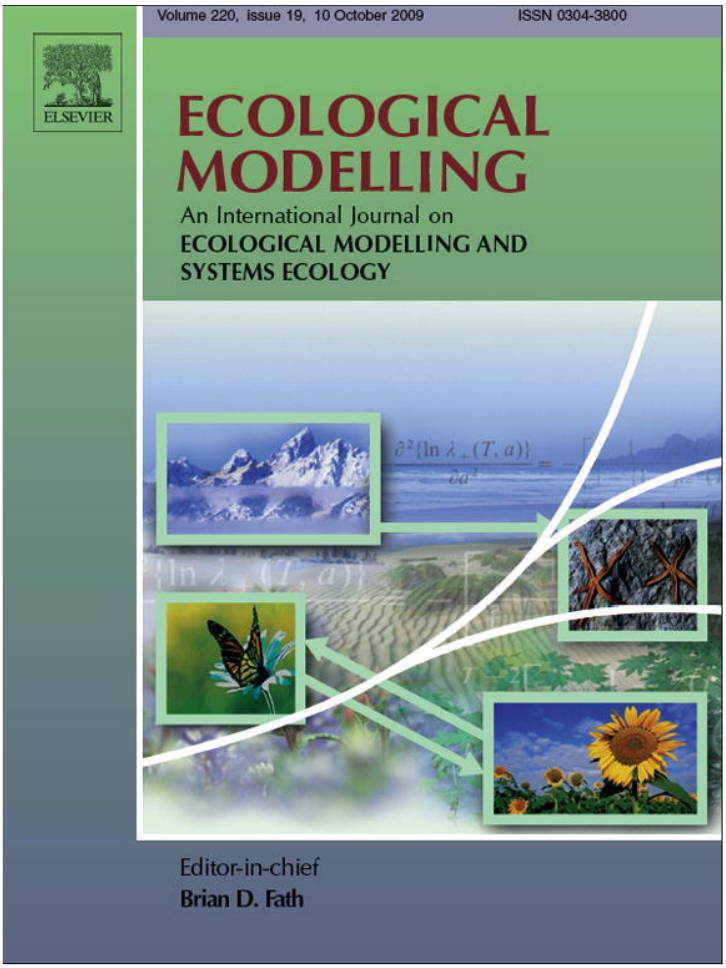

Ecological Modelling 220 (2009) 2365–2379

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

The effect of species response form on species distribution model prediction and inference

Truly Santika*, Michael F. Hutchinson

The Fenner School of Environment and Society, W.K. Hancock Building 43, Biology Place, The Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 21 January 2009
Received in revised form 22 May 2009
Accepted 2 June 2009
Available online 14 July 2009

Keywords:
Autocovariate logistic
BIOCLIM
Classification and regression tree
Generalized additive modelling
Logistic regression
Species response curve
Spatial autocorrelation

ABSTRACT

Ecological theory and current evidence support the validity of various species response curves according to a variety of environmental gradients. Various methods have been developed for building species distribution models but it is not well known how these methods perform under various assumptions about the form of the underlying species response. It is also not well known how spatial correlation in species occurrence affects model performance. These effects were investigated by applying an environmental envelope method (BIOCLIM) and three regression-based methods: logistic regression (LR), generalized additive modelling (GAM), and classification and regression tree (CART) to simulated species occurrence data. Each simulated species was constructed as a sum of responses with varying weights. Three basic species response curves were assumed: Gaussian (bell-shaped), Beta (skew) and linear. The two non-linear responses conform to standard ecological niche theory. All three responses were applied in turn to three simulated environmental variables, each with varying degrees of spatial autocorrelation. GAM produced the most consistent model performance over all forms of simulated species response. BIOCLIM and CART were inclined to underrate the performance of variables with a linear response. BIOCLIM was less sensitive to data density. LR was susceptible to model misspecification. The use of a linear function in LR underestimated the performance of variables with non-linear species response and contributed to increased spatial autocorrelation in model residuals. Omission of important environmental variables with non-linear species response also contributed to increased spatial autocorrelation in model residuals. Adding a spatial autocovariate term to the LR model (autologistic model) reduced the spatial autocorrelation and improved model performance, but did not correct the misidentification of the dominant environmental determinant. This is to be expected since the autologistic approach was designed primarily for prediction and not for inference. Given that various forms of species response to environmental determinants arise commonly in nature: (1) higher order functions should always be tested when applying LR in modelling species distribution; (2) spatial autocorrelation in species distribution model residuals can indicate that environmental determinants with non-linear response are missing from the model; and (3) deficiencies in LR model performance due to model misspecification can be addressed by adding a spatial autocovariate to the model, but care should be taken when interpreting the coefficients of the model parameters.

1. Introduction

The prediction of species distribution is an important aspect of conservation biology. Habitat suitability models of species based on a set of environmental factors provide meaningful information for the management of endangered species (Palma et al., 1999; Engler et al., 2004), human–wildlife conflicts (Le Lay et al., 2001) and species reintroductions (Yáñez and Floater, 2000; Schadt et al., 2002). They can also be used to assess the impacts of climate and land-use change on species distribution (Guisan and Theurillat, 2000; Dirnbock et al., 2003).

Numerous methods have been developed for building species habitat suitability models. Guisan and Zimmermann (2000) presented a comprehensive review and classified the methods into two categories: (1) regression-based methods; and (2) environmental envelope methods. Regression methods relate species response to single or multiple environmental predictors. These methods include frequently used approaches such as logistic regression (LR; Hosmer and Lemeshow, 1989), generalized additive modelling (GAM; Hastie and Tibshirani, 1990), and classification and regression tree (CART; Breiman et al., 1984). The environmental envelope method identifies the locations of species habitat

* Corresponding author. Tel.: +61 2 6125 4758; fax: +61 2 6125 0746.
E-mail address: truly.santika@anu.edu.au (T. Santika).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.06.004

suitability based on calculating minimal rectilinear envelopes in multi-dimensional bioclimatic space. The method was pioneered with the development of BIOCLIM (Nix, 1986; Busby, 1991; Houlder et al., 1999).

The use of regression methods to predict species presence–absence distributions has been considered inadequate without incorporating the effect of spatial autocorrelation (SAC) on species distribution (Lennon, 2000; Kühn, 2007). SAC occurs when the values of variables sampled at nearby locations are not independent from each other. SAC can lead to autocorrelated model residuals, thus violating the assumption of independent identically distributed errors of most standard regression procedures (Anselin, 2002). Autocorrelated residuals can also arise if predictors do not fully reflect the actual controls on the species distribution (Augustin et al., 1996).

Several approaches that deal with SAC exist. The autocovariate logistic approach (autologistic; Augustin et al., 1996) extends the usual regression model by adding a spatial autocovariate term into the model. Such an approach was derived from a statistical analysis of lattice systems by Besag (1974). In principle, there is no restriction on the type of model that can be used with an autocovariate. However, autocovariates have mainly been applied to the LR method (Augustin et al., 1996; Luoto et al., 2002; Betts et al., 2006; McPherson and Jetz, 2007; Syartinilia and Tsuyuki, 2008) and the GAM method (Knapp et al., 2003; Segurado and Araujo, 2004) in species distribution modelling studies.

Species occurrence is to a certain extent determined by the species' underlying response to each environmental determinant. Niche theory, as applied to both plants and animals, assumes an approximately symmetric bell-shaped curve, in which the species fundamental niche has a central maximum with declining values toward higher and lower levels (Swan, 1970; Austin, 1985). However, factors such as competition, predation and disturbance, can place pressure on a species, causing the response curve to alter from symmetry to skewed and non-unimodal responses in the realized niche (Austin and Meyers, 1996; Oksanen and Minchin, 2002; Olden and Jackson, 2002). Moreover, given that species respond to each environmental factor in different ways, the reliability of a modelling method in predicting species spatial distribution depends on its ability to take the various underlying forms of response jointly into account.

Many studies have been carried out to assess the performance of species distribution modelling methods (Manel et al., 1999; Robertson et al., 2003; Segurado and Araujo, 2004; Araujo et al., 2005; Elith et al., 2006). These methods have mainly been assessed using real species distribution data (i.e. the model is calibrated and evaluated using real data derived from field studies). These studies generally arrived at similar conclusions about the predictive performance of the modelling methods, suggesting that novel methods with the ability to fit complex species occurrence–environmental relationship tend to perform better than the simpler methods. However, it is not well known how the ability of modelling method to fit species occurrence–environmental relationship affects model accuracy. Studies have also shown that methods with flexibility to fit complex species response form to environmental determinants such as GAM yielded model residuals with lower SAC compared with the residuals of model produced by simpler methods such as LR with a linear function (Segurado et al., 2006; Dormann et al., 2007). It is also not well known how the ability of modelling method to fit species occurrence–environmental relationship affects the strength of SAC in model residuals. Simulated species data offers a way to investigate these particular issues.

Simulated data, with known properties, has increasingly been used to evaluate the performance of modelling methods. Hirzel et al. (2001) employed simulated species data to compare the performance of logistic regression with a quadratic fitting function and Ecological Niche Factor Analysis (ENFA; Hirzel et al., 2002). The simulated species data were generated using 11 real environmental variables, three presumed species response shapes to each environmental variable, and weights assigned to each environmental variable with most of the weights assigned to variables with a linear and a Gaussian species response.

In this study, the approach for generating simulated species data introduced by Hirzel et al. (2001) was adapted to investigate the performance of species distribution modelling methods under various assumed underlying species response forms. Four modelling methods were assessed. These were the environmental envelope method BIOCLIM (DIVA version) and three regression-based methods: LR with a linear, quadratic and cubic fitting functions, GAM with smoothing spline function, and the CART. The effectiveness of the autologistic approach in dealing with spatially autocorrelated residual in LR was also assessed.

The use of simulated data in this study is essential to permit the evaluation of the outcome of analysis against the predefined "truth". The simulated species data were generated by using three artificial environmental variables with varying degrees of SAC; three presumed species responses to the environmental variables: Gaussian (bell-shaped), Beta (skew) and linear; and numerous weighting combinations representing the importance of each variable in determining spatial occurrences of each simulated species. Following Hirzel et al. (2001), it was assumed that the three artificial variables combine in an additive way to determine overall species occurrence.

The simulation approach used in this study systematically examined various weighting combinations assigned to each variable. This extends the original construction proposed by Hirzel et al. (2001) that used a single set of prescribed weights to achieve certain properties of the simulated data. By using the various weighting combinations, this study addresses several questions: (1) how does species response form affect the predictive performance of the fitted species distribution model; (2) how does model structure affect SAC in model residuals and predictive performance; and (3) how does the density of calibration data affect SAC in model residuals and model performance.

2. Materials and methods

The study has five main stages:

(1) generating presence and absence data of the simulated species;
(2) fitting species distribution models using four modelling methods;
(3) assessing the performance of the modelling methods;
(4) assessing the impact of SAC in model residuals; and
(5) assessing the impact of incorporating spatial autocovariate into LR model.

Probability or habitat suitability maps for the simulated species were generated using three artificial environmental variables, three assumed species responses and numerous variable weights. Species presence–absence maps were then simulated by employing a fixed threshold for determining the probability of species presence and absence. These maps were assumed to represent the true presence–absence status of species and were later used to test the performance of models generated by modelling methods (i.e. simulated evaluation data sets). To build species distribution models, perturbed versions of the true species presence–absence maps were used. This was achieved by adding error components to the original species probability maps (i.e. simulated calibration data sets).

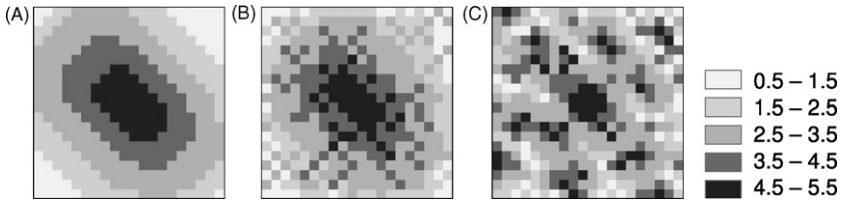

Fig. 1. Three simulated patterns with spatial realization of 60 × 60 cells, representing three spatially dependent continuous environmental variables. The numbers for each level are the same for each pattern. The degree of spatial dependence for each variable as quantified by global Moran's I is 0.900, 0.802 and 0.721 for variables *A*, *B* and *C*, respectively.

The simulated calibration data sets were submitted to each modelling method to generate species presence–absence prediction maps. These were used to assess the predictive performance, goodness-of-fit, and the residuals of the models. The predictive performance of the models was evaluated using the simulated evaluation data sets. The goodness-of-fit of the LR and GAM models was tested using a standard goodness-of-fit indicator the Akaike's Information Criterion (AIC; Akaike, 1974). Finally, the residuals of the models were assessed by examining the significance of SAC in model residuals for each type of species response. Procedures for each stage are explained in detail in the following sections.

2.1. Generating presence–absence of simulated species

2.1.1. Calculating probability of occurrence of the simulated species

Adopting the construction of Hirzel et al. (2001), for each simulated species *s*, the probability of occurrence p_{sij} was calculated as:

$$p_{sij} = \sum_{m \in \{A,B,C\}} w_{sm} H_{smij} \quad \text{with} \quad \sum_{m \in \{A,B,C\}} w_{sm} = 1 \tag{1}$$

where H_{smij} denotes the habitat suitability score of cell (i,j) representing the shape of the response of species *s* to the simulated environmental variable *m*; w_{sm} denotes the weight or importance assigned to the variable *m* for species *s*.

2.1.2. The three artificial environmental variables

Three environmental variables *A*, *B* and *C*, each were assigned a spatial pattern over a 3600 cells in a square lattice. Each cell within each pattern had a value ranging from 0.5 to 5.5. The values were split into five levels to give the patterns shown in Fig. 1. The numbers of grid cells for each level are the same for each simulated environmental variable.

Variables *A*, *B* and *C* were constructed to have various strengths of spatial clustering or SAC. The strength of SAC reduces from variable *A* to variable *C*, as quantified by the global Moran's I coefficient (Moran, 1950; Fortin and Dale, 2005). Using the eight nearest neighbours' rule to determine the connectivity between cells, values of Moran's I of 0.900, 0.802 and 0.721 were obtained for variables *A*, *B* and *C*, respectively. Moran's I coefficient of a variable typically ranges from −1 to 1, with −1 indicating a strong negative SAC (i.e. the value of sites that are close to each other are more dissimilar than those that are far apart) and 1 indicating a strong positive SAC (i.e. the value of sites that are close to each other are more similar than those that are far apart).

2.1.3. Assumed shapes of species response curves

The relationship between species occurrence and the environmental gradient was constructed to take one of three shapes: Gaussian (bell-shaped), Beta (skew), and linear. A Gaussian species response favours intermediate level conditions and gradually avoids low and high values of the environmental variable in a symmetric fashion. A Beta species response also favours intermediate values but the response to either high or low values is sharply reduced. This may occur under limiting conditions, such as competition, predation and disturbance which drives the response curve to shift from symmetry to sharply skewed in the realized niche (Austin and Meyers, 1996; Oksanen and Minchin, 2002). The linear response is a simple non-unimodal response that can arise with environmental variables that are not closely aligned with controlling processes (Austin and Meyers, 1996). It has also been posed by Olden and Jackson (2002).

The Gaussian species response can be modelled using the following function (Hirzel et al., 2001; Oksanen and Minchin, 2002; Olden and Jackson, 2002):

$$H.gaussian_{smij} = \exp\left[\frac{-(g_{smij} - \mu_{sm})^2}{2\sigma_{sm}^2}\right] \tag{2}$$

where $H.gaussian_{smij}$ represents the Gaussian habitat suitability score of cell (i,j) given the value g_{smij} of variable *m* at cell (i,j) for species *s*; μ_{sm} denotes the mean or optimum condition of variable *m* for species *s*; and σ_{sm} denotes the standard deviation or tolerance of variables *m* for species *s*.

The skewed Beta species response to an environmental variable can be modelled using the following function (Austin, 1976; Minchin, 1987; Austin et al., 1994; Oksanen and Minchin, 2002):

$$H.beta_{smij} = c_{sm}(g_{smij} - k_{1sm})^{\alpha_{sm}}(k_{2sm} - g_{smij})^{\gamma_{sm}} \tag{3}$$

where $H.beta_{smij}$ represents the Beta habitat suitability score of cell (i,j) given the value g_{smij} of variable *m* at cell (i,j) for species *s*; k_{1sm} and k_{2sm} denote the endpoints of the range of occurrences of species *s* within variable *m*; c_{sm} denotes the scaling parameter adjusting the response height to fit the observations of variable *m* for species *s*; and α_{sm} and γ_{sm} denote the parameters for determining the shape of response curve (i.e. location of the optimum, skewness and kurtosis) of variable *m* for species *s*. If $\alpha_{sm} > \gamma_{sm}$, the shape of the species response curve would lean towards the left, i.e. positive-skew. However, if $\alpha_{sm} < \gamma_{sm}$, the shape of the species response curve would lean towards the right, i.e. negative-skew.

Linear species responses can be modelled using the following function (Hirzel et al., 2001; Olden and Jackson, 2002):

$$H.linear_{smij} = \frac{g_{smij} - \min_{sm}}{\max_{sm} - \min_{sm}} \tag{4}$$

where $H.linear_{smij}$ represents the linear habitat suitability score of cell (i,j) given the value g_{smij} of variable *m* at cell (i,j) for species *s*; and $\min_{sm}$ and $\max_{sm}$ represent the lower and upper tolerable bound of variable *m* for species *s*.

For all simulated species, the parameter values for determining the shape of each species response were fixed for each type of response. Parameter values of $\mu = 3$ and $\sigma = 0.6$ were used for the Gaussian response; $c = 0.0375$, $k_1 = 1$, $k_2 = 5$, $\alpha = 1$, and $\gamma = 3$ were used for the Beta response; and min = 1 and max = 5 were used for the linear response. The three species response curves are shown in Fig. 2.

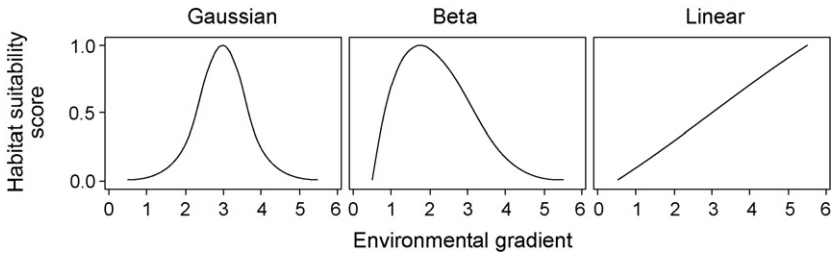

Fig. 2. Gaussian, Beta and linear species environmental response curves. The curves were generated using parameter values of $\mu = 3$ and $\sigma = 0.6$ for the Gaussian response; $c = 0.0375$, $k_1 = 1$, $k_2 = 5$, $\alpha = 1$, and $\gamma = 3$ for the Beta response; and min = 1 and max = 5 for the linear response.

2.1.4. Assigning weights for the simulated species

Each simulated species was constructed by assigning three possible combinations for the Gaussian, Beta, and linear responses to the set of simulated spatially dependent environmental variables *A*, *B* and *C*. These were: (1) Gaussian, Beta, and linear responses (G-B-L); (2) Beta, linear, and Gaussian responses (B-L-G); and (3) linear, Gaussian, and Beta responses (L-G-B), each assigned to variables *A*, *B* and *C*, respectively.

These combinations were applied with all possible weights, ranging from 0 to 1 with an increment of 0.1, for each of the environmental variables *A*, *B* and *C* (Fig. 1). These weights represented the contribution of each variable in determining spatial occurrence of each simulated species. In each case the three weights added to 1, thus giving 66 weighting combinations for w_A, w_B and w_C corresponding to the weights of the variables *A*, *B* and *C* (see Eq. (1)). Except for 6 cases with two equal weights, there were 20 weighting combinations with one dominant weight for each environmental variable. Given the different type of responses and weighting combinations assigned to variables *A*, *B* and *C*, a total of 198 (66×3) simulated species were generated.

2.1.5. Generating simulated species presence–absence for calibration and evaluation sets

The habitat suitability maps of the simulated species were translated into records of species presence–absence by applying a threshold of 0.5 to each probability score. The average number of species occurrences obtained for each map varied between 1800 and 1900 records (over $60 \times 60 = 3600$ cells in total). These data represent the true simulated species presences and absences data and constituted the evaluation datasets.

The model calibration datasets were generated by perturbing the original habitat suitability maps. This was done by adding an error component (sampled from a normal distribution error with mean zero and standard deviation 0.25) to the cells of the original habitat suitability maps whose scores lay between 0.2 and 0.8. The size of the perturbations was set so that the areas with extremely high habitat suitability would remain highly favoured while the areas with extremely low habitat suitability would remain less favoured. Species occurrence in the areas with intermediate level habitat suitability was more uncertain. A threshold of 0.5 was applied to the perturbed habitat suitability scores to obtain the simulated species presence–absences for model calibration. The average number of presences for each calibration dataset varied between 1700 and 1800 records (over $60 \times 60 = 3600$ cells in total).

In a real modelling situation, data used for model calibration are normally incomplete or even scarce. Thus it is worthwhile to investigate how modelling methods perform with sampled data. This was achieved by randomly sampled the full calibration data (3600 presence–absence records) to obtain subsets of 720 (20% of the full data), 360 (10% of the full data) and 72 (2% of the full data) presence–absence observations. For each of the 66 simulated species and each of the 720, 360 and 72 presence–absence observations, 50 different calibration samples were drawn.

2.2. Fitting the species distribution models

The simulated species presence–absence data were submitted to each modelling method to fit predictive models of simulated species distributions. Four modelling methods were used.

2.2.1. BIOCLIM

The original BIOCLIM method (Nix, 1986; Busby, 1991; Houlder et al., 1999) defines the ecological niche of a species as a bounding hyper-box that includes all species records in bioclimatic space. It creates a rectilinear envelope in environmental space, defined by pre-determined lower and upper percentiles of the species occurrence with respect to each environmental variable. The BIOCLIM method implemented in DIVA-GIS software package (Hijmans et al., 2005) extends the original version by assessing records for all percentile ranges to construct species habitat suitability scores.

2.2.2. Logistic regression (LR)

Logistic regression (Hosmer and Lemeshow, 1989) is a special form of the generalized linear modelling (GLM; McCullagh and Nelder, 1983). Assuming that the probability of presence p given factors $X_1, \ldots, X_n$ is to be modelled, the logistic model assumes that the log of the odds (i.e. logit of the probability of presence p) is linear, i.e.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

where $\beta_0, \beta_1, \ldots, \beta_n$ denote the set of parameters to be estimated.

To allow further flexibility in fitting the relationship between the probability of presences and a set of predictors, a higher order polynomial fitting function was introduced in LR (Ferrer-Castan et al., 1995; Guisan et al., 1999; Thuiller, 2003), i.e.

$$\log\left(\frac{p}{1-p}\right) = \beta + f(X_1) + \cdots + f(X_n) \quad (5)$$

where $f(X_1), \ldots f(X_n)$ denote polynomial functions for each variable, which were allowed to take linear (i.e. $f(X_q) = \delta_{q_1} X_q$, with $q \in \{1, \ldots, n\}$), quadratic (i.e. $f(X_q) = \delta_{q_2} X_q^2 + \delta_{q_1} X_q$, with $q \in \{1, \ldots, n\}$), or cubic form (i.e. $f(X_q) = \varphi_{q_3} X_q^3 + \varphi_{q_2} X_q^2 + \varphi_{q_1} X_q$, with $q \in \{1, \ldots, n\}$). $\beta, \delta_{q_1}, \delta_{q_2}, \varphi_{q_1}, \varphi_{q_2}, \varphi_{q_3}$ with $q \in \{1, \ldots, n\}$ denote the set of parameters to be estimated. This study assessed the performance of LR method using the three polynomial functions. The LR method with the linear, quadratic and cubic function was denoted as LR-L, LR-Q and LR-C respectively. Statistical analyses for the LR method were performed using the glm function in standard R library (R Development Core Team, 2007).

2.2.3. Autocovariate model

The autocovariate model addresses SAC by estimating how much the response variable at any one site reflects response values at surrounding sites. This is achieved for the GLM model by adding a distance-weighted function of neighbouring response values. This additional predictor is known as the autocovariate.

Applying the autocovariate approach to the LR model (i.e. autologistic) transforms the standard form of Eq. (5) to:

$$\log\left(\frac{p}{1-p}\right) = \beta + f(X_1) + \ldots + f(X_n) + \rho R$$

where ρ is the estimated coefficient of the autocovariate R.

The autocovariate R at any site u is calculated as:

$$R_u = \frac{\sum_{v \in r_u} \omega_{uv} y_v}{\sum_{v \in r_u} \omega_{uv}}$$

where y_v is the observed value at site v surrounding site u, ω_{uv} is the weight of site v in relation to site u, and r_u is the number of neighbouring sites considered to be influential for site u (Augustin et al., 1996; Gumpertz et al., 1997). The weight of site v in relation to site u can be defined as a function of geographical distance (Augustin et al., 1996; Osborne et al., 2001; Segurado et al., 2006) or environmental distance between the two sites (Augustin et al., 1998; Ferrier et al., 2002). In this study, the reciprocal of Euclidean distance between sites was employed to define the neighbouring weight. The spatial autocovariate term was calculated utilizing autocov_dist function in spdep R library (Bivand, 2008).

The application of autologistic model is straightforward when presence–absence is surveyed at every site. However, when data are available only from a sample of sites, the autocovariate cannot be calculated directly as the species presence–absence pattern in neighbouring sites is not known. A solution to this problem is to estimate the probabilities of occurrence for the unsurveyed sites using the LR method. Once these estimated probabilities were obtained, the autocovariate for each site can be calculated. Then, the autologistic model can be constructed using these autocovariates and the estimated probabilities for each site can be obtained.

Augustin et al. (1996) recommended applying a Gibbs sampler algorithm to optimize the performance of autocovariate models on sampled data. The application of Gibbs sampler can be computer intensive. Many studies that use autocovariates do not use Gibbs sampler, yet these studies showed that the use of autocovariates alone improved predictive performance of the LR models and reduced SAC in the model residuals (Sanderson et al., 2005; Piorecky and Prescott, 2006). In the present study, the performance of the autocovariate logistic on sampled calibration data was assessed without the application of the Gibbs sampler routine.

2.2.4. Generalized additive modelling (GAM)

Generalized additive modelling (Hastie and Tibshirani, 1990) is a semi-parametric form of LR. It uses smooth functions instead of the usual regression coefficients used in LR. GAM was fitted using cubic splines $f(X_1), \ldots, f(X_n)$ as the smooth functions, i.e.

$$\log\left(\frac{p}{1-p}\right) = \beta + f(X_1) + \ldots + f(X_n)$$

Statistical analyses for the GAM method were performed using the gam function in the gam *R* library (Hastie, 2006).

2.2.5. Classification and regression tree (CART)

Classification and regression tree (CART; Breiman et al., 1984) is a non-parametric approach based on recursive partitions of dimensional space defined by dividing the predictor variables into groups that are as homogenous as possible for the response variable. A recursive algorithm is used to split the data into successive binary branches that at each stage yield the maximum reduction in residual deviance or improvement in the model overall fit. Any split that does not improve the overall fit by a prescribed factor, namely the complexity parameter, is not attempted. This procedure is called pruning. Although the main aim of pruning is to save computing time, it can also improve model accuracy by removing tree branches reflecting noise in the data (Han and Kamber, 2000, p. 373).

In this study, each simulated species was submitted to CART models by applying in turn complexity parameters of 0.01, 0.025, 0.05, 0.075, and 0.1. The model that yielded the maximum accuracy over these range of complexity parameters was chosen for further analysis. CART method was implemented using the rpart function in rpart R library (Therneau and Atkinson, 2008).

2.3. Assessing the performance of the modelling methods

2.3.1. Assessing the predictive performance of the fitted models

The first assessment for testing the reliability of modelling methods in responding to various species response forms was to evaluate the predictive performance of the fitted models against the independent simulated evaluation datasets. This predictive performance was assessed in two ways: (1) by examining the predictive performance of the univariate models (i.e. models including only one predictor variable); and (2) by examining the predictive performance of the full models (i.e. models which incorporate all three predictor variables) produced by each modelling method.

The assessment of the predictive performance of a univariate model can be used to infer which variables exert the greatest influence on species distribution. When modelling real data, the relative importance of the environmental variables for species distribution and the underlying species responses to the environmental variables are not known explicitly. Therefore, it is not possible to test the performance of univariate models in delivering correct inferences on the contributions of variables in determining species occurrence. However, given that in this simulation study the relative importance of each variable and the responses of the simulated species to each environmental variable were known, the success of each modelling method in identifying the relative contributions of the contributing variables was able to be assessed. Predictive accuracies of the univariate models should conform with the weight of the variable assigned for each simulated species. The correctness of the estimated univariate model predictive accuracy provided an indication of the sensitivity of the modelling method to species response form. For each modelling method, 594 ($66 \times 3 \times 3$) univariate models for the complete data (3600 data points) and 29,700 ($66 \times 3 \times 3 \times 50$) models for each of the sampled calibration data (sampled data with 720, 360 and 72 data points) were evaluated.

The analysis of the full models assessed whether predictive performance varied with respect to the form of species response applied to the dominant variable (i.e. the variable with the highest contribution or weight). A good method would therefore be the one that was able to accurately predict occurrences for all simulated species, regardless of the type of species response of the dominant variable. For each modelling method, 198 (66×3) full models for the complete data (3600 data points) and 9900 ($66 \times 3 \times 50$) models for each of the sampled calibration data (sampled data with 720, 360 and 72 data points) were evaluated.

Regression methods such as LR, GAM and CART require records of both presences and absences for model calibration. Therefore, all presence and absence information of the simulated calibration data was used to fit these models. The environmental envelope method BIOCLIM only requires presence records to calibrate the simulated species distribution model. Therefore, only presence records of the simulated calibration data were needed to run this model. The performance of the models produced by all modelling methods was evaluated using both presence and absence records of the simulated evaluation data.

Two indicators for measuring the predictive performance were used. These were the kappa accuracy index (denoted by κ; Cohen, 1960) and the Area under the Receiver Operating Characteristic curve (AUC; Hanley and McNeil, 1982).

The range of possible values of κ is from −1 to 1. Positive values indicate higher model predictive performance, with unity representing perfect agreement (i.e. both model and evaluation (actual) datasets agree in their classification in every case). Because κ has various values depending on the choice of probability threshold used to define species presence and absence, Liu et al. (2005) recommended the use of κ value that maximizes the model predictive performance across various thresholds as a representation of the overall accuracy of the model. This is denoted as κ_{max} which will be used henceforth throughout this paper.

The Area under the Receiver Operating Characteristic curve is a threshold-independent accuracy measure. Receiver Operating Curve is a graphical method that represents the relationship between false positive (1–specificity) and sensitivity as a function of probability thresholds ranging from 0 to 1. If all predictions were expected to occur by chance alone, the relationship would be a 45° line. Good model performance is characterized by a curve that maximizes sensitivity for low values of (1–specificity) or when the curve passes close to the upper left corner of the plot. The area between the 45° line and the curve measures the ability of the model to correctly classify a species as present or absent. The calculation of AUC for each model was performed using the roc.area function in the verification R library (NCAR Research Application Program, 2008).

2.3.2. Assessing goodness-of-fit of the LR and GAM models

The second assessment for testing the performance of modelling method in responding to various species response form was to examine the goodness-of-fit of the LR and GAM models. A standard regression goodness-of-fit measure, the Akaike's Information Criterion (Akaike, 1974) based on a maximum likelihood estimate (MLE) was used to assess the models' fitness. The performance of LR and GAM was assessed by examining the trend of the relationship between the AIC of the univariate models produced by each method against the predefined variable weight for the simulated species.

2.3.3. Assessing residuals of models generated by the regression methods

The third assessment for testing the performance of modelling methods in responding to various species response forms was to examine the strength of SAC in the deviance residuals (McCullagh and Nelder, 1983; Pierce and Schafer, 1986) of the fitted univariate models and the fitted full models generated by the regression-based methods. The deviance residuals of univariate regression models were assessed by: (1) testing whether predictors with a particular type of species response left out of a model inflated the strength of SAC in the model residuals; (2) testing whether the level of SAC in model residuals due to missing important explanatory variables was higher for some modelling methods than others; and (3) testing whether the density of data used for model building affected the significance of SAC being detected in model residuals.

The strength of SAC of the residuals of the models was measured by global Moran's I (Fortin and Dale, 2005). The eight nearest neighbours' contiguity rule was employed for defining the links between sites. The Moran's I coefficients were calculated by the moran function in the spdep R library (Bivand, 2008). The strength of SAC in the residuals of the full regression models was used to measure the impact of the inability of the regression methods to respond to a particular type of species response. A Moran's I of approximately zero indicates the lack of SAC. A good modelling method would therefore be one that is able to yield model residuals with Moran's I values concentrated around zero, i.e. randomly distributed residuals, for any type of species response applied to the dominant variable in the simulated species.

2.3.4. Assessing the impact of incorporating spatial autocovariate into logistic regression

The concluding aim of this study was to assess whether the inclusion of a spatial autocovariate term into the logistic regression model could improve the predictive ability of the LR model and reduce the SAC of the model residuals. The performance of the autologistic model was tested by adding the spatial autocovariate term to the LR model with each of the three polynomial functions. To differentiate among the three models, the application of autocovariate regression to LR model with linear, quadratic and cubic functions was denoted by ACL-L, ACL-Q and ACL-C, respectively.

The impact of spatial autocovariate term was investigated using the full calibration set (3600 data points) and random calibration samples of 720 data points (20% of the full data). For the full calibration set, all neighbours within a 1.5 unit distance from the central point were used to define the neighbouring links between sites. For the random calibration samples, all neighbours within a 1.5, 3, and 5 unit distances from the central point were used to define the neighbouring links.

3. Results

3.1. Model predictive performance

3.1.1. Predictive performance of the univariate models

Fig. 3 shows the predictive performance of the univariate models based on AUC generated using the full calibration set (3600 data points) in relation to the dominant contributing environmental variable for simulated species with B-L-G set of responses. Similar results were obtained for κ_{max} and the other two sets of responses (G-B-L and L-G-B). All modelling methods correctly identified the species with a dominant Beta response, as shown in first row of Fig. 3. Row 3 of Fig. 3 shows that the LR-L method did not identify the simulated species with a dominant Gaussian response. BIOCLIM tended to underrate the contributions of variables with a linear response, as shown in row 2 of Fig. 3. This is to be expected since BIOCLIM is based on an assumption that all of the modelled environmental responses are bell-shaped. The impact is less serious than the deficiency in LR-L described above, but illustrates a shortcoming in BIOCLIM when the simulated species responds to important environmental variables in a linear way.

3.1.2. Predictive performance of the full models

The performance of the full models generated by each modelling method based on AUC as a function of calibration size for simulated species with B-L-G set of responses is shown in Fig. 4. Similar results were obtained for κ_{max} and the other two sets of responses (G-B-L and L-G-B). For the full samples calibration data (3600 data points), as shown in Fig. 4, the GAM method and the higher order LR methods yielded good predictive performance for all types of species responses of the dominant variables. BIOCLIM and CART methods had poorer predictive performance for simulated species whose dominant predictor had a linear response compared with those whose dominant predictors had a Gaussian response. Discrepancies between the predictive performance for the two different species were more significant for models fitted by BIOCLIM compared with the discrepancies for the models fitted by CART, confirming the finding from the univariate analysis that BIOCLIM, and to an extent CART, performs less well when the dominant environmental response is linear.

As the density of the calibration data decreased, the performance of BIOCLIM, CART, GAM and the higher order LR methods became comparable. While data density had minimal impact on the performance of BIOCLIM, it caused the predictive ability of CART, GAM and the higher order LR to reduce substantially with

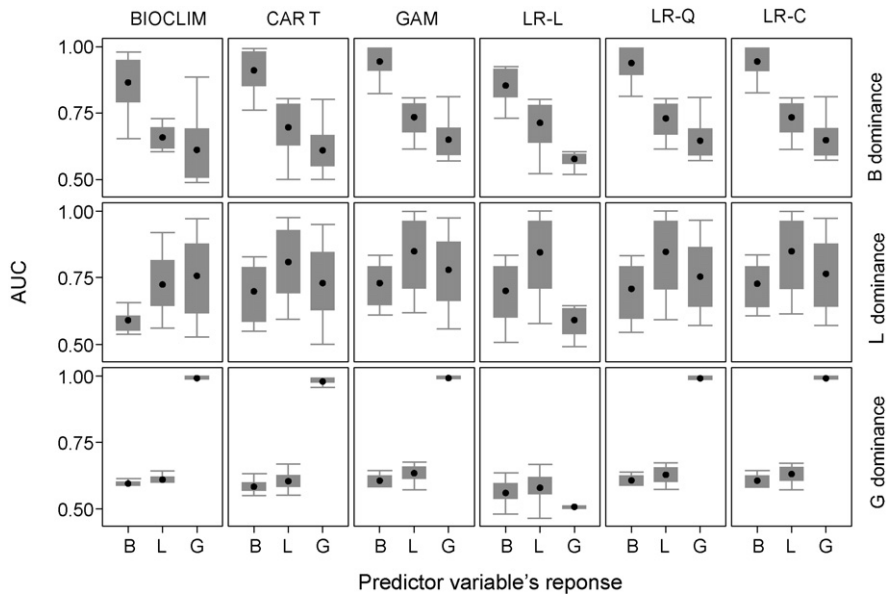

Fig. 3. The predictive performance, as measured by AUC, of the univariate models generated by each modelling method using full sample calibration data for simulated species with the B-L-G set of responses. The results are grouped by the dominant simulated variable for each type of species response. Each boxplot summarized the results for 20 univariate models.

sparse calibration data. Among the modelling methods studied, BIOCLIM produced the least reduction rate and variability in the predictive performance as density of the calibration data decreased, as shown in Fig. 5 for simulated species with the B-L-G set of responses.

3.2. Goodness-of-fit of the LR and GAM models

Fig. 6 shows that the AIC of the univariate models generated by LR-Q and LR-C declined as the weights of variables increased for all types of species response. This indicated the consistency of LR-Q and LR-C in providing correct inferences about the contributions of variables. The AIC of GAM univariate models also showed a decline as the weight of variable increased. However LR-L showed almost no response as the variable weights increased for species with a Gaussian response.

3.3. Spatial autocorrelation in regression model residuals

3.3.1. Spatial autocorrelation in univariate model residuals

Fig. 7 shows the SAC in univariate model residuals for the four regression-based methods generated using the full calibra-

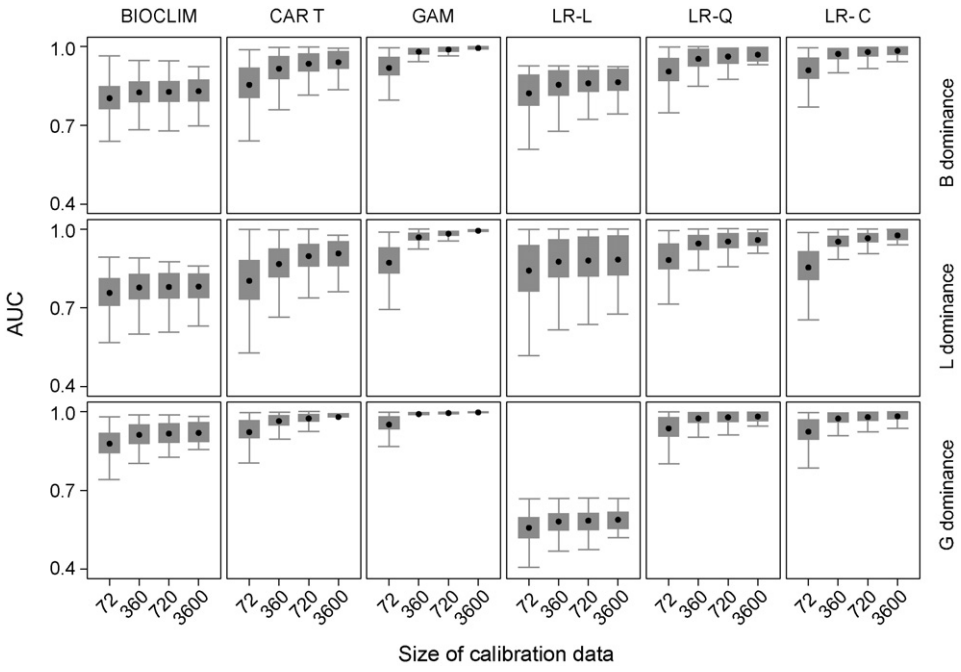

Fig. 4. The predictive performance, as measured by AUC, of the full models generated by each modelling method as a function of calibration size for simulated species with the B-L-G set of responses. Labels on vertical axis denote the dominant simulated variable for each type of species response. Each boxplot for the full calibration data summarized the results for 20 full models and each boxplot for the sampled calibration data summarized the results for 1000 (20×50) full models.

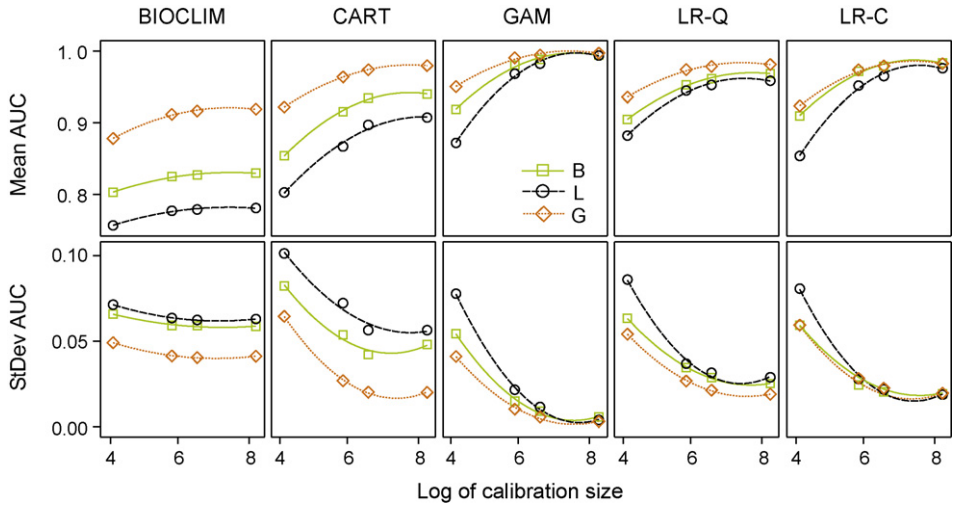

Fig. 5. Quadratic trend in the mean and standard deviation of predictive performance (measured by AUC) on log calibration size for the simulated species with the B-L-G set of responses. BIOCLIM produced the least reduction rate and variability in the predictive performance as calibration size decreased.

tion samples (3600 data points) with respect to the dominance of each predictor variable for simulated species with B-L-G set of response. Similar results were obtained for the G-B-L and L-G-B sets of response. The largest SAC in model residuals occurred in row 3 of Fig. 7 for all four methods when the dominant predictor with a Gaussian response was omitted from the model. Similarly, the SAC in model residuals showed a moderate increase in first row of Fig. 7 when the dominant predictor with a Beta response was omitted from the model.

Moreover, the LR method, showed an increase in SAC in model residuals when the functional form of true species response was misspecified (i.e. linear function was used to fit the non-linear species response). This can be seen for the residuals of the LR-L model for simulated species whose dominant variables respond in a Gaussian way as shown in the row 3 of Fig. 7.

3.3.2. Spatial autocorrelations in full model residuals

The strength of SAC in the residuals of full regression models as a function of calibration size for simulated species B-L-G are shown in Fig. 8. Similar results were obtained for the other two sets of responses (G-B-L and L-G-B). Analysis using the complete calibration data (3600 data points) suggests that misspecification of the functional form of true species response in LR-L model substantially increased the SAC in the model residuals. The strength of SAC in the model residuals for species whose dominant variables possessed a Beta and Gaussian response were substantially higher compared with model residuals for species whose dominant variables possessed a linear response. The strength of SAC as a consequence of misspecification of the functional form of true species response in LR-L model residuals became less significant for less dense calibration data.

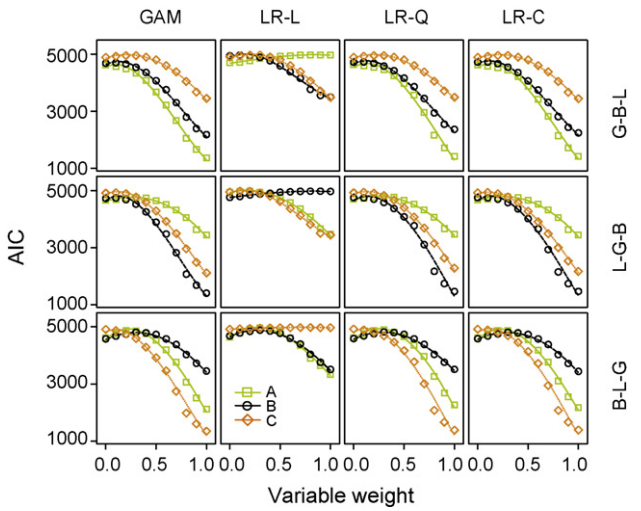

Fig. 6. The relationship between the mean goodness-of-fit of the univariate models, as measured by AIC, generated by the GAM and LR methods using full sample calibration data and variable weights by variable *A*, *B* and *C* for the simulated species G-B-L, L-G-B, and B-L-G. Solid lines represent cubic fits.

3.4. The effect of incorporating spatial autocovariate into logistic regression

The three full LR models were extended by including an autocovariate term. The resulting predictive performance of the full models measured by AUC before and after the inclusion of the spatial autocovariate term for simulated species B-L-G is shown in Fig. 9. Similar results were obtained for κ_{max} and the other two sets of responses. For the complete calibration data (3600 data points), the predictive performance of models for species predominantly affected by variables with Gaussian response improved dramatically after the spatial autocovariate term was added to the LR-L models (first row and third column of Fig. 9a). Consequently, the SAC in model residuals was dramatically reduced for the ACL-L (i.e. autologistic approach applied to LR with linear function) compared with LR-L (second row and third column of Fig. 9a).

The incorporation of the spatial autocovariate variable into the LR-L full model was found to be less effective with limited calibration data. Predictive performance of the LR-L model showed no improvement for models with a dominant Gaussian response based on 720 samples (first row and third column of Fig. 9b). The strength of SAC in the model residuals reduced mildly after adding the autocovariate term (second row and third column of Fig. 9b).

The resulting predictive performance of the univariate models measured by AUC before and after the inclusion of the spatial autocovariate term with respect to the dominance of each predictor variable for simulated species B-L-G is shown in Fig. 10. Similar results were obtained for κ_{max} and the other two sets of responses. Adding autocovariates to the LR univariate models using complete calibration data substantially improved predictive performance (rows 1, 2, and 3 of Fig. 10a) and reduced the SAC in the residuals of the models (rows 4, 5, and 6 of Fig. 10a). For the sam-

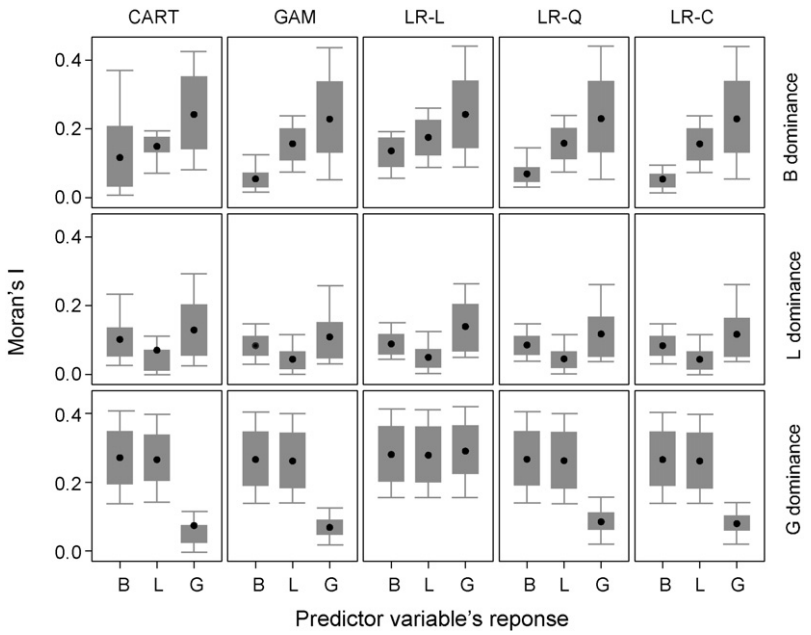

Fig. 7. The strength of SAC in the residuals of the univariate models (measured by Moran's I) generated by each regression-based method using full sample calibration data for the simulated species B-L-G grouped by the dominant variable. This shows the impact of omitting important explanatory variables with Gaussian response. Each boxplot summarized the results for 20 univariate model residuals.

pled calibration data (720 data points), the autocovariates offered a mild improvement in model predictive performance (rows 1, 2, and 3 of Fig. 10b) and reduction in the strength of SAC in the LR model residuals (rows 4, 5, and 6 of Fig. 10b).

The significance of estimated variable coefficients for the LR and ACL full models for simulated species with B-L-G set of responses grouped by the dominant variable are shown in Fig. 11. Similar results were obtained for the G-B-L and L-G-B sets of responses. Incorporating the spatial autocovariate term into the LR-L full model using full calibration samples (3600 data points) did not change markedly the significance of the estimated coefficients of variables *A*, *B* and *C* (corresponding to B, L and G types of responses, respectively) for the species with a dominant linear or beta response (first and second row of Fig. 11a). However, the significance of the coefficient of variable *C* (with G type of response) for species with a dominant Gaussian response declined markedly, while the coefficient of the corresponding autocovariate was highly significant (third row of Fig. 11a). This indicated that the good predictive performance of the ACL-L model on species with a dominant Gaussian response was solely due to the full

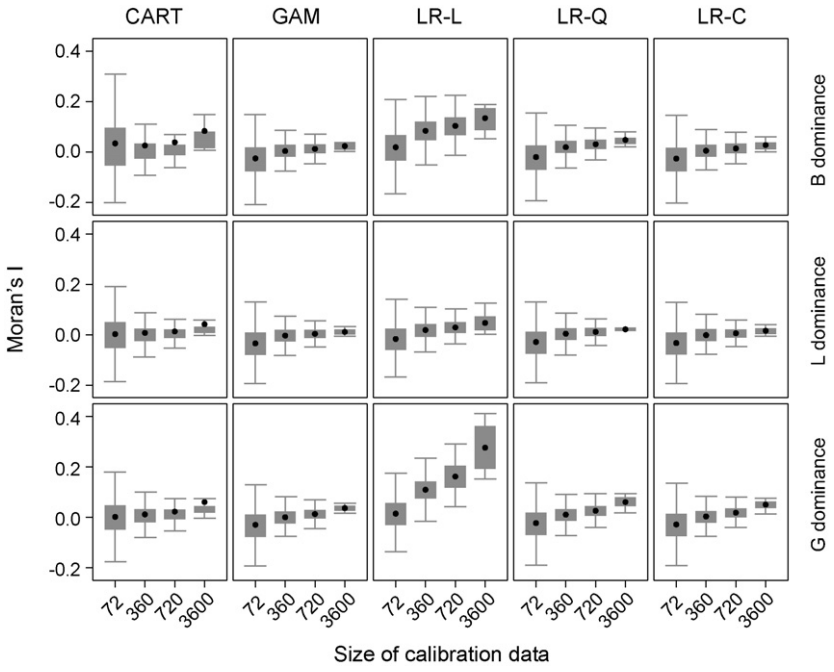

Fig. 8. The strength of SAC in the residuals of full models (measured by Moran's I) generated by each regression-based method as a function of calibration size for simulated species B-L-G. Labels on vertical axis denote the dominant simulated variable for each type of species response. Each boxplot for the full calibration data summarized the results for 20 full model residuals and each boxplot for the sampled calibration data summarized the results for 1000 (20×50) full model residuals.

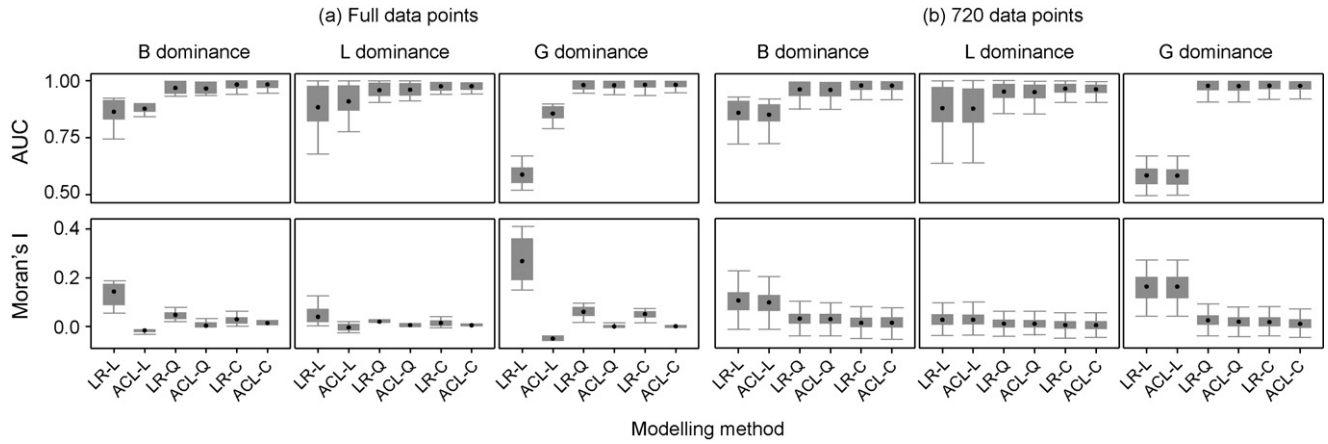

Fig. 9. The predictive performance of the LR and ACL full models (measured by AUC) and the strength of SAC in the residuals of the models (measured by Moran's I) for the B-L-G set of simulated species using calibration data with: (a) full samples (3600 data points); and (b) 720 data points. Each boxplot for the full calibration data summarized the results for 20 full models, while each boxplot for the sampled calibration data summarized the results for 1000 (20×50) LR models and 3000 ($20 \times 50 \times 3$) ACL models.

spatial coverage of the calibration data. This was confirmed by the decline in the ACL-L performance on less dense data. For sparse data the autocovariate was unable to make up for the deficiency in the LR-L model for species with a dominant Gaussian response.

The estimated coefficients of the autocovariate variable in higher order LR models (quadratic and cubic) for the full data coverage were significant, particularly for simulated species with a dominant Gaussian response (row 3 of Fig. 11a). The estimated coefficient of the spatial autocovariate variable was significant whether or not

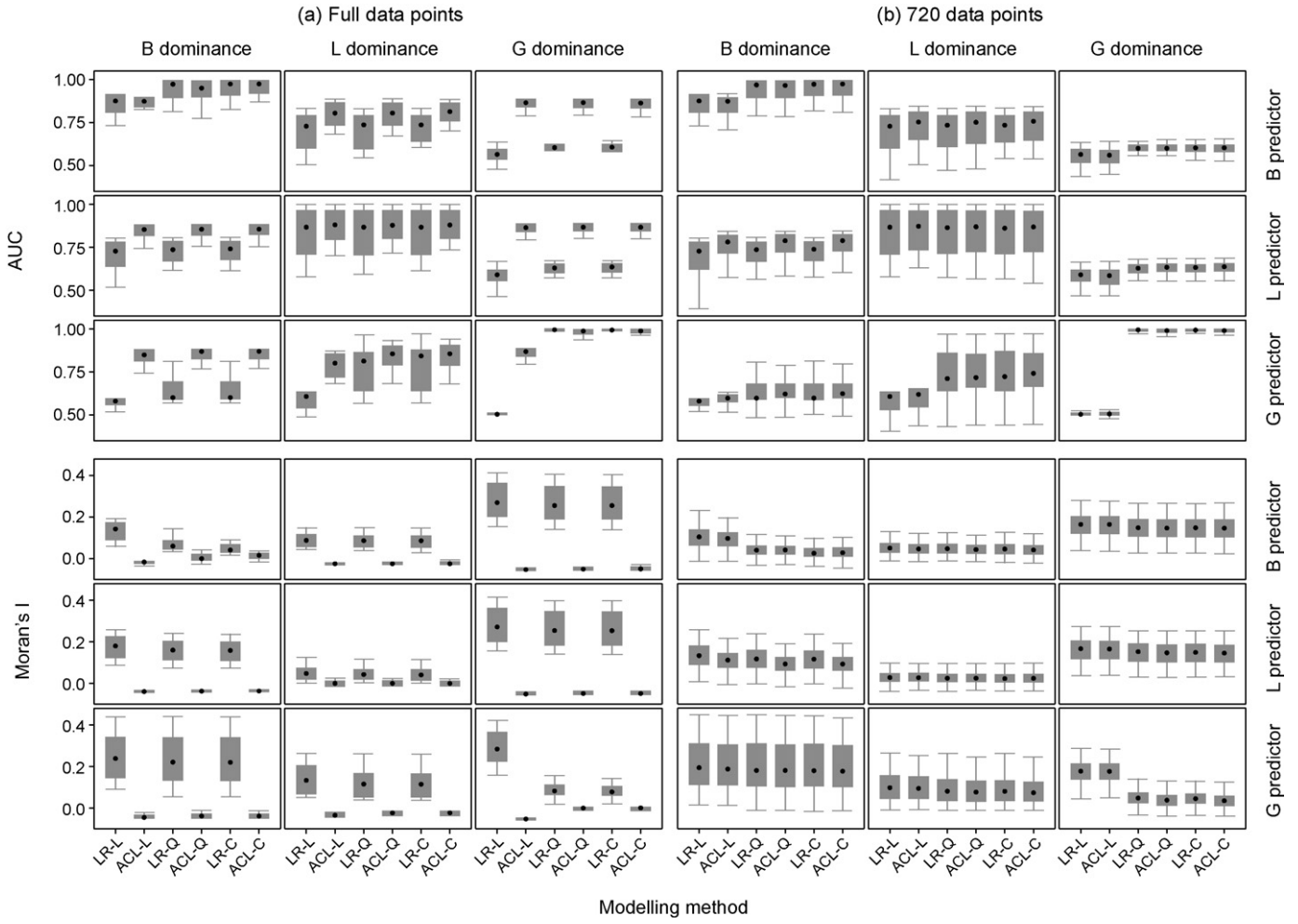

Fig. 10. The predictive performance of the LR and ACL univariate models (measured by AUC) and the strength of SAC in the residuals of the models (measured by Moran's I) for each simulated species B-L-G using calibration data with: (a) full samples (3600 data points); and (b) 720 data points. Each boxplot for the full calibration data summarized the results for 20 univariate models, while each boxplot for the sampled calibration data summarized the results for 1000 (20×50) LR models and 3000 ($20 \times 50 \times 3$) ACL models.

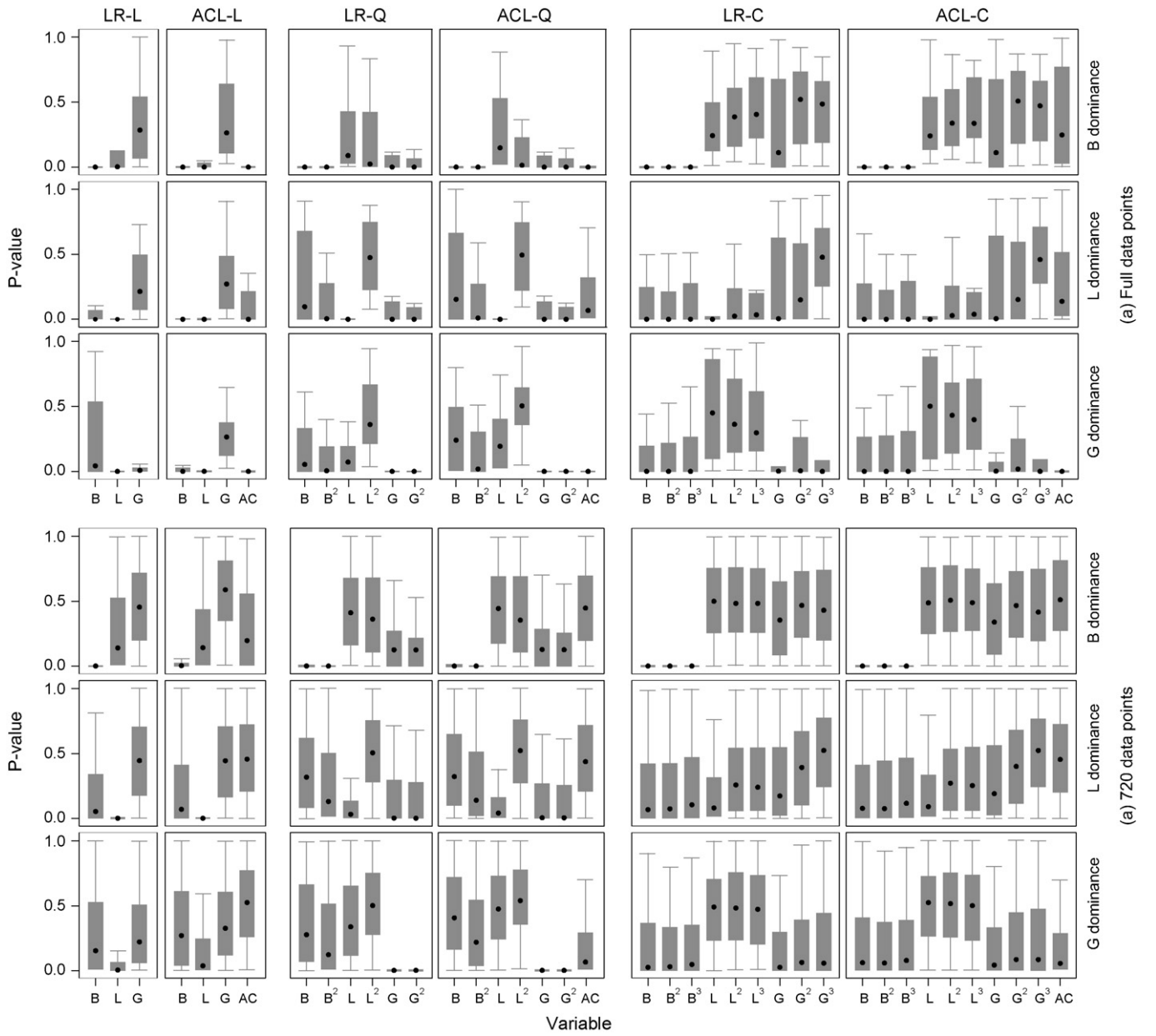

Fig. 11. The significance (*P*-value) of estimated variable coefficients for the LR and ACL full models for simulated species B-L-G grouped by the dominant variable, using calibration data with (a) full samples (3600 data points); and (b) 720 data points. Lower *P*-value indicates that there is evidence to reject the null hypothesis that the coefficient is zero. AC denotes the spatial autocovariate variable.

the LR model alone has captured adequately the effect of the environmental determinants. With sparser data (720 data points), the estimated coefficients of the spatial autocovariate term became less significant (row 3 of Fig. 11b).

4. Discussion

Simulated data have increasingly been used for evaluating the performance of species distribution modelling methods (Hirzel et al., 2001; Austin et al., 2006; Meynard and Quinn, 2007). Simulated data provide a more reliable tool for assessing the performance of modelling methods compared with real data sampled in the field. Statistical methods for modelling species distribution are difficult to evaluate using real data because real data are confounded by unknown driving factors. The level of complexity of real data may be different from one species to another. Some modelling methods may respond to particular data characteristics better than other methods. When real species distribution data with particular property are used to assess the performance of modelling methods, the methods that respond to this property better than other methods may have superior performance than the other methods.

Data density also affects the reliability of real species data for evaluating the performance of modelling methods. The density of data for model building has been shown to contribute to modelling uncertainties (Stockwell and Peterson, 2002; Barry and Elith, 2006; Hernandez et al., 2006; Wisz et al., 2008). As shown in this study, each modelling method responds to calibration data density in different manner in terms of predictive performance.

In this study, the data were constructed so that:

(1) spatial distributions of the simulated species were determined by spatially correlated environmental variables;
(2) the response of the simulated species to each environmental variable took a variety of forms (i.e. linear and non-linear);

(3) the environmental determinants were combined in an additive manner (i.e. there was no multiplicative effect or interactions between the environmental determinants);
(4) no species biological distance-related process (e.g. species dispersal, active spatial aggregation) were considered;
(5) data errors were likely to occur at sites where habitat suitability score for the simulated species were at intermediate level (i.e. high degree of uncertainty); and
(6) the simulated species were assumed to have the same level of rarity (prevalence or proportion of species presence over all sampled sites) of 0.5.

Although the simulated data used in this study do not reflect all aspects of actual species distribution found in nature, it does represent a range of realistic species distribution behaviour and incorporates the possibility of error arising during field sampling. The simulated species data proved to be an effective instrument in exploring how modelling methods perform in response to various underlying forms of species responses to spatially correlated environmental variables and the density of calibration data.

4.1. Model performance with respect to species response form and data density

The simulation results obtained from full samples calibration data (3600 data points) showed that among the four modelling methods studied, GAM was able to maintain the consistency of model predictive performance across various forms of species response of the simulated species. With sparse data this method was however inclined to underrate the performance of variables possessing linear response.

The LR method was prone to failing to capture the underlying species response forms in relation to environmental factors. This occurred, particularly when LR with a linear fitting function was applied to a non-linear species response. Under such conditions, LR-L severely underrated the significance of variables with a non-linear species response. This can yield false inferences pertaining to the true significant environmental factors related to species distribution.

The use of the LR method for predicting spatial distributions of plants and animals is common. When applying LR method, the common practice is to apply linear, quadratic or cubic polynomial fitting functions. Austin (2007) however recently noted that modellers using the LR method for predicting species presence–absence distributions often do not recognize the need: (1) to specify the type of function of species response; and (2) to investigate the possibility of using higher order polynomial functions. The application of the LR-L method without justification that a species responds to a set of environmental candidates in purely linear way can lead to serious bias.

Unlike LR-L which responds poorly to variables with a non-linear response, the simulation study using complete sample calibration data (3600 data points) showed that BIOCLIM and CART methods had a tendency to underrate the significance of variables with a linear response compared with the variables with a non-linear response studied. The discrepancy between the performance of variables with non-linear response and the performance of variables with linear response was larger for BIOCLIM than for CART, suggesting that BIOCLIM was more susceptible to variation in types of species response than the CART method. Because BIOCLIM and CART methods were able to rank univariate model performance appropriately according to the true contributions of the variables, they may not cause a serious bias. Nonetheless, users of BIOCLIM and CART should be aware of such effects when an assessment of univariate model predictive performance is used as a basis for selecting important explanatory variables.

In practice, species distributions are usually determined from relatively sparse sample data. BIOCLIM was less sensitive to density of calibration data. The predictive ability of models generated from CART, GAM and higher order LR reduced substantially under limited calibration data (72 data points or 2% of the full data points). A study carried out by Stockwell and Peterson (2002) found a similar conclusion about the sensitivity of the performance of higher order LR models on calibration size. Wisz et al. (2008) also recently reported that novel methods with capabilities to incorporate complex species response forms such as multivariate adaptive regression splines (MARS; Friedman, 1991), a modified version of GAM, performed well with large calibration datasets but less well with limited calibration data.

4.2. The causes of spatial autocorrelation in regression model residuals

While the former applies particularly to LR-L method, the later can occur with all regression-based methods. This study showed that model misspecification increased the strength of SAC in regression model residuals. Model misspecification was caused by:

(1) incorrect specification of functional form of species response curve (i.e. using a linear function to fit a non-linear species response); and/or
(2) omission of important predictors with a Gaussian response.

The occurrence of SAC in model residuals caused by using the incorrect functional form has been described in spatial analysis and regression modelling literature, mainly in the context of ordinary least square (OLS) application (Haining, 1990, p. 332; Cliff and Ord, 1981, p. 211). The occurrence of SAC in model residuals caused by omission of important predictors has also been highlighted in spatial analysis literature (Cliff and Ord, 1981, p. 197; McMillen, 2003). However, the fact that omitting important predictors with a non-linear response contributed to increased SAC in model residual is not explicitly well known.

Besides the model misspecifications described above, SAC can also occur due to omission of species biological distance-related processes, such as speciation, extinction, dispersal, growth rate and species interaction (Bahn et al., 2007; Miller et al., 2007). Given the work of this paper, omission of important predictors with a non-linear response can be considered to fall into the same category as omission of species biological distance-related processes, because in both situations the model omits the important environmental determinants which account for SAC for species distribution.

It should be noted that the simulated species data used in this study were constructed from artificial environmental variables with some degree of SAC. This mimics natural phenomena where environmental factors are likely to be spatially structured. Spatially autocorrelated residuals caused by model misspecification may not occur if the simulated species are generated from a randomly distributed artificial variable with no spatial dependence.

The strength of SAC observed in the residuals of regression-based models can vary depending on the amount of information carried by calibration data. For the LR method with a linear fitting function, SAC in the model residuals was significant when species data were densely sampled, but insignificant when species data were sparse. This explains why the commonly used sub-sampling procedure can be successful in eliminating the effect of SAC when fitting linear functions with the LR model. Segurado et al. (2006) have reported that by systematically sub-sampling the study area, the effect of SAC in the LR-L model can be reduced. Although this approach lessened the effect of SAC in the LR-L model, it may not correct the misinterpretations of the dominant predictors found by the model.

4.3. Autocovariate logistic in species distribution modelling

When SAC occurs in model residuals, a common approach to eliminate the effect is to add a spatial autocovariate term to the model. When applied to the LR model, this is known as the autologistic approach. Although the autocovariate can in principle be added to LR model with any form of fitting function, it has mainly been applied to LR with a linear function models (Sanderson et al., 2005; Jewell et al., 2007).

The autologistic model was first introduced in species presence–absence modelling by Augustin et al. (1996) to improve the performance of an earlier LR-L model for red deer distribution in the Grampian Region in Scotland (Buckland and Elston, 1993). The authors noted that the SAC in the earlier LR-L model residuals occurred because the available covariates did not fully reflect the actual conditions for the red deer distribution. The autologistic approach was shown to improve the predictive performance of the LR-L model and reduce the strength of SAC in the model residuals. Given that the authors applied the autocovariates specifically to LR with a linear function model to real species data, it is not clear whether the autocovariates essentially addressed the spatially correlated residuals due to model misspecification or addressed species biological distance-related processes.

Concern regarding the possible misuse of the autocovariates with LR with a linear function models was raised by Austin (2002). The author pointed out that the autocorrelation found to be reduced in several earlier papers by incorporating a spatial autocovariate term in LR with linear function was probably due to model misspecification.

The performance of the autologistic approach has so far mainly been assessed on its ability to address the effect of species dispersal (Wintle and Bardos, 2006; Dormann et al., 2007). To our knowledge, this was the first study to focus specifically on the effect of model misspecification. Moreover, because the performance of the autologistic approach was evaluated on full data coverage and sampled calibration data, the practical utility of the approach could be assessed.

In this study, it was found that with complete sample calibration data (3600 data points), the predictive performance of models for species predominantly affected by variables with a Gaussian response improved dramatically after a spatial autocovariate term was added to the LR with linear fitting function (LR-L) models. The level of SAC in the model residuals was markedly reduced and the predictive performance of the model improved substantially. This suggests that the autocovariate also has a prominent effect on models that have omitted important non-linear response predictors. Thus, with full data coverage, the autocovariates successfully make up for the lack of model performance due to model misspecification, whether it was due to misspecifying the functional form in LR-L or omission of non-linear environmental determinants from the regression model. However, it is important to note that the improved model performance was solely due to the information obtained from the complete coverage of the calibration data via the autocovariate. The significance of the predictor based on the Gaussian variable declined.

The results of analysis based on the complete data coincided with the findings from an earlier study by Dormann et al. (2007). The autocovariate approach is superior in terms of improving model prediction and eliminating the effect of SAC in LR model residuals when information regarding species presence–absence is available from each neighbouring sites. However, care should be taken when interpreting the parameter coefficients based on the autologistic model. Augustin et al. (1998) noted that the autocovariate approach may be suitable for improving the prediction use of regression rather than for parameter estimation or inferential purposes.

In the more realistic application where calibration data are relatively sparse, the basic autocovariate approach offered a mild improvement in the performance of LR model and reduction in the strength of SAC in the model residuals. The performance of the autocovariate approach on sampled data may perform better with the application of Gibbs sampler routine, the approach that was not attempted in the present study. The Gibbs sampler routine extends the basic autologistic approach by iteratively updating the probability of species occurrence for each unsurveyed site based on the probability value of the neighbouring sites. Improved prediction has been found in studies using Gibbs sampler in autologistic modelling (Hoeting et al., 2000; McPherson et al., 2004).

Despite the omission of the Gibbs sampler, the present analysis has provided an insight into the scope of problems that the basic autocovariate model can address. The autocovariate model performs at best when there is full data coverage (i.e. when presence–absence information is available at each neighbouring site). In the more practical applications with sampled data, the basic autocovariates slightly improved the LR model performance where the deficiency was due to model misspecification, as shown in Fig. 10b.

5. Conclusion

This study assessed the performance of four modelling methods in terms of their abilities to respond to various underlying species responses to environmental variables. Using simulated species distribution data, it was found that GAM produced the most consistent model performance among the four non-spatial methods, regardless of the forms of species response of the simulated species. BIOCLIM and CART had a tendency to underrate the significance of variables with a linear response compared with variables with a non-linear response. The BIOCLIM model was least sensitive to data density.

Logistic regression was susceptible to failing to accurately identify the importance of determinant variables for the simulated species distribution. This occurred particularly when LR method applied an incorrect function of the true species–environmental relationship, i.e. when a linear function was used to fit a non-linear species response. The misspecification of the functional form in LR had crucial implications: (1) an underestimation of variables with non-linear properties; and (2) an increased spatial autocorrelation in model residuals. Omission of important environmental predictors with a non-linear response in the simulated species distribution model also contributed to an increase spatial autocorrelation in the model residuals. Given that non-linear species response to environmental determinants arises commonly in nature, it can be concluded that: (1) higher order functions should always be tested when applying LR in modelling species distribution; and (2) spatially correlated regression model residuals can indicate missing non-linear environmental determinants for species distribution in the model.

The use of the autologistic approach was able to improve model prediction and reduce the strength of spatial autocorrelation in the model residuals in the presence of full coverage of calibration data, but this did not help the model to identify the true dominant Gaussian predictor. This is to be expected since autologistic approach was designed primarily for improving the prediction in regression and not for model estimation. However, improved prediction in the presence of full coverage of data has limited practical value. In the presence of sparse data, the more usual application, when the predictors in the LR were misspecified, the basic autocovariate offered only a mild improvement in the performance and a mild reduction in strength of spatial autocorrelation in the model residuals.

Acknowledgements

We sincerely thank Adrian Manning, Mike Austin, Sue Holzknecht and two anonymous reviewers for their valuable comments on an earlier draft of this manuscript. The first author was supported by an Australian National University Postgraduate Research Scholarship.

References

Akaike, H., 1974. New look at statistical-model identification. IEEE Trans. Automat. Contr. AC19, 716–723.
Anselin, L., 2002. Under the hood: issues in the specification and interpretation of spatial regression models. Agricult. Econ. 17, 247–267.
Araujo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species-climate impact models under climate change. Global Change Biol. 11, 1504–1513.
Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1996. An autologistic model for the spatial distribution of wildlife. J. Appl. Ecol. 33, 339–347.
Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1998. The role of simulation in modelling spatially correlated data. Environmetrics 9, 175–196.
Austin, M.P., 1976. Nonlinear species response models in ordination. Vegetatio 33, 33–41.
Austin, M.P., 1985. Continuum concept, ordination methods and niche theory. Ann. Rev. Ecol. Syst. 16, 39–61.
Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19.
Austin, M.P., Meyers, J.A., 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecol. Manag. 85, 95–106.
Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. Vegetatio 5, 215–228.
Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D., Luoto, M., 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. Ecol. Model. 199, 197–216.
Bahn, V., Krohn, W.B., O'Connor, R.J., 2007. Dispersal leads to spatial autocorrelation in species distributions: a simulation model. Ecol. Model. 213, 285–292.
Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423.
Besag, J., 1974. Spatial interaction and statistical-analysis of lattice systems. J. Roy. Statist. Soc. Ser. B 36, 192–236.
Betts, M.G., Diamond, A.W., Forbes, G.J., Villard, M.A., Gunn, J.S., 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. Ecol. Model. 191, 197–224.
Bivand, R., 2008. Spatial Dependence: Weighting Schemes, Statistics and Models. R package version 0.4-20.
Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA.
Buckland, S.T., Elston, D.A., 1993. Empirical-models for the spatial-distribution of wildlife. J. Appl. Ecol. 30, 478–495.
Busby, J.R., 1991. BIOCLIM—A bioclimate analysis and prediction system. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, pp. 64–68.
Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion Limited, London, 266 pp.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.
Dirnbock, T., Dullinger, S., Grabherr, G., 2003. A regional impact assessment of climate and land-use change on alpine vegetation. J. Biogeogr. 30, 401–417.
Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.
Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.
Ferrer-Castan, D., Calvo, J.F., Esteveselma, M.A., Torresmartinez, A., Ramirezdiaz, L., 1995. On the use of 3 performance-measures for fitting species response curves. J. Veg. Sci. 6, 57–62.
Ferrier, S., Drielsma, M., Manion, G., Watson, G., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales II. Community-level modelling. Biodiver. Conserv. 11, 2309–2338.
Fortin, M.J., Dale, M.R.T., 2005. Spatial Analysis: A Guide for Ecologist. Cambridge University Press.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 19, 1–67.
Guisan, A., Theurillat, J.P., 2000. Equilibrium modeling of alpine plant distribution: how far can we go? Phytocoenologia 30, 353–384.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modeling of plant species distribution. Plant. Ecol. 143, 107–122.
Gumpertz, M.L., Graham, J.M., Ristaino, J.B., 1997. Autologistic model of spatial pattern of *Phytophthora* epidemic in bell pepper: effects of soil variables on disease presence. J. Agricult. Biol. Environ. Stat. 2, 131–156.
Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press.
Han, J., Kamber, M., 2000. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 143, 29–36.
Hastie, T., 2006. gam: Generalized Additive Models. R package version 0.98.
Hastie, T.J., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.
Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.
Hijmans, R., Guarino, L., Jarvis, A., O'Brien, R., Mathur, P., Bussink, C., Cruz, M., Barrantes, I., Rojas, E., 2005. DIVA-GIS Version 5.2. University of California.
Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.
Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036.
Hoeting, J.A., Leecaster, M., Bowden, D., 2000. An improved model for spatially correlated binary responses. J. Agric. Biol. Environ. Stat. 5, 102–114.
Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley Interscience Publication, Chichester, UK.
Houlder, D., Hutchinson, M., Nix, H.A., MacMahon, J., 1999. ANUCLIM Version 5.0 User Guide. Centre for Resource and Environmental Studies, Australian National University, Canberra.
Jewell, K.J., Arcese, P., Gergel, S.E., 2007. Robust predictions of species distribution: spatial habitat models for a brood parasite. Biol. Conserv. 140, 259–272.
Knapp, R.A., Matthews, K.R., Preisler, H.K., Jellison, R., 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. Ecol. Appl. 13, 1069–1082.
Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Divers. Distrib. 13, 66–69.
Le Lay, G., Clergeau, P., Hubert-Moy, L., 2001. Computerized map of risk to manage wildlife species in urban areas. Environ. Manage. 27, 451–461.
Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101–113.
Liu, C.R., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.
Luoto, M., Kuussaari, M., Toivonen, T., 2002. Modelling butterfly distribution based on remote sensing data. J. Biogeorgr. 29, 1029–1037.
Manel, S., Dias, J.M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120, 337–347.
McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman & Hall, London, 511 pp.
McMillen, D.P., 2003. Spatial autocorrelation or model misspecification? Int. Regional Sci. Rev. 26, 208–217.
McPherson, J.M., Jetz, W., 2007. Effect of species' ecology on the accuracy of occurrence models. Ecogaphy 30, 135–151.
McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.
Meynard, C.N., Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J. Biogeogr. 34, 1455–1469.
Miller, J., Franklin, J., Aspinall, R., 2007. Incorporating spatial dependence in predictive vegetation models. Ecol. Model. 202, 225–242.
Minchin, P.R., 1987. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio 71, 145–156.
Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.
NCAR Research Application Program, 2008. Verification: Forecast Verification Utilities. R package version 1.26.
Nix, H.A., 1986. A Biogeographic Analysis of Australian Elapid Snakes. Australian Government Publications Service, Canberra.
Oksanen, J., Minchin, P.R., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecol. Model. 157, 119–129.
Olden, J.D., Jackson, D.A., 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biol. 47, 1976–1995.
Osborne, P.E., Alonso, J.C., Bryant, R.G., 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. J. Appl. Ecol. 38, 458–471.
Palma, L., Beja, P., Rodrigues, M., 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. J. Appl. Ecol. 36, 812–824.

Pierce, D.A., Schafer, D.W., 1986. Residuals in generalized linear models. J. Am. Stat. Assoc. 81, 977–986.
Piorecky, M.D., Prescott, D.R.C., 2006. Multiple spatial scale logistic and autologistic habitat selection models for northern pygmy owls, along the eastern slopes of Alberta's Rocky Mountains. Biol. Conserv. 129, 360–371.
R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Robertson, M.P., Peter, C.I., Villet, M.H., Ripley, B.S., 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecol. Model. 164, 153–167.
Sanderson, R.A., Eyre, M.D., Rushton, S.P., 2005. Distribution of selected macroinvertebrates in a mosaic of temporary and permanent freshwater ponds as explained by autologistic models. Ecography 28, 355–362.
Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Cervený, J., Koubek, P., Huber, T., Stanisa, C., Trepl, L., 2002. Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. J. Appl. Ecol. 39, 189–203.
Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.
Segurado, P., Araujo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43, 433–444.
Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Model. 148, 1–13.
Swan, J.M.A., 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology 51, 89–102.
Syartinilia, Tsuyuki, S., 2008. GIS-based modeling of Javan Hawk-Eagle distribution using logistic and autologistic regression models. Biol. Conserv. 141, 756–769.
Therneau, T.M., Atkinson, B., 2008. rpart: Recursive Partitioning. R package version 3.1-4.1.
Thuiller, W., 2003. BIOMOD—optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biol. 9, 1353–1362.
Wintle, B.A., Bardos, D.C., 2006. Modeling species–habitat relationships with spatially autocorrelated observation data. Ecol. Appl. 16, 1945–1958.
Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773.
Yáñez, M., Floater, G., 2000. Spatial distribution and habitat preference of the endangered tarantula *Brachypelma klaasi* (Araneae: Theraphosidae) in Mexico. Biodivers. Conserv. 9, 795.