Research

Ecography 37: 637–647, 2014
doi: 10.1111/j.1600-0587.2013.00564.x

Subject Editor: Carsten Dormann. Accepted 15 November 2013

Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records

Guilherme de Oliveira, Thiago Fernando Rangel, Matheus Souza Lima-Ribeiro, Levi Carina Terribile and José Alexandre Felizola Diniz-Filho

G. de Oliveira (guilhermeoliveirabio@yahoo.com.br), Centro de Ciências Agrárias, Ambientais e Biológicas (CCAAB), Univ. Federal do Recôncava da Bahia (UFRB), 44,380-000, Campus Cruz das Almas, Rua Rui Barbosa, 710, Centro, Cruz das Almas, BA, Brasil. – M. S. Lima-Ribeiro and L. C. Terribile, Laboratório de Macroecologia, Univ. Federal de Goiás (UFG), Campus Jataí, 75801-615, Jataí, GO, Brasil. – T. F. Rangel and J. A. F. Diniz-Filho, Laboratório de Ecologia Teórica e Síntese, Depto de Ecologia, ICB, Univ. Federal de Goiás (UFG), Cx.P. 131, 74001-970, Goiânia, GO, Brasil.

Most species data display spatial autocorrelation that can affect ecological niche models (ENMs) accuracy-statistics, affecting its ability to infer geographic distributions. Here we evaluate whether the spatial autocorrelation underlying species data affects accuracy-statistics and map the uncertainties due to spatial autocorrelation effects on species range predictions under past and future climate models. As an example, ENMs were fitted to *Qualea grandiflora* (Vochysiaceae), a widely distributed plant from Brazilian Cerrado. We corrected for spatial autocorrelation in ENMs by selecting sampling sites equidistant in geographical (GEO) and environmental (ENV) spaces. Distributions were modelled using 13 ENMs evaluated by two accuracy-statistics (TSS and AUC), which were compared with uncorrected ENMs. Null models and the similarity statistics *I* were used to evaluate the effects of spatial autocorrelation. Moreover, we applied a hierarchical ANOVA to partition and map the uncertainties from the time (across last glacial maximum, pre-insustrial, and 2080 time periods) and methodological components (ENMs and autocorrelation corrections). The GEO and ENV models had the highest accuracy-statistics values, although only the ENV model had values higher than expected by chance alone for most of the 13 ENMs. Uncertainties from time component were higher in the core region of the Brazilian Cerrado where *Q. grandiflora* occurs, whereas methodological components presented higher uncertainties in the extreme northern and southern regions of South America (i.e. outside of Brazilian Cerrado). Our findings show that accounting for autocorrelation in environmental space is more efficient than doing so in geographical space. Methodological uncertainties were concentrated in outside the core region of *Q. grandiflora*'s habitat. Conversely, uncertainty due to time component in the Brazilian Cerrado reveals that ENMs were able to capture climate change effects on *Q. grandiflora* distributions.

Ecological niche models (ENMs; see Araújo and Peterson 2012, Peterson and Soberón 2012, and Warren 2012 for recent discussions of concepts and terminology) are being increasingly used in several areas of the biological sciences (Zimmermann et al. 2010), including conservation biology (Diniz-Filho et al. 2009a, 2010, Araújo et al. 2011, de Oliveira et al. 2012), biogeography and phylogeography (Richards et al. 2007, Werneck et al. 2011, Collevatti et al. 2013), and evolutionary biology (Peterson et al. 1999, Anderson et al. 2002, Graham et al. 2004). These models can be employed to extrapolate observed patterns and predict species' geographical distributions using climatic (e.g. temperature, precipitation, temperature seasonality) and environmental (e.g. soil pH, relief, altitude) spatial layers under a GIS-based approach (Anderson et al. 2002, Colwell and Rangel 2009).

However, sampling the species occurrence records throughout its distribution usually contain biases (e.g. species' sampled near rivers, roads, conservation units, and cities) and thus these records are not evenly distributed in geographical or environmental spaces, usually containing many gaps (the Wallacean shortfall; Reddy and Dávalos 2003, Bini et al. 2006, Hortal et al. 2007, Phillips et al. 2009). At the same time, environmental variables are often geographically structured (autocorrelated) such that closer regions have more similar climates than distant ones (Legendre 1993). Thus, in most cases, both species occurrences and environmental data are spatially autocorrelated, which is a well-known cause of bias when evaluating statistical performance of modelling methods (Legendre 1993, Diniz-Filho et al. 2003, 2008), including ENM models (Segurado et al. 2006, Dormann 2007,

Dormann et al. 2007, De Marco et al. 2008). Further, the geographical and environmental ranges from where pseudo-absences are sampled, their respective spatial distances, and proportion (prevalence) in relation to the species presences, also affect ENM predictions (Thuiller et al. 2004, Chefaoui and Lobo 2008, Jiménez-Valverde et al. 2009, VanDerWal et al. 2009, Lobo et al. 2010). Thus, sampling procedures used to assemble data on species distribution used in ENM should minimize autocorrelation both in geographical and environmental space, therefore generating the smallest possible prediction bias on ENMs.

Presence–absence methods using pseudo-absence records are becoming increasingly popular when using ENMs, either because of the lack of species' absence data in available online datasets and/or recording species' absences on the field is not usually done (Stokland et al. 2011). The most widespread methodology for selecting pseudo-absence records is to randomly sample locations from regions where no occurrence records exist for a focal species (background sampling; Stockwell and Peters 1999). A derived strategy consists in sampling pseudo-absences as a function of geographical and environmental factors (Hirzel et al. 2001, Zaniewski et al. 2002). At the same time, the predictive power of ENMs is usually evaluated by cross-validation methods where species presences and/or absences (or pseudo-absences) are split into training (i.e. used to fit a model) and testing data (i.e. data against which the model is evaluated) (Fielding and Bell 1997).

In the procedures described above, training and testing datasets are also frequently assumed to be geographically independent, and the data can therefore be randomly split. However, spatial autocorrelation in the full dataset violates the independence assumption, and the ENMs' measures of accuracy, consequently, tend to be inflated (Araújo and Guisan 2006, Veloz 2009). In addition, aggregation of species presence records could have different statistical properties than the pseudo-absence records generated, because presences tend to be clustered in geographical space (as previously discussed), whereas pseudo-absences are usually randomly sampled from the entire background (Hijmans 2012). Conversely, if sampling biases in both presence and pseudo-absence records were corrected in the same way, the accuracy measures would become more reliable (Phillips et al. 2009).

In this study we used two strategies to overcome the problems caused by autocorrelation in spatial and environmental space due to aggregation in presence and pseudo-absences records. To illustrate our approach, we used as model organism *Q. grandiflora*, a typical woody tree species from the Brazilian Cerrado (Ratter et al. 2003). This species is a fine example for dealing with spatial autocorrelation problems because it occurs over a broad spatial scale and in different habitats, ranging from open vegetation to more forested areas (Costa and Araújo 2001). Moreover, the presence records currently available for this species have important gaps across its wide geographic distribution. In short, we sampled the range of *Q. grandiflora* looking for a fixed number of equidistant points in geographical and in environmental spaces, based on an autocorrelation analysis of environmental suitability in the occurrence points. We then investigated the influence of the sampling protocol on ENM predictions by testing the accuracy of 13 ENMs (six presence-only and seven presence–absence methods). Thus, the effects of autocorrelation in ENMs, generated both by spatial aggregation of sampling points and spatial patterns in environmental variables, were evaluated by testing the following hypotheses: 1) the spatial autocorrelation in species data inflates the ENMs' accuracy-statistics; and 2) accounting for spatial autocorrelation in both presence and pseudo-absence data using equidistant points yields ENMs with higher accuracy-statistics.

Furthermore, we also analysed how the uncertainties from modelling components, such as different ENM methods and strategies to correct for spatial autocorrelation, affect the ENM predictions through time and highlighted their implications to biogeography and conservation studies. The effects of climate changes on species geographical ranges are often analysed from ENMs and their consequent ecological implications though time have been recently used to address many relevant issues (see synthesis in Peterson et al. 2011). Hopefully, the partitioning and mapping of modelling uncertainties can give further insights about the approaches mentioned above and allow a better evaluation of which regions across the species geographic range have less uncertainty regarding different methodological components or are climatically more affected due to the temporal dynamics of climate change.

Material and methods

Data collection: presence and pseudo-absences records and climate layers

We retrieved 238 presence records for *Q. grandiflora* from the following online museum collections: 'Florescer' (Flora Integrada da Região Centro-Oeste, <www.florescer.unb.br/>), 'JABOT – Banco de Dados da Flora Brasileira' (JBRJ 2012), and Species Link (<http://splink.cria.org.br/>). All records were examined for species-name's synonymies and nomenclature errors (Giovanni et al. 2012). We sampled absence points from background region (i.e. localities where species were not recorded), following the methodology described in the next section. We mapped species presences and pseudo-absences in 6818 grid cells of $0.5° \times 0.5°$ of latitude and longitude covering the entire Neotropical region (Fig. 1a).

We obtained the climate layers and characterised the environmental space for ENMs using the climatic simulations for the pre-industrial time period, simulated for the middle of the eighteenth century, and stabilized across a 200 years' time period, derived from the CCSM4 (<http://cmip-pcmdi.llnl.gov/cmip5>) coupled atmosphere-ocean general circulation model (AOGCM). We downloaded monthly simulations for four climate variables from AOGCM outputs (i.e. precipitation and mean, maximum, and minimum temperatures), which were downscaled to the same grid and used to compute 19 bioclimatic variables using the same methodology as the WorldClim database described by Hijmans et al. (2005, <www.worldclim.org/bioclim>).

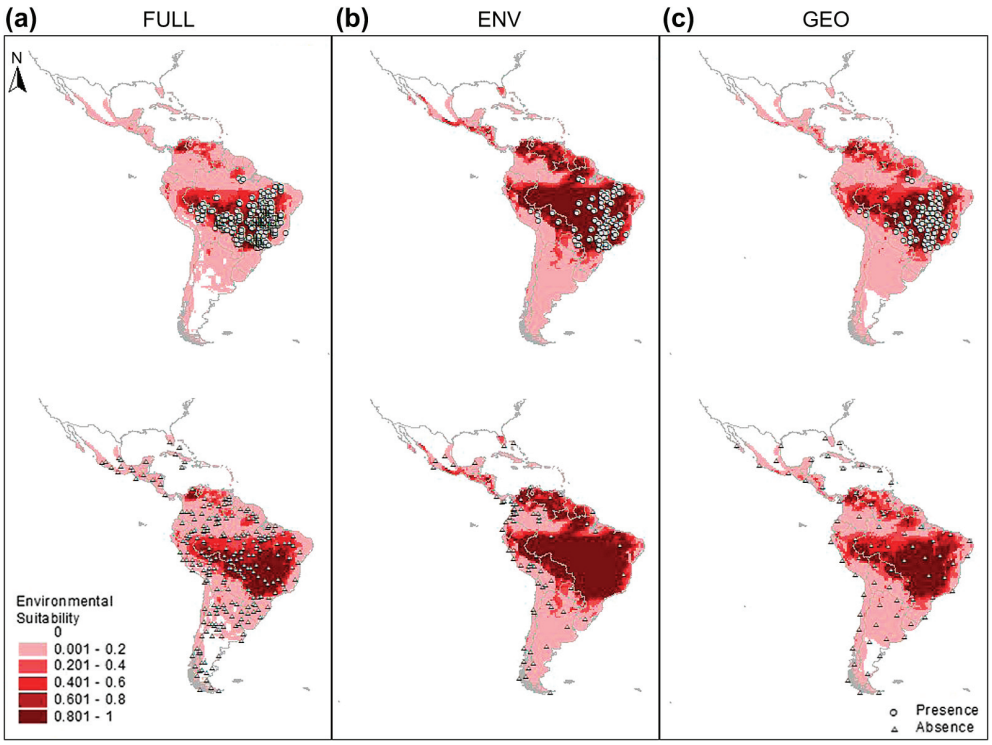

Figure 1. Maps of predicted distributions of *Qualea grandiflora* across Neotropical region using 13 ENMs. They were separated according to *Q. grandiflora*'s presences/pseudo-absences types into (a) using all 238 cells retrieved as presences and selecting 238 random cells as absences across the entire Neotropics (FULL), and using 72 environmentally (b; ENV) and geographically (c; GEO) equidistant cells from those 238 occurrence records, as presences, and selecting other 72 environmentally and geographically equidistant cells, respectively, across entire Neotropics, as pseudo-absences. All maps have the same color scale and show increasing environmental suitability from 0 to 1.

We selected five bioclimatic variables (i.e. annual mean temperature, annual temperature range, precipitation of the wettest month, precipitation of the driest month, and precipitation of the warmest quarter), using a factor analysis based on the correlation matrix to minimise collinearity problems when building the ENMs (i.e. selecting the variables with the highest loadings in the first five Varimax rotated eigenvectors; Terribile et al. 2012). Along with these variables, we also included subsoil pH (30–100 cm; from the Harmonized World Soil Database – ver. 1.1, FAO/IIASA/ISRIC/ISS-CAS/JRC 2009) as a constraint variable to improve the ENM predictions. This inclusion was mainly because *Q. grandiflora* is a tree species, and in the Brazilian Cerrado pH is known to improve the predictive performance of ENMs (Collevatti et al. 2012).

Dealing with spatial autocorrelation

We used two strategies to correct for spatial autocorrelation in the presence and pseudo-absence data for *Q. grandiflora* using sampling data based on 1) geographical equidistant points (GEO) and 2) environmental equidistant points (ENV). First, we calculated the Mahalanobis distances (D^2) in environmental space (E-space) (based on the six variables defined above) between each of the 238 occurrence records and their centroid and mapped these distances in geographical space, as a way to synthesize spatial variation across multiple dimensions of E-space. Because spatial autocorrelation can be viewed as redundant information related to the similarity among spatial distribution of points (or values of a variables measured in these points), the overall idea is to obtain a geographically effective sample size that takes autocorrelation in data into account. Sampling procedures can be used to select, among the available data, a corresponding number of equidistant points, increasing statistical independence of sampling units. Although there may be loss of statistical power using a heuristic procedure that sub-sample the dataset, such strategy is still useful to understand how autocorrelation affects the model performance (Hawkins et al. 2007).

We started by fitting a simultaneous autoregressive (SAR) model of the form

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}$$

to model the D^2 distances to species' environmental centroid (i.e. the vector **y**) (SAR model was done in SAM software, Rangel et al. 2006, 2010). In this SAR model, **W** is a matrix of spatial weights given by the inverse of squared geographical distances among occurrences and ρ is the autoregressive coefficient, which measures the amount of spatial autocorrelation in data. This coefficient ρ is then used to calculate the effective degrees of freedom, v*, by

applying the empirical formulae provided by Griffith (2003) and given by

$$v^* = n\,[1 - (1/(1 - \exp(-1.92349))(n - 1/n)\,(1 - \exp(-2.12373\rho + 0.20024(\rho\hat{}1/2))))] - 2$$

Based on the estimated autoregressive coefficient ρ, we found that our data effectively contain 72 degrees of freedom. Notice that Mahalanobis distances were calculated in relation to the centroid of the species in E-space, so that they do not indicate direction. However, regardless of direction, short distances indicate statistical redundancy among sampling sites, which is important to reduce the degrees of freedom and reduce the effect of spatial autocorrelation. Because distances in **W** matrix are squared, short distance relationships will have a much higher weight in the calculation of autoregressive coefficient ρ than large distance relationships. Although it is possible to use other strategies to explore this reduction (e.g. average ρ for environmental variables, marginality axes of ENFA), the advantage of using the Mahalanobis distance is that environmental space is not evaluated independently of the species' occurrences.

We then calculated the pairwise geographical distances (based on geographical coordinates, forming the G-space) and environmental distances (based on the six environmental variables, forming the E-space), among the total 238 occurrence points. Next, we independently select 72 (v^*) points that are most equidistant in G and E-spaces. The algorithm that searches the combination of most equidistant points starts by selecting the most distant pair of points, given a definition of space (G or E). The algorithm proceeds by iteratively adding the point that is most distant to all previous selected points. The iteration ends when 72 (v^*) points have been selected. Notice that the same algorithm was independently used to select equidistant points in both G and E-space (Fig. 1b, c) (see also Hawkins et al. 2007). We then repeated the same protocol using the background cells (i.e. those cells for which no records of *Q. grandiflora* were found) across the entire Neotropical region. Thus the algorithm selected the species' 72 environmentally (Fig. 1b) and geographically equidistant (Fig. 1c) pseudo-absence records, thereby maintaining a proportional prevalence of 0.5 in the presence/pseudo-absence data. These independent presence/pseudo-absence data were then used to build the ENMs (see below).

Finally, to compare methodologies (see below) we established a general model (FULL model) with all 238 occurrence records and randomly selected 238 pseudo-absence records (Fig. 1a) across the Neotropical background area, thereby also maintaining the same species prevalence of 0.5 in sub-sampled datasets. Moran's I correlograms plots (Fig. 2) show the spatial autocorrelation in the models' residuals (FULL, GEO, ENV, and one with 72 randomly chosen presences records) reveals that our sampling strategy was effective in removing spatial autocorrelation in suitability maps projected by ENMs.

Ecological niche models

Ensemble methodologies for defining the species' ecological niche models (Araújo and New 2007) were implemented following Diniz-Filho et al. (2009b, 2010) and Terribile et al. (2012). We randomly divided presence/pseudo-absence data into 75% for calibration and 25% for evaluation and repeated this process 50 times. The environmental suitability of *Q. grandiflora* was determined by the proportion of all 50 cross-validated presences, projected as presences in each cell of the Neotropical region for each ENM based on thresholds established by the receiver operating characteristic (ROC) curve (sensu Terribile et al. 2012).

Thirteen different ENMs were used. These included six presence-only or presence-background methods (i.e. BIOCLIM, Euclidian, Gower, Mahalanobis distances, genetic algorithm for rule set production – GARP, and maximum entropy – MAXENT) and seven presence–absence methods (i.e. generalized linear modeling – GLM, Random forest, generalized additive modeling – gam, factorial discriminant analysis – FDA, multiple adaptive regression splines –MARS, environmental niche factor analysis – ENFA, and neural network). Franklin (2009) and Peterson et al. (2011) present general descriptions of methods. Notice that, for accuracy-statistics comparisons reasoning, in both types of ENMs (presence-only and presence–absence) we used pseudo-absences data, in presence-only ENMs pseudo-absences records were used as background. These methods were run with the BioEnsembles software computational platform (Diniz-Filho et al. 2009b, Terribile et al. 2012, Collevatti et al. 2013).

Data analysis

Comparison of accuracy-statistics

For each ENM and each type of sampling procedure applied to the presence/pseudo-absence records (FULL, ENV, or GEO), we calculated two types of accuracy-models statistics: true skill statistics (TSS, Allouche et al. 2006), using a threshold of 0.5 level, and the area under the receiver operating characteristic (ROC) curve, known as the AUC (Fielding and Bell 1997). We used a factorial analysis of variance (ANOVA) to compare differences in TSS and AUC values using three factors: 1) different accuracy-statistics (TSS and AUC), 2) presence/pseudo-absence sampling strategies (FULL, ENV, and GEO), and 3) the 13 ENMs. This comparison indicates that correcting for spatial autocorrelation can affect the accuracy of the ENMs and that accuracy-statistics and ENMs can interact and their accuracy values can depend of the interaction between these two factors.

Moreover, we evaluated whether ENMs accuracy-statistics values for ENV and GEO presence/pseudo-absence records can be reached by chance alone using a null model. We selected 72 random presences from the full 238 occurrence records and 72 random pseudo-absences from the whole Neotropical background and modelled random presence/pseudo-absence records following the protocol described above. We repeated this procedure 200 times. Finally, by comparing the values of TSS and AUC statistics from ENV and GEO presence/pseudo-absence records with those from 200 random models, we established the probability that equal or greater values of TSS and AUC statistics could be reached by

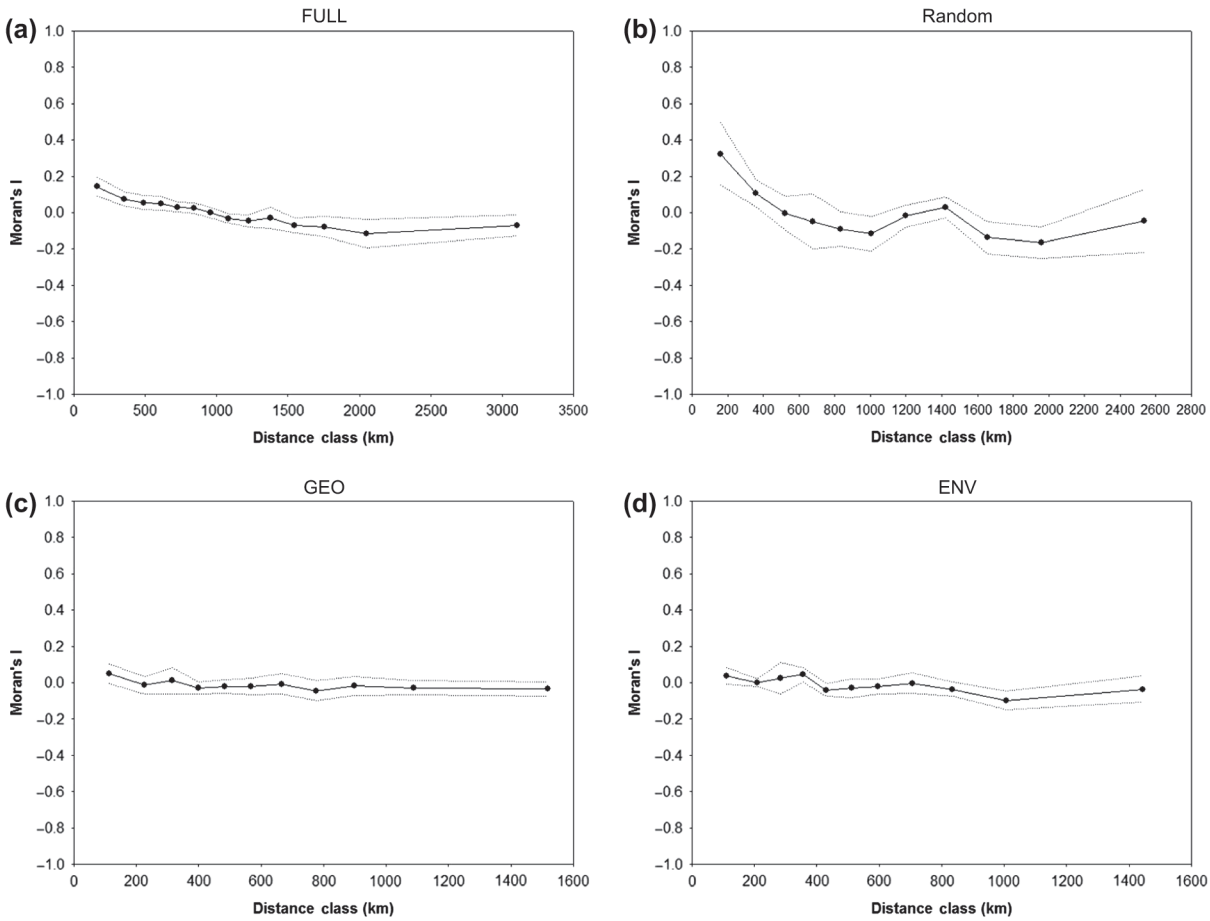

Figure 2. Moran's I spatial correlograms of the (a; FULL), 72 random points (b; Random), (c; GEO), and (d; ENV) models' residuals. Solid line is the average from 13 ENMs results and dashed lines are the standard deviation from this mean value.

chance independently of the spatial autocorrelation in the occurrence and environmental data.

Evaluating the influence of spatial structure on model accuracy

As suggested by Veloz (2009), spatial autocorrelation in data can falsely inflate accuracy measures of ENMs. Consequently, models that do not account for spatial autocorrelation tend to reach higher accuracy-statistic values (e.g. TSS, AUC) than those that consider spatial autocorrelation but with lower similarity (as an *I* statistic; Veloz 2009). Therefore, as suggested by Veloz (2009), we calculated the similarity statistic *I* using ENMTOOLS (Warren et al. 2008), which is a measure of overlap, using the environmental suitability from ENMs between models with controlled spatial autocorrelation (i.e. ENV and GEO) and models with no correction of spatial autocorrelation (i.e. FULL).

Moreover, we also tested the ability of ENV- and GEO-procedures to correct for spatial autocorrelation and to avoid undesirable autocorrelation effects on ENM predictions. Following Veloz (2009) analysis protocol, we assumed that the FULL procedure would generate the most accurate model. Under this assumption, models fit with lower levels of spatial autocorrelation (ENV and GEO) should have higher similarity index than models with no control of spatial autocorrelation. To test this hypothesis, we used the same 200 models from random selections of presences and pseudo-absences described above (i.e. models with no control of spatial autocorrelation). The statistical significance was then obtained by computing how often *I*-values equal to or higher than those from the ENV and GEO models could be reached by chance only.

Partitioning and mapping sources of uncertainty in model predictions

We measured the magnitude of uncertainties in ENM projections caused by the choice of methodological strategies to control for spatial autocorrelation, as well as differences in climate projections through the time. We projected *Q. grandiflora*'s environmental suitability for the last glacial maximum (LGM, 21 000 yr ago/21 ky BP), pre-industrial, and future (2080–2100, end-of-century; hereafter, EOC) climate scenarios using the same environmental variables and AOGCM used to build the models for pre-industrial period (see Terribile et al. 2012 for details). Next, following the overall reasoning proposed by Diniz-Filho et al. (2009b), we performed an ANOVA for each of the 6818 grid cells using environmental suitability as the response variable, while nesting spatial autocorrelation component (FULL, ENV, and GEO) within ENMs component (13 ENMs), which were also nested into time component (pre-industrial, LGM, and EOC) (Terribile et al. 2012). This hierarchical

ANOVA disentangles the variance due to species dynamics across changing environments through the pre-industrial, LGM, and EOC times (the time component, which contain ecological meanings; e.g. the climate change effects on species distribution) from ENM methods and spatial autocorrelation components (which act as sources of methodological uncertainties).

Results

The consensus maps (i.e. the environmental suitability resulted from the average between the 13 ENM's environmental suitability, and weighted by their TSS values) showed slight differences between geographical ranges predicted by the FULL, ENV, and GEO models (Fig. 1). Nonetheless, the FULL-model yielded the narrowest potential distribution (Fig. 1a), whereas the ENV-model predicted the widest geographical range (Fig. 1b). All three maps predicted a wide distribution of *Q. grandiflora* in the core region of the Brazilian Cerrado and diverge in their predictions across the Suriname, Guyana, Venezuela, and northern Colombia (Fig. 1) (see Supplementary material Appendix 1 for the functional relationship between FULL, GEO, and ENV-models' points and the environmental variables).

The TSS-statistics values were generally higher for the ENV-model (Fig. 3a) (except for the GARP, ENFA, and Neural Networks methods, Supplementary material Appendix 2). The GEO- and FULL-models presented similar TSS-statistics, although some variability among the ENM methods exists. This pattern remained for the AUC-statistics, but it showed slight differences for Random Forest and ENFA estimates (Supplementary material Appendix 2). In accordance with these results, eight from 13 TSS and AUC statistics-values for the ENV-model were significantly higher than expected by chance; this result did not occur with the GEO-model (Table 1). The factorial ANOVA (Table 2) showed a significant interaction among the factors and indicated that the inflation of the model's accuracy depended on the combination of ENM methods, the strategy to correct autocorrelation (FULL, GEO or ENV), and the accuracy-statistics (AUC or TSS) used.

The prediction from the GEO-model showed a significantly higher niche overlap with the FULL-model than with the ENV-model measured by *I*-statistic (Fig. 3c) except for the Bioclim algorithm (Supplementary material Appendix 2).

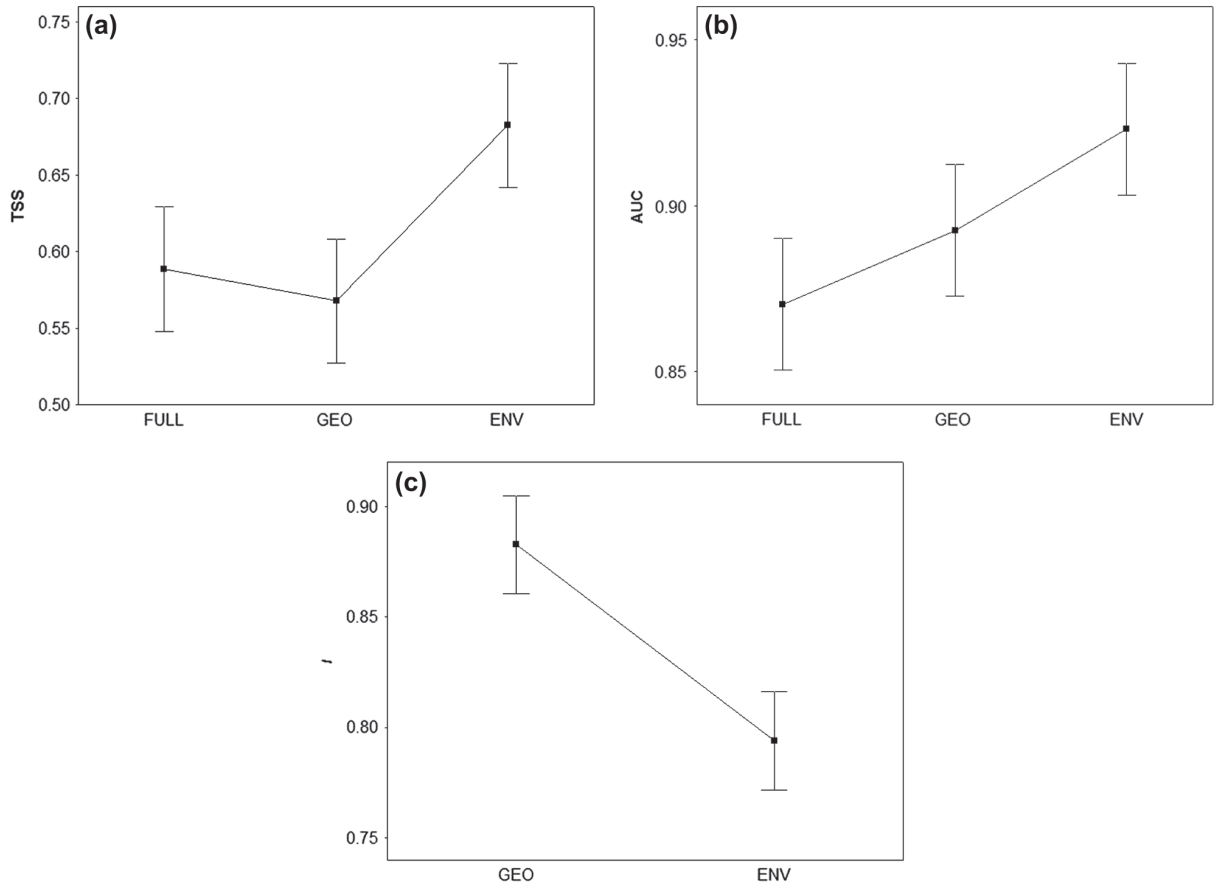

Figure 3. Mean values for (a) true skill statistics (TSS), (b) area under the receiver operating characteristic curve (AUC), and (c) similarity statistics (*I*), using 13 ENMs, and separated by three types of species data: first, using all 238 cells retrieved as presences and selecting 238 random cells as pseudo-absences across the entire Neotropics (FULL), and using 72 environmentally (b; ENV) and geographically (c; GEO) equidistant cells from those 238 occurrence records, as presences, and selecting other 72 environmentally and geographically equidistant cells, respectively, across entire Neotropics, as pseudo-absences. Bars indicate confidence interval at 0.95 level.

Table 1. Probability values comparing the values of true skill statistics (TSS), area under the curve of a receiver operating characteristic (AUC), and similarity statistics (*I*) of each ecological niche models (ENMs) between species presence and pseudo-absence based on geographic and environmental equidistance with 200 random selection of species presence and pseudo-absence. Values under 0.05 probability-threshold value are italic-bolded.

	Geographic equidistance – GEO			Environmental equidistance – ENV		
ENM	TSS	AUC	*I*	TSS	AUC	*I*
Bioclim	0.61	0.15	0.36	0.08	***0.005***	0.29
Euclidean distance	0.35	0.25	0.98	***0.005***	***0.005***	1
Gower distance	0.57	0.195	0.77	***0.005***	***0.005***	1
Mahalanobis distance	0.08	0.18	0.915	***0.005***	***0.005***	1
GLM	0.35	0.295	0.35	***0.005***	***0.005***	1
Maxent	0.11	0.495	0.57	***0.005***	***0.005***	1
Random forest	0.56	1	0.82	***0.005***	1	0.995
GARP	***0.015***	0.38	0.17	0.52	0.81	0.99
GAM	0.135	0.37	0.505	***0.005***	***0.005***	1
FDA	0.395	0.435	0.975	***0.005***	0.125	1
MARS	0.18	0.43	0.48	***0.005***	0.14	1
ENFA	0.205	***0.035***	0.91	0.695	***0.03***	1
Neural networks	0.39	0.695	0.875	0.09	0.755	1

Nevertheless, these *I*-statistic values for both ENV- and GEO-models can be expected by chance alone when we compare it with our null models with no correction for spatial autocorrelation on species presence and pseudo-absence records (Table 1).

The hierarchical ANOVA (Table 3) showed the time component with the highest median and amplitude proportion of the total sum of squares, the second highest values was for the spatial autocorrelation component, and the lowest values was for ENMs component. Divergences in predictions through time were concentrated in central Brazil, the core region of the Brazilian Cerrado in which the species potential distribution was also mapped (Fig. 4a). Conversely, predictive conflicts from methodological components occurred in regions where the focal species is not expected to occur (i.e. outside the predicted potential distribution), in southern Colombia and Ecuador (ENMs, Fig. 4b), as well as in southern South America in Argentina, Chile, Bolivia, Paraguay, and Uruguay (spatial autocorrelation; Fig. 4c).

Discussion

Our application can furnish interesting insights on how spatial autocorrelation in environmental data and aggregation of records can affect overall measures of ENM accuracy, and we believe they can be well understood based on our current theoretical knowledge on these methods. The most important overall finding is that taking autocorrelation into account and sampling equidistant points in environmental space (rather than in geographical space) improves model accuracy estimates. Moreover, Varela et al. (in press), using virtual species, also showed that environmental filtering reduces the effects of spatial autocorrelation underlying species records and, as consequence, improve the predictions from ENMs. Our results and those from Varela et al. (in press) results are complementary because both studies showed that the effects of spatial autocorrelation are best corrected in the environmental space.

Does spatial autocorrelation inflate ecological niche models' accuracy-statistics?

Contrary to the results of Veloz (2009), who detected the inflation of ENMs' accuracy-statistics of non-corrected species presence records, our findings show higher accuracy-statistics only for ENMs built using species data corrected for spatial autocorrelation, particularly for ENV-models, which reached values higher than expected by chance. Moreover, Veloz (2009) predictions could be dependent on the accuracy-statistic method (i.e. using only AUC instead of compute also TSS) and the spatial correction method (i.e. using correction in geographical space rather than environmental space), as shown by our findings in the factorial

Table 2. Degree of freedom, proportions, in percentage, of the sum of squares (SS), mean squares (MS), F- and probability-values from the factorial ANOVA, using, as response variable the accuracy values, and as factors, the 13 ecological niche models (ENMs), the two types of accuracy statistics: TSS and AUC (Sta), and the three types of spatial autocorrelation species data: FULL, GEO, and ENV (SAut).

Factors	DF	SS (%)	MS	F	p
Ecological niche models	12	20.246	0.0395	27.963	<0.00001
Statistics	1	66.496	1.5558	1101.98	<0.00001
Spatial autocorrelation	2	3.9492	0.0462	32.716	<0.00001
Sta × SAut	2	0.983	0.0115	8.156	0.00198
ENM × Sta	12	1.9276	0.0038	2.66	0.01993
ENM × SAut	24	4.9494	0.0048	3.419	0.0019
Residuals	24	1.4489	0.0014		

Table 3. Proportions, in percentage, of the median and the range (maximum and minimum) of total sum of squares (SS) from the nested ANOVA calculated for each 6818 cell grid covering the Neotropical region. The hierarchy of nesting is shown in brackets.

	SS (%)	
Source	Median	Min–Max
Time	32.27	0.00–95.09
Ecological niche models [time]	25.79	2.04–82.82
Spatial autocorrelation [ecological niche models]	30.34	0.19–75.15
Residuals	2.92	0.07–28.57

ANOVA (Table 2). The strong interaction between the strategy to correct for spatial autocorrelation and accuracy-statistic suggests that geographical correction decreases only TSS statistic values, but not AUC values (which was computed by Veloz 2009) (Supplementary material Appendix 3). The ENV model, on the other hand, improved both TSS and AUC accuracy-statistics in relation to GEO and FULL models, but TSS was improved in a higher proportion than AUC (Supplementary material Appendix 3). Thus, our model suggests that environmental correction increased all statistics values independent of the type of statistics used (AUC or TSS).

Furthermore, the highest values of similarity statistics (i.e. the overlap with the uncorrected model – FULL) were those from geographically corrected models (GEO) compared with those from environmentally corrected ones (ENV) (Fig. 3c). Therefore, selecting geographically equidistant sampling sites may not be the best strategy to solve problems of spatial autocorrelation in ENM predictions or to improve models accuracies. According to statistic *I*, the GEO-model retained part of the spatial autocorrelation contained in the FULL-model (Fig. 2a). Our results also showed that the similarity statistic *I*, between GEO and FULL-model, as proposed by Veloz (2009), did not indicate inflation from spatially structured models on accuracy-statistics. In fact, this similarity of ENMs predictions among these models indicates the degree of retained autocorrelation, which, in turn, was reflected in the values of similarity-index *I*. Our findings do not support Veloz's (2009) assumption that FULL-model is more accurate and suggest that his strategy to verify the effects of spatial autocorrelation on the accuracy-statistics values may need to be reconsidered.

How correcting for spatial autocorrelation improved ecological niche models' accuracy-statistics

Potential distributions were remarkable different between models that accounted for spatial autocorrelation (GEO- and ENV-models) and the model that did not account (FULL-model). Also, our findings show that the ENV-model had higher values of accuracy-statistics when compared with both GEO- and FULL-models. Models that used presence/pseudo-absence records with minimum spatial structure in E-space were the most accurate in predicting species' ecological niche and, as a consequence, its potential distribution through time. Increased accuracy is a consequence of maximizing environmental distances between sampling sites, which decreases the model's over-fit in environmental space and generates the most widespread projected potential distribution. Even ENMs that are recognised to underestimate the species potential distribution models, restricting ENMs predictions, such as Maxent (see comparison between GARP and Maxent in Peterson et al. 2007), presented higher values of accuracy-statistics and more widespread predictions for species' geographical range (Supplementary material Appendix 2).

Conversely, the GEO-model had highlighted effects on ENMs' accuracy-statistics when compared to the FULL-model, improving them, but it was dependent on the ENM and statistics used. Moreover, its improvement is far less than those from ENV-models. Our findings show that GEO, although an intuitive strategy to correct for spatial autocorrelation in species data (i.e. selecting occurrence records that are geographically equidistant; Phillips et al. 2009, Veloz 2009, Hijmans 2012), is not necessarily the best strategy to improve the accuracy of ENM predictions. The geographically equidistant records may retain the same spatial-structure pattern as the FULL-model even with a lower power due to the reduction of presence and pseudo-absence records. Although not addressed here, our findings still suggest that adding spatial terms (i.e. spatial eigenvectors, autoregressive, or contagious terms) as predictors in the ENMs (De Marco et al. 2008, Václavík

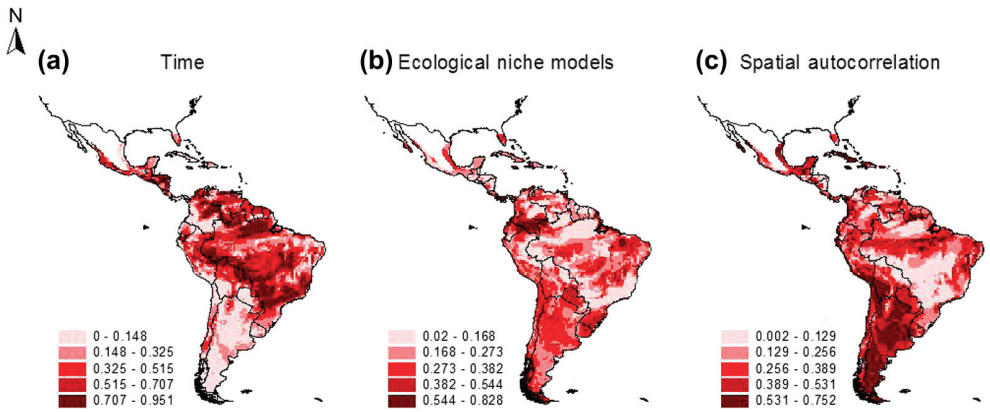

Figure 4. Maps of the proportion of the total of sum of squares assigned to (a) time components, (b) ecological niche models components, and (c) spatial-autocorrelation components.

et al. 2012), may not entirely solve the problems caused by spatial autocorrelation, because accounting for autocorrelation in the geographic space alone does not deal with distribution of records in environmental space, which is the actual source of information used in ENMs. Thus, we propose that correction for spatial autocorrelation on species records may be more effective if independent records for analyses are selected directly from environmental (rather than geographical) space.

Are methodological uncertainties a real problem?

The ENMs' predictive ability can vary widely due to many factors (Araújo and New 2007), such as the differences inherent to algorithms' performance, variation in the outputs of AOGCMs and spatial autocorrelation in data as addressed here. These factors generate uncertainties that can mask the evidence of climate effects on species distributions. For instance, models that seek to hind or forecast climatic effects in geographical ranges may be unreliable if methodological uncertainties (i.e. variances from ENMs and spatial autocorrelation components) are greater than uncertainties due to climate changes projections. Thus, disentangling the nested effects of each uncertainty component is an essential step of the modelling process. By using a hierarchical ANOVA, we were able to estimate the magnitude of climate effects on species range expressed in the variance of the time component, apart from the methodological uncertainties due to the ENMs and the strategies to correct spatial autocorrelation.

Thus, the highest variance from the time component (Table 3) suggests that the predicted distributions for *Q. grandiflora* can be reliably used to address biogeographical questions or to investigate the effects of climate change on its geographical range through the time. However, the methodological components (i.e. ENMs' algorithms and strategies for correcting spatial autocorrelation) also reached a high proportion of the sum of squares from ANOVA. Although this result indicates that the methodological components (mainly spatial autocorrelation) introduced an important source of uncertainty in the ENM predictions, uncertainty is not evenly distributed across the geographical space. That is, the regions that presented higher methodological uncertainties are located outside the core area of *Q. grandiflora* distribution (i.e. Brazilian Cerrado). Thus, methodological uncertainties are not as important as expressed in median proportional values using the entire Neotropical grid cells, and are consequence of differences in model's predictive power outside the core area of species distribution. Conversely, the highest variances from the time components match the Brazilian Cerrado limits, thus reinforcing the property of models to correctly predict the climate change effects on *Q. grandiflora* regardless of methodological uncertainties. Nevertheless, several other sources of methodological uncertainties (Araújo and New 2007, Diniz-Filho et al. 2009b, Buisson et al. 2010), including different AOGCMs, different greenhouse gas emission scenarios, and different environmental variables used to build the ENMs were not included in our analyses.

Concluding remarks

Using *Q. grandiflora* as case study, our findings suggest that spatial autocorrelation underlying species and environmental data did not falsely inflate the ENMs' measures of accuracy. In fact, this spatial autocorrelation actually can decreased the accuracy-statistics values (TSS and AUC) in comparisons with models in which species data were corrected for spatial autocorrelation in both the presence and pseudo-absence data. In addition, autocorrelation was best accounted by using environmentally equidistant sampling sites, as opposed to geographical equidistant sites. Accounting for autocorrelation in environmental space can prevent model over-fit and provided higher accuracy (as measured by TSS and AUC).

Finally, our example also showed that partitioning and mapping different sources of uncertainty in ENMs revealed that, despite methodological components have influenced ENM predictions, the pure effects of different ENMs and autocorrelation corrections are located in regions outside the core habitat region. Thus, for *Q. grandiflora*, tests for biogeographical hypotheses (often for past climates) and evaluations of conservation status in global changing scenarios (often for future climates) would not be strongly affected by the methodological differences in ENMs. Thus, we recommend that spatial autocorrelation in environmental space should always be accounted for, while the effects of different sources of uncertainties may be assertively interpreted through a hierarchical ANOVA, providing better understanding of how the climate change affects species distribution through the time, and where these effects are geographically spatialized.

Acknowledgements – We thank Carsten Dormann for comments and suggestion. Our research program has been continuously supported by grants to the research network GENPAC (Geographical Genetics and Regional Planning for natural resources in Brazilian Cerrado), provided by CNPq/MCT/CAPES (projects no. 564717/2010-0, 564718/2010, 563727/2010-1 and 563624/2010-8). We also thank the World Climate Research Programmer's Working Group on Coupled Modeling for providing CMIP5, and the climate-modeling group from NCAR for producing and making available CCSM.

References

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Anderson, R. et al. 2002. Using niche–based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. – Oikos 98: 3–16.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Araújo, M. B. and Peterson, T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

Araújo, M. B. et al. 2011. Climate change threatens European conservation areas. – Ecol. Lett. 14: 484–492.

Bini, L. M. et al. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. – Divers. Distrib. 12: 475–482.
Buisson, L. et al. 2010. Uncertainty in ensemble forecasting of species distribution. – Global Change Biol. 16: 1145–1157.
Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecol. Model. 210: 478–486.
Collevatti, R. G. et al. 2012. A coupled phylogeographical and species distribution modelling approach recovers the demographical history of a Neotropical seasonally dry forest tree species. – Mol. Ecol. 21: 5843–5863.
Collevatti, R. G. et al. 2013. Drawbacks in palaeodistribution modelling: the case of South American seasonally dry forests. – J. Biogeogr. 40: 345–358.
Colwell, R. K. and Rangel, T. F. 2009. Hutchinson's duality: the once and future niche. – Proc. Natl Acad. Sci. USA 106: 19651–19658.
Costa, A. A. and Araújo, G. M. 2001. Comparação da vegetação arbórea de cerradão e de cerrado na Reserva do Panga, Uberlândia, Minas Gerais. – Acta Bot. Brasilica 15: 63–72.
De Marco, P. et al. 2008. Spatial analysis improves species distribution modelling during range expansion. – Biol. Lett. 4: 577–580.
de Oliveira, G. et al. 2012. Conserving the Brazilian semiarid (Caatinga) biome under climate change. – Biodivers. Conserv. 21: 2913–2926.
Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. – Global Ecol. Biogeogr. 12: 53–64.
Diniz-Filho, J. A. F. et al. 2008. Model selection and information theory in geographical ecology. – Global Ecol. Biogeogr. 17: 479–488.
Diniz-Filho, J. A. F. et al. 2009a. Conservation biogeography and climate change in the Brazilian Cerrado. – Natureza Conservação 7: 100–112.
Diniz-Filho, J. A. F. et al. 2009b. Partitioning and mapping uncertainties in ensemble of forecasts of species turnover under climate change. – Ecography 32: 897–906.
Diniz-Filho, J. A. F. et al. 2010. Defying the course of ignorance: perspectives in insect macroecology and conservation biology. – Insect Conserv. Divers. 3: 172–179.
Dormann, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. – Global Ecol. Biogeogr. 16: 129–138.
Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.
FAO/IIASA/ISRIC/ISS–CAS/JRC 2009. Harmonized World Soil Database (version 1.1). – FAO, IIASA-Rome, Laxenburg.
Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
Franklin, J. 2009. Mapping species distribution: spatial inference and prediction. – Cambridge Univ. Press.
Giovanni, R. et al. 2012. The real task of selecting records for ecological niche models. – Natureza Conservação 10: 139–144.
Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 498–503.
Griffith, D. A. 2003. Spatial autocorrelation and spatial filtering – gaining understanding through theory and scientific visualization. – Springer.
Hawkins, B. A. et al. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. – Ecography 30: 375–384.
Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. – Ecology 93: 679–688.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.
Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. – Conserv. Biol. 21: 853–863.
JBRJ 2012. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Jabot. – Banco de dados da Floresta Brasileira, <www.jbrj.gov.br/jabot> accessed 23 July 2012.
Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Comm. Ecol. 10: 196–205.
Legendre, P. 1993. Spatial autocorrelation – trouble or new paradigm. – Ecology 74: 1659–1673.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
Peterson, A. T. and Soberón, J. 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. – Natureza Conservação 10: 102–107.
Peterson, A. T. et al. 1999. Conservatism of ecological niches in evolutionary time. – Science 285: 1265–1267.
Peterson, A. T. et al. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. – Ecography 30: 550–560.
Peterson, A. T. et al. 2011. Ecological niches and geographical distributions. – Princeton Univ. Press.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Rangel, T. F. L. V. B. et al. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. – Global Ecol. Biogeogr. 15: 321–327.
Rangel, T. F. et al. 2010. SAM: a comprehensive application for spatial analysis in macroecology. – Ecography 33: 46–50.
Ratter, J. A. et al. 2003. Analysis of the floristic composition of the brazilian cerrado vegetation III: comparison of the woody vegetation of 376 areas. – Edinburgh J. Bot. 60: 57–109.
Reddy, S. and Dávalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Richards, C. L. et al. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. – J. Biogeogr. 34: 1833–1845.
Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. – J. Appl. Ecol. 43: 433–444.
Stockwell, D. and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 2: 143–158.
Stokland, J. N. et al. 2011. Species distribution modelling-effect of design and sample size of pseudo-absence observations. – Ecol. Model. 222: 1800–1809.
Terribile, L. C. et al. 2012. Areas of climate stability in the Brazilian Cerrado: disentangling uncertainties through time. – Natureza Conservação 10: 152–159.
Thuiller, W. et al. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.

Václavík, T. et al. 2012. Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). – J. Biogeogr. 39: 42–55.
VanDerWal, J. et al. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? – Ecol. Model. 220: 589–594.
Varela, S. et al. in press. Environmental filters reduce the effects of sampling bias and improve predictions of species distribution models. – Ecography.
Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.
Warren, D. L. 2012. In defense of 'niche modeling'. – Trends Ecol. Evol. 27: 497–500.
Warren, D. L. et al. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. – Evolution 62: 2868–2883.
Werneck, F. P. et al. 2011. Revisiting the historical distribution of seasonally dry tropical forests: new insights based on palaeodistribution modelling and palynological evidence. – Global Ecol. Biogeogr. 20: 272–288.
Zaniewski, A. E. et al. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. – Ecol. Model. 157: 261–280.
Zimmerman, N. E. et al. 2010. New trends in species distribution models. – Ecography 33: 985–989.

Supplementary material (Appendix ECOG-00564 at <www.oikosoffice.lu.se/appendix>). Appendix 1–3.