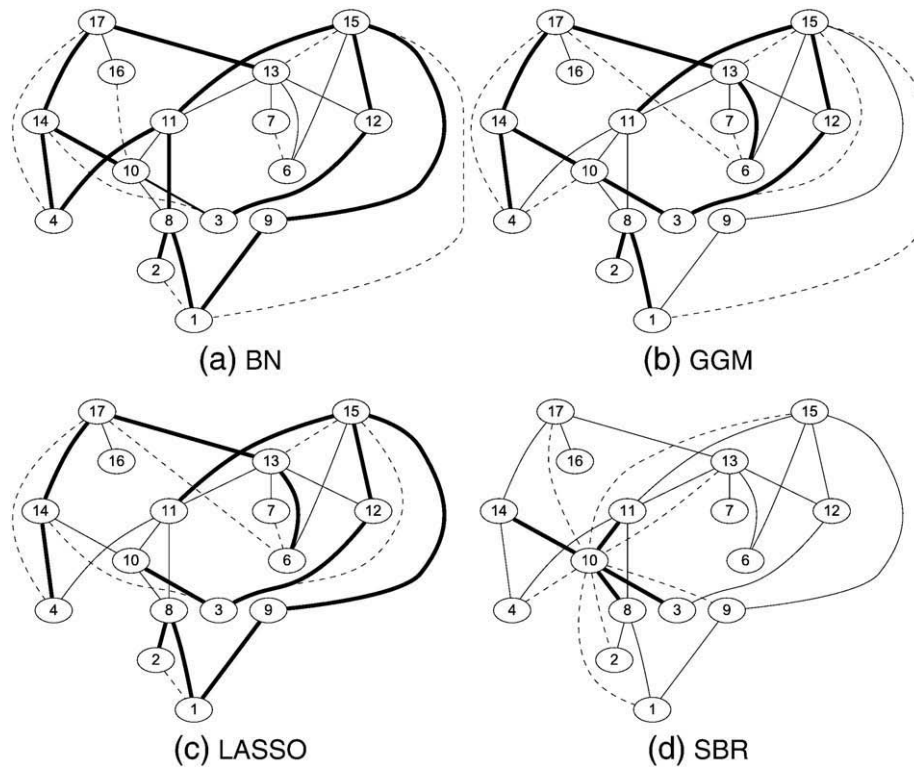


Fig. 2. An example of a network recovered by GGM, BN, LASSO and SBR. Thick edges represent edges that were identified correctly (true positives), thin edges represent edges that were not found (false negatives) and dashed edges are spurious edges (false positives). The threshold was chosen so that the false positive rate was constant at 5%, resulting in 7 false positive edges.

get the MATLAB code for our own implementations of the extensions in [Section 3.2](#).

4. Results

4.1. Simulation results

All four network recovery methods succeeded in recovering some of the true network structure, even when spatial autocorrelation was not incorporated ([Fig. 2](#)), though the methods varied in their performance ([Fig. 3](#)). Sparse Bayesian regression recovered networks that were significantly worse than those recovered by the other methods, having significantly lower AUC and TFP5 scores ($t_9 > 3$,

$p < 0.01$) except for the comparison with BN using the AUC score ($t_9 = 2.19$, $p = 0.06$). All t statistics and p -values have been calculated using a two-sided paired t -test, and the significance level was set at $p = 0.05$ (see [Appendix A](#) for information on where to find detailed p -values for all comparisons). Analysis of the inferred interaction strengths indicates that poor performance of SBR is a consequence of recovering networks that have too few links (i.e. are too sparse). This is the result of SBR being over-regularised (see [Section 3.1.3](#) for a discussion of this phenomenon).

For the three network recovery methods where this was applied (LASSO, BN, SBR), incorporating spatial autocorrelation resulted in improved performance, especially for those methods that performed less well in the simple model ([Fig. 3](#)). In particular, although

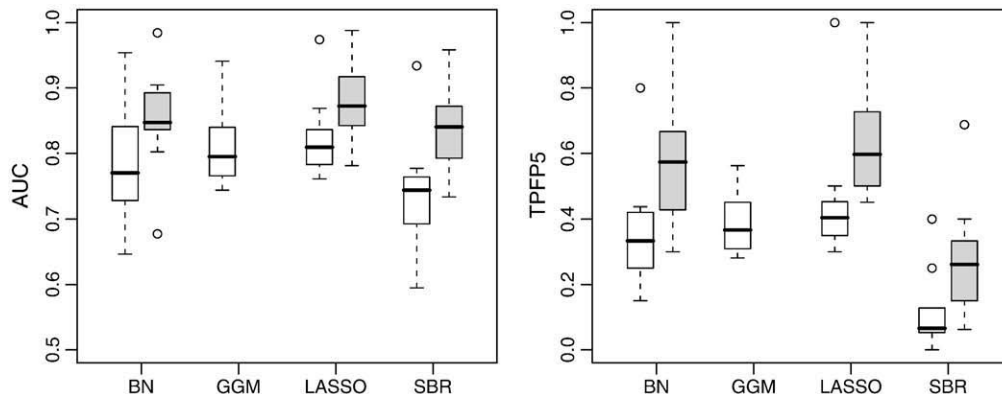


Fig. 3. AUC and TFP5 performance measures for continuous simulation data. Shaded boxes represent models which include spatial autocorrelation. The expected random performance scores are AUC = 0.5 and TFP5 = 0.05. See [Section 3.1](#) for an explanation of the abbreviations BN, GGM, LASSO, and SBR.

incorporating spatial autocorrelation improved the performance of SBR, it was still significantly worse than the other two methods ($t_9 > 3$, $p < 0.01$) except in the case of BN with the AUC score again ($t_9 = 0.68$, $p = 0.52$).

Adding an observation process to discretise the simulation datasets gave qualitatively similar results (beyond an expected drop in AUC and TFP5 scores). Consequently, we do not report this analysis further here, but see [Appendix A](#) for more information.

4.2. Application results

We recovered three consensus networks for the warbler data: for data sets with birds only ([Fig. 4](#)), with birds and spatial autocorrelation ([Fig. 5](#)) and with birds, spatial autocorrelation and bio-climate covariates ([Fig. 6](#)). Comparison of the recovered consensus networks with the *a priori* network predicted from the literature and expert judgement revealed small but statistically significant relationships ([Fig. 7](#)). We also identified small but significant relationships between the interaction score for the recovered consensus networks and both the phylogenetic and ecological distances ([Table 3](#)). Increasing model complexity (i.e. sequentially adding autocorrelation and bio-climate covariates) generally led to both stronger correlations with the predicted network structure and sparser networks ([Fig. 8](#)). Our predictions in [Section 2.3](#) were therefore corroborated.

Network characterisation identified that the degree distribution of the consensus networks was consistent across all threshold values, with all networks showing an exponential distribution. Both the

clustering coefficient and the mean shortest path length varied greatly as the threshold level changed and are therefore not considered a useful description of these networks. See [Appendix A](#) for more information about the network characterisation.

5. Discussion

As expected, we found that warblers in Europe form a well connected network, with most well known interactions (e.g. several *Acrocephalus* warblers: *A. arundinaceus*/*A. melanopogon*/*A. schoenobaenus*/*A. scirpaceus* ([Schäfer et al., 2006](#); [Rolando & Palestini, 1991](#)), and a triangle of interacting *Sylvia* warblers: *S. borin*/*S. atricapilla*/*S. communis* ([Elle, 2003](#); [Garcia, 1983](#))) accurately described by the better consensus network structures.

Given the general expectation that climate alone shapes distributions at large scales, it might seem surprising that the chosen bioclimate variables were not more strongly connected to species distributions. We believe there are two primary reasons for the relatively low effect of climate variables: firstly, our discretised climate data is likely to be too crude to capture all the meaningful climate variation, reducing the association with these parameters. Secondly, there is growing evidence to suggest that the importance of climate and abiotic variables has previously been overstated (e.g. [Watts & Worner, 2008](#)) largely because processes like the biotic interactions included in our models have previously been neglected ([Davis et al., 1998](#); [Beale et al., 2008](#); [Holt & Barfield, 2009](#); [La Sorte et al., 2009](#)). It would clearly be valuable to develop the methods

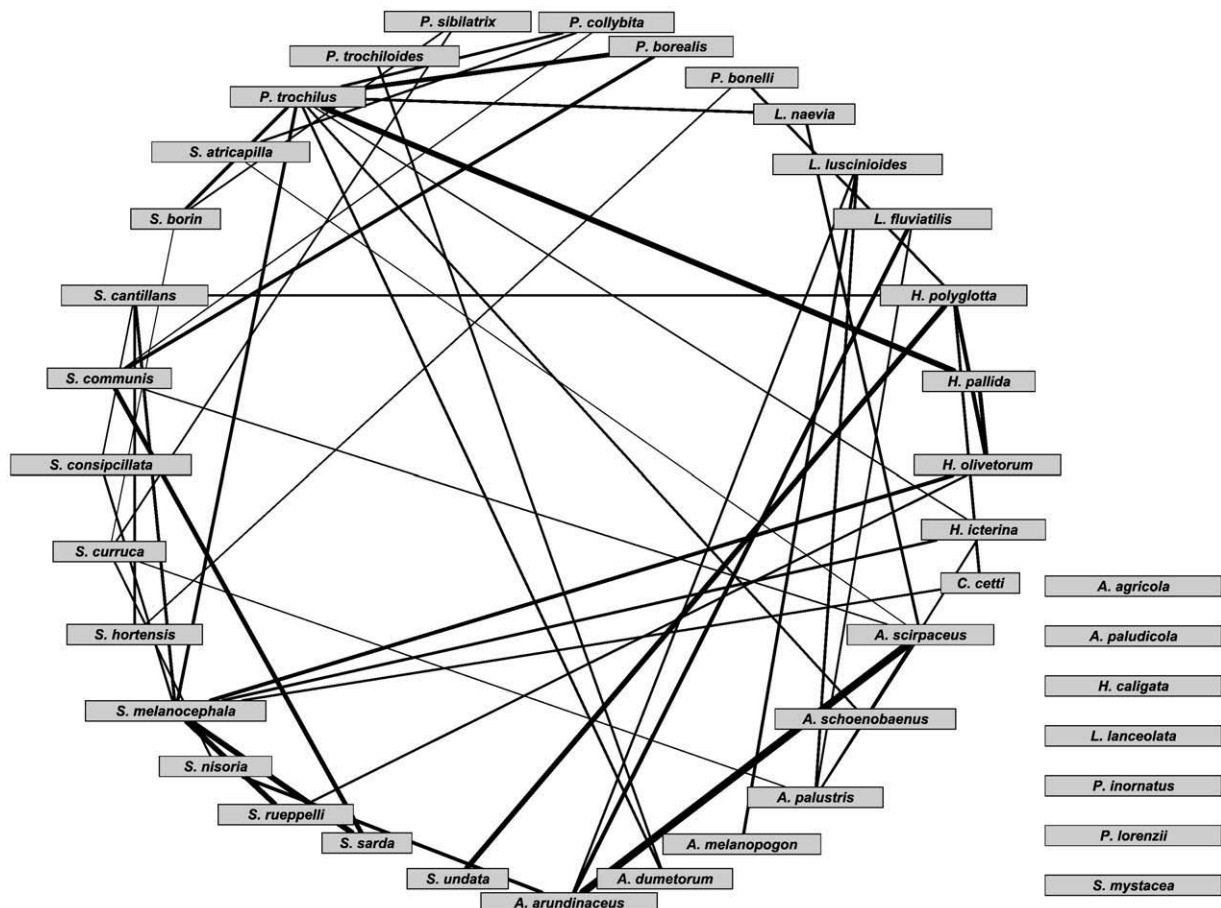


Fig. 4. Consensus network for the warbler data (without spatial autocorrelation or bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the consensus network, which corresponds to a p -value of 0.01. The thickness of an edge represents the strength of the interaction. The boxes on the right show unconnected species.

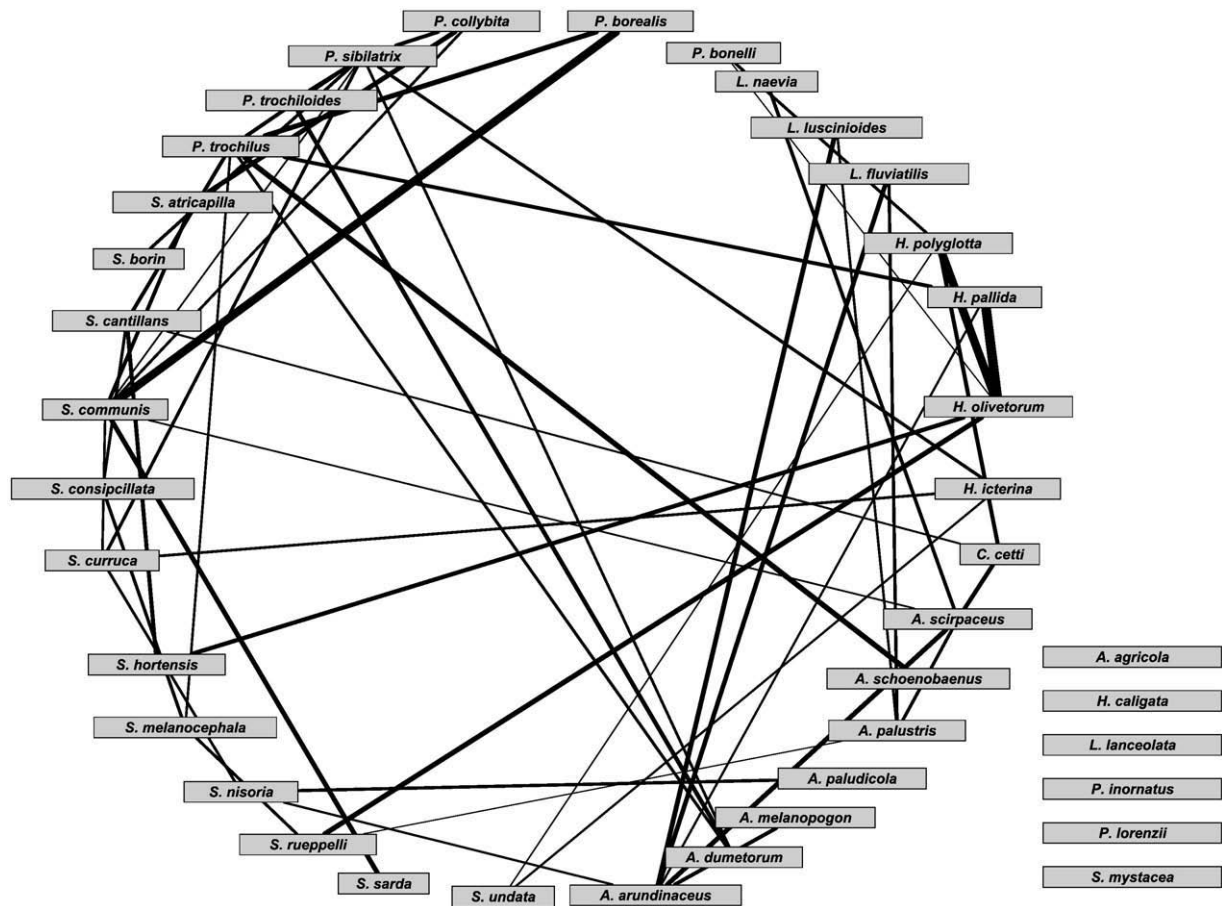


Fig. 5. Consensus network for the warbler data, with spatial autocorrelation included (but without bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p -value of 0.01. The thickness of an edge represents the strength of the interaction. The boxes on the right show unconnected species.

further to include both continuous variables and binary variables in the same analyses. Defining appropriate probability distributions is rather straightforward. However, these distributions depend on parameters, and integrating them out in the likelihood is analytically intractable. To address this difficulty, one can either seek approximate solutions based on variational calculus, or resort to an extended sampling scheme with MCMC. A development of these ideas and a comparative evaluation study provides an interesting and challenging project for future work.

To quantify the network reconstruction accuracy, we have applied various evaluation criteria (described in Section 2.3). We found that the correlations between the interaction scores obtained from the network reconstruction methods and those used for evaluation – phylogenetic distances and ecological similarities – were significant (Table 3). Likewise, the reconstruction assessment scores obtained on the basis of an overall *a priori* network structure elicited from expert judgement – AUC and TFP5 (Fig. 7) – were significantly better than random. We note that the correlations are weak (Table 3) and the AUC and TFP5 scores (Fig. 7) are significantly below the score of a perfect reconstruction (AUC = TFP5 = 1.0). This is over-pessimistic in that the scores are based on evaluation criteria which themselves are noisy and distorted characterisations of the unknown true species interaction network: we found that the correlation coefficients and network reconstruction scores for these criteria are also weak (see Appendix A). This is a general problem when trying to assess the network reconstruction on real data, for which the true interaction network is unknown. The fact that the reconstructed networks show weak yet consistently significant agreement with the various

evaluation criteria indicates that the machine learning methods investigated in our study have reconstructed genuine patterns of the (unknown) species interaction network.

To compensate for the lack of gold standard for the warbler data, we have extended our study by applying the network reconstruction methods to simulated data, for which the underlying network is known. Our results are consistent with related studies in molecular systems biology (Werhli et al., 2006). The global network reconstruction in terms of AUC scores typically lies in the range between 0.75 and 0.9, which is considerably better than random (0.5), but not perfect (1.0). In terms of TFP5 scores, we can expect to reconstruct about 60% of the true species interactions at a false prediction rate of 5%. Aiming for a perfect reconstruction would be an unrealistic target, given the noise in the data, the limited data set size, and the fact that all reconstruction models investigated in our study are simplifications of the complex ecological processes.

Our comparative evaluation of different network reconstruction methods has found that SBR performed significantly worse than the other methods (Fig. 3) and discovered a much smaller proportion of edges than the other methods (illustrated e.g. in Fig. 2). We provide a mathematical explanation in Section 3.1.3. We have also shown that including spatial autocorrelation effects lead to a clear and significant improvement in the network reconstruction accuracy on simulated data (Fig. 3). The evaluation on the warbler data was more difficult due to the lack of a gold standard. In general, more complex models, which included spatial autocorrelations and bio-climate covariates, resulted in stronger matches between the predicted species interactions and the prior network derived from expert

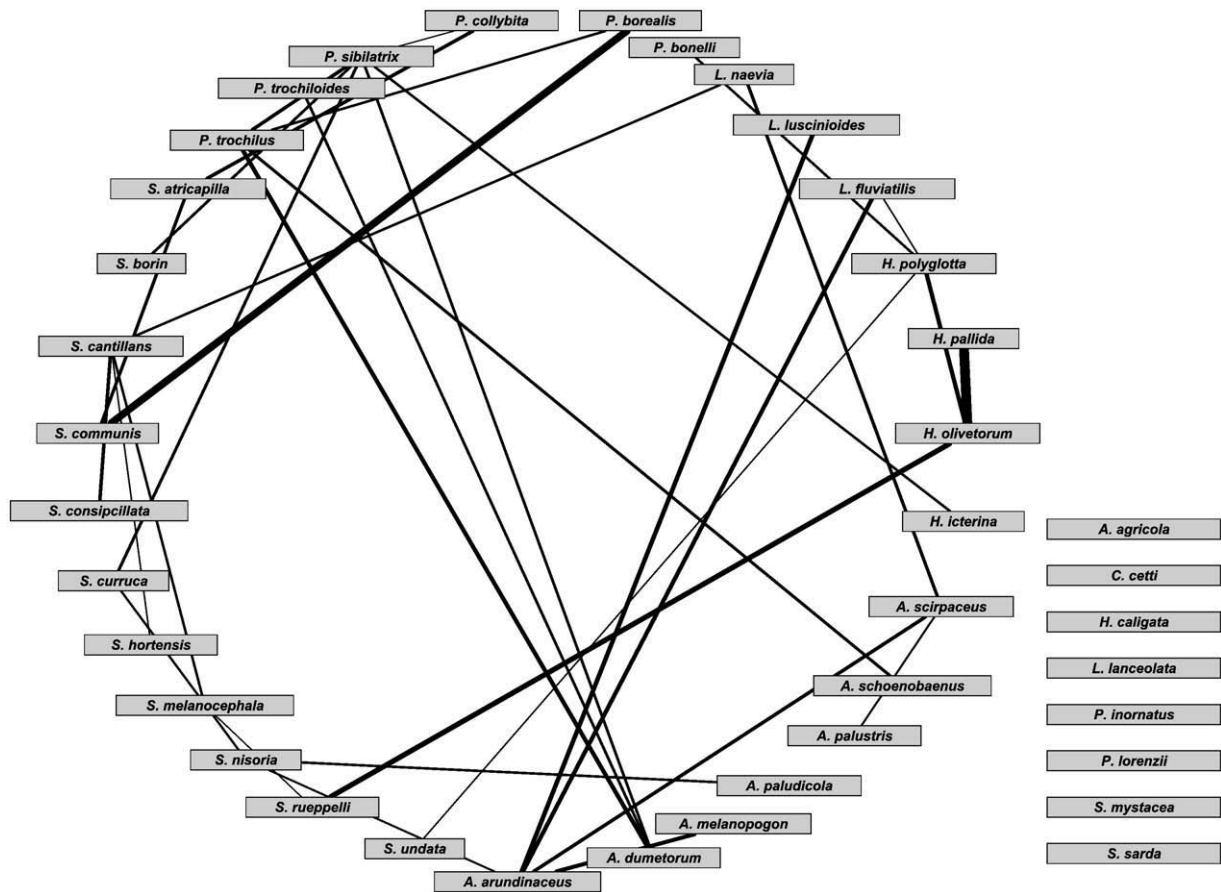


Fig. 6. Consensus network for the warbler data, with spatial autocorrelation and bio-climate covariates. The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p -value of 0.01. The thickness of an edge represents the strength of the interaction. The boxes on the right represent unconnected species.

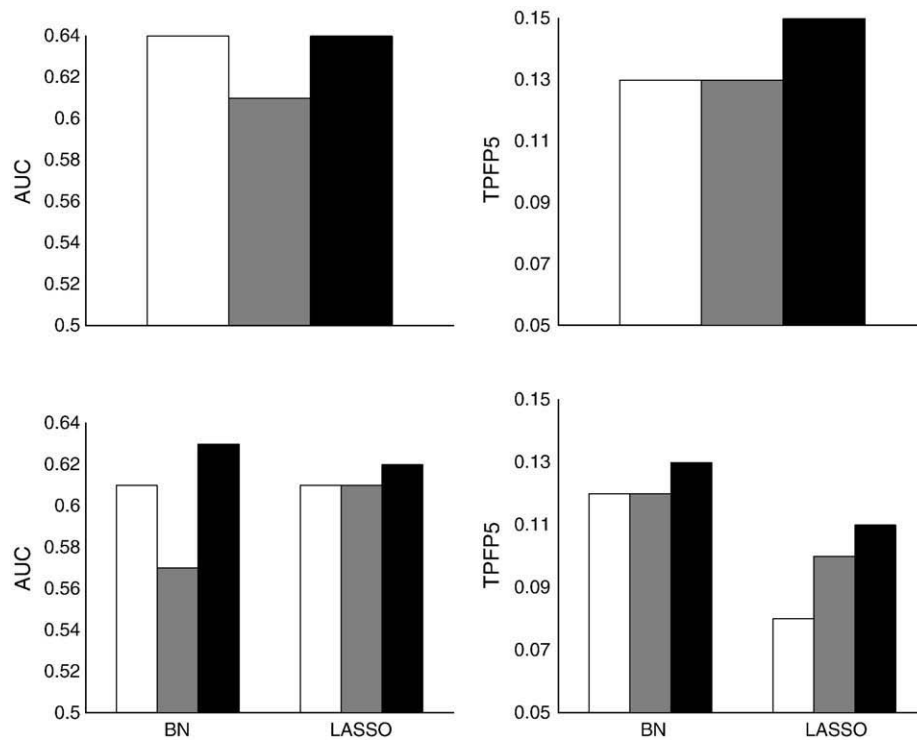


Fig. 7. Comparison of recovered consensus networks with the *a priori* interaction network: AUC scores on the left and TFP5 scores on the right. White bars show the birds only dataset, grey bars the birds and spatial autocorrelation, black bars the birds, spatial autocorrelation and bio-climate covariate dataset. The top row shows the results for consensus networks, while the bottom row shows the results for BN and LASSO individually. Note that the AUC and TFP5 scores tend to increase as the model complexity increases. The vertical position of the horizontal axis indicates the expected performance of a random predictor.

Table 3

Results of comparison between recovered consensus networks with the *a priori* interaction network, phylogenetic distance and ecological distance. For comparisons with the *a priori* network (second column), we show the regression coefficient of a logistic regression. Other results (third and fourth column) are Pearson's correlation coefficients, all with 95% confidence intervals shown in brackets. Confidence intervals that do not include zero indicate that the correlation is significant.

Recovered network	<i>A priori</i> net	Phylogenetic dist.	Ecological dist.
Basic dataset	−0.98 (0.32, −2.28)	−0.11 (−0.18, −0.04)	−0.13 (−0.20, −0.06)
Spatial autocorrelation	−1.40 (−0.03, −3.16)	−0.12 (−0.19, −0.05)	−0.15 (−0.22, −0.08)
Spatial autocorrelation	−1.60	−0.14	−0.14
Bio-climate covariates	(−0.03, −3.16)	(−0.21, −0.07)	(−0.22, −0.07)

judgement (Fig. 7). We also found that the absolute value of the correlations between predicted species interaction strengths and both phylogenetic and ecological distance scores increased as a consequence of including spatial autocorrelations and bio-climate covariates (Table 3). This suggests that accounting for additional sources of variation removed spurious interactions and led to a more plausible network structure.

The reconstructed warbler interaction networks have shown an exponential rather than a power law degree distribution (see Appendix A). This finding is consistent with Dunne et al. (2002) and contributes to the ongoing discussion about the global characteristics of species interaction networks. The networks inferred in our study suggest a number of novel strong interactions that may exist among the warblers. This leads to the formulation of new hypotheses: do *S. curruca* and *S. nisoria* interact, and is the relationship between *H. icterina* and *P. sibilatrix* real? Investigation of the mechanisms behind these interactions may prove valuable.

6. Conclusion

We have carried out one of the first studies to address the problem of reconstructing species interaction networks from species abundance data. To this end, we have applied and adapted four machine learning methods recently developed in the field of computational molecular systems biology. We have applied these models and their adaptations to a subset of the European bird atlas data (warblers), and have discovered both interactions that are known from the literature, and significant correlations with interaction scores based on phylogenetic distances and ecological similarities.

We have complemented our study with an evaluation of the network reconstruction on simulated data, for which a proper gold-standard is known. The reconstruction performance was considerably better than random, but we note that perfect reconstruction is unlikely given limited data and the complexity of the ecological processes involved. The machine learning methods investigated in our study therefore do not provide a mechanism for hypothesis validation. However, our findings suggest that they do offer a useful tool for hypothesis generation, which can enrich and complement traditional methods based on fieldwork and experimental analysis.

The comparative evaluation of different network reconstruction methods has deepened our insight into their relative performance. However, we have found that for a successful application in ecology, the network reconstruction methods currently applied in molecular systems biology need to be modified and improved. We have incorporated a mechanism for taking spatial autocorrelation into account, and we have expanded the models so as to include exogenous bio-climate variables.

Future model improvement should focus on the explicit inclusion of ecological prior knowledge, along the lines of Werhli and Husmeier (2007), and the inclusion of latent variables to allow for unobserved effects. We have investigated the adaptation of the model proposed in Grzegorzczak et al. (2008a) to include latent variables in Bayesian networks. While our preliminary results on the simulated data were encouraging, the application of this scheme to the warbler data suffered from convergence and mixing problems of the MCMC simulations, which calls for further methodological improvements.

The true value of our study lies in demonstrating that even using large-scale spatial datasets, relevant patterns in ecological networks can be identified using the machine learning methods described here. This suggests that these methods have the potential to contribute novel important tools for gaining deeper insight into the structure and stability of ecosystems, managing biodiversity, and predicting the impact of climate change.

Acknowledgements

This work was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). We thank Jonathan Yearsley for providing the initial community simulation model and Marco Grzegorzczak for providing the Bayesian Network code and being available for questions. We would also like to thank Prof. Samuel Kaski for supporting Ali Faisal on this project.

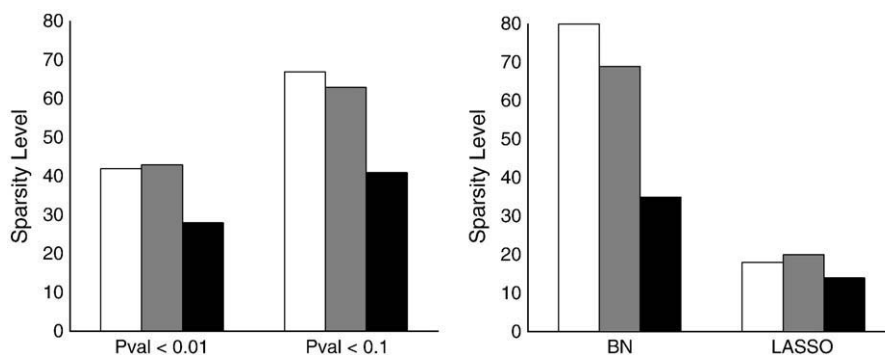


Fig. 8. Sparsity of the recovered networks. White bars show the birds only dataset, grey bars the birds and spatial autocorrelation, black bars the birds, spatial autocorrelation and bio-climate covariate dataset. The left figure shows the results for consensus networks at two different thresholds, while the right figure shows the results for BN and LASSO individually at a threshold with *p*-value < 0.01.

Appendix A. Available supplementary material

We have made supplementary information available at <http://www.bioss.ac.uk/students/frankd.html>. This document contains additional results that were left out of the main paper due to space constraints. The material contained in the supplementary information includes:

- Description of a latent variable model for capturing unobserved environmental effects in the data (supplementary Section S2.1), and preliminary results (supplementary Section S4.3).
- Description of an observation process for discretising the simulated datasets (supplementary Section S2.2) as well as network reconstruction performance results (supplementary Section S4.1).
- Experimental results showing a strong correlation between interaction probabilities and interaction strengths, motivating our choice to use consensus networks of the two measures (supplementary Section S3).
- Simulation results showing that consensus networks outperform individual networks (supplementary Section S4.2).
- Description and detailed results of significance tests for comparisons of network reconstruction performance between the different methods (supplementary Section S4.4).
- A comparison between using 4 vs 8 neighbours in the spatial autocorrelation model, showing that 4 neighbours are sufficient (supplementary Section S4.5).
- Description of the *a priori* warbler interactions obtained from the literature and using expert judgement (supplementary Section S5.1).
- Details of the phylogeny used to calculate the phylogenetic distance between bird species (supplementary Section S5.2).
- Details of the information used for obtaining the ecological distance measure for bird species (supplementary Section S5.3).
- Comparison of the ecological measures used to evaluate the reconstructed Warbler networks (supplementary Section S5.4).
- Description of how the *p*-values used to determine the threshold on interaction strengths were obtained (supplementary Section S5.5).
- Detailed characterisation of the warbler networks we recovered in terms of degree distribution, clustering coefficients and network diameter (supplementary Section S5.7).

Appendix B. *A priori* network, phylogenetic distance and ecological distance measure

In this appendix, we describe the *a priori* network from the literature, as well as the phylogenetic and ecological distance measures that we used to evaluate the Warbler bird networks recovered in Figs. 4–6.

***A priori* Network.** To construct the *a priori* network, we used two sources: knowledge from the literature, and expert judgement.

First, we searched the ecological literature using ISI Web of Knowledge⁴ (accessed on 10/5/09). For each species, we searched for all articles using the complete scientific name. If more than 100 articles were returned, we refined the search adding the terms ‘interaction’ or ‘competition’. We studied all abstracts and identified papers containing information about interspecific interactions for detailed reading. We identified 30 interactions using this method.

For the remaining 711 pairwise interactions we used our expert judgement to answer the question: In areas where these species occur in close proximity, is it plausible that one of the species would become more abundant or expand into different habitats if the other species were absent? In cases where we considered this likely we recorded an interaction in the network.

The final network can be found at <http://www.bioss.ac.uk/students/frankd.html>.

Phylogenetic Distance. To calculate the phylogenetic distances between warbler species, we first needed to get general information on warbler phylogeny. To that end, we searched the taxonomic literature (e.g. Alstroem et al., 2006) and ‘Tree of Life’ servers (such as The Tree of Life Web Project in Maddison et al., 2007). A conservative consensus tree was generated depicting relationships between the 39 warbler species.

As path lengths were unavailable we computed a range of distances using the method advocated by Grafen (1989) with values of ρ of 1, 0.6 and 0.3. Although correlations between the phylogenetic distance and recovered interaction scores were not qualitatively different when these different distances were assumed, they are arbitrary choices nonetheless. Consequently, we repeated the correlation analysis using Kendall’s τ as a measure of rank correlation that is unaffected by assumed branch lengths. Again, results were qualitatively similar. For the correlation analyses we used only data from the upper triangle of the distance matrices.

Ecological Distance. Ecological trait data for the length, mass, wing span, clutch size, migrant status and preferred habitat of each of the 39 species was collected. From the habitat and migration status data we generated indicator variables identifying species with shared habitat and shared migration strategy. We combined these indicator variables with the morphological data and clutch size, centred and scaled each variable and calculated the Euclidian distance. As with the phylogenetic distance analysis, we used only data from the upper triangle of the distance matrix in correlation analyses.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.ecoinf.2010.06.005](https://doi.org/10.1016/j.ecoinf.2010.06.005).

References

- Albert, R., Barabási, A.L., 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74 (1), 47–97.
- Alstroem, P., Ericson, P., Olsson, U., Sundberg, P., 2006. Phylogeny and classification of the avian superfamily *Sylvioidae*. *Mol. Phylogenet. Evol.* 38 (2), 381–397.
- Araujo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species–climate impact models under climate change. *Global Change Biology* 11 (9), 1504–1513.
- Augustin, N.H., Muggleston, M.A., Buckland, S.T., 1996. An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.* 33 (2), 339–347.
- Barabási, A., Albert, R., Schiffer, P., 1999. The physics of sand castles: maximum angle of stability in wet and dry granular media. *Physica A* 266 (1–4), 366–371.
- Batjes, N.H., 1996. Development of a world data set of soil water retention properties using pedotransfer rules. *Geoderma* 71 (1–2), 31–52.
- Beale, C.M., Lennon, J.J., Gimona, A., 2008. Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proc. Natl Acad. Sci.* 105 (39), 14908–14912.
- Beisner, B.E., Haydon, D.T., Cuddington, K., 2003. Alternative stable states in ecology. *Front. Ecol. Environ.* 1 (7), 376–382.
- Blüthgen, N., Menzel, F., Blüthgen, N., 2006. Measuring specialization in species interaction networks. *BMC Ecol.* 6 (1), 9.
- Chickering, D.M., 1995. A transformational characterization of equivalent Bayesian network structures. *Int. Conf. Uncertainty Artif. Intell.* 11, 87–98 (UAI).
- Cohen, J.E., Schoenly, K., Heong, K.L., Justo, H., dArida, G., Barrión, A.T., Litsinger, J., 1994. A food-web approach to evaluating the effect of insecticide spraying on insect pest population-dynamics in a Philippine irrigated rice ecosystem. *J. Appl. Ecol.* 31, 747–763.
- Dale, M.R.T., Fortin, M.J., 2002. Spatial autocorrelation and statistical tests in ecology. *Ecoscience* 9 (2), 162–167.
- Davis, A.J., Jenkinson, L.S., Lawton, J.H., Shorrocks, B., Wood, S., 1998. Making mistakes when predicting shifts in species range in response to global warming. *Nature* 391 (6669), 783–785.
- de Silva, E., Stumpf, M.P.H., 2005. Complex networks and simple models in biology. *J. R. Soc. Interface* 2 (5), 419.
- Dondelinger, F., 2008. Inferring ecological networks from species abundance data: evaluation on simulated data. School of Informatics, University of Edinburgh, Master’s thesis.
- Dunne, J., Williams, R., Martinez, N., 2002. Food-web structure and network theory: the role of connectance and size. *Proc. Natl Acad. Sci.* 99 (20), 12917–12922.
- Edwards, D.M., 2000. Introduction to Graphical Modelling. Springer Verlag, New York.

⁴ Found at <http://www.isiwebofknowledge.com/>.

- Elle, O., 2003. Quantification of the integrative effect of ecotones as exemplified by the habitat choice of Blackcap and Whitethroat (*Sylvia atricapilla* and *S. communis*, Sylviidae). *J. Ornithol.* 144 (3), 271–283.
- Faisal, A., 2008. Inferring ecological networks from species abundance data: application to the European bird atlas data. School of Informatics, University of Edinburgh, Master's thesis.
- Friedman, N., Koller, D., 2003. Being Bayesian about network structure. *Mach. Learn.* 50, 95–126.
- Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. *J. Comp. Biol.* 7, 601–620.
- Garcia, E., 1983. An experimental test of competition for space between blackcaps *Sylvia atricapilla* and garden warblers *Sylvia borin* in the breeding season. *J. Anim. Ecol.* 52 (3), 795–805.
- Gaston, K.J., 2003. The structure and dynamics of geographic ranges. Oxford University Press.
- Geiger, D., Heckerman, D., 1994. Learning Gaussian networks. Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, In. pp. 235–243.
- Grafen, A., 1989. The phylogenetic regression. *Philos. Trans. R. Soc. B* 326, 119–157.
- Grandvalet, Y., Canu, S., 1999. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. *Adv. Neural Inf. Process. Syst.* 11, 445–451.
- Grzegorzczak, M., Husmeier, D., 2008. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Mach. Learn.* 71 (2–3), 265–305.
- Grzegorzczak, M., Husmeier, D., Edwards, K., Ghazal, P., Millar, A., 2008a. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics* 24 (18), 2071–2078.
- Grzegorzczak, M., Husmeier, D., Werhli, A., 2008b. Reverse engineering gene regulatory networks with various machine learning methods. In: Emmert-Streib, F., Dehmer, M. (Eds.), *Analysis of Microarray Data: A Network-Based Approach*. Wiley-VCH, Weinheim, pp. 101–142.
- Hagemeyer, 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.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 6, 422–433.
- Heckerman, D., 1999. A tutorial on learning with Bayesian networks. In: Jordan, M.I. (Ed.), *Learning in Graphical Models*. MIT Press, Cambridge, Massachusetts, pp. 301–354.
- Heckerman, D., Geiger, D., Chickering, D.M., 1995. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* 20, 245–274.
- Henneman, M.L., Memmott, J., 2001. Infiltration of a Hawaiian community by introduced biological control agents. *Science* 293 (5533), 1314–1316.
- Holt, R.D., Barfield, M., 2009. Trophic interactions and range limits: the diverse roles of predation. *P. Roy. Soc. B* 276 (1661), 1435–1442.
- Huntley, B., Collingham, Y.C., Willis, S.G., Green, R.E., 2008. Potential impacts of climatic change on European breeding birds. *PLoS ONE* 3 (1).
- Husmeier, D., Dybowski, R., Roberts, S., 2005. Probabilistic modeling in bioinformatics and medical informatics. *Advanced Information and Knowledge Processing*, Springer, New York.
- Ings, T.C., Montoya, J.M., Bascompte, J., Bluthgen, N., Brown, L., Dormann, C.F., Edwards, F., Figueroa, D., Jacob, U., Jones, J.L., Lauridsen, R.B., Ledger, M.E., Lewis, H.M., Olesen, J.M., van Veen, F.J.F., Warren, P.H., Woodward, G., 2009. Review: Ecological networks beyond food webs. *J. Anim. Ecol.* 78, 253–269.
- La Sorte, F.A., Lee, T.M., Wilman, H., Jetz, W., 2009. Disparities between observed and predicted impacts of climate change on winter bird assemblages. *Proc. R. Soc. B* 276 (1670), 3167.
- Lande, R., Engen, S., Saether, B., 2003. Stochastic population dynamics in ecology and conservation. Oxford University Press.
- Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74 (6), 1659–1673.
- Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. *Ecography* 23, 101–113.
- Losos, J.B., 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol. Lett.* 11 (10), 995–1003.
- Luce, R., Perry, A., 1949. A method of matrix analysis of group structure. *Psychometrika* 14 (2), 95–116.
- MacKay, D.J.C., 1992. Bayesian interpolation. *Neural Comput.* 4, 415–447.
- MacKay, D.J.C., 1996. Hyperparameters: optimize, or integrate out. In: Heidbreder, G. (Ed.), *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publisher, Santa Barbara, pp. 43–59.
- Maddison, D., Schulz, K., Maddison, W., 2007. The tree of life web project. *Zootaxa* 1668, 19–40.
- Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. *Int. Stat. Rev.* 63, 215–232.
- Memmott, J., 1999. The structure of a plant–pollinator food web. *Ecol. Lett.* 2 (5), 276–280.
- 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.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. *Science* 298 (5594), 824.
- Murray Jr, B.G., 1988. Interspecific territoriality in *Acrocephalus*: a critical review. *Ornis Scand.* 19 (4), 309–313.
- Needham, C., Bradford, J., Bulpitt, A., Westhead, D., 2007. A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.* 3 (8), 1409–1416.
- New, M., Hulme, M., Jones, P., 1999. Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate* 12 (3), 829–856.
- Opgen-Rhein, R., Strimmer, K., 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* 1 (37).
- Park, T., Casella, G., 2008. The Bayesian Lasso. *J. Am. Stat. Assoc.* 103 (482), 681–686.
- Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeogr.* 19 (2), 117–134.
- Proulx, S., Promislow, D., Phillips, P., 2005. Network thinking in ecology and evolution. *Trends Ecol. Evol.* 20 (6), 345–353.
- Rogers, S., Girolami, M., 2005. A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics* 21 (14), 3131–3137.
- Rolando, A., Palestini, C., 1991. The effect of interspecific aggression on territorial dynamics in *Acrocephalus* warblers in a marsh area of north-western Italy. *Bird Study* 38 (2), 92–97.
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P., 2005. Protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529.
- Schäfer, J., Strimmer, K., 2005a. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21 (6), 754–764.
- Schäfer, J., Strimmer, K., 2005b. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4 (1) Article 32.
- Schäfer, T., Ledebur, G., Beier, J., Leisler, B., 2006. Reproductive responses of two related coexisting songbird species to environmental changes: global warming, competition, and population sizes. *J. Ornithol.* 147 (1), 47–56.
- Schmitz, O.J., Krivan, V., Ovadia, O., 2004. Trophic cascades: the primacy of trait-mediated indirect interactions. *Ecol. Lett.* 7 (2), 153–163.
- Sinclair, A.R.E., Byrom, A.E., 2006. Understanding ecosystem dynamics for conservation of biota. *J. Anim. Ecol.* 75, 64–79.
- 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.
- Snow, D.W., Perrins, C.M., 1998. The Birds of the Western Palearctic Concise Edition. Oxford University Press, Oxford.
- Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L., Williams, S.E., 2004. Extinction risk from climate change. *Nature* 427 (6970), 145–148.
- Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T., Prentice, I.C., 2005. Climate change threats to plant diversity in Europe. *Proc. Nat. Acad. Sci.* 102 (23), 8245.
- Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* 58 (1), 267–288.
- Tipping, M., 2001. Sparse Bayesian learning and the relevance vector machine. *JMLR* 1 (3), 211–244.
- Tipping, M., Faul, A., 2003. Fast marginal likelihood maximisation for sparse Bayesian models. In: Bishop, C.M., Frey, B.J. (Eds.), *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, vol.9.
- Tirelli, T., Pozzi, L., Pessani, D., 2009. Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy). *Ecol. Inf.* 4 (4), 234–242.
- Tong, A.H.Y., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C.W.V., Fields, S., Boone, C., Cesareni, G., 2002. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295 (5553), 321–324.
- Valiente, G., 2002. Algorithms on trees and graphs. Springer Verlag.
- van Someren, E.P., Vaes, B.L.T., Steegenga, W.T., Sijbers, A.M., Dechering, K.J., Reinders, M.J.T., 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.
- Vázquez, D.P., Simberloff, D., Jun, 2002. Ecological specialization and susceptibility to disturbance: conjectures and refutations. *Am. Nat.* 159 (6), 606–623.
- Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. *Nature* 393 (6684), 440–442.
- Watts, M., Worner, S., 2008. Comparing ensemble and cascaded neural networks that combine biotic and abiotic variables to predict insect species distribution. *Ecol. Inf.* 3 (6), 354–366.
- Werhli, A., Husmeier, D., 2007. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* 6 (1) Article 15.
- Werhli, A.V., Grzegorzczak, M., Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* 22, 2523–2531.
- Werner, E.E., Peacor, S.D., 2003. A review of trait-mediated indirect interactions in ecological communities. *Ecology* 84 (5), 1083–1100.
- Williams, P.M., 1995. Bayesian regularization and pruning using a Laplace prior. *Neural Comput.* 7, 117–143.
- Williams, R., Martinez, N., 2000. Simple rules yield complex food webs. *Nature* 404 (6774), 180–183.