Technical University of Denmark

DTU

Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range

Thorson, James T.; Scheuerell, Mark D.; Shelton, Andrew O.; See, Kevin E.; Skaug, Hans J.; Kristensen, Kasper

DTU Library
Technical Information Center of Denmark

Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range

James T. Thorson[1]*, Mark D. Scheuerell[2], Andrew O. Shelton[3], Kevin E. See[4], Hans J. Skaug[5] and Kasper Kristensen[6]

[1]*Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2725 Montlake Blvd. E, Seattle, WA 98112, USA;* [2]*Fish Ecology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2725 Montlake Blvd. E, Seattle, WA 98112, USA;* [3]*Conservation Biology, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2725 Montlake Blvd. E, Seattle, WA 98112, USA;* [4]*Quantitative Consultants, Inc., Boise, ID 83707, USA;* [5]*Department of Mathematics, University of Bergen, PO Box 7800, 5020 Bergen, Norway; and* [6]*Technical University of Denmark, Charlottenlund Slot Jægersborg Allé 1, 2920 Charlottenlund, Denmark*

Summary

1 Predicting and explaining the distribution and density of species is one of the oldest concerns in ecology. Species distributions can be estimated using geostatistical methods, which estimate a latent spatial variable explaining observed variation in densities, but geostatistical methods may be imprecise for species with low densities or few observations. Additionally, simple geostatistical methods fail to account for correlations in distribution among species and generally estimate such cross-correlations as a post hoc exercise.

2 We therefore present spatial factor analysis (SFA), a spatial model for estimating a low-rank approximation to multivariate data, and use it to jointly estimate the distribution of multiple species simultaneously. We also derive an analytic estimate of cross-correlations among species from SFA parameters.

3 As a first example, we show that distributions for 10 bird species in the breeding bird survey in 2012 can be parsimoniously represented using only five spatial factors. As a second case study, we show that forward prediction of catches for 20 rockfishes (*Sebastes* spp.) off the U.S. West Coast is more accurate using SFA than analysing each species individually. Finally, we show that single-species models give a different picture of cross-correlations than joint estimation using SFA.

4 Spatial factor analysis complements a growing list of tools for jointly modelling the distribution of multiple species and provides a parsimonious summary of cross-correlation without requiring explicit declaration of habitat variables. We conclude by proposing future research that would model species cross-correlations using dissimilarity of species' traits, and the development of spatial dynamic factor analysis for a low-rank approximation to spatial time-series data.

Key-words: factor analysis, Gaussian process, Gaussian random field, geostatistics, habitat envelope model, hierarchical model, joint species distribution models, mixed-effects model, spatial factor analysis

Introduction

The spatial distribution of organisms is one of the basic characteristics of populations and communities (Elton 1927). Species distributions are studied in invasion biology to explore the causes and consequences of non-native species, in spatial ecology as indicative of colonization and extinction processes (Hanski 1998), and in macroecology as an impetus for exploring general life-history patterns and principles (Brown 1999). The well-documented relationship between occupancy and abundance across species implies that monitoring occupancy is a useful proxy for detecting changes in population abundance and viability (Gaston *et al.* 2000), and co-occurrence of species is frequently used to screen for facultative and obligatory mutualisms. Species distributions are also interpreted to plan conservation actions (Johnson, Seip & Boyce 2004) or infer ecological dynamics (e.g. community assembly, Gotelli & McCabe 2002), and parsimonious estimates of species distributions (and their correlations among species) is an ongoing research topic (Gaston & Rodrigues 2003; Kissling *et al.* 2012).

There exists a well-developed literature regarding the estimation of species distribution and range using detection/non-detection and count data. Methods for estimating species distribution include 'hierarchical' models that explicitly separate measurement and biological process models (MacKenzie *et al.* 2005; Royle & Wikle 2005), regression-based methods that correlate detection and count data to measured variables

(Goetz *et al.* 2014) and sample-based methods that often will not propagate uncertainty explicitly (Gaston 1991). There is also recent interest in methods that use data from multiple species simultaneously within 'joint species distribution models', JSDM (Clark *et al.* 2013; Pollock *et al.* 2014). As one example, Latimer *et al.* (2009) estimated an unmeasured (latent) variable representing unmeasured spatial variation for each of four invasive species, while estimating cross-correlations representing the impact of one species on the probability of encountering the other. However, methods such as this require estimating as many latent variables as there are species, and will likely not be parsimonious (or even computationally possible) for large numbers of species. Other JDSMs also include measured habitat variables and include pairwise correlations among residuals for all species at a given location.

Species will have a nonzero cross-correlation whenever the occurrence of one implies an increased (positive cross-correlation) or decreased (negative cross-correlation) probability that the other will occur at the same location. Positive cross-correlations can arise whenever species have similar environmental, dispersal or biotic requirements, or when there is some direct or indirect positive interaction between these species (Ovaskainen, Hottola & Siitonen 2010; Kissling *et al.* 2012). Given that two species have ranges that are not statistically independent (i.e. have some positive or negative cross-correlation), the spatial distribution of one will be informative about the distribution of the other (Harris 2015). This implies that multispecies information can be a useful way to leverage limited sampling information for low-density and difficult-to-detect species (Ovaskainen & Soininen 2011). Additionally, most communities are composed of a few species with high densities and many species with low densities. Given this, we may have sparse data regarding species occupancy and range for the majority of species in a community. Finally, many species will have ranges that fluctuate over short- and long-term cycles, and quantifying range shifts over time requires that ecologists be capable of accurately estimating species ranges using limited data from short time-intervals.

For these reasons, it is important to develop tools for estimating species ranges that utilize information in multispecies data sets when this can improve predictive performance for low-density and otherwise data-poor species. Previous analysis of multispecies data has generated an extensive literature of multivariate statistical techniques (see, e.g. McCune, Grace & Urban 2002). One such technique is factor analysis, which decomposes the variance in a multivariate data set into measurement variance (which is independent for each variable) and variation explained by a reduced set of unobserved (latent) 'factors' (Rencher 2002). Each measured variable has a loading onto each factor, and this loading represents the degree to which a measured variable can be explained by a given factor. Multiple variables have loadings onto each factor, such that correlations among measured variables are explained by these estimated factors (where two correlated variables will have similar loadings onto at least one factor). Factor analysis has subsequently been expanded to time-series analysis of ecological communities (i.e. dynamic factor analysis, Zuur, Tuck &

Bailey 2003), but we know of no previous ecological studies that have modified factor analysis for use with spatial data, for example estimating distribution models for multiple species simultaneously.

We therefore demonstrate spatial factor analysis (SFA) as a new tool for joint species distribution modelling. Spatial factor analysis has previously been discussed in biomedical, environmental monitoring and statistical contexts (Wang & Wall 2003; Hogan & Tchernis 2004), but has not previously been used for modelling species distributions. Spatial factor analysis uses a reduced number of unobserved spatial factors to represent unobserved environmental or biological variables in a large number of species. Similar to geostatistical methods, each factor is estimated as a random field such that nearby locations are, on average, more similar than geographically distant locations. We use SFA to illustrate how relatively few spatial factors can describe the distributions of 10 bird species from the western USA, despite very different nesting habitats and foraging behaviours. We then apply SFA to data for 10 years for an assemblage of demersal fish species off the U.S. West Coast and show that SFA has greater predictive accuracy than a single-species geostatistical approach without the need to account for unmeasured covariates. We conclude by deriving an analytical form of spatial correlation and demonstrate its equivalence to Pearson's sample-based approach.

Materials and methods

OVERVIEW

We seek to characterize the co-occurrence of multiple species simultaneously using count data **C** from spatially referenced sampling in a way that (i) is parsimonious and (ii) allows inference about species cross-correlations. Measured environmental variables are typically used when explaining species distribution and often include correlations in model residuals among species estimated either explicitly (Clark *et al.* 2013; Pollock *et al.* 2014) or implicitly (Dunstan, Foster & Darnell 2011; Ovaskainen & Soininen 2011). As an alternative to using measured environmental variables, we define a SFA model, which estimates a low-rank approximation to the spatial distribution of multiple species simultaneously. Instead of using measured environmental variables, SFA estimates one or more latent variables that vary over space (each representing unobserved environmental and biotic effects), without requiring specification of environmental variables *a priori*. Latent spatial variables have been used previously in a several species distribution models (Latimer *et al.* 2006), sometimes in conjunction with measured environmental variables (Latimer *et al.* 2009; Shelton *et al.* 2014; Harris 2015) or phylogenetic relationships (Kaldhusdal *et al.* 2015), and are estimated here as a Gaussian random field (GRF) (Thorson *et al.* In press).

Spatial factor analysis estimates *K* GRFs to approximate the distribution of *J* species, where the number of random fields used in this approximation can range between 1 (a single distribution for all species) and *J* (a different distribution for each species, and estimating all cross-correlations among species). Parameters representing the value of GRFs at sampled locations are estimated as random effects using maximum likelihood, where integration across random effects is approximated using the Laplace approximation (Skaug & Fournier 2006) via Template Model Builder (Kristensen 2014). Model selection tools can

then be used to select the most parsimonious number of GRFs for a given application. Spatial factor analysis estimates the distribution for each species as a linear function of these GRFs, and the matrix of coefficients for these linear functions is called the loading matrix (in analogy to conventional factor analysis). We also estimate residual variability in survey data for each species and observation. In this way, SFA separately estimates the effect of process error (i.e. spatial variation in species densities, including their correlations among species) and measurement error (residual variability caused by the sampling process).

SPATIAL FACTOR ANALYSIS MODEL

The number of individuals observed for species j of J total species and each location i of I total locations is assumed to follow a Poisson distribution with log-normal overdispersion $\varepsilon_{i,j}$:

$$C_{i,j} \sim \text{Poisson}(\exp\{\Lambda_{i,j} + \varepsilon_{i,j}\}) \qquad \text{eqn 1}$$

where $\Lambda_{i,j}$ is the log-expected count of species j in sample i, and $\varepsilon_{i,j}$ accounts for sampling variation in excess of a Poisson distribution (e.g. as caused by variation in densities at a fine spatial scale), where $\varepsilon_{i,j} \sim \text{Norm}\left(0, \sigma_j^2\right)$ and σ_j^2 is the variance of overdispersion for species j. We use a lognormal-Poisson mixture distribution for samples, rather than the more common negative-binomial distribution (Lindén & Mäntyniemi 2011), so that the magnitude of overdispersion σ_j^2 is directly comparable with the magnitude of variation explained by spatial variables (discussed below).

We observe counts, $C_{i,j} = c_{i,j}$, for each sampling location i and species j. The vector of log-expected values $\Lambda_{i,\bullet}$ for all J species at the i-th sampling location depends upon the K latent spatial fields:

$$\Lambda_{i,\bullet} = \boldsymbol{\alpha} + \mathbf{L}\Omega_{i,\bullet}^T \qquad \text{eqn 2}$$

such that

$$\Lambda_{i,j} = \alpha_j + \sum_k l_{jk}\Omega_{i,k} \qquad \text{eqn 3}$$

where α_j is the average log-count for species j, $\Omega_{i,\bullet}$ is the value of $\mathbf{\Omega}$ for all K latent spatial fields at location i, and $\mathbf{\Omega}$ is a matrix where each column represents one of the latent spatial fields:

$$\mathbf{\Omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_k) \qquad \text{eqn 4}$$

The loading matrix $\mathbf{L}$ is a $J \times K$ matrix representing the linear relationship between spatial fields $\mathbf{\Omega}$ and the logarithm of expected counts Λ for each species:

$$\mathbf{L} = \begin{bmatrix} l_{1,1} & 0 & \ldots & 0 \\ l_{2,1} & l_{2,2} & \ldots & 0 \\ l_{3,1} & l_{3,2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ l_{J,1} & l_{J,2} & \ldots & l_{J,K} \end{bmatrix} \qquad \text{eqn 5}$$

Each factor $\boldsymbol{\omega}_k$ is estimated as a GRF that has marginal variance of one:

$$\boldsymbol{\omega}_k \sim \text{MN}(0, \mathbf{\Sigma}) \qquad \text{eqn 6}$$

where MN is a multivariate normal distribution over the I locations with mean 0 and covariance matrix $\mathbf{\Sigma}$. Covariance is in turn derived from a stationary and isotropic correlation function with a Matérn covariance function with smoothness $\nu = 1$:

$$r(x, x') = \text{Matérn}(|x, x'|) \qquad \text{eqn 7}$$

where x and x' are two locations and $|x,x'|$ is the distance between these two points.

The loading matrix $\mathbf{L}$, the range of the Matérn covariance function κ and the variance of the log-normal overdispersion σ_i^2 are estimated as fixed effects, while the spatial fields $\mathbf{\Omega}$ and log-normal overdispersion parameters $\varepsilon_{i,j}$ are estimated as random effects. Additionally, we implement restricted maximum-likelihood (REML) estimation by treating the mean for each species $\boldsymbol{\alpha}$ as a random effect with a 'flat' prior (Harville 1974). We use R-INLA (Lindgren & Rue 2015) to compute the three components of the precision matrix necessary for the stochastic partial differential equation approximation to GRFs (Lindgren, Rue & Lindström 2011). We then pass these matrices to Template Model Builder, which computes the marginal likelihood of fixed effects using the Laplace approximation (Skaug & Fournier 2006) to integrate across GRFs, overdispersion parameters and the intercept vector $\boldsymbol{\alpha}$ given the joint distribution of fixed and random effects. Template Model Builder computes both the marginal likelihood and its first derivatives with respect to fixed effects, and these are then used by a conventional nonlinear optimizer in the R statistical environment (R Core Development Team 2013) to maximize the marginal likelihood (see Thorson *et al.* (In press) for details). All code for estimating the SFA model is distributed as an R package *SpatialFA* and is available at the first author's GitHub repository (https://github.com/James-Thorson/spatial_factor_analysis).

Interpretation of the estimated spatial factors $\mathbf{\Omega}$ is complicated because the loading matrix $\mathbf{L}$ has a particular structure pre-specified to ensure identifiability. Specifically, the condition that the upper-right corner of the loading matrix is 0 is analogous to a similar condition in dynamic factor analysis (Zuur, Tuck & Bailey 2003). However, interpretation can be simplified by rotating the loading matrix and spatial factors. This rotation is also advocated for conventional factor analysis (Rencher 2002) and dynamic factor analysis (Holmes, Ward & Scheuerell 2014). We have chosen to use varimax rotation:

$$\begin{aligned} \mathbf{\Omega}' &= \mathbf{H}\mathbf{\Omega} \\ \mathbf{L}' &= \mathbf{L}\mathbf{H}^{-1} \end{aligned} \qquad \text{eqn 8}$$

where $\mathbf{H}$ is the varimax rotation matrix, and $\mathbf{\Omega}'$ and $\mathbf{L}'$ are easier to interpret because $\mathbf{L}'$ will tend to be more 'sparse' (have many small values and a few big values) than $\mathbf{L}$.

CALCULATING BETWEEN-SPECIES CORRELATIONS

We next seek to estimate the magnitude of spatial association among all species. We calculate this as a function of elements l_{jk} of the loading matrix $\mathbf{L}$, noting that the expected value and variance of log-catches across the entire spatial domain are:

$$\begin{aligned} E[\Lambda_{\bullet,j}] &= \alpha_j \\ \text{Var}[\Lambda_{\bullet,j}] &= \sum_k l_{jk}^2 \text{Var}[\Omega_{\bullet,k}] = \sum_k l_{jk}^2 \end{aligned} \qquad \text{eqn 9}$$

while the covariance between species is (as shown in the Appendix S1):

$$\text{Cov}[\Lambda_{\bullet,j}, \Lambda_{\bullet,j'}] = \sum_k l_{jk} l_{j'k} \qquad \text{eqn 10}$$

so that

$$\text{Corr}[\Lambda_{\bullet,j}, \Lambda_{\bullet,j'}] = \frac{\text{Cov}[\Lambda_{\bullet,j}, \Lambda_{\bullet,j'}]}{\sqrt{\text{Var}[\Lambda_{\bullet,j}]}\sqrt{\text{Var}[\Lambda_{\bullet,j'}]}} = \frac{\sum_k l_{jk} l_{j'k}}{\sqrt{\sum_k l_{jk}^2}\sqrt{\sum_k l_{j'k}^2}} \qquad \text{eqn 11}$$

This provides a closed-form solution to the expected correlation in the log-expected count between species j and species j' as estimated from the SFA. We note that this solution is similar to Eq. 4 of Pollock *et al.* (2014), given that spatial factors are defined to have variance of

one and zero covariance. However, our derivation is defined via expectations for random fields, while Pollock *et al.* (2014) derives a similar solution via sample statistics of measured covariates (which have no expectation). For comparison, we also calculate the sample correlation between the log-predicted count $\hat{\Lambda}_{\bullet,j}$ for each pair of species j and j':

$$
\text{Sample.Corr}(\hat{\Lambda}_{\bullet,j}, \hat{\Lambda}_{\bullet,j'})
$$

$$
= \frac{\sum\limits_{i=1}^{I}\left(\hat{\Lambda}_{i,j} - \frac{1}{I}\sum\limits_{i=1}^{I}\hat{\Lambda}_{i,j}\right)\left(\hat{\Lambda}_{i,j'} - \frac{1}{I}\sum\limits_{i=1}^{I}\hat{\Lambda}_{i,j'}\right)}{\sqrt{\sum\limits_{i=1}^{I}\left(\hat{\Lambda}_{i,j} - \frac{1}{I}\sum\limits_{i=1}^{I}\hat{\Lambda}_{i,j}\right)^2 \sum\limits_{i=1}^{I}\left(\hat{\Lambda}_{i,j'} - \frac{1}{I}\sum\limits_{i=1}^{I}\hat{\Lambda}_{i,j'}\right)^2}} \quad \text{eqn 12}
$$

while expecting that the analytic and sample correlation will be similar for the SFA.

FIRST CASE STUDY – BIRDS IN THE WESTERN U.S.A

We first use data from the US Breeding Bird Survey (https://www.pwrc.usgs.gov/bbs/) to demonstrate how SFA can explain many different spatial patterns with relatively few spatial factors. Specifically, we used data from the summer of 2012 and limit the data to the western U.S.A. (Washington, Oregon, California, Idaho, Nevada, Montana, Wyoming and Colorado). We choose ten frequently encountered species from a variety of taxonomic families with contrasting spatial ranges and habitat requirements, including species commonly found in woodland, grassland, sagebrush and coastal habitats (Table 1). These species typically occur in different bird conservation regions, or ecoregions, across the western U.S.A. (Fig. 1), as defined by the North American Bird Conservation Initiative (NABCI), and we chose them (rather than species within a single taxonomic or functional group) to demonstrate the application of SFA for species with varied spatial distributions (Babcock *et al.* 1998). We fit the SFA model using 1 through 10 factors and use marginal AIC (as calculated from the maximum marginal likelihood of the SFA model) to select the most parsimonious model. We then show the spatial factors for the AIC-selected model, as well as the

analytic and sample correlation for both the SFA and when running each species individually. We also compare the predicted density of each species with counts at the same set of sampling locations in the summer of the following year (2013), as a measure of predictive accuracy. Specifically, we compute log-expected counts Λ_i (using only data from 2012) for each sample location in 2013. We then compare these predictions with true counts c_i in 2013 and calculate the rank (Spearman) correlation between predictions (using data in year 2012) and true counts (in year 2013).

SECOND CASE STUDY – PACIFIC ROCKFISHES

We next apply SFA to 10 years of data (2003–2012) from a multi-species survey of marine fishes off the U.S. West Coast, and restrict our analysis to 20 species of Pacific rockfishes (*Sebastes* spp.) to demonstrate the application of SFA for species within a single taxonomic group (and hence presumably with more similar spatial distributions than in the bird case study). This survey uses a stratified random design to allocate approximately 650 bottom trawl tows annually along the entire coast and identifies all fishes caught to species so that it yields a count c_i for potentially hundreds of species simultaneously. We analyse data for each year independently, which allows us to estimate species distributions using data from a particular year (e.g. 2003), and predicting catches for those species in the next year (e.g. 2004). This forward prediction is the gold standard for model evaluation and gives us 9 years of forward prediction with which to evaluate model performance. For each year, we only fit the model to data for species that have at least 25 observations of the species, and this leaves between 16 and 20 species (where the precise number varies somewhat from year to year). For each year, we fit the SFA model using from 1 to 8 spatial factors ($1 \leq K \leq 8$), and use marginal AIC to select the best-fitting number of factors. We take this AIC-selected model, fitted to data for all 16–20 species for that year, and again compare predictions of log-expected catch with true catches c_i to calculate the rank (Spearman) correlation between predictions (using data in year t) and true catches (in year $t + 1$). This

Table 1. List of species used in the two case studies

| Selected birds in Western N. America		Pacific rockfishes	
Scientific	Common	Scientific	Common
Pipilo chlorurus	Green-tailed Towhee	*Sebastes alutus*	Pacific ocean perch
Oreoscoptes montanus	Sage Thrasher	*S. babcocki*	Redbanded rockfish
Amphispiza bilineata	Black-throated Sparrow	*S. crameri*	Darkblotched rockfish
Calypte anna	Anna's hummingbird	*S. diploproa*	Splitnose rockfish
Selasphorus rufus	Rufous hummingbird	*S. elongatus*	Greenstriped rockfish
Ardea alba	Great egret	*S. entomelas*	Widow rockfish
Circus cyaneus	Northern harrier	*S. flavidus*	Yellowtail rockfish
Melanerpes formicivorus	Acorn woodpecker	*S. helvomaculatus*	Rosethorn rockfish
Dryocopus pileatus	Pileated woodpecker	*S. jordani*	Shortbelly rockfish
Ammodramus savannarum	Grasshopper sparrow	*S. paucispinis*	Bocaccio
		S. pinniger	Canary rockfish
		S. zacentrus	Sharpchin rockfish
		S. melananostictus aleutianus	Rougheye complex
		S. saxicola	Stripetail rockfish
		S. aurora	Aurora rockfish
		S. melanostomus	Blackgill rockfish
		S. chlorostictus	Greenspotted rockfish
		S. goodei	Chilipepper
		S. semicinctus	Halfbanded rockfish
		S. levis	Cowcod
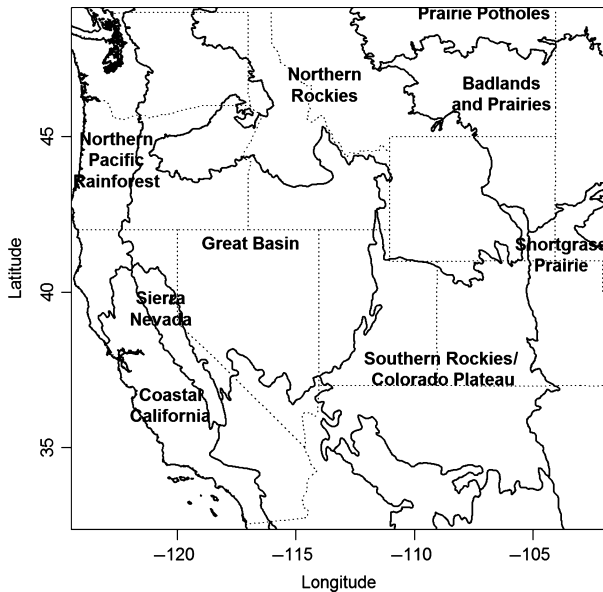
Fig. 1. Bird conservation regions across the western U.S.A., as defined by the North American Bird Conservation Initiative (Babcock *et al.* 1998).

is repeated for each of 9 years and, in total, we have 163 predictive evaluations of the SFA model.

MODEL EVALUATION

Finally, we compare the SFA model with a conventional geostatistical model that analyses data for each species separately. This model approximates single-species catches using a species-specific spatial field:

$$C_i \sim \text{Poisson}(\exp\{\Lambda_i + \varepsilon_i\}, \theta)$$
$$\Lambda_i = \alpha + \omega_i$$
$$\omega \sim \text{MVN}(0, \Sigma)$$
$$\varepsilon_i \sim \text{Norm}(0, \sigma^2)$$

eqn 13

where the covariance of the spatial field is as defined previously. This single-species model is fitted to data for all rockfish and bird species in each case study and each year, and its predictions of log-expected catches Λ_i for each sample location in the next year is compared with observed catches c_i. Predictions and true counts are again used to calculate the rank (Spearman) correlation between predictions and true catches in that year. This rank correlation is then used as a basis for comparing the relative performance of SFA and conventional geostatistical approaches to species distribution models. Future research could also explore model goodness-of-fit and parsimony using predictive scores (Gelman, Hwang & Vehtari 2014) or other criteria. We use rank correlation to evaluate model performance because we are primarily interested in the ability of spatial models to broadly reconstruct areas of high and low density, and we hypothesize that good performance as measured by predictive scores will require more detailed treatment of residual errors for count data (e.g. using zero inflation and heteroskedastic variance inflation) that we use here. Finally, we also evaluate model fit by calculating the per cent deviance explained (PDE) for both single-species and SFA models. To do so, we fitted a 'null' model that does not include spatial variation (i.e. where **L = 0**). We then calculate PDE as the deviance of the 'null' model, minus the deviance of the SFA or the combination single-species models, and then divided by the deviance of the 'null' model.

Results

We found that the spatial distributions of 10 different bird species based on counts from across the western USA could be explained with only five spatial factors (Fig. 2a) and that these five spatial factors line up well with bird conservation regions previously defined by the NABCI (i.e. comparing Figs 1 and 2a). Factors 1 and 3 (see Fig. 2b for species loadings) are (negatively) associated with the Great Basin region, while Factor 2 is (negatively) associated with the Southern Rockies. Factor 4 is associated with the northern Pacific rain forests that dominate the coast of Washington and Oregon, whereas Factor 5 is associated with the low-precipitation areas in California and Nevada. The subsequent species loadings (Fig. 2b) result in positive cross-correlations between (i) the three sagebrush obligates: green-tailed towhee, sage thrasher and black-throated sparrow, and (ii) the primarily coastal species: Anna's hummingbird, Rufous hummingbird and great egret. The PDE for the SFA model was 19·4%, compared with 17·7% for the combination of single-species models. The rank correlation of both models was similarly close (median rank correlation $\rho = 0·362$ for the SFA model, and $\rho = 0·365$ for the single-species models), indicating little improvement in predictive importance for SFA for these species relative to single-species distribution models.

Spatial factor analysis also revealed several interpretable spatial patterns in catches of Pacific rockfishes in 2003 (Fig. 3a,b). AIC model selection indicated that a 6-factor model was far more parsimonious (i.e. approx. 1600 AIC units less) than either (i) a single-factor SFA model, or (ii) the conventional geostatistical approach, where each species has an independent spatial factor (Table 2). Among spatial factors (Fig. 3a), several can be easily interpreted, including a factor (negatively) associated with a band of intermediate depth running north–south along the coast (Factor 2), a factor discriminating between northern and southern rockfishes (Factor 3), a factor (negatively) associated with rockfishes primarily off the southern Oregon coast (Factor 4) and a factor associated with elevated densities in nearshore California environments (Factor 5). Different linear combinations of these factors (**LΩ**) can generate a wide diversity of spatial distributions (Fig. 4), where most species (except aurora) are negatively associated with Factor 2, but factors 1 and 3–6 generally include both positive and negative loadings (Fig. 3b) and hence discriminate among species. For Pacific rockfishes in 2003, the SFA model had a PDE of 17·6%, compared with a PDE of 7·2% for the combination of single-species models.

Calculation of between-species correlations (Fig. 5) shows that the analytic computations (upper-left panel of Fig. 5a,b) show a strong resemblance to the sample correlation in **Ω** (upper-right panel of 5a,b). As expected from the loading matrix **L'** for the rockfish example, aurora has a spatial distribution that differs strongly from most other species (in particular, canary and yellowtail). Meanwhile, the standard single-species geostatistical approach implicitly assumes that all species are statistically independent (correlation = 0 for off-diagonal; lower-left panel in Fig. 5a,b). However, the
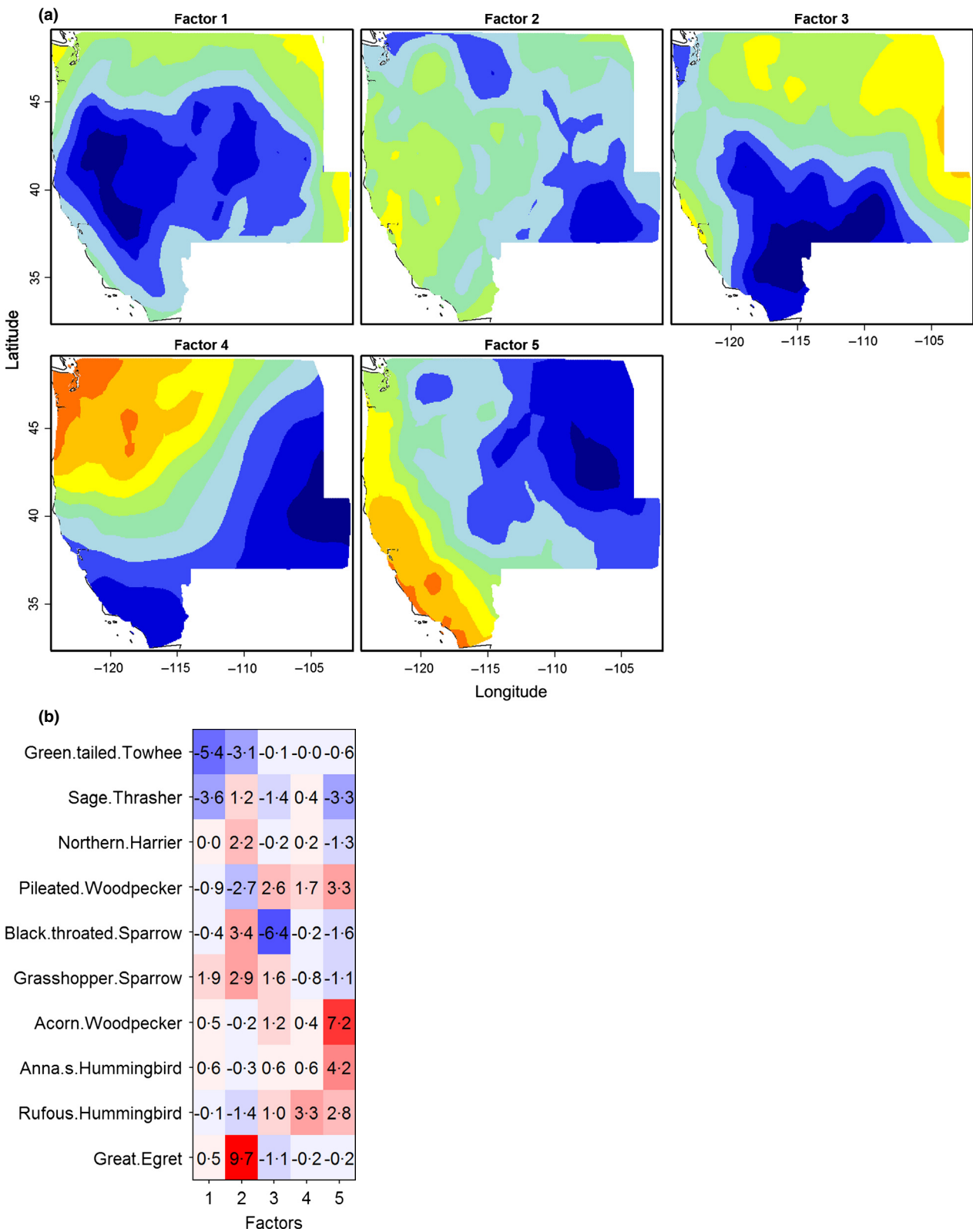
Fig. 2. Results for the spatial factor analysis model applied to breeding bird survey data for 10 species in 2013 (see main text for list): (a) the estimated factors after varimax rotation ($\mathbf{\Omega}'$, where red signifies high and blue signifies low values), (b) the varimax-rotated loadings matrix ($\mathbf{L}'$).

single-species estimates of $\mathbf{\omega}_j$ for all species 1 through J are still correlated among species. Interestingly, the sample correlations for the single-species models are very different than the sample or analytic correlations for SFA in the rockfish example (Fig. 5a), but less so for the North American bird example.
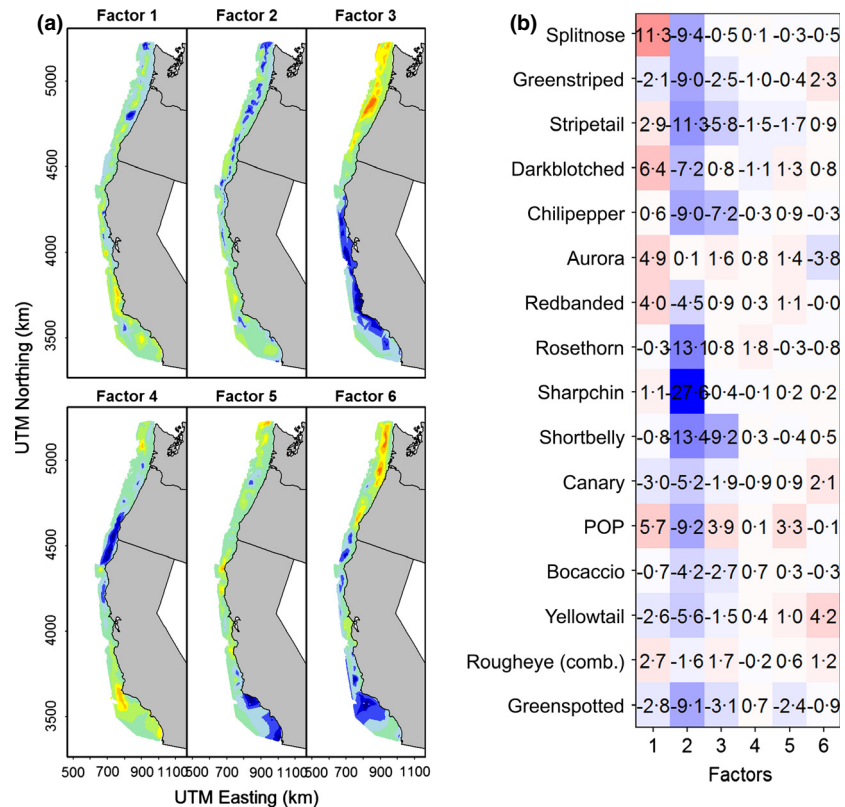
Fig. 3. Results for the spatial factor analysis model applied to Pacific rockfish data for 16 species in 2003 (see main text for list): (a) the estimated factors after varimax rotation ($\mathbf{\Omega}'$, where red signifies high and blue signifies low values), (b) the varimax-rotated loadings matrix ($\mathbf{L}'$).

When using cross-validation in 2003 to evaluate model performance for the Pacific rockfish case study (i.e. comparing model predictions when fitting to data in 2003 to the true catches in 2004; Table 2), the median rank correlation across all 16 species was maximized by a 5-factor model ($\rho = 0.337$), but it was essentially the same as that for the AIC-ranked best model with six factors ($\rho = 0.335$). Notably, both of these rank correlations were substantially higher than the median rank correlation across all 16 species for the single-species geostatistical models ($\rho = 0.296$). Across all 163 combinations of species and year (Fig. 6), SFA generally provides greater rank correlation (median: 0.329) than the single-species models (median: 0.285) and also decreases the number of species and years with a correlation <0.2. We therefore conclude that AIC identifies a number of factors that has reasonable performance for predictive purposes. We also conclude that SFA has better predictive performance than the single-species models for Pacific rockfishes. We hypothesize that differences in predictive performance between case studies arise because the Pacific rockfishes generally have a strong positive correlation (and hence a large amount of mutual information), while the bird species generally have fewer pairs with strong positive or negative correlation (and hence less mutual information).

Table 2. Summary of goodness-of-fit for rockfish data in 2003 when predicting data in 2004, that is model selection criteria (Akaike Information Criterion, applied to marginal likelihood), and predictive accuracy (in-bag: Pearson correlation between model fits and data used to fit the model; out-of-bag: Pearson correlation between model predictions and next year's data) for single-species models and the spatial factor analysis model using 1–7 factors

Factors	Model selection ΔAIC	Correlation Predicting 2003 data ('in-bag prediction')	Predicting 2004 data ('out-of-bag prediction')
Single-species models			
NA	1620·7	0·529	0·296
Spatial factor analysis			
1	1273·0	0·374	0·207
2	509·0	0·451	0·29
3	142·9	0·511	0·297
4	17·2	0·509	0·333
5	0·4	0·513	0·337
6	0	0·513	0·335
7	8·1	0·511	0·331

Discussion

We have developed a new spatial analogue of conventional factor analysis, SFA, and shown how it can be used to estimate joint distributions for multiple species using a small number of spatial factors. In the style of principal components analysis or other dimension-reducing techniques that simplify large data sets (McCune, Grace & Urban 2002), SFA incorporates data from a variety of species, summarizes the numerous underlying landscape factors that drive their distributions and presents the results as a reduced series of maps. By analysing all species simultaneously, we have shown that the improved parsimony of SFA (relative to individual species distribution models) in some cases translates to improved precision in predictions of
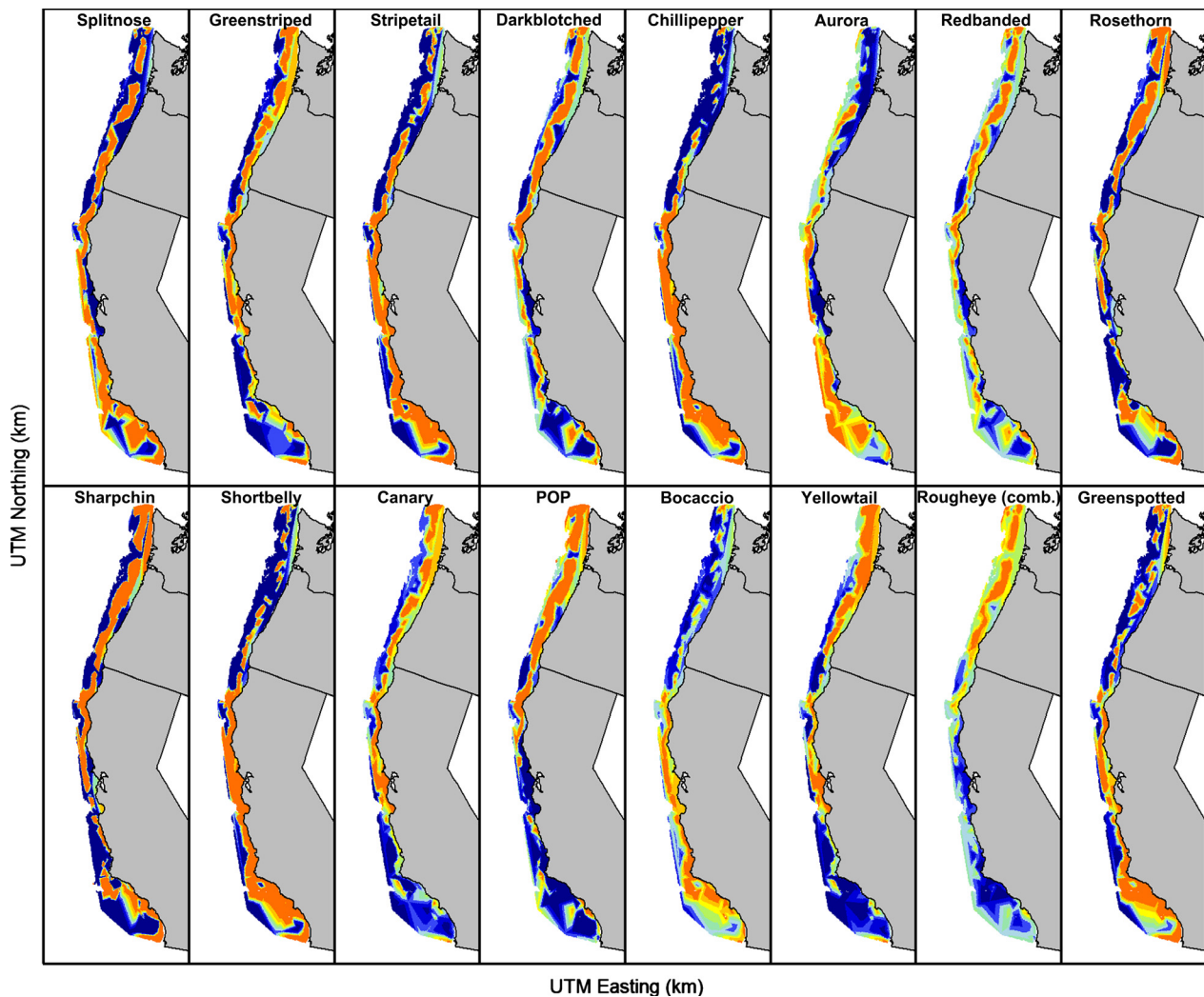
Fig. 4. Expected log-abundance (standardized to have mean zero for each species, that is, **LΩ**$_i$, and having the same colour-scale for all panels) for all 16 rockfishes that have 25 or more encounters in 2003 (red is positive values, and blue is negative values).

future survey data. This confirms results from previous studies that share information among species to achieve improved predictions or explanatory power (Ovaskainen & Soininen 2011; Clark *et al.* 2013; Pollock *et al.* 2014). SFA has the additional benefit of providing an analytical form for cross-correlation, which complements previous work for calculating cross-correlation due to measured covariates and correlated residuals (Pollock *et al.* 2014).

Spatial maps of species abundance and diversity form the basis for biogeographical studies and conservation planning (Graham & Hijmans 2006; Spalding *et al.* 2007; Tittensor *et al.* 2010). Examination of species distribution maps is typically the first step in designing protected areas or reserves that will protect the greatest level of biodiversity (Margules & Pressey 2000). A variety of techniques exist for estimating species distribution and density maps, ranging from expert opinion to logical and statistical models (Johnson, Seip & Boyce 2004; Graham & Hijmans 2006), although it remains difficult to synthesize all available information or determine which method is optimal for a given planning task. For

example, simple range maps may suffice for determining simple presence or absence, but they are often insufficient for conservation planning because they assume a uniform distribution within the region of interest (Gaston & Rodrigues 2003; Williams *et al.* 2014). By contrast, SFA (like other joint species distribution models) produces a rank-reduced, comprehensive map of multispecies densities. This allows for easy identification of dominant spatial patterns in densities that could, in turn, be incorporated into formal spatial planning. Furthermore, errors in distribution maps are rarely acknowledged, which can plague conservation decision-making (Tulloch *et al.* 2013). SFA differs from other JSDMs by not including measured covariates to predict individual species (Clark *et al.* 2013; Pollock *et al.* 2014) or species archetypes (Dunstan, Foster & Darnell 2011; Hui *et al.* 2013). Bahn & McGill (2007) highlighted the good performance of spatial models like SFA when imputing between sampled locations, but we hypothesize that JSDMs that include measured covariates will often perform better than SFA when extrapolating outside the sampled domain.
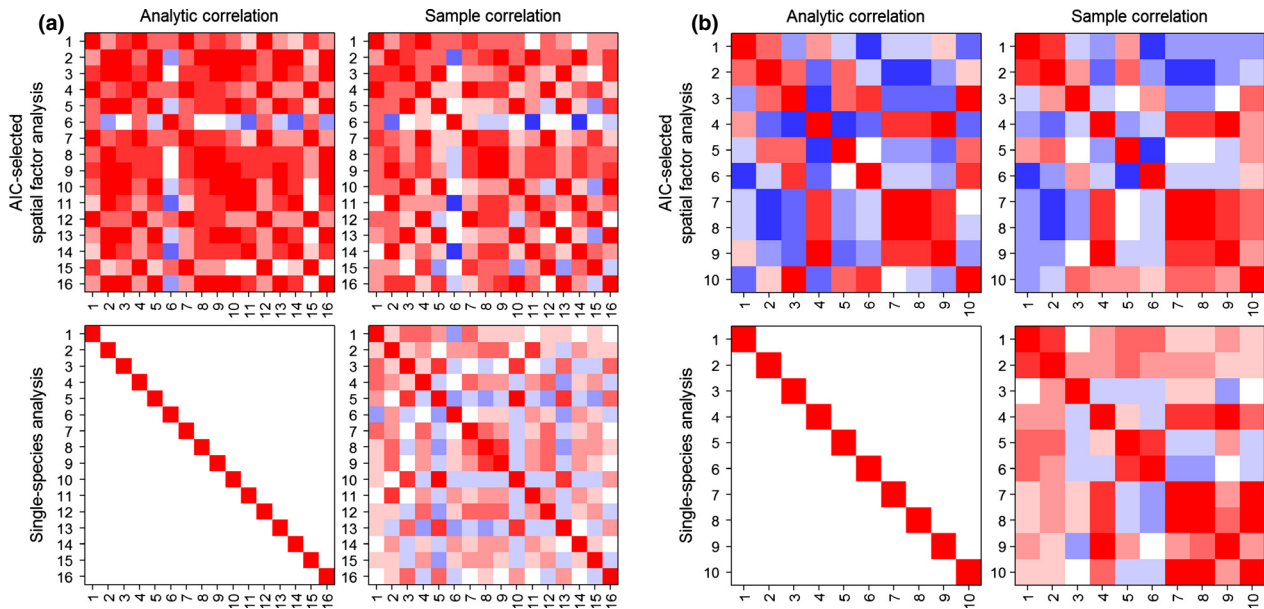
Fig. 5. Estimated correlations among species using rockfish data from 2003 (a) and breeding bird survey data from 2013 (b), where red is a correlation of 1, white is a correlation of 0, and blue is a correlation of −1 (shading is used on a gradient between these colours). The top row shows the estimated correlations from the spatial factor analysis, while the bottom row shows correlations from each single-species geostatistical model. The left column shows analytic estimates of correlations, while the right column for each shows the sample correlation (see Methods section for details).
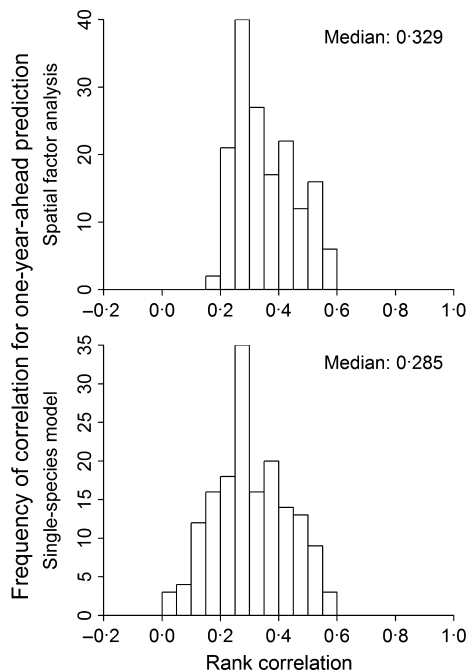

Fig. 6. Distribution of Spearman correlations between AIC-selected spatial factor analysis estimates of spatial distribution from each year of the Pacific rockfish case study, and the realized catches in the next year. Also shown is the same distribution when using single-species geostatistical models to predict next year's catches.

In the USA and elsewhere, marine spatial planning efforts are underway to balance a variety of stakeholder interests such as commercial fishing and ecotourism (White, Halpern & Kappel 2012; Rassweiler *et al.* 2014). Spatial factor analysis could prove to be a valuable tool in these endeavours for two reasons. First, many species distributions are estimated from statistical relationships with habitat or other descriptions of the environment (Guisan & Thuiller 2005; Stelzenmüller, Ellis & Rogers 2010). Spatial factor analysis does not require any pre-measured environmental covariates with which to estimate species abundance (although covariates could be easily added) and instead allows the data to highlight any perceived biogeographical features (e.g. Point Conception in the California Transition Zone). Secondly, conservation planning is usually focused on static notions of abundance patterns rather than the dynamic processes that generated them (Pressey *et al.* 2007). Spatial factor analysis provides a better prediction of 1-year ahead spatial patterns than an *ad hoc* combination single-species models for Pacific rockfishes and hence may be more suitable for identifying changes in their distribution over time.

We have also shown that SFA provides an analytical estimate of cross-correlation, where cross-correlation has been frequently discussed in ecology, for example regarding community assembly rules (Gotelli & McCabe 2002). The rank reduction used by SFA will probably be parsimonious in many applications and will likely improve estimates of the cross-correlation matrix in these cases. The ability to provide an analytic estimate of cross-correlation enables future modifications, such as using species-level covariates to predict or inform species' cross-correlations. One potential avenue of particular interest would be using species traits to generate one or more trait- or phylogeny-based dissimilarity matrix (Kissling *et al.* 2012; Oke, Heard & Lundholm 2014). This dissimilarity matrix could then be used as a covariate when modelling species cross-correlations. In this way, information regarding species traits could be used directly when modelling the

spatial distribution of species within a given community. This strategy could potentially be used in a hypothesis-testing design, to determine whether the similarity of species traits is a better predictor of spatial distributions than expected at random.

We note that single-species geostatistical models provided a very different picture of species co-occurrence than did the SFA model. In particular, the sample cross-correlation from single-species models had many more negative elements than did the SFA when applied to data for Pacific rockfishes, and would lead to inference that different environmental associations and/or negative interactions were stronger in the single-species than the SFA model. The interpretation of spatial cross-correlations has a long history in ecology. For example, spatial cross-correlations have been used to infer drivers of bird population eruptions (Bock & Lepthien 1976), causes of fungal co-occurrence in different habitats (Ovaskainen, Hottola & Siitonen 2010) and providing evidence for species assembly rules (Gotelli & McCabe 2002). Given that researchers have advocated increased use of multispecies distribution models to infer correlations in species distribution (e.g. Kissling *et al.* 2012), the fact that estimated structure of cross-correlations depend strongly on the method used to estimate it is an important result. Which model provides a better picture of species interactions will require future simulation testing of the SFA model. However, the SFA model has better performance at predicting future catches, so we hypothesize that it will have superior performance in other characteristics as well.

Future research can combine GRFs and measured environmental variables in joint species distribution models (Latimer *et al.* 2009; Ovaskainen, Hottola & Siitonen 2010; Clark *et al.* 2013; Pollock *et al.* 2014; Harris 2015), although we have not done so in this analysis. This decision was made to ensure that our model is a strict analogue of conventional factor analysis (which does not include measured covariates). Nevertheless, it would be easy to modify the equations for the expected value of each species to include covariates, and cross-correlations between random fields and measured covariates could be calculated following Pollock *et al.* (2014). We recommend that future joint species distribution models include measured covariates (as is possible in our R package *SpatialFA*) because these are likely to improve predictive accuracy, particularly when predicting species densities in regions with low sampling intensity or outside the sampling domain of the original data set (Shelton *et al.* 2014). However, in many cases the covariates may not be available for the entire spatial domain (e.g. the presence of biogenic habitats in marine systems; Krigsman *et al.* 2012) or an appropriate parametric relationship between the covariate and species may be unknown. In such cases, the flexibility provided by SFA is preferable to estimating an *ad hoc* functional form. A low-rank approximation to cross-correlations may also be important for JSDMs that already estimate cross-correlations in residuals (Clark *et al.* 2013; Pollock *et al.* 2014). We note that improved precision could also be achieved in some cases by pooling data across multiple time periods and hence increasing sample sizes (e.g. by combining data among all years 2003–2012 in the rockfish example). However, combining data from multiple years will only be appropriate if spatial distributions do not change over time, or changes can be modelled as an annual offset without changes in the relative distribution among areas.

Finally, we advocate further research regarding joint species distribution models that estimate changes over time (Kissling *et al.* 2012). Dynamic spatial models for multiple species could generalize non-spatial approaches for inferring competitive interactions (Ives *et al.* 2003), and the spatial approach is likely important when densities vary spatially and interactions are local (Thorson *et al.* In press). In the case that distributions change among years, we suspect that a fully spatiotemporal approach is more appropriate than pooling all years in a static distribution model. However, dynamic joint species distribution models will require efficient use of available data, which may be benefited by the rank-reduced method illustrated by the SFA model. Therefore, a dynamic SFA model would be an appropriate next step for research when combining data from multiple species and years to estimate changes in species distributions and co-occurrence. This model has been explored outside ecology (Calder 2007; Lopes *et al.* 2012), but remains an important tool to test for joint species distribution modelling.

Data accessibility

The bird case study analyses data for 10 species and is obtainable from the Breeding Bird Survey Website (https://www.pwrc.usgs.gov/bbs/). The fish case study analyses data for 20 species over 10 years and is obtainable upon request from the Fisheries Resource and Monitoring Division of the Northwest Fisheries Science Center (beth.horness@noaa.gov).

References

Babcock, K., Baxter, C., Brown, S., Ford, B., Johnson, R., Leach, J. *et al.* (1998). *A Proposed Framework for Delineating Ecologically-Based Planning, Implementation, and Evaluation Units for Cooperative Bird Conservation in the U.S.* North American Colonial Waterbird Conservation Plan, Puebla, Mexico.

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Bock, C.E. & Lepthien, L.W. (1976) Synchronous eruptions of boreal seed-eating birds. *American Naturalist*, **110**, 559–571.

Brown, J.H. (1999) Macroecology: progress and prospect. *Oikos*, **87**, 3–14.

Calder, C.A. (2007) Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environmental and Ecological Statistics*, **14**, 229–247.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K.(2013). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, **24**, 990–999.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. *Ecological Modelling*, **222**, 955–963.

Elton, C.S. (1927). *Animal Ecology*. University of Chicago Press, Chicago, IL.

Gaston, K.J. (1991) How large is a species' geographic range? *Oikos*, **61**, 434–438.

Gaston, K.J. & Rodrigues, A.S. (2003) Reserve selection in regions with poor biological data. *Conservation Biology*, **17**, 188–195.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J., Gregory, R.D., Quinn, R.M. & Lawton, J.H. (2000) Abundance–occupancy relationships. *Journal of Applied Ecology*, **37**, 39–59.

Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

Goetz, S.J., Sun, M., Zolkos, S., Hansen, A. & Dubayah, R. (2014) The relative importance of climate and vegetation properties on patterns of North American breeding bird species richness. *Environmental Research Letters*, **9**, 034013.

Gotelli, N.J. & McCabe, D.J. (2002) Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model. *Ecology*, **83**, 2091–2096.

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.

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

Hanski, I. (1998) Metapopulation dynamics. *Nature*, **396**, 41–49.

Harris, D.J. (2015) Generating realistic assemblages with a Joint Species Distribution Model. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12332

Harville, D.A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Hogan, J.W. & Tchernis, R. (2004) Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, **99**, 314–324.

Holmes, E.E., Ward, E.J. & Scheuerell, M.D. (2014) *Analysis of Multivariate Time-Series Using the MARSS Package*. http://mirror.ufs.ac.za/cran/web/packages/MARSS/vignettes/UserGuide.pdf (accessed 23 January 2015).

Hui, F.K., Warton, D.I., Foster, S.D. & Dunstan, P.K. (2013) To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology*, **94**, 1913–1919.

Ives, A.R., Dennis, B., Cottingham, K.L. & Carpenter, S.R. (2003) Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, **73**, 301–330.

Johnson, C.J., Seip, D.R. & Boyce, M.S. (2004) A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, **41**, 238–251.

Kaldhusdal, A., Brandl, R., Müller, J., Möst, L. & Hothorn, T. (2015) Spatio-phylogenetic multispecies distribution models. *Methods in Ecology and Evolution*, **6**, 187–197.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J. *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Krigsman, L.M., Yoklavich, M.M., Dick, E.J. & Cochrane, G.R. (2012) Models and maps: predicting the distribution of corals and other benthic macro-invertebrates in shelf habitats. *Ecosphere*, **3**, 1–16.

Kristensen, K. (2014). *TMB: General Random Effect Model Builder Tool Inspired by ADMB*. https://github.com/kaskr/adcomp.

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

Latimer, A.M., Banerjee, S., Sang Jr, H., Mosher, E.S. & Silander Jr, J.A. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters*, **12**, 144–154.

Lindén, A. & Mäntyniemi, S. (2011) Using negative binomial distribution to model overdispersion in ecological count data. *Ecology*, **92**, 1414–1421.

Lindgren, F. & Rue, H. (2015). Bayesian spatial and spatiotemporal modelling with r-inla. *Journal of Statistical Software*, **63**, 1–25.

Lindgren, F., Rue, H. & Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Lopes, H.F., Schmidt, A.M., Salazar, E., Gómez, M. & Achkar, M. (2012) Measuring the vulnerability of the Uruguayan population to vector-borne diseases via spatially hierarchical factor models. *The Annals of Applied Statistics*, **6**, 284–303.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2005) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, Burlington, MA.

Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–253.

McCune, B., Grace, J.B. & Urban, D.L. (2002). *Analysis of Ecological Communities*. MjM software design, Gleneden Beach, OR.

Oke, O.A., Heard, S.B. & Lundholm, J.T. (2014) Integrating phylogenetic community structure with species distribution models: an example with plants of rock barrens. *Ecography*, **37**, 614–625.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–2521.

Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A. & McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.

Pressey, R.L., Cabeza, M., Watts, M.E., Cowling, R.M. & Wilson, K.A. (2007) Conservation planning in a changing world. *Trends in Ecology & Evolution*, **22**, 583–592.

R Core Development Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rassweiler, A., Costello, C., Hilborn, R. & Siegel, D.A. (2014) Integrating scientific guidance into marine spatial planning. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20132252.

Rencher, A.C. (2002) *Methods of Multivariate Analysis*, 2nd edn. Wiley-Interscience, Hoboken, New Jersey.

Royle, J.A. & Wikle, C.K. (2005) Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics*, **12**, 225–243.

Shelton, A.O., Thorson, J.T., Ward, E.J. & Feist, B.E. (2014) Spatial semiparametric models improve estimates of species abundance and distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**, 1655–1666.

Skaug, H. & Fournier, D. (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, **51**, 699–709.

Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M.A.X. *et al.* (2007) Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. *BioScience*, **57**, 573–583.

Stelzenmüller, V., Ellis, J.R. & Rogers, S.I. (2010) Towards a spatially explicit risk assessment for marine management: assessing the vulnerability of fish to aggregate extraction. *Biological Conservation*, **143**, 230–238.

Thorson, J.T., Skaug, H., Kristensen, K., Shelton, A.O., Ward, E.J., Harms, J. & Benante, J. (In press) The importance of spatial models for estimating the strength of density dependence. *Ecology*. doi:10.1890/14-0739.1

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V. & Worm, B. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature*, **466**, 1098–1101.

Tulloch, V.J., Possingham, H.P., Jupiter, S.D., Roelfsema, C., Tulloch, A.I. & Klein, C.J. (2013) Incorporating uncertainty associated with habitat data in marine reserve design. *Biological Conservation*, **162**, 41–51.

Wang, F. & Wall, M.M. (2003) Generalized common spatial factor model. *Biostatistics*, **4**, 569–582.

White, C., Halpern, B.S. & Kappel, C.V. (2012) Ecosystem service tradeoff analysis reveals the value of marine spatial planning for multiple ocean uses. *Proceedings of the National Academy of Sciences*, **109**, 4696–4701.

Williams, R., Grand, J., Hooker, S.K., Buckland, S.T., Reeves, R.R., Rojas-Bracho, L., Sandilands, D. & Kaschner, K. (2014) Prioritizing global marine mammal habitats using density maps in place of range maps. *Ecography*, **37**, 212–220.

Zuur, A.F., Tuck, I.D. & Bailey, N. (2003) Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 542–552.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Derivation of covariance between species.