Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 344–352 doi: 10.1111/2041-210X.12163

The fourth-corner solution – using predictive models to understand how species traits interact with the environment

Alexandra M. Brown[1]*, David I. Warton[1], Nigel R. Andrew[2], Matthew Binns[2], Gerasimos Cassis[3] and Heloise Gibb[4]

[1]*School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052, Australia;* [2]*Centre for Behavioural and Physiological Ecology, Discipline of Zoology, School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia;* [3]*Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, NSW 2052, Australia; and* [4]*Department of Zoology, La Trobe University, Melbourne, Vic. 3068, Australia*

Summary

1. An important problem encountered by ecologists in species distribution modelling (SDM) and in multivariate analysis is that of understanding why environmental responses differ across species, and how differences are mediated by functional traits.

2. We describe a simple, generic approach to this problem – the core idea being to fit a predictive model for species abundance (or presence/absence) as a function of environmental variables, species traits and their interaction.

3. We show that this method can be understood as a model-based approach to the fourth-corner problem – the problem of studying the environment–trait association using matrices of abundance or presence/absence data across species, environmental data across sites and trait data across species. The matrix of environment–trait interaction coefficients is the fourth corner.

4. We illustrate that compared with existing approaches to the fourth-corner problem, the proposed model-based approach has advantages in interpretability and its capacity to perform model selection and make predictions.

5. To illustrate the method we used a generalized linear model with a LASSO penalty, fitted to data sets from four different studies requiring different models, illustrating the flexibility of the proposed approach.

6. Predictive performance of the model is compared with that of fitting SDMs separately to each species, and in each case, it is shown that the trait model, despite being much simpler, had comparable predictive performance, even significantly outperforming separate SDMs in some cases.

Key-words: environment–trait association, fourth-corner problem, LASSO, multivariate analysis, predictive modelling, RLQ analysis, species distribution model

Introduction

The ecological literature is rich in methods predicting species distribution and abundance based on environmental variables, commonly known as species distribution models ('SDM', Elith & Leathwick 2009). When modelling the distribution of a set of co-occurring species, a common approach is to model each species separately as a function of environmental variables. This approach allows different species to be modelled as having different environmental responses, but it fails to model how and why species differ in these responses. The traits of a species, its behaviours, physiology and morphology, help define how each species responds to the environment (McGill *et al.* 2006). Yet, there are currently few methods in the literature for modelling how environmental response is mediated by species traits – typically, previous researchers (e.g. Thuiller *et al.* 2004; Pottier, Marrs & Bedecarrats 2007; Kleyer *et al.* 2012 'OMI-GAM') have used an indirect, two-stage approach, whether modelling species separately then using traits to characterize differences in properties of the fitted species models, or constructing community-aggregated trait measures and relating those to the environment (*e.g.* Kühn *et al.* 2006; Suding *et al.* 2008). But the effect of traits on environmental response is direct (McGill *et al.* 2006) and what is needed is an approach to SDM ling which can incorporate traits directly. This article describes such a method and relates it to the fourth-corner problem – thus bridging a gap between the SDM ling and multivariate analysis literatures in ecology.

Within the multivariate analysis literature in ecology, the problem of associating species traits and environmental variables using species abundance or presence/absence data is known as the fourth-corner problem (Legendre, Galzin &

*Correspondence author. E-mail: alexandrabrown@fas.harvard.edu

Harmelin-Vivien 1997). The fourth-corner problem can be thought of as a three table problem, which takes matrices of environmental data (**R**), species abundance or presence/absence data (**L**) and species trait data (**Q**), and uses these three tables to infer how species traits relate to the environment (**D**), as illustrated in Fig. 1a. For example, Tatibouet (1981) collected presence/absence data to study bird communities in rural and urban environments, to determine how bird species with different traits varied in prevalence along this range (available in the ADE4 package, Chessel, Dufour & Thiulouse 2004). Two strategies for the fourth-corner problem have previously been proposed (Doledec *et al.* 1996; Legendre, Galzin & Harmelin-Vivien 1997). These will be briefly reviewed below and applied to the bird data set of Tatibouet (1981) in Fig. 2 (using a subset of environmental and trait variables for simplicity).

RLQ analysis is an exploratory ordination approach to the fourth-corner problem (Doledec *et al.* 1996). Pairs of ordinations are jointly constructed via a generalized singular value decomposition, which can be understood as jointly relating each of the environmental variables and species traits to species abundance or presence/absence. This can provide a broad qualitative overview of how traits and environmental variables are associated, as opposed to specific details. When applying RLQ analysis to Tatibouet's (1981) bird data, Doledec *et al.* (1996) describe the trait–environment relationship as a rural to urban gradient (Fig. 2a). The urban end is characterized by variables and traits such as building presence and aerial feeding, while the rural end includes field presence and breeding on the ground (Fig. 2a). We can get a sense from Fig. 2a, of which species traits tend to occur in which environments, but we cannot extract specific details on the significance or quantitative nature of those trends.

A hypothesis testing approach to the fourth-corner problem has also been proposed (Legendre, Galzin & Harmelin-Vivien 1997). In this approach, the fourth-corner problem is considered as a matrix algebra problem, and **R**, **L** and **Q** are used to determine a matrix **D** (the matrix product of **R**'**LQ**) which summarizes the association between environmental variables and species traits. This matrix is constructed in different and somewhat *ad hoc* ways depending on the properties of **R**, **Q** (Legendre, Galzin & Harmelin-Vivien 1997) and **L** (Dray & Legendre 2008). Permutation tests are then used to test hypotheses about the environment–trait association, although it remains unclear precisely how to undertake permutation testing in a valid way – current proposals (Dray & Legendre 2008; ter Braak, Cormont & Dray 2012) break the connection between species and traits and/or sites and environmental variables as well as the environment–trait connection. This hypothesis testing approach to the fourth-corner problem will tell the user which combinations of environmental variables and species traits are significantly associated with each other. However, it does not return information on the strength of the association, only its significance. For example, when applying the method to the bird data of Tatibouet (1981) as in Fig. 2b, presence or absence of small buildings is significantly associated with feeding in foliage, but we do not know the impact this has on species nor the strength of the association.

In this article, we describe and evaluate a model-based approach to the fourth-corner problem, essentially, an approach to SDM ling inspired by the fourth-corner literature. It is similar to a method proposed by Jamil *et al.* (2013) and Pollock, Morris and Vesk (2012) independently of the current authors (Brown 2010). The main points of distinction in our article are as follows: we propose the model in a more general form, applicable with any SDM ling approach; we evaluate its predictive performance on four multivariate data sets; we make new connections to the fourth-corner problem, connecting the SDM ling and multivariate analysis literatures in ecology. This intuitive new method complements the ordination (Doledec *et al.* 1996) and hypothesis testing (Legendre, Galzin & Harmelin-Vivien 1997; Dray & Legendre 2008) approaches already in the literature, by allowing the user to quantify the strength and nature of environment–trait associations (*interpretability*), to identify which environmental and trait variables are important to species abundance (*model selection*), and to predict species abundance in new scenarios (*prediction*), including under scenarios of environmental change.

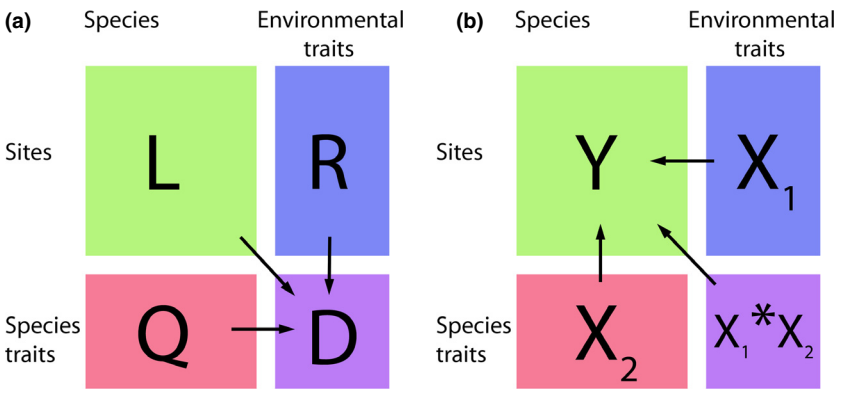

Fig. 1. Graphical representation of the fourth-corner problem and its solution. (a) The problem as posed by Legendre, Galzin & Harmelin-Vivien (1997), where the goal is to combine abundance (**L**), trait (**Q**) and environment (**R**) data in some way, to determine a matrix describing the trait–environment relationship (**D**). (b) The proposed model-based solution to the fourth-corner problem, where the goal is to predict abundance (**Y**) as a function of predictor variables for environment ($\mathbf{X}_1$), species traits ($\mathbf{X}_2$) and their interaction ($\mathbf{X}_1$*$\mathbf{X}_2$). The matrix of coefficients for the interaction between $\mathbf{X}_1$ and $\mathbf{X}_2$ is the fourth corner.

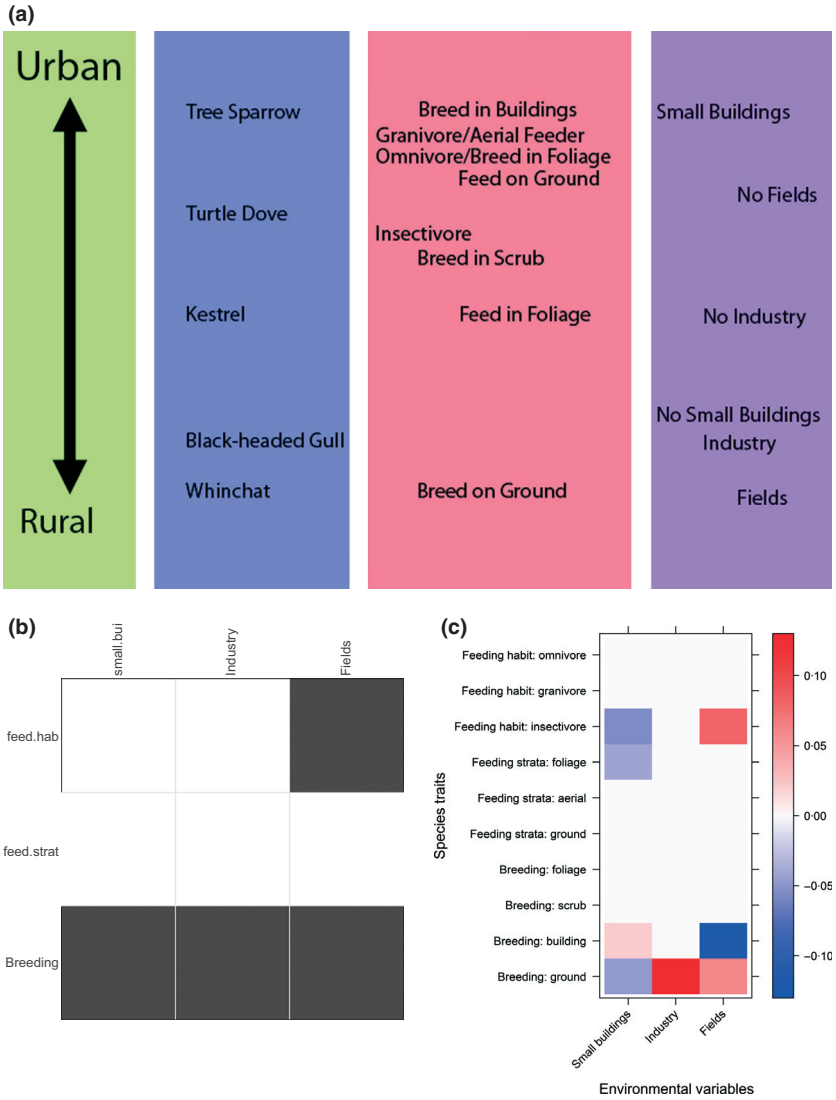

Fig. 2. Fourth-corner analysis results for the bird data of Tatibouet (1981). (a) RLQ ordination. The first ordination axis is plotted, to show how different environmental, species trait and species abundance variables are related to the urban–rural gradient. (b) Fourth-corner hypothesis testing (Legendre, Galzin & Harmelin-Vivien 1997). Highlighted terms are statistically significant after multiple testing (Holm's step-down method). (c) Fourth-corner modelling results. Standardized coefficients for all environment–trait interaction terms are presented, from a generalized linear models (GLM)-LASSO model. Brighter squares show stronger associations than paler ones, positive associations are red and negative associations are blue.

Model-based approach

The approach proposed in this article is to fit a single predictive model for abundance (or presence/absence) of all species at all sites simultaneously, as a function of three different types of explanatory variable: the environmental variables measured at each site, species traits and trait–environment interactions (Fig. 1b). This re-expresses the problem in Fig. 1a such that **R** and **Q** are now considered as predictor variables and the trait–environment interaction terms in the model represent the fourth corner.

This approach can be applied using any type of predictive model capable of estimating interactions between predictors, making it quite versatile. Pollock, Morris & Vesk (2012) and Jamil *et al.* (2013) used a generalized linear mixed model. In this article, we will use generalized linear models ('GLM', Nelder & Wedderburn 1972), which have recently been shown to have close links to maximum entropy (Phillips, Anderson & Schapire 2006; Renner & Warton 2013), but the method could equally well be applied using other SDM ling techniques that can model interactions between predictors.

When fitting a regression model, the type of model to fit depends on the properties of the abundance or presence/absence matrix (**Y**). For presence/absence data, one possible approach would be to use logistic regression (Warton & Hui 2011). In this case, the model for the probability of presence at the *i*th site for the *j*th species could be written as:

$$\text{logit}(p_{ij}) = b_0 + b_1'\text{env}_i + b_2'\text{trait}_j + b_3'(\text{env} * \text{trait})_{ij} \quad \text{eqn 1}$$

If abundance was measured as counts, Poisson or negative binomial regression could be used (O'Hara & Kotze 2010).

Two important variations of the method are worthy of mention at this point. First, if using a GLM framework for analyses and using environmental and trait variables which are both quantitative, it is sensible to start by including all quadratic terms in the model, not just the interaction term:

$$\text{logit}(p_{ij}) = b_0 + b_1'\text{env}_i + b_2'\text{trait}_j + b_3'\text{env}_i^2 + b_4'\text{trait}_j^2 + b_5'(\text{env} * \text{trait})_{ij} \quad \text{eqn 2}$$

Secondly, if the primary purpose is to use traits to explain *differences in environmental response* across species, then different species should be allowed to have different total

abundances (via species-specific intercept terms):

$$\text{logit}(p_{ij}) = b_0 + b_1'\text{env}_i + b_2'\text{spp}_j + b_3'(\text{env} * \text{trait})_{ij} \qquad \text{eqn 3}$$

Such a species term is straightforward to add to the analysis on most software, as a factor giving the species identity of each observation. We specified the factor as a fixed effects term, Jamil *et al.* (2013) chose to include it as a random effect.

Notice that while the data are multivariate, in the sense that responses are likely to be correlated across species, the model proposed above does not account for this, instead treating abundances across all species at all sites as a single univariate response. Correlation in such species-by-site data sets is a very difficult problem to handle from a model-based perspective, because there are often a large number of species relative to the number of sites, and because biotic interactions are often unknown or complex in nature (Wisz *et al.* 2013). Pollock, Morris & Vesk (2012) and Jamil *et al.* (2013) tried to address this using a random site effect, but that approach can only handle a very restrictive form of correlation – correlation that is introduced through differences in total abundance across sites. Ives & Helmus (2011) introduced a vector of random effects that are correlated via their phylogeny. Neither approach can reasonably be expected to adequately capture the nature of biotic interactions at sites – which depend on complex ecological interactions (Wisz *et al.* 2013) and are not explained by phylogeny nor total abundance – yet, a model-based approach to inference (e.g. MCMC) assumes such correlation has been appropriately accounted for in the model. Instead, we advocate design-based inference (Manly 2006), which enables valid inferences that are robust to correlation between species, even when such correlation has not been incorporated into the fitted model. In the context of hypothesis testing, this is standard in the multivariate literature and is achieved by resampling rows of observations or residuals (as in Wang *et al.* 2012). In the context of predictive modelling, this can be achieved by cross-validation (where *sites* are the unit that is assigned to training/test groups), as below.

MODEL SELECTION

When analysing fourth-corner data, a key decision that needs to be made is which environmental and trait variables to include in the analyses. An advantage of framing the fourth-corner problem in terms of a predictive model is that then this decision is phrased as a 'model selection' problem, for which there are many solutions in the literature (Hastie, Tibshirani & Friedman 2009). The approach we used in this article was to apply the least absolute shrinkage and selection operator (LASSO), where cross-validation was used to choose the LASSO parameter (Hastie, Tibshirani & Friedman 2009) leaving out 10% of observations as test data, averaging over 50 replicate runs. This was implemented within the statistical package R (R Development Core Team 2011) using our own code (Appendix S1), but the 'GLMNET' package (Friedman, Hastie & Tibshirani 2008) is a viable alternative for the presence/absence data.

The LASSO, reviewed concisely in Hastie, Tibshirani & Friedman (2009), is a method of penalized likelihood which imposes a constraint (specifically, an L_1 constraint) on estimates of model parameters. The effect of this constraint is to shrink some terms to zero, making it an attractive approach to the model selection problem (Hastie, Tibshirani & Friedman 2009) which has seen widespread use recently in MAXENT software (Phillips, Anderson & Schapire 2006). MAXENT is widely used in SDM because it has been shown to have high predictive performance (Elith *et al.* 2006), which can now be understood to be largely due to its use of a LASSO penalty (Gastón & García-Viñas 2011; Renner & Warton 2013).

Cross-validation is a standard approach for estimating the LASSO penalty parameter; in fact, it is the only method implemented in the 'GLMNET' package (Friedman, Hastie & Tibshirani 2008). It is important when implementing cross-validation to ensure observations in different validation groups are independent. In the case of multi-species modelling, this means that it is important to keep abundances from all species at a site in the same validation group, and to sort observations into different validation groups using (independent) sites only.

APPLICATION TO BIRD PRESENCE/ABSENCE DATA

Here, we apply the proposed model-based approach to the bird data set of Tatibouet (1981). We fitted a logistic regression model to the data as in eqn (3), using all trait and environmental variables from Fig. 2b. We used a LASSO penalty estimated via cross-validation as above and include code for the analyses of Fig. 2c in the Appendix S1.

The fourth-corner results are summarized in Fig. 2c, using the level plot function from the lattice package on R (Sarkar 2008) to visualize the table of environment–trait interaction coefficients. The model-based approach to fourth-corner analyses has potential advantages in interpretation, prediction and model selection, as described below:

Interpretation

The nature and strength of the environment–trait interactions are indicated, respectively, by the sign and magnitude of the interaction coefficients. For example, there is a weak, positive association between fields being present and whether or not a bird breeds on the ground, and there is a strong, negative interaction between fields and whether or not a bird breeds on buildings (Fig. 2c). We can go further and use this model to construct interaction plots such as in Fig. 3, to visualize the interaction between breeding habits and field presence in affecting species abundance.

Prediction

The fourth-corner model can be used to make predictions under new scenarios. For example, the estimated probability of the Garden Warbler being present at site one is 0·17. Site one contains fields but has no small buildings or industry,

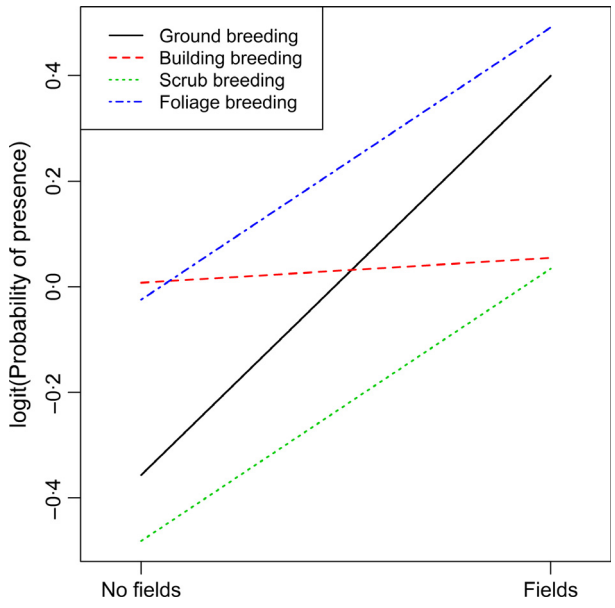

Fig. 3. Interaction plot between field presence and breeding location. The lines represent changes in prevalence according to presence/absence of fields. Birds of most breeding types were more likely to be found in fields, especially ground-breeders, less so for building-breeders.

which suits this species, characterized as an insectivore that breeds in scrub. If the fields were replaced with small buildings on site 1, the predicted probability of the Garden Warbler would more than halve (0·07).

Model selection

The LASSO penalty automatically sets to zero any terms in the model which do not explain any variation in species response. For the bird data, for example, all interaction terms with feeding strata were zero, suggesting that these variables had little interaction with the environment in predicting the presence of bird species.

The nonzero terms in the model (Fig. 2c) corresponded well with those interactions deemed statistically significant in fourth-corner hypothesis testing (Fig. 2b), although not perfectly. The two methods provided qualitatively similar results, with the four statistically significant interactions (Fig. 2b) all corresponding to nonzero coefficients in the fourth-corner model (Fig. 2c), including the three with largest magnitude. Two weak interactions detected in the fourth-corner model were not regarded as statistically significant in fourth-corner testing: that between the presence of small buildings and each of feeding habit and feeding strata.

Differences in results reflect differences in analysis goals – not only due to the distinction between hypothesis testing and predictive modelling, but also due to the distinction between the study of marginal and partial effects. The hypothesis testing method tests for marginal effects (ignoring the effect of all other variables), whereas the fourth-corner model estimates partial effects (controlling for all other variables), and so is interpreted slightly differently.

Evaluation

Fourth-corner models have now been shown to be useful from an interpretational standpoint, for explaining variation in species response. But how much of species-to-species variation in environmental response is actually mediated by measured traits? In this section, we specifically look at the role of traits in predicting species' distribution and abundance, by comparing models with species traits to those without. Models without traits can be understood as separate SDMs by species – assuming different environmental responses for different species, without relating interspecific variation in response to traits in any way. Specifically, three competing methods are compared:

1 'trait*env': the fourth-corner model proposed above, with an interaction between species traits and environmental variables explaining species-to-species variation in response. The fitted model has the form:

$$g(p_{ij}) = b_0 + b_1'\mathrm{env}_i + b_2'\mathrm{spp}_j + b_3'(\mathrm{trait} * \mathrm{env})_{ij}$$

for some link function $g(.)$.

2 'spp*env': different environmental responses are fitted for different species, as in SDM.

$$g(p_{ij}) = b_0 + b_1'\mathrm{env}_i + b_2'\mathrm{spp}_j + b_3'(\mathrm{spp} * \mathrm{env})_{ij}$$

3 'trait*env + spp*env': a hybrid model which still uses species traits to explain variation in environmental response across species, but which also includes an environment by species term to account for any residual species-to-species variation not explained by traits:

$$g(p_{ij}) = b_0 + b_1'\mathrm{env}_i + b_2'\mathrm{spp}_j + b_3'(\mathrm{trait} * \mathrm{env})_{ij} + b_4'(\mathrm{spp} * \mathrm{env})_{ij}$$

These three models were fitted using the same approach as previously: GLMs with a LASSO term included for model selection. Four data sets with varying properties were used for the comparison, and their properties are summarized in Table 1. Two data sets were presence–absence and analysed via logistic regression, the remaining two consisted of over-dispersed counts that were analysed using negative binomial regression using purpose-written *R* code (Appendix S1). Orthogonal linear and quadratic terms were included in the model when predictor variables were quantitative via the poly function on *R*. All models included a LASSO term, whose penalty was chosen to minimize predictive error on a hold-out sample.

The three model approaches (trait*env, spp*env and trait*env+spp*env) were compared in terms of their performance when predicting to a 'test' or hold-out data set consisting of a random sample of 30% of sites. Deviance was calculated as the measure of predictive performance and reported as% deviance explained, as in Hui *et al.* (2013). Predictive deviance is a measure of predictive accuracy which has a number of advantages: it is a consistent metric that can be applied across presence/absence and count data; it has strong theoretical foundations, being an estimator of Kullback–Leibler distance; it has the

Table 1. Summary of properties of the four data sets used in evaluations. Nonzero values– the total number of presence points (nonzero abundance) in the data set, across all species and sites. Singleton species– the number of species only found in one site

Dataset	Urban Birds	Eucalypt Ants	Boreal Ants	Grassland Hemiptera
Source	Tatibouet (1981)	Gibb & Cunningham (2013)	H. Gibb (unpublished)	M. Binns & A. Brown (unpublished)
Location	Lyon, France	NSW, Australia	Sweden	Eastern Australia
Variables measured				
L	Birds	Ants	Ants	Hemiptera
R	Urban/rural indicators	Vegetation structure and cover	Vegetation structure and cover	Climate
Q	Behaviour/habits	Morphological measurements	Morphological measurements	Morphological measurements
Variable type				
L	Presence/absence	Presence/absence	Counts	Counts
R	Categorical	Quantitative	Quantitative	Quantitative
Q	Categorical	Quantitative	Quantitative	Quantitative
Data set size				
L	51 × 40	20 × 60	34 × 8	36 × 64
R	51 × 3	20 × 7	34 × 3	36 × 4
Q	40 × 3	60 × 6	8 × 3	64 × 5
Nonzero values	492	323	91	120
Singleton species	0	19	1	49

same form as the criterion minimized on training data in the original model fit. We additionally computed ROC curves for presence/absence, but these led to similar results (see Appendix S1). Analyses were repeated 50 times for different choices of hold-out sample to make inferences about predictive performance for each data set.

Across the four data sets, the proportion of deviance explained by environmental variables varied considerably (3–21%, Fig. 4) when predicting to a hold-out sample. However, the key point is that most of the species-to-species variation in environmental response was captured by trait variables – in fact, the trait model had a significantly better fit for the Eucalypt Ants and Grassland Hemiptera data sets. In the one instance where the trait model had significantly weaker predictive performance (Urban Birds), the hybrid model was able to soak up the variation in environmental response not explained by species traits.

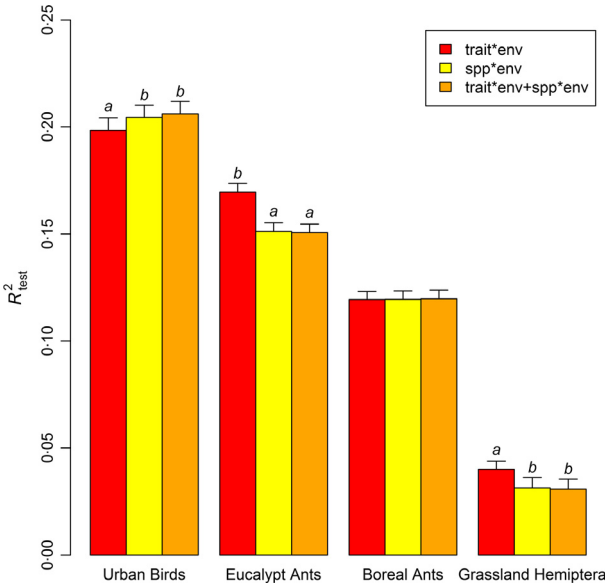

Fig. 4. Predictive performance for the fourth-corner model (trait*env), separate species distribution models (spp*env) and a hybrid of the two approaches (trait*env+spp*env), across the four data sets of Table 1. Predictive performance is measured as proportion of deviance explained (R^2_{test}) when predicting to a hold-out sample, averaged across fifty different choices of hold-out sample (with 95% confidence interval). The fourth-corner model had comparable predictive performance overall, in fact, significantly better performance in two data sets (for which lower case letters indicate multiple comparison results).

Discussion

There has been a recent trend towards the development of model-based approaches to analysing multispecies data (Dunstan, Foster & Darnell 2011; Ives & Helmus 2011; Ovaskainen & Soininen 2011; Warton, Wright & Wang 2012), and in this article, we have showed that such a method can be used to solve the 'fourth-corner problem' of understanding how species traits interact with their environment, by explicitly incorporating traits into the model. This fourth-corner model can be understood as a competitor to standard SDMs which have reasonable predictive performance (Fig. 4), but whose real value is in explaining why environmental response varies across species, which standard SDMs cannot. Fourth-corner models can also be understood as complementing existing fourth-corner methods, offering greater interpretation of the nature of trait–environment interactions, as well as the predictive and model selection capacities of model-based approaches (Ives & Helmus 2011; Warton, Wright & Wang 2012). Pollock, Morris & Vesk (2012) showed that this type of model can be helpful in explaining species variation in environmental response in Eucalypt communities at a broad scale. Here, we have added that the general idea of incorporating traits can apply under any SDM ling framework and is not limited to hierarchical models, that the approach loses relatively little species-specific information, as it can have comparable predictive performance to separate species models, and we have linked the method to the fourth-corner literature.

The fourth-corner model proposed here can be understood as combining the strengths of a few other recently proposed methods. Shipley, Vile & Garnier (2006) proposed a maximum entropy approach for using traits to explain biodiversity patterns, evaluated recently (Merow, Latimer & Silander 2011). A key difference from fourth-corner modelling is that Shipley, Vile & Garnier (2006) did not make use of environmental data (**R**), and modelled **L** and **Q** only. In contrast, Ives & Helmus (2011), Ovaskainen & Soininen (2011) and others proposed methods of modelling abundance as a function of environmental variables only (**L** against **R** with no **Q**). The fourth-corner model bridges these two types of method and forms a connection between the literatures on multivariate analysis in ecology (Legendre, Galzin & Harmelin-Vivien 1997) and SDM ling (Franklin 2010; Pollock, Morris & Vesk 2012).

A key difference between our model and other environment–trait approaches (Doledec *et al*. 1996; Legendre, Galzin & Harmelin-Vivien 1997) is that previously traits (**Q**) have been treated as the response variable, or traits and environment jointly (**R** and **Q**). In contrast, we consider the species matrix (**L**) as the response, which is consistent with SDM ling practice, and indeed with most of the multivariate analysis literature. The argument for treating **L** as the response variable is that this reflects typical study designs – one typically visits a set of pre-specified locations (fixed **R**) and observes species whose abundance or presence/absence is random (random response **L**). In contrast, if one were to sample a subset of species based on their abundance, **L** would no longer be random so our analysis approach would no longer be valid. In such a setting, treating **Q** as the response variable might more closely reflect the study design.

A different type of approach to studying the environment–trait relationship is to compute a community-aggregated trait measure at each site (such as the matrix product **LQ**) and treat it as the response to be related to environmental variables (**R**) (e.g. Kühn *et al*. 2006; Suding *et al*. 2008). This rephrases the problem such that rather than studying how traits mediate change in environmental response, one instead studies how the aggregated trait measure relates to the environment. The success or otherwise of this method depends how well an aggregate trait measure has been chosen, with potential loss of information (on traits and on uncertainty) in the aggregation step. Preliminary simulations we have undertaken suggest there is often some loss of efficiency from using this two-stage approach as compared to a fourth-corner model, and arguably, a two-stage approach lacks the intuitive appeal of directly modelling the processes by which traits influence response of species to their environment.

One issue that needs to be carefully addressed in multispecies models such as that considered here is correlation between species and the potential for pseudo-replication. Essentially, model-based approaches implemented here and elsewhere (Pollock, Morris & Vesk 2012; Jamil *et al*. 2013) can be understood as univariate approaches – the model does not account for correlation between species in any meaningful way in the estimation step. Capturing the correlation structure between a large number of species at a small number of sites is a very difficult task, and unfortunately, we are not aware of any satisfactory solutions in the ecological literature for this purpose, beyond cases involving only a few species (Wisz *et al*. 2013). While this situation is far from perfect, our approach is still valid in the sense that one can still achieve unbiased estimates of parameters without modelling correlation (Liang & Zeger 1986). The model cannot however achieve unbiased estimates of uncertainty, and for this reason, it is critical to use a design-based approach to inference and not standard model-based tools (AIC, MCMC, etc.), such that inferences one makes about which environment–trait interactions are important are robust to failure to correctly model species interactions. In the Appendix S1, we describe some short simulations which verify that model-based approaches to model selection perform poorly relative to cross-validation of sites.

While GLMs were used in this article, other methods of predictive modelling can be applied. To illustrate this point, the Urban Birds data set has been reanalysed using a classification tree in the Appendix S1. Classification and regression trees automatically incorporate interactions between predictors in the fitted model, so including both environmental and trait variables as predictors would by default estimate 'fourth-corner' interactions without specifying them explicitly as inputs. As a further example, Stoklosa, Gibb & Warton (2014) reanalysed the Eucalypt Ants data set using generalized estimating equations with forward selection of interaction terms. Gabriel (1998) proposed bilinear regression, a method related to our GLMs which could also be used, but additionally, offering the possibility of assuming the matrix of interaction coefficients has reduced rank, which could be especially useful for ordination purposes. Maximum entropy (Phillips, Anderson & Schapire 2006) could also be applied to fourth-corner data, although results would be expected to be similar to those reported here, because it has been shown in related work that MAXENT is mathematically equivalent to a Poisson GLM (Renner & Warton 2013). If analysing grid cell data then one might expect strong spatial structure, in which case a model which explicitly accounts for spatial relationships between observations would be advisable (Zuur, Ieno & Elphick 2010).

Similarly, while the LASSO combined with cross-validation was used for model selection in this article, other methods of model selection could in principle be used. However, information criteria such as AIC (Burnham & Anderson 2002) or BIC should be used with caution – because of the pseudo-replication issue discussed above, these methods of model-based inference require the inter-species correlation to have been adequately modelled. Cross-validation, where sites are randomly allocated to training/test groups, offers a design-based approach to model selection that is robust to failure of inter-species correlation assumptions, provided that observations are independent across sites. The simulation in the Appendix S1 verifies the advantages of our model selection approach over BIC.

The four multivariate data sets our method was evaluated on were all small in size (20–51 replicates) as compared to data sets typically used for SDM ling. Differences in sample size are characteristic of these two literatures – while essentially,

community-level SDM ling is just a multivariate analysis, operationally, multivariate data sets often seem quite small in comparison. Small data sets often arise in experiments or field surveys, and the main challenge they pose is whether or not sampling has been sufficient to detect patterns (as we were able to in Figs 2 and 4). Large data sets often arise from digitized atlas records, and the main challenges they pose are computational – an atlas data set with 8000 transects and 300 species requires a model to be simultaneously fitted to a response vector with 2·4 million entries. In such instances, a fourth-corner model remains applicable, but computationally efficiency might need to be considered when choosing software and model selection approach. The approach taken in this article (GLM with LASSO penalties) is well suited to big data problems.

One interesting opportunity afforded by fourth-corner models is the possibility of studying intraspecific variability in traits (Violle *et al*. 2012) and its influence on abundance patterns. In principle, there is no reason why the trait variables used in analyses need to be species means – for example, site means could be computed for each species, where possible. One option would be to include two types of trait predictors – a species mean and also local deviations from this mean, to disentangle selection effects operating within vs. across species.

While the proposed model-based approach can be used for model selection (*i.e*. for identifying which are the key environmental and trait variables), care should be taken when choosing the environmental variables and species traits to include in the model. Model selection methods tend to perform best when given a relatively small number of variables (and hence a small number of candidate models) to select from Burnham & Anderson (2002). It is especially important to try to keep the number of explanatory variables small in fourth-corner models because of the use of interaction terms – a model with ten environmental terms and ten trait terms would then have 100 interaction terms, which is usually too many to reliably identify the important interactions when using just a single data set. Hence, one should carefully consider prior to analysis which variables to include and attempt to compile a 'shortlist'. Particularly, useful for shortlisting variables is prior knowledge from the literature, and diagnostic tools for collinearity (Zuur, Ieno & Elphick 2010).

While we have focussed in this article on the problem of predictive modelling, some interesting relations between methods can be seen if our model was used in a hypothesis testing context. Specifically, a hypothesis test for significance of fourth-corner terms from our model can be mathematically related to the hypothesis testing approach of Legendre, Galzin & Harmelin-Vivien (1997), a finding which offers some new insights into their method and how it can be extended (Appendix S1). In the special case, where **L** is presence–absence and **R** and **Q** consist of a factor (or several factors), a likelihood-ratio test for association using the method of Legendre, Galzin & Harmelin-Vivien (1997) is mathematically equivalent to a test for interaction in a Poisson regression model for **L** as a function of **R** and **Q** (Appendix S1). The two methods diverge in their treatment of more complicated situations, and the model-based approach proposed here offers a unified framework that quite naturally generalizes the hypothesis testing method of Legendre, Galzin & Harmelin-Vivien (1997) to more complicated settings such as the analysis of overdispersed counts or the analysis of multiple environmental and species trait variables. For example, the same fourth-corner modelling framework was used for each of the four data sets in this article, but three different types of test would have been required if using the hypothesis testing approach of Legendre, Galzin & Harmelin-Vivien (1997). The model-based approach also offers insight into the problems that have been encountered constructing a valid resampling scheme for fourth-corner testing (Legendre, Galzin & Harmelin-Vivien 1997; Dray & Legendre 2008) – the key problem is that we want to test the significance of an interaction term, which is a notoriously difficult problem with no exact solution (Manly 2006). We are currently developing methods of resampling-based hypothesis testing for GLM, with the intention of better addressing this issue.

Acknowledgements

This research was funded by the Australian Research Council Discovery Projects funding scheme (project number DP0985886) awarded to DIW, NRA and HG.

References

ter Braak, C.J.F., Cormont, A. & Dray, S. (2012) Improved testing of species traits–environment relationships in the fourth-corner problem. *Ecology*, **93**, 1525–1526.

Brown, A. (2010) *The fourth corner problem: a model-based approach to grassland Hemipteran assemblages*. Honours thesis, University of New South Wales, Sydney, New South Wales.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York, New York.

Chessel, D., Dufour, A.B. & Thiulouse, J. (2004) The ade4 package -I- one-table methods. *R News*, **4**, 5–10.

Doledec, S., Chessel, D., ter Braak, C.J.F. & Champely, S. (1996) Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, **3**, 143–166.

Dray, S. & Legendre, P. (2008) Testing the species traits-environment relationships: the fourth-corner problem revisited. *Ecology*, **89**, 3400–3412.

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

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, **40**, 677–697.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Franklin, J. (2010) *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press, Cambridge.

Friedman, J., Hastie, T. & Tibshirani, R. (2008) Regularisation paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Gabriel, K.R. (1998) Generalised bilinear regression. *Biometrika*, **85**, 689–700.

Gastón, A. & García-Viñas, J.I. (2011) Modelling species distributions with penalised logistic regressions: a comparison with maximum entropic models. *Ecological Modelling*, **222**, 2037–2041.

Gibb, H. & Cunningham, S.A. (2013) Restoration of trophic structure in an assemblage of omnivores, considering a revegetation chronosequence. *Journal of Applied Ecology*, **50**, 449–458.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn. Springer, New York, New York.

Hui, F.K.C., 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. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, **81**, 511–525.
Jamil, T., Ozinga, W.A., Kleyer, M. & ter Braak, C.J. (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, **24**, 988–1000.
Kleyer, M., Dray, S., Bello, F., Lepš, J., Pakeman, R.J., Strauss, B., Thuiller, W. & Lavorel, S. (2012) Assessing species and community functional responses to environmental gradients: which multivariate methods? *Journal of Vegetation Science*, **23**, 805–821.
Kühn, I., Bierman, S.M., Durka, W. & Klotz, S. (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. *New Phytologist*, **172**, 127–139.
Legendre, P., Galzin, R. & Harmelin-Vivien, M.L. (1997) Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology*, **78**, 547–562.
Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data-analysis using generalized-linear models. *Biometrika*, **73**, 13–22.
Manly, B.F.J. (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Chapman and Hall, Boca Raton, FL.
McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006) Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution*, **20**, 178–185.
Merow, C., Latimer, A.M. & Silander, J.A. (2011) Can entropy maximization use functional traits to explain species abundances? A comprehensive evaluation. *Ecology*, **92**, 1523–1537.
Nelder, J.A. & Wedderburn, R.W. (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.
O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.
Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.
Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.
Pollock, L.J., Morris, W.K. & Vesk, P.A. (2012) The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, **35**, 716–725.
Pottier, J., Marrs, R.H. & Bedecarrats, A. (2007) Integrating ecological features of species in spatial pattern analysis of a plant community. *Journal of Vegetation Science*, **18**, 223–230.
R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
Renner, I.W. & Warton, D.I. (2013) Equivalence of maximum entropy modelling and poisson regression for species distribution modelling in ecology. *Biometrics*, **69**, 274–281.
Sarkar, D. (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York.
Shipley, B., Vile, G. & Garnier, E. (2006) From plant traits to plant communities: a statistical mechanistic approach to biodiversity. *Science*, **314**, 812–814.
Stoklosa, J., Gibb, H. & Warton, D.I. (2014) Fast forward selection for generalized estimating equations with a large number of predictor variables. *Biometrics*, **70**, 110–120.
Suding, K.N., Lavorel, S., Chapin, F.S., Cornelissen, J.H., Diaz, S., Garnier, E. *et al.* (2008) Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. *Global Change Biology*, **14**, 1125–1140.
Tatibouet, F. (1981) *Approche écologique d'un établissment humain (environement et structure)*. Example de la communauté urbained de Lyon, University of Lyon, Lyon.
Thuiller, W., Lavorel, S., Midgley, G., Lavergne, S. & Rebelo, T. (2004) Relating plants traits and species distributions along bioclimatic gradients for 88 Leucadendron taxa. *Ecology*, **85**, 1688–1699.
Violle, C., Enquist, B.J., McGill, B.J., Jiang, L., Albert, C.H., Hulshof, C., Jung, V. & Messier, J. (2012) The return of the variance: intraspecific variability in community ecology. *Trends in Ecology & Evolution*, **27**, 244–252.
Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012) mvabund- an R package for model-based analysis of multivariate abundance data. *Ecology and Evolution*, **3**, 471–474.
Warton, D.I. & Hui, F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, **92**, 3–10.
Warton, D.I., Wright, S.T. & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.
Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F. *et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, **88**, 15–30.
Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.

Received 16 October 2013; accepted 14 January 2014
Handling Editor: Nigel Yoccoz

Supporting Information

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

Appendix S1. Proof of equivalence of tests using the model-based approach and the method of Legendre et al. (1997), for binary **L** and factors **R** and **Q**.