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]

[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*

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 typi-

[*] Corresponding author. Tel: +44 (0)131 650 7536, Fax: +44 (0)131 650 4901

Email addresses: ali.faisal@tkk.fi (Ali Faisal), frankd@bioss.ac.uk (Frank Dondelinger), dirk@bioss.ac.uk (Dirk Husmeier), c.beale@macaulay.ac.uk (Colin M. Beale)

cally 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 operator characteristics (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.

Key words: network reconstruction, warbler interactions, spatial autocorrelation, bio-climate variables

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 and Memmott, 2001). Altering pressures to which ecosystems are exposed can drive them to alternative states (Beisner et al., 2003) or catastrophic failure (Sinclair and 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 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 mutuality relationships are unimportant when these have recently been identified as perhaps overwhelming (Werner and 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 and Barabási, 2002), clustering coefficient (Watts and 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 and 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 networks structure 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 Hagemeijer and Blair (1997) (Hagemeijer, 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. Material and Methods

2.1. 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 and 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 $[\frac{r_i}{2}, n_i]$. 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:

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

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 omega 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.

2.2. 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 and Barfield, 2009a), 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., 2008a). 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 (Hagemeijer and 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 Jr, 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., 2008a). 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 WISE.AWC, Batjes 1996) using a bucket model (following Prentice et al. 1992) and interpolated to 50km 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. 2008b; 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., 2008b; Huntley et al., 2008). Other biologically relevant climate variables could also be used but are usually strongly correlated with one or other of these and have little affect on the strength of associations realised (Beale et al., 2008b). 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 (supplementary Section S2.1) 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 (we provide this network and relevant literature in supplementary Section S5.1). 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 (details of the phylogeny used are provided in supplementary Section S5.2). 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 and Perrins (1998)) and summarised these variables to generate a measure of ecological distance (see supplementary Section S5.3 for 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 are themselves linked to the focal species. We measured the clustering coefficient: Luce and 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.

2.3. Units

Table **??** 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

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

Symbol/Quantity	Equation	Unit
$\mathbf{x}_r$, x_i, X_i	2, 16, 19, 20	Discrete presence/absence value of species over 50 km^2 area
$\hat{\mathbf{y}}_g$, $\mathbf{y}_g$	2, 4, 7, 20	Discrete presence/absence value of species over 50 km^2 area
$\hat{\mathbf{w}}_g$, $\mathbf{w}_g$, $\mathbf{w}$, v	2, 4, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20	Dimensionless weight parameters
a	19, 20	Spatial Autocorrelation: Discrete presence/absence value of species from averaging over 4 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

Table 1: Units for the different quantities in the paper, along with the equations where they are used (if any). Note that we have only given units for the discrete bird atlas data (see section 2.2). When applied to the continuous simulation data, presence/absence values are replaced by population densities, and the area of each location is assumed to be the same and to have no impact on the population densities.

Gaussian Models (Schäfer and Strimmer, 2005a,b), LASSO regression (Tibshirani, 1996; van Someren et al., 2006), sparse Bayesian regression (Tipping and Faul, 2003; Rogers and Girolami, 2005) and Bayesian Networks (Friedman et al., 2000; Werhli and 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 a description of the important features of the methods we trial, 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 supplementary Section S2.3).

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 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):

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

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 (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 and 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 $\{\mathbf{x}_r\}$ are good predictors of abundance profile $\mathbf{y}_g$:

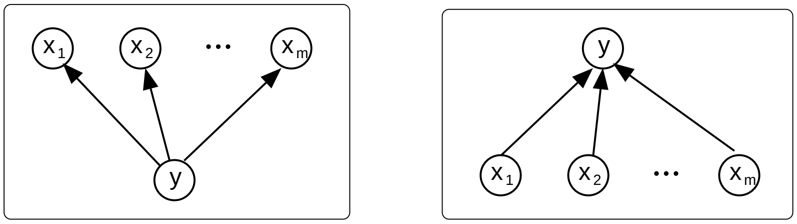


Figure 1: Schematic of the approach of partial correlation (left) and sparse regression (right). Left: Conditional on y, the species abundance profiles $x_1, x_2, \ldots, x_m$ are independent, and the partial correlation coefficients will be small. Right: The approach of sparse regression aims to find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

$$\hat{\mathbf{y}}_g \;=\; \sum_r w_{gr}\mathbf{x}_r \tag{3}$$

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

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

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

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:

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

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 $\sum_r w_{gr}^2 < C$, 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 $\lambda^{-1}\mathbf{I}$: $P(\mathbf{w}_g) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$.

A disadvantage of ridge regression is that the set of interaction parameters $\{w_{gr}\}$ 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 $w_{gr} \to 0$. 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:

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

which can be interpreted as a Bayesian *maximum a posteriori* estimate under a Laplacian prior on $\mathbf{w}_g$, 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 $w_{gr} \to 0$ 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:

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

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 equation (4) just needs to be replaced by:

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

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 equation (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 equations (5) and (7) are equivalent to the *maximum a posteriori* estimates:

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

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:

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

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:

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

with separate hyperparameters for species r. This scheme is similar to equation (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[1]. 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 equation 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:

[1] In statistics this is called a type-II maximum likelihood estimation.

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

Which for an individual weight gives:

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

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:

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

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:

$$\text{SBR: } \frac{d}{dw_{gr}}\left\{-logP(w_{gr})\right\} \propto \frac{1}{w_{gr}} \tag{15}$$

$$\text{LASSO: } \frac{d}{dw_{gr}}\left\{-logP(w_{gr})\right\} \propto const \tag{16}$$

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 $X_1, X_2, ..., X_n$ be a set of random variables represented by the nodes $i \in \{1, ..., n\}$ 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

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

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

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

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

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

If certain regulatory conditions, discussed in Heckerman (1999), are satisfied and the data are complete, then the integral in (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 and 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 (18) is analytically intractable and is therefore approximated with Markov Chain Monte Carlo (MCMC) (Madigan and York, 1995; Friedman and Koller, 2003; Grzegorczyk and 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), incorporates 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 Grzegorczyk 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 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. Where possible, we applied an autoregressive approach similar to that of Augustin 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:

$$a = \frac{\sum_{i=1}^{N} \omega_i x_i}{\sum_{i=1}^{N} \omega_i} \tag{20}$$

where N is the number of neighbours that we're considering (usually $N = 4$), 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:

$$\hat{\mathbf{y}}_g \; = \; \sum_r w_{gr}\mathbf{x}_r + va \tag{21}$$

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 (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 by 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 (supplementary Section S4.2). 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. (For a comparison with other combination methods, e.g. based on the harmonic mean, see supplementary Section S4.2.) 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 ($\rho = 0.92$), as discussed in more detail in supplementary Section S3.

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 operator characteristics (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 operator characteristics 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 TPFP5 score). A good network reconstruction method is characterised by both a high AUC score and a high TPFP5 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 get the MATLAB code for our own implementations of the extensions in Section 3.2.

Method	**Software**	**Package**	**Description**
GGM	R	GeneNet	The software implementing Graphical Gaussian models is described in Schäfer and 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 Genlab 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 and 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 Grzegorczyk 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`

Table 2: Network reconstruction software used

4. Results

4.1. Simulation results

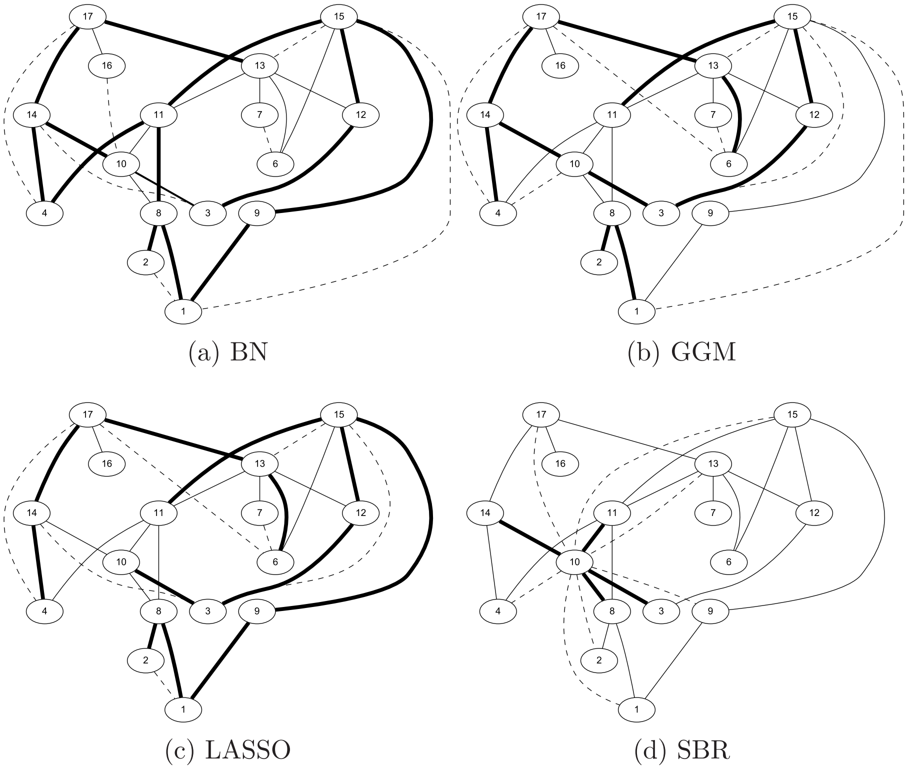
All four network recovery methods succeeded in recovering some of the true network structure, even when spatial autocorrelation was not incorporated (Fig. 2), though the methods varied in their performance (Fig. 3). Sparse Bayesian regression recovered networks that were significantly worse than those recovered by the other methods, having significantly lower AUC and TPFP5 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 $p = 0.05$. Supplementary tables S2-S4 give a full overview of all p-values. 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 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 (described in supplementary Section S2.2) gave qualitatively similar results (beyond an expected drop in AUC and TPFP5 scores). Consequently, we do not report this analysis further here; full details are presented in supplementary Section S4.1.

4.2. Application results

We recovered three consensus networks for the warbler data: for data sets with birds only, with birds and spatial autocorrelation and with birds, spatial autocorrelation and bio-climate covariates. The first two can be found in the supplementary material (Supplementary Figs. S9 and S10); here we just present the third (Fig. 4). Comparison of the recovered consensus networks



(a) BN (b) GGM

(c) LASSO (d) SBR

Figure 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.

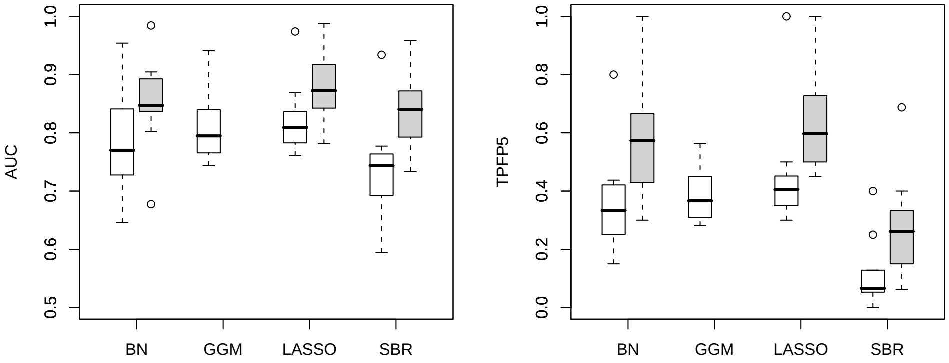


Figure 3: AUC and TPFP5 performance measures for continuous simulation data. Shaded boxes represent models which include spatial autocorrelation. The expected random performance scores are AUC=0.5 and TPFP5=0.05. See Section 3.1 for an explanation of the abbreviations BN, GGM, LASSO, and SBR.

with the *a priori* network predicted from the literature and expert judgement revealed small but statistically significant relationships (Fig. 5). 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. 6). Our predictions in Section 2.2 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. Further details on the network characterisation can be found in supplementary Section S5.7.

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:

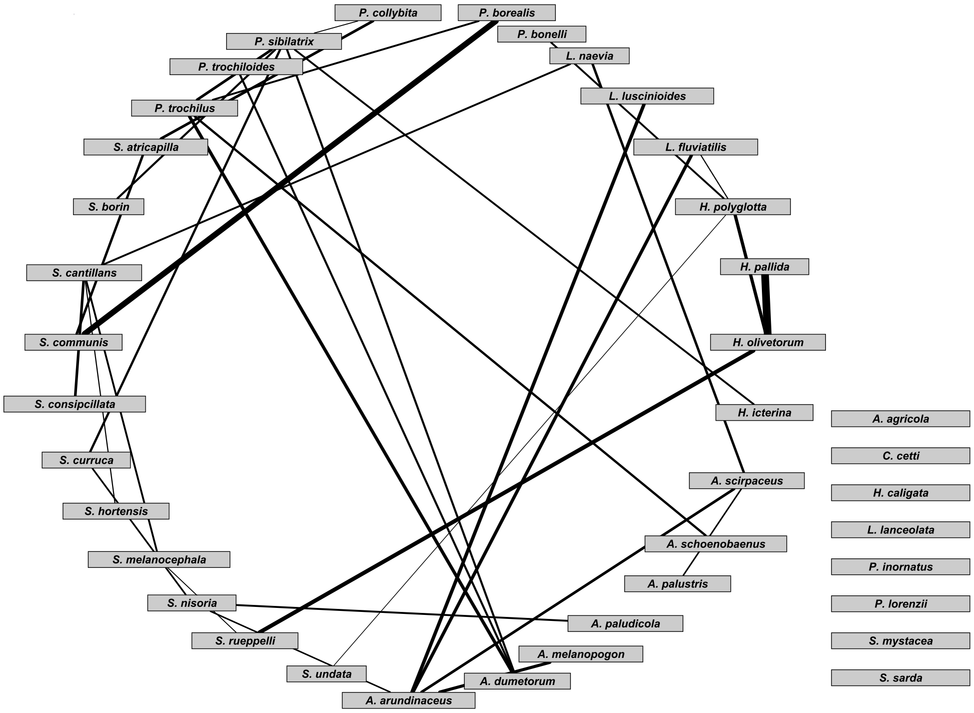


Figure 4: An example consensus network for the warbler data, with spatial autocorrelation and bio-climate covariates. The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p-value of 0.01. See supplementary Section S5.5 for a description of how these p-values were calculated. The thickness of an edge represents the strength of the interaction. The boxes on the right represent unconnected species. Equivalent plots of consensus networks for the other datasets are also available (Supplementary Figs. S9 and S10).

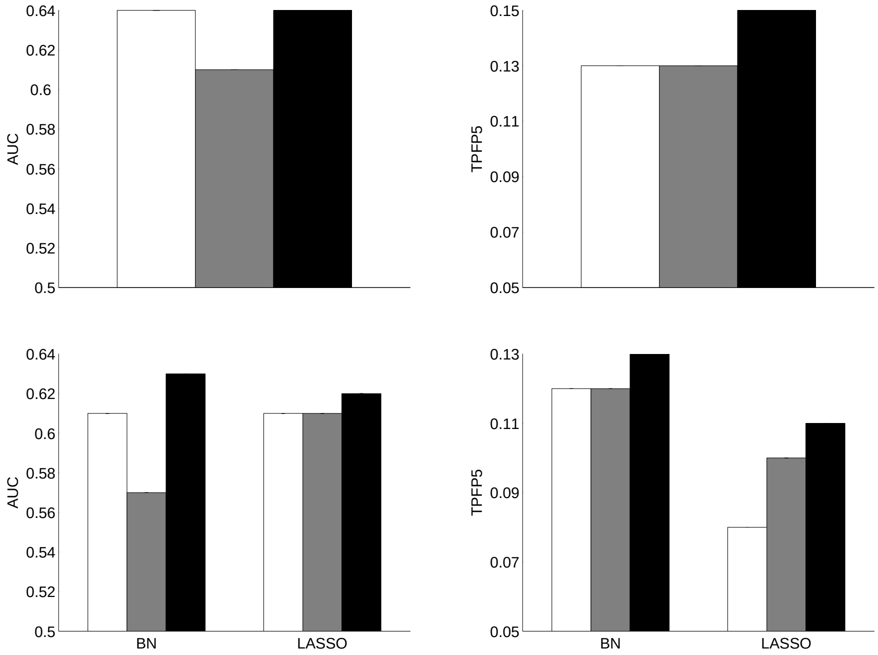


Figure 5: Comparison of recovered consensus networks with the *a priori* interaction network: AUC scores on the left and TPFP5 scores on the right. White bars show the birds only dataset, grey bars the birds and spatial autocorrelation, black bars the birds, spatial autocorrelation and bio-climate covariate 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 TPFP5 scores tend to increase as the model complexity increases. The vertical position of the horizontal axis indicates the expected performance of a random predictor.

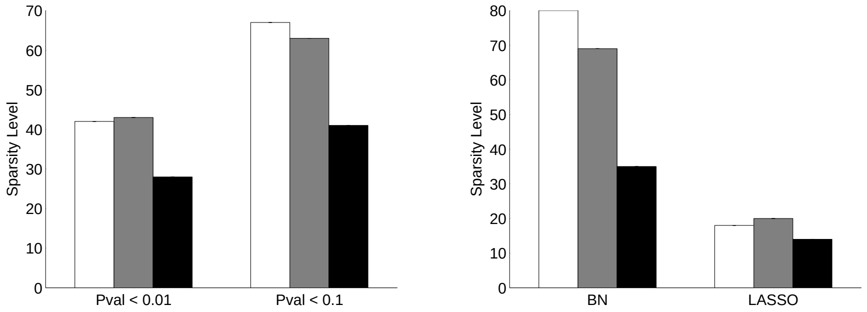


Figure 6: Sparsity of the recovered networks. White bars show the birds only dataset, grey bars the birds and spatial autocorrelation, black bars the birds, spatial autocorrelation and bio-climate covariate dataset. The left figure shows the results for consensus networks at two different thresholds, while the right figure shows the results for BN and LASSO individually at a threshold with p-value < 0.01.

Recovered Network	*A priori* 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 and Bio-climate Covariates	-1.60 (-0.03, -3.16)	-0.14 (-0.21, -0.07)	-0.14 (-0.22, -0.07)

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.

A. arundianceus/A. melanopogon /A. schoenobaenus/A. scirpaceus (Schäfer et al., 2006; Rolando and Palestrini, 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 and Worner 2008) largely because processes like the biotic interactions included in our models have previously been neglected (Davis et al., 1998; Beale et al., 2008b; Holt and Barfield, 2009b; La Sorte et al., 2009). It would clearly be valuable to develop the methods 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 2.2). 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 TPFP5 (Fig. 5) – were significantly better than random. We note that the correlations are weak (Table 3) and the AUC and TPFP5 scores (Fig. 5) are significantly below the score of a perfect reconstruction (AUC = TPFP5 = 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: Supplementary tables S7 and S8 demonstrate that the correlation coefficients and network reconstruction scores for these criteria are also weak. 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 TPFP5 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 including spatial autocorrelation effects leads 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 judgement (Fig. 5). 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 (Supplementary Figs. S12 and S13). 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. currucca* 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 autocorrelations 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 (see supplementary Sections S2.1 and S4.3 for a preliminary investigation). We have investigated the adaptation of the model proposed in Grzegorczyk et al. (2008a) to include latent variables in Bayesian networks. While our preliminary results on the simulated data were encouraging, as shown in supplementary Figure S6, 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 Grzegorczyk 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.

References

Albert, R., Barabási, A. L., 2002. Statistical mechanics of complex networks. Reviews of Modern Physics 74 (1), 47–97.

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., Mugglestone, 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., 2008a. Opening the climate envelope reveals no macroscale associations with climate in European birds. Proc. Natl. Acad. Sci. 105 (39), 14908–14912.

Beale, C. M., Lennon, J. J., Gimona, A., 2008b. Opening the climate envelope reveals no macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences 105 (39), 14908.

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 ecology 6 (1), 9.

Chickering, D. M., 1995. A transformational characterization of equivalent Bayesian network structures. International Conference on Uncertainty in Artificial Intelligence (UAI) 11, 87–98.

Cohen, J. E., Schoenly, K., Heong, K. L., Justo, H., dArida, G., Barrion, 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. Journal of the Royal Society Interface 2 (5), 419.

Dondelinger, F., 2008. Inferring ecological networks from species abundance data: evaluation on simulated data. Master's thesis, School of Informatics, University of Edinburgh.

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.

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. Master's thesis, School of Informatics, University of Edinburgh.

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. In: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA., pp. 235–243.

Grandvalet, Y., Canu, S., 1999. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. Adv. Neural. Inf. Process. Syst. 11, 445–451.

Grzegorczyk, 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.

Grzegorczyk, 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.

Grzegorczyk, 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.

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.

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. Machine Learning 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., 2009a. Trophic interactions and range limits: the diverse roles of predation. P. Roy. Soc. B 276 (1661), 1435–1442.

Holt, R. D., Barfield, M., 2009b. Trophic interactions and range limits: the diverse roles of predation. In: Proc. R. Soc. B. Vol. 276. pp. 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. I., 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. Proceedings of the Royal Society 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.

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. Ecology Letters 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. Journal of Biogeography 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., Palestrini, 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., Williams, S. E., 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., 2004. Biodiversity conservation: uncertainty in predictions of extinction risk/Effects of changes in climate and land use/Climate change and extinction risk (reply). Nature 430 (6995).

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., Prentice, I. C., 2005. Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences 102 (23), 8245.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. 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). Ecological Informatics 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. The American Naturalist 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. Ecological Informatics 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., Grzegorczyk, 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.

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

Supplementary Material

Ali Faisal[a], Frank Dondelinger[b,c], Dirk Husmeier[b], Colin M. Beale[d]

[a] *Helsinki Institute for Information Technology, Department of Information and Computer Science, Helsinki University of Technology, Finland*

[b] *Biomathematics and Statistics Scotland, Edinburgh, United Kingdom*

[c] *Institute for Adaptive Neural Computation, School of Informatics, University of Edinburgh, United Kingdom*

[d] *Macaulay Land Use Research Institute, Aberdeen, United Kingdom*

1. Introduction

This document contains the supplementary material for the paper "Inferring Species Interaction Networks from Species Abundance Data". Please refer to the main paper for a discussion of the background and motivation for the work. We also present and discuss the main findings there. The supplementary material contains extensions to the methods that we used, as well as some additional findings that had to be omitted from the main paper due to space constraints.

In Section 2.1, we describe an extension to the Bayesian network method (see Section 2.1.4 in the main paper) that allows including unobserved factors in

Email addresses: `ali.faisal@tkk.fi` (Ali Faisal), `frankd@bioss.ac.uk` (Frank Dondelinger), `dirk@bioss.ac.uk` (Dirk Husmeier), `c.beale@macaulay.ac.uk` (Colin M. Beale)

the network inference. Section 2.2 describes how we extended the simulation model that was used to generate the synthetic data with an observation process that discretises the data by deciding for each location whether the presence of a species was observed or not. In Section 3 we investigate the difference between edge strengths and confidence values for edges in regression. Section 4 presents some additional experiments on the synthetic data. Section 5 gives additional information on how the networks inferred from the bird atlas data were evaluated, and presents the recovered networks and their characteristics in more detail.

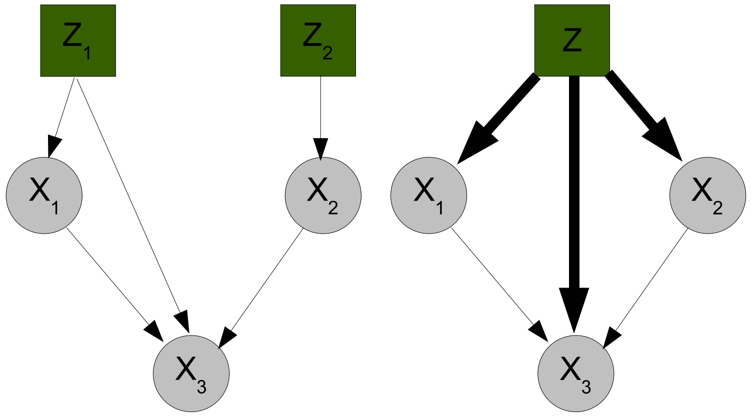
2. Methods

2.1. Latent variable model allowing for unobserved factors

We want to extend the Bayesian network approach to allow for unobserved factors in the environment, e.g. related to climate change or the availability of natural resources. This can be achieved by including additional so-called latent variables in the model. Inference can be carried out with the allocation sampler described in Nobile and Fearnside (2007) and Grzegorczyk et al. (2008), which is based on the following iterative procedure: Given the network structure, new values for the latent variables are inferred (imputation step). Then, given the complete data (real data, and imputed values for the latent variables), the network structure is modified with a standard structure MCMC step (Madigan and York, 1995). This procedure is iterated, and leads to a Markov chain which (on convergence) samples both the network structure and the allocation of the latent variables from the posterior distribution.

Ideally, the interactions between the latent variables and the species are treated as flexible (Fig. 1 a). To reduce the computational complexity, we keep them fixed, i.e. they were enforced to be connected to all species. It is easy to prove that for discrete values, this is equivalent to a model with a single latent variable and a flexible number of discretisation levels (Fig. 1 b); this is the model described in Grzegorczyk et al. (2008).

While the application of this scheme to the simulated data led to encouraging results (Section 4.3), the MCMC simulations did not properly converge for the warbler data. The reason is that a straightforward adaptation of the method proposed in Grzegorczyk et al. (2008) introduces a separate latent variable for each spatial location, leading to a model that is significantly more complex



(a) General Latent Variable Model

(b) Restricted Latent Variable Model

Figure 1: (a) Unrestricted latent variable model, here with two latent variables and three observed ones. (b) Alternative model with a single completely connected latent variable; this is effectively a mixture model. Zs are latent variables, Xs are observed variables. Thin edges are learnt, thick edges are fixed.

than explored in the original application. Our future work therefore aims to simplify the model complexity and explore alternative inference schemes based on variational learning.

2.2. Observation Process for Simulation Data

The simulation described in Section 3.1 of the main paper produces continuous values for the population densities. In order to transform these into presence/absence data similar to the Bird Atlas data, we implement an observation process. We assume that the probability $P(x_g)$ of missing (i.e. not observing) a population of density x_g is modelled by a Gaussian $N(\mu, \sigma^2)$. Then the probability of observing a species with density x_g is $P(X < x_g)$, i.e. the cumulative distribution function:

$$P(X < x_g) = \frac{1}{2}\left(1 + erf\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right) \tag{1}$$

We can then sample a discrete value for x_g from a binomial distribution, using $P(X < x_g)$ as the parameter. Mean and variance of the Gaussian distribution are fitted so that the distribution of ones and zeros over all locations and species is the same as in the real data set.

2.3. Implementations

Table 1 shows which software we used for the different network reconstruction methods described in Section 2.1 in the main paper, as well as where to get the MATLAB code for our own implementations of the extensions in Section 2.2 of the main paper.

Method	**Software**	**Package**	**Description**
GGM	R	GeneNet	The software implementing Graphical Gaussian models is described in Schäfer and Strimmer (2005) 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 Genlab 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 and 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 Grzegorczyk 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`

Table 1: Network reconstruction software used

3. Investigation into LASSO Weights versus Confidence Values for Edges

3.1. Motivation

When using LASSO linear regression to reconstruct an interaction network, we have two options. One is to use the weights found during the regression and interpret them as edge strengths between the target variable and the other variables in the network (we will refer to this as "the weight method"). The other is to obtain confidence values for the presence of an edge ("the confidence value method"). Obtaining the weights is straightforward, and only requires one regression per variable. However, it is potentially biased towards edges that have a strong effect, and may ignore edges with a small (but consistent) effect.

To obtain confidence values, we use a method that is essentially an approximation of a full Bayesian approach to regression. Rather than obtaining the probability that an edge is zero from a posterior distribution of the weights, we follow Friedman et al. (2000) and approximate this value by 'sampling' data from the original dataset[1] using bootstrapping and subsampling. In bootstrap sampling, we sample data points with replacement until the sample size is the same as the size of the original dataset. In subsampling, we sample without replacement until we have obtained a dataset that is half the size of the original dataset.

For each dataset sampled in this way, we run a LASSO regression. Then we record the non-zero weights. After we have done this for a large number of samples, we average over the results. This gives the confidence value for the occurrence of each edge, independent of the strength of that edge. The drawback is that it requires many more runs of the regression algorithm than just calculating the weights once.

We wanted to find out if the difference between using confidence values and using the weights was substantial enough to warrant the extra computational cost. For that reason, we used two synthetic datasets: A simple network model without cycles (in other words, a DAG) from which we generated data using a linear regression model, and a more complex ecological simulation based on Lotka-Volterra interactions between species in a food web (see Section 3.1 of

[1] This should not be confused with sampling from a posterior distribution.

the main paper).

3.2. Simple Network Model

To simulate data from the simple network model based on linear regression, we first sample a network from the niche model described in Section 3.1 of the main paper. If the model is not a DAG, we remove edges until acyclicity has been restored. For each remaining edge, we draw an interaction strength from the Gaussian distribution $N(0, 1)$.

Then we identify species without any parents in the network and draw their population numbers from the Gaussian distribution $N(0, 1)$[2]. For each of the remaining species, we do a standard regression:

$$\hat{y_g} \;=\; \sum_r w_{gr} x_r + N(0, 0.1) \tag{2}$$

where r ranges over all species x_r that are parents of species y_g, and w_{gr} is the weight of the edge linking x_r and y_g. The $N(0, 0.1)$ factor adds a small amount of observational noise. We repeat this process, drawing new population numbers each time to generate different data points.

3.3. Results

Simple Network Model. We generated data from 10 random networks using the simple linear regression model, and for each network we generated 100 bootstrap/subset replica. Figure 2 shows the results. We started off by computing the confidence values straightforwardly: For each sampled dataset, every weight that was not set to 0 by the LASSO regression was counted as detecting an edge. The results of this basic approach are shown in the unshaded boxes in Figure 2.

Using a two-sided paired t-test, we determined that while the difference in TPFP5 values was not significant, the difference in AUC values between the two sampling methods and the weight method was significant ($p < 0.01$). It is surprising to see the weight method outperform the confidence value methods,

[2] This allows for negative population numbers, but this is not a problem since LASSO regression does not assume that population numbers have to be positive.

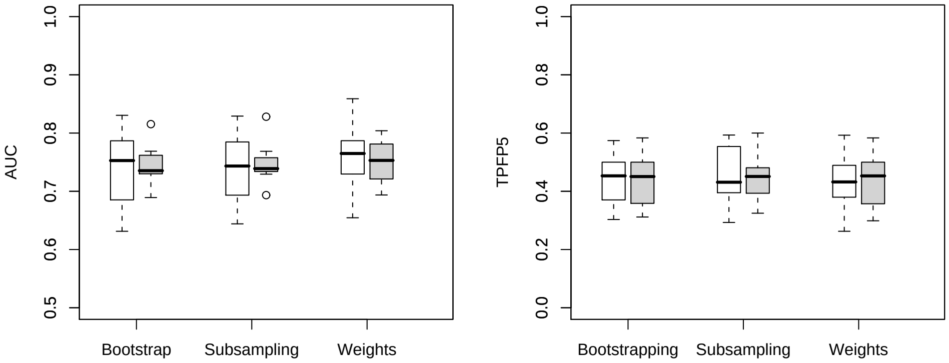


Figure 2: AUC and TPFP5 performance measures for the LASSO reconstruction of the simple network model. Shaded boxes show the result when thresholding is applied.

as we would expect confidence values to produce equally good if not better results.

The reason for this discrepancy becomes apparent once we change the procedure for estimating confidence values slightly. Instead of treating all non-zero weight values in each sampled dataset as evidence of an edge, we only keep those above a certain threshold (arbitrarily set at 0.1). To be fair in our comparison, we also apply the threshold to the weight method. When we do this, we notice that the AUC values between the confidence value methods and the weight method are no longer significantly different ($p > 0.3$).

The problem is that the selection process which sets some weights to zero is not a very conservative process. This means that some weights may never or rarely get set to zero, despite having a very low value. A threshold artificially removes those weights, and thus reduces the variance in the performance. This evens out the difference between the weight method and the confidence value methods.

Ecological Network Simulation Model. We also want to compare the different methods using the simulation model described in Section 3.1 of the main paper. We use the same datasets that were used in the rest of this study.

Since we have already established that thresholding is needed to remove the variance due to small but persistent weights in the confidence value methods, we also use this method here. Figure 3 shows the results on the ecological simulation data. A two-sided paired t-test shows that all differences in AUC

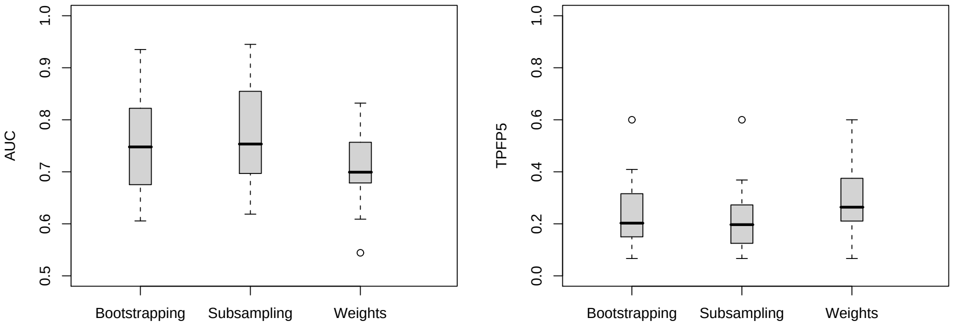


Figure 3: AUC and TPFP5 performance measures for the LASSO reconstruction of the ecological network simulation model.

values are significant ($p < 0.01$), but none of the differences in TPFP5 values are ($p > 0.08$).

Interestingly, the significant difference in AUC now shows an increased performance for the confidence value methods. However, one must remember that the model does not include any spatial autocorrelation (cf. Section 2.2.1 of the main paper), which is by necessity, as sampling destroys the spatial structure. But this also means that sampling reduces the spatial autocorrelation, because we only sample a subset of the total number of nodes, so some of the neighbours of a selected location are left out. This explains why we see a slight increase in performance in AUC. It is reasonable that it would not be mirrored in the TPFP5 score, because this score relies on edges with high edge weights, which will be found in any case.

4. Additional Results on Simulated Data

In this section, we present additional results on the simulated data that could not be included in the main paper due to space restrictions. We show the results for discretised data (Section 4.1), a study of different types of consensus networks (Section 4.2) and the results of a latent variable model for Bayesian networks (Section 4.3). We also list the significance of all results in a separate section (Section 4.4).

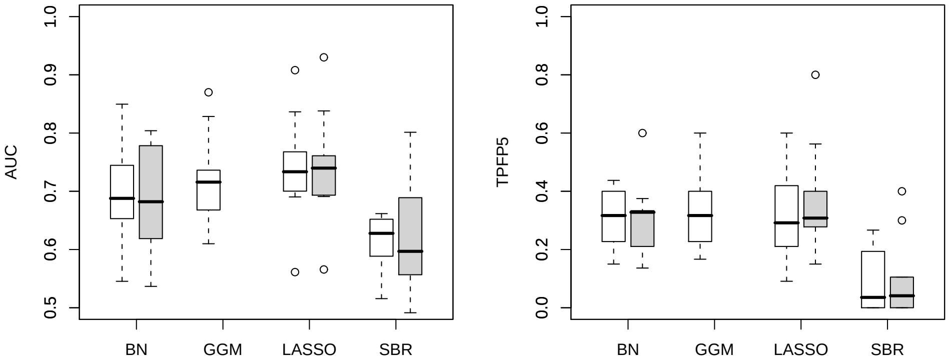


Figure 4: AUC and TPFP5 performance measures for discretised simulation data. Shaded boxes show the result when spatial autocorrelation is included in the model.

4.1. Discrete Data

To better simulate a real dataset, we discretise the continuous data from the simulation model using the observation process described in Section 2.2. The results of applying our network reconstruction methods on the discrete data can be seen in Figure 4.

As expected, the performance decreased when compared to the continuous data (see Section 4.1 in the main paper), due to the information loss inherent in the discretisation process. The AUC scores dropped around 0.1 for all methods, and the TPFP5 scores showed a similar drop, except in the case of SBR, which stayed about the same. This is because discretisation mostly hinders the identification of the more subtle interactions, which SBR had not even detected in the continuous case. Apart from SBR, there is no significant difference in the scores between methods for discrete data.

To finish our investigation, we looked at the effect of including spatial autocorrelation for the discretised data. The results are shown in Figure 4 (shaded boxes).

Unfortunately, none of the scores improved significantly when including spatial autocorrelation in the discrete case. This is likely due to the information loss in the observation process, which makes it harder to estimate spatial autocorrelation effects reliably. Our future work aims to reduce the information loss by applying more complex spatial-temporal models, e.g. along the lines of the Markov random field model proposed in Wei and Li (2007).

4.2. Consensus Networks

As described in the main paper, it is useful to combine outputs of different network reconstruction methods into one single recovered network. We call this a consensus network, because it captures the consensus between the various methods, whilst simultaneously allowing the strengths of the different methods to be combined. There are several different ways in which we can combine these methods:

- Arithmetic Mean: Edge strengths produced by regression methods are scaled to the range $[0, 1]$ (posterior probabilities obtained by Bayesian nets are left unchanged), then we take the arithmetic mean of the scaled strengths and probabilities obtained by all methods and use this as indication of the confidence we have in each edge.
- Harmonic Mean: This is the same as the previous method, but instead of using the arithmetic mean, we calculate the harmonic mean, which is generally more appropriate for rates.
- Thresholded: In this method, we use the posterior probabilities obtained by Bayesian nets as a threshold. All edges with probability less than 0.1 are removed. Then the remaining edges are evaluated based on the interaction strengths found in regression.

Note that some of these methods potentially confuse confidence values (probabilities) with interaction strengths, but for methods where both were available we found a very strong Spearman rank correlation between the two ($\rho = 0.92$), so this is not problematic. As a base line, we used the mean of the AUC or TPFP5 scores obtained from the different network reconstruction methods in isolation. A consensus method works if it produces a better score than the mean score of the individual methods.

Figure 5 shows the results using the discretised dataset with spatial autocorrelation modelled. This most closely mirrors the experiments on the bird data; however, results using continuous data and data without modelling the spatial autocorrelation were similar. As can be seen, the only method performing better than our baseline is the arithmetic mean. For AUC the difference is significant (using a two-sided paired t-test, $p = 0.03$) while the harmonic mean does not perform significantly different (though only barely, $p = 0.05$) and the thresholded approach performs significantly worse ($p = 10^{-3}$). For the TPFP5

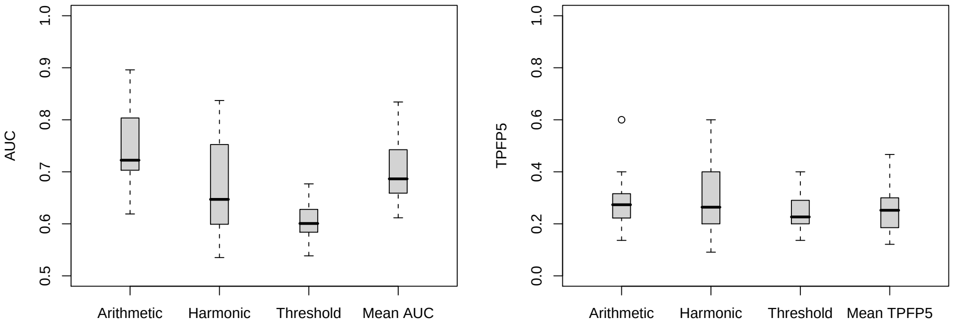


Figure 5: AUC and TPFP5 performance measures for different types of consensus networks. This figure only shows the results for discrete data with an spatial autocorrelation model. Results for other datasets were similar.

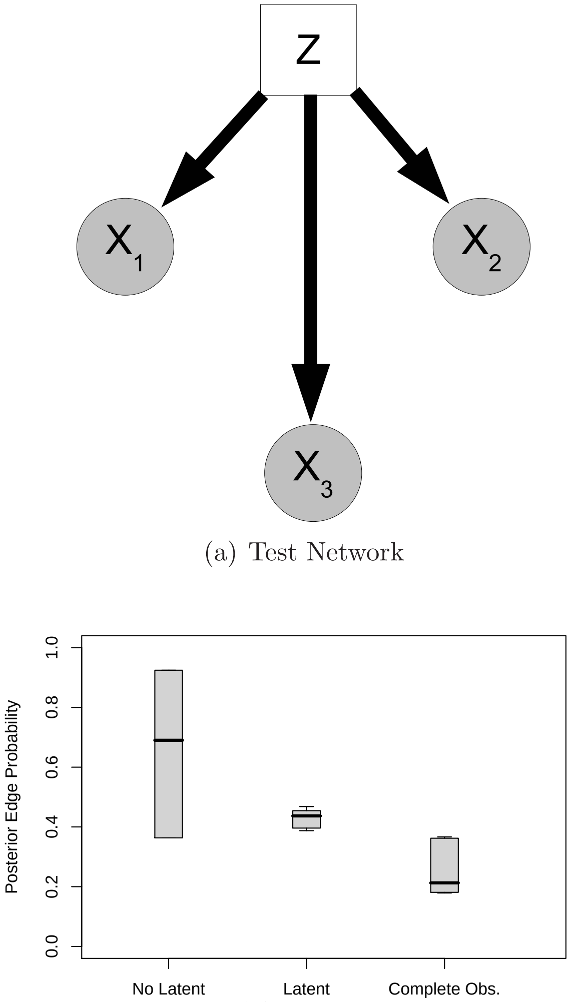
score, none of the three consensus methods performs significantly different from the baseline of taking the mean of the scores, although the arithmetic mean comes closest ($p = 0.06$ versus $p = 0.25$ and $p = 0.58$ for harmonic mean and thresholded approach, respectively).

These results show that the arithmetic mean performs best when it comes to combining different network reconstruction methods. On the basis of this investigation, we have used the arithmetic mean to construct consensus networks for the bird atlas data.

4.3. Allowing for Unobserved Effects

As explained in Section 2.1, we may want to take account of unobserved effects that act on the different species. While there are no explicit environmental factors (other than noise) in the simulation model, it is easy to model an unobserved effect by adding a species that acts directly on all other species, and removing the presence/absence data for that species reconstructing the network. To assess the helpfulness of this approach, we tested it on a small network consisting of three observed nodes and one unobserved node, with no interactions between the observed nodes (Fig. 6a). Under these circumstances, the latent variable model should produce fewer spurious interactions than a model without latent variables. In the Bayesian network model, this means that the posterior probability of edges between observed nodes should be lower when using the latent variable model.

Figure 6b shows the performance of the Latent Variable Model, compared with



(a) Test Network

(b) Results

Figure 6: (a) The network used to test the performance of the latent variable model, consisting of one fully connected species Z, and three unconnected species X_1, X_2, X_3. (b) Boxplot showing the posterior probabilities of spurious edges found using Structure MCMC with one fully-connected missing species, Structure MCMC with a latent variable, and Structure MCMC with a complete dataset (no missing species).

(a) Continuous, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.06	**0.02**	0.06
GGM		1	0.28	**0.00**
LASSO			1	**0.00**
SBR				1

(b) Discrete, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.09	0.06	**0.00**
GGM		1	0.08	**0.00**
LASSO			1	**0.00**
SBR				1

(c) Continuous, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.21	0.52
LASSO		1	**0.00**
SBR			1

(d) Discrete, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.08	0.16
LASSO		1	**0.01**
SBR			1

Table 2: Significance values obtained using a two-sided paired t-test when comparing different methods based on the AUC scores of the reconstructed networks. Significant results (with threshold $p = 0.05$) are marked in bold.

the baseline of using simple Structure MCMC with a missing species and the optimal scenario of having complete data. As can be seen, the Latent Variable Model succeeds in reducing the median probability of spurious edges, although not quite to the level of having complete knowledge of the data.

4.4. Significance Tests

This section gives an overview of the significance of the differences between the network reconstruction methods (Tables 2 and 3), as well as between methods that include spatial autocorrelation and those that do not (Table 4). We have used two-sided paired t-tests everywhere, pairing up results on data simulated from the same network. The threshold for statistical significance is set at $p = 0.05$. For an interpretation of the significant results, see Section 4 in the main paper.

(a) Continuous, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.55	**0.02**	**0.00**
GGM		1	0.38	**0.00**
LASSO			1	**0.00**
SBR				1

(b) Discrete, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.22	0.58	**0.00**
GGM		1	0.71	**0.00**
LASSO			1	**0.01**
SBR				1

(c) Continuous, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.17	**0.00**
LASSO		1	**0.00**
SBR			1

(d) Discrete, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.06	**0.01**
LASSO		1	**0.01**
SBR			1

Table 3: Significance values obtained using a two-sided paired t-test when comparing different methods based on the TPFP5 scores of the reconstructed networks. Significant results (with threshold $p = 0.05$) are marked in bold.

(a) Continuous Data

	AUC	TPFP5
BN	**0.01**	**0.00**
LASSO	**0.00**	**0.00**
SBR	**0.00**	**0.04**

(b) Discrete Data

	AUC	TPFP5
BN	0.58	0.80
LASSO	0.51	0.07
SBR	0.65	0.84

Table 4: Significance values obtained using a two-sided paired t-test when network reconstruction methods with spatial autocorrelation model to those without. Significant results (with threshold $p = 0.05$) are marked in bold.

5. Application to the European bird atlas data

5.1. A priori network construction

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 Knowledge [3] (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`.

5.2. Phylogenetic distance analysis

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 in Figure 7.

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 none the less. Consequently, we repeated the correlation analysis using Kendall's τ as a measure of rank correlation that

[3] Found at `http://www.isiwebofknowledge.com/`.

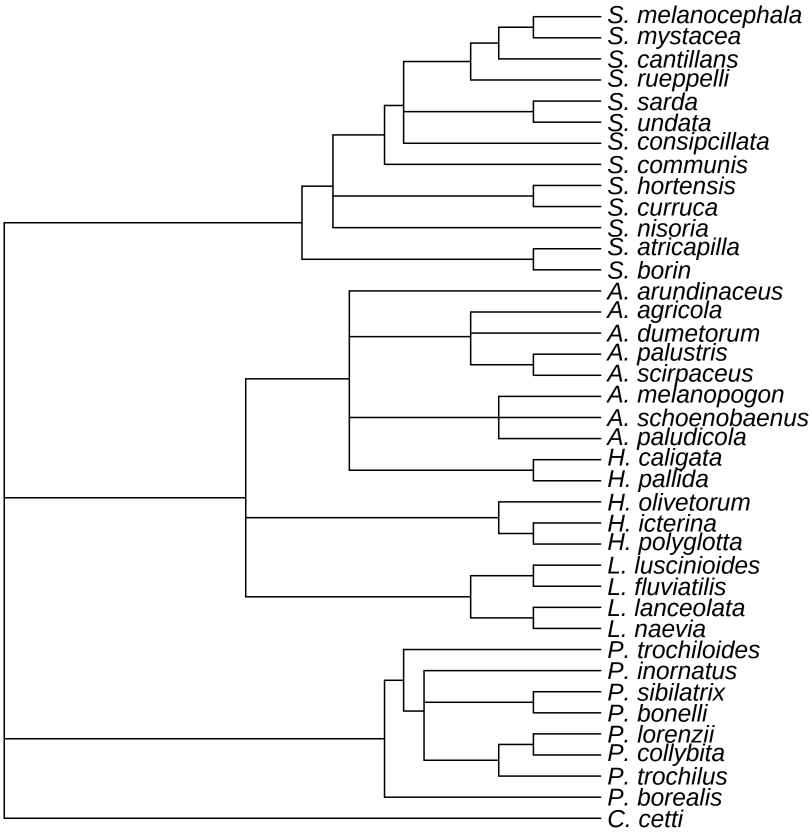


Figure 7: Phylogenetic tree for the warbler species in our study.

is unaffected by assumed branch lengths. Again, results were qualitatively similar; they can be found in Table 5. For the correlation analyses we used only data from the upper triangle of the distance matrices.

Network	$\boldsymbol{\rho}$	**Correlation**
Basic	1	-0.12 (-0.18, -0.04)
	0.6	-0.11 (-0.18, -0.04)
	0.3	-0.12 (-0.19, -0.05)
	Kendall's τ	-0.08
Spat. Autocorr	1	-0.12 (-0.19, -0.05)
	0.6	-0.12 (-0.19, -0.05)
	0.3	-0.12 (-0.19, -0.05)
	Kendall's τ	-0.08
Spat. Autocorr. and Bio-Climate Covariates	1	-0.14 (-0.21, -0.07)
	0.6	-0.14 (-0.21, -0.07)
	0.3	-0.12 (-0.22, -0.07)
	Kendall's τ	-0.09

Table 5: Correlation coefficients of reconstructed networks with the phylogenetic tree whose branch lengths have been generated with different values of ρ, or with Kendall's τ. Numbers in brackets show the confidence intervals at 95%. None of the confidence intervals includes zero, indicating that the correlations are significant.

5.3. Ecological distance analysis

Ecological trait data for each of the 39 species is presented in Table 6. 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.

Species	**Length**	**Mass**	**Wingspan**	**Clutch**	**Migrant status**	**Preferred Habitat**
Acrocephalus agricola	13	11	16	4.5	Long Distance	Resident
Acrocephalus arundinaceus	20	33	26	4.5	Long Distance	Resident
Acrocephalus dumetorum	13	12	18	5	Long Distance	Resident
Acrocephalus melanopogon	12	12	16	4.5	Short Distance	Resident
Acrocephalus paludicola	13	12	18	5	Long Distance	Resident
Acrocephalus palustris	13	13	20	4.5	Long Distance	Shrub
Acrocephalus schoenobaenus	13	12	19	5	Long Distance	Shrub
Acrocephalus scirpaceus	13	13	19	4	Long Distance	Resident
Cettia cetti	14	13.5	17	4.5	Resident	Resident
Hippolais icterina	14	13	22	4.5	Long Distance	Broad-leaf Forest
Hippolais olivetorum	15	18	25	3.5	Long Distance	Broad-leaf Forest
Hippolais pallida	13	11	20	2.5	Long Distance	Shrub
Hippolais polyglotta	13	13	18	4	Long Distance	Broad-leaf Forest
Locustella fluviatilis	13	17	20	6	Long Distance	Shrub
Locustella luscinioides	14	18	20	5	Long Distance	Resident
Locustella naevia	13	14	17	5.55	Long Distance	Shrub
Phylloscopus collybita collybita	10	9	18	5.49	Short Distance	Broad-leaf Forest
Phylloscopus bonelli bonelli	12	9	18	5	Long Distance	Broad-leaf Forest
Phylloscopus borealis	11	10	19	5.5	Long Distance	Pine Forest
Phylloscopus trochiloides	10	8	18	5.49	Long Distance	Pine Forest
Phylloscopus sibilatrix	12	10	22	5.77	Long Distance	Broad-leaf Forest
Phylloscopus trochilus	11	10	19	5.93	Long Distance	Broad-leaf Forest
Sylvia atricapilla	13	21	22	4.56	Short Distance	Shrub
Sylvia borin	14	19	22	4.32	Long Distance	Shrub

Sylvia cantillans	12	11	17	4	Long Distance	Garrigue
Sylvia communis	14	16	20	4.64	Long Distance	Shrub
Sylvia conspicillata	12	10	15	4.5	Short Distance	Garrigue
Sylvia curruca	13	12	18	4.67	Long Distance	Shrub
Sylvia hortensis	15	21	22	5	Long Distance	Broad-leaf Forest
Sylvia melanocephala	14	13	16	4.5	Resident	Garrigue
Sylvia nisoria	16	25	25	4.5	Long Distance	Shrub
Sylvia rueppelli	14	14	20	5	Long Distance	Broad-leaf Forest
Sylvia sarda	12	10	16	4	Resident	Garrigue
Sylvia undata	12	10	16	4	Resident	Garrigue
Sylvia mystacea	14	10	17	4.5	Short Distance	Shrub
Hippolais caligata	12	10	20	3.5	Long Distance	Shrub
Phylloscopus inornatus	10	7	17	4	Long Distance	Pine Forest
Phylloscopus lorenzii	10	8	18	5.5	Short Distance	Broad-leaf Forest
Locustella lanceolata	12	12	15	5	Long Distance	Shrub

Table 6: Ecological traits for the warbler birds.

	A priori net	Phylogenetic Dist.	Ecological Dist.
A priori net	1	0.38 (0.73, 0.02)	0.08 (0.21, -0.05)
Phylogenetic Dist.		1	0.28 (0.21, 0.34)
Ecological Dist.			1

Table 7: Results of comparison between the ecological measures represented by the *a priori* interaction network, phylogenetic distance and ecological distance. *A priori* comparisons made with logistic regression are the regression coefficient, other results are Pearson's correlation coefficients, all with 95% confidence intervals

5.4. Comparison of Ecological Measures

We have three different ecological indicators that we can compare our reconstructed networks to: The a priori network, the phylogenetic distance and the ecological distance. The correlation of these indicators with the reconstructed networks that we present in the main paper is always significant, but also far from perfect correlation. This can be explained by the fact that these measures are not a true gold standard. In fact, each measure captures different aspects of the true relationships between species. In Table 7 we present the correlation coefficients between the three ecological measures and show that they are also small but (mostly) significant.

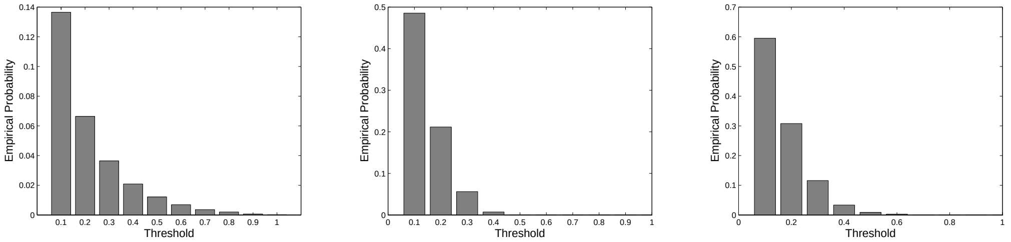
Another way to compare the ecological indicators is by taking the a priori network as a gold standard, and calculating the AUC and TPFP5 values for the phylogenetic and ecological distance measures. In effect, we are treating these distance measures as inverse edge scores. The results are shown in Table 8. Again, The scores are better than random expectation (AUC=0.5, TPFP5=0.05), but far from perfect (AUC=TPFP5=1.0). This indicates that the various measures capture relevant, but only partial aspects of the unknown true interaction network.

5.5. Thresholding on Edge Interactions

To produce a single, interpretable network from the edge interaction strengths, we need to set a threshold to discard edges with low values. Recall that the "interaction strengths" are of different nature: marginal posterior probabilities for Bayesian networks, and regularised regression coefficients for LASSO. We would like to map them to p-values, which are more commonly used in

	AUC	TPFP5
Phylogenetic Distance	0.79	0.37
Ecological Distance	0.67	0.22

Table 8: Comparison between the ecological measures by computing AUC and TPFP5 scores for phylogenetic and ecological distance measures, using the *a priori* interaction network as a gold standard.



(a) BN

(b) LASSO

(c) Consensus

Figure 8: Distribution of edge strengths/posterior probabilities under the null hypothesis, averaged over 15,210 random species interactions from permuted data.

statistics. To this end, we carried out a randomisation test. The rows and columns of the original warbler data were permuted ten times, and on each of these replications we carried out the same inference as for the original data. Since the permutation destroys all genuine associations among the species, the distribution of "interaction strengths" represents the null hypothesis of no species interaction. From this distribution, the p-value is easily computed as the probability of exceeding a given threshold.

Figure 8 shows the null distributions obtained for Bayesian networks (left panel), LASSO (centre panel), and the consensus network (right panel). Table 9 shows the "interaction strengths" corresponding to p-values of 0.1 and 0.01. Note that the p-values are used as descriptive measures, and no Bonferroni correction (which would be too conservative) was carried out.

p-value	**BN**	**LASSO**	**Consensus**
0.1	0.2	0.3	0.4
0.01	0.5	0.4	0.5

Table 9: Mapping from p-value thresholds to edge strengths/posterior probabilities.

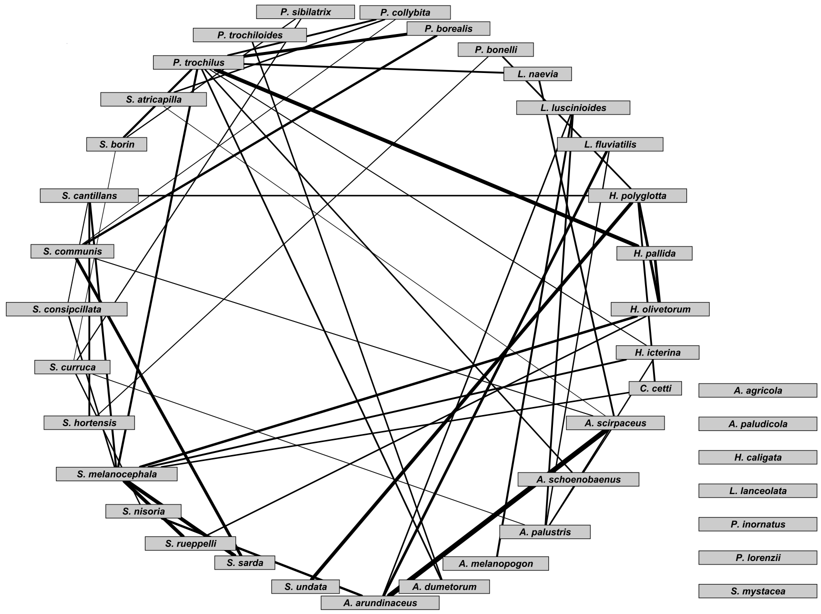


Figure 9: Consensus network recovered from the basic dataset (without spatial autocorrelation or bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

5.6. Recovered Networks

Figures 9-11 shows the consensus networks that were recovered from the warbler data. We get three different networks: one for the basic dataset, one for a dataset where we have modelled spatial autocorrelation as described in Section 2.2.1 of the main paper, and one for a dataset where we have included both spatial autocorrelation and two bio-climate covariates: temperature and availability of water. Details on how the sparsity and the correlation with the ecological measures vary for the different networks can be found in the main paper (Section 4.2).

5.7. Network Characterisation

Studies have shown that molecular regulatory networks have degree distributions that approximately follow a power-law (Wagner, 2001; Guelzim et al.,

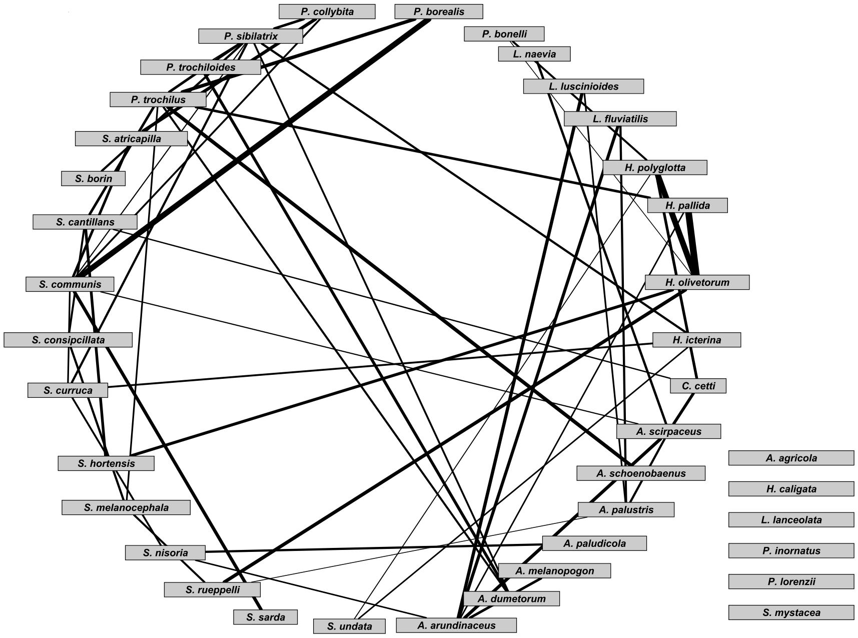


Figure 10: Consensus networks recovered from the dataset with spatial autocorrelation included (but without bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

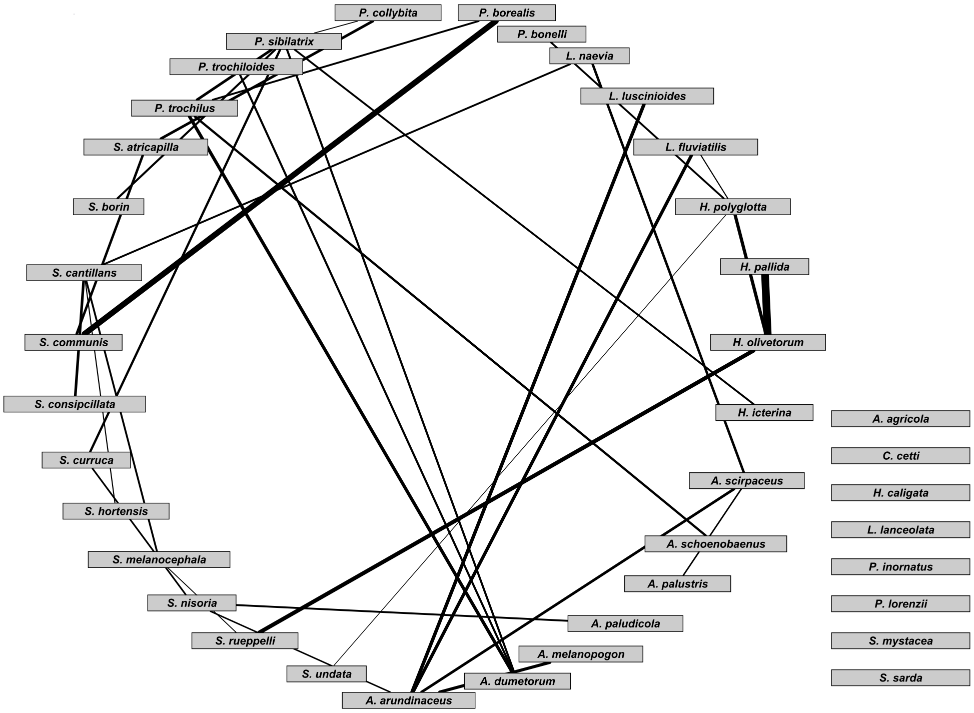


Figure 11: Consensus networks recovered from the dataset with both spatial autocorrelation and bio-climate covariates included. The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

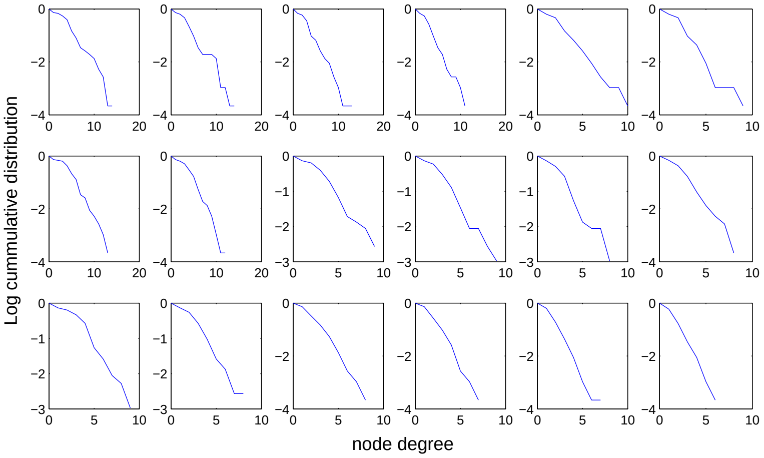


Figure 12: Cumulative degree distribution for the consensus networks on the **log-linear scale** as the threshold varies. (Top) Basic bird data, (Middle) Bird data with spatial autocorrelation model added, (Bottom) Birds with spatial autocorrelation and bio-climate covariates. From left to right the thresholds are set at p-values 0.2, 0.15, 0.1, 0.05, 0.02, and 0.01.

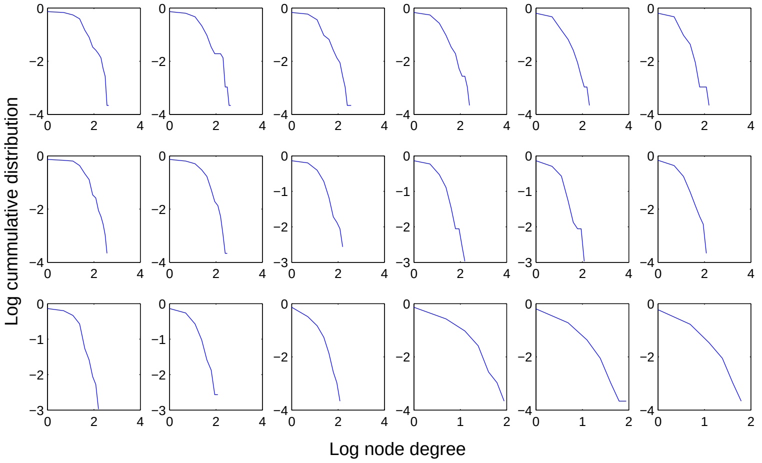
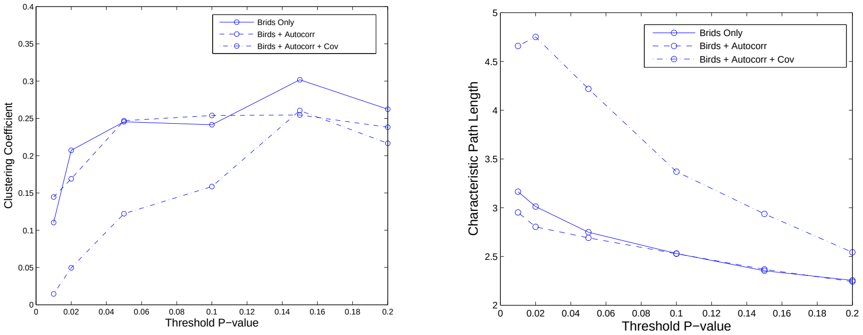


Figure 13: Cumulative degree distribution for the consensus networks on the **log-log scale** as the threshold varies. (Top) Basic bird data, (Middle) Bird data with spatial autocorrelation model added, (Bottom) Birds with spatial autocorrelation and bio-climate covariates. From left to right the thresholds are set at p-values 0.2, 0.15, 0.1, 0.05, 0.02, and 0.01.



(a) Clustering Coefficient

(b) Network Diameter

Figure 14: Variation of the clustering coefficient and network diameter for the consensus networks as the threshold varies.

2002; May, 2006). Loosely speaking, this means that there are many nodes with only one or few connections, but also some nodes with many more connections than the average degree. Studies on food webs generally agree that the degree distribution is not Poisson (Proulx et al., 2005), however they disagree on whether the degree distributions are best fit by a power-law or by some other distribution. The existence of a variety of distributions has been shown, including power-law, truncated power-law and exponential (Dunne et al., 2002; Jordano et al., 2003; Laird and Jensen, 2006). In our study we observe that the distributions are closer to linear on the log-linear plot of the cumulative degree distribution (Fig. 12), than on the log-log plot (Fig. 13). Linearity on the log-log plot would be characteristic of a power-law distribution, but linearity on the log-linear plot shows that the network exhibits a near exponential distribution. The data also displays the insensitivity of this behaviour to varying the threshold.

Figure 14 shows the variation of the clustering coefficient and the network diameter (characteristic path length) as the threshold varies. There is no discernable trend, which may mean that these particular statistics are not useful characterisations of the types of networks that we are considering.

Acknowledgement

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 Grzegorczyk 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.

References

Alstroem, P., Ericson, P., Olsson, U., Sundberg, P., 2006. Phylogeny and classification of the avian superfamily *Sylvioidea*. Mol. Phylogenet. Evol. 38 (2), 381–397.

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.

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

Grafen, A., 1989. The phylogenetic regression. Phil. Trans. R. Soc. B 326, 119–157.

Grzegorczyk, M., Husmeier, D., Edwards, K., Ghazal, P., Millar, A., 2008. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics 24 (18), 2071–2078.

Guelzim, N., Bottani, S., Bourgine, P., Képès, F., 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat. Genet. 31 (1), 60–63.

Jordano, P., Bascompte, J., Olesen, J., 2003. Invariant properties in coevolutionary networks of plant-animal interactions. Ecol. Lett. 6 (1), 69–81.

Laird, S., Jensen, H., 2006. The Tangled nature model with inheritance and constraint: Evolutionary ecology restricted by a conserved resource. Ecol. Complex. 3 (3), 253–262.

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.

May, R., 2006. Network structure and the biology of populations. Trends Ecol. Evol. 21 (7), 394–399.

Nobile, A., Fearnside, A., 2007. Bayesian finite mixtures with an unknown number of components: the allocation sampler. Stat. Comput. 17 (2), 147–162.

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.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4 (1), Article 32.

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.

Wagner, A., 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol. 18 (7), 1283–1292.

Wei, Z., Li, H., 2007. A Markov random field model for network-based analysis of genomic data. Bioinformatics 23 (12), 1537.