Spatial analysis improves species distribution modelling during range expansion

Paulo De Marco Jr,
José Alexandre Felizola Diniz-Filho*
and Luis Mauricio Bini

Departamento de Biologia Geral, ICB, Universidade Federal de Goiás, 74001-970 Goiânia, GO, Brazil
Author for correspondence (diniz@icb.ufg.br).

Species distribution models (SDMs) assume equilibrium between species' distribution and the environment. However, this assumption can be violated under restricted dispersal and spatially autocorrelated environmental conditions. Here we used a model to simulate species' ranges expansion under two non-equilibrium scenarios, evaluating the performance of SDM coupled with spatial eigenvector mapping. The highest fit is for the models that include space, although the relative importance of spatial variables during the range expansion differs in the two scenarios. Incorporating space to the models was important only under colonization-lag non-equilibrium, under the expected scenario. Thus, mechanisms that generate range cohesion and determine species' distribution under climate changes can be captured by spatial modelling, with advantages compared with other techniques and in line with recent claims that SDMs have to account for more complex dynamic scenarios.

1. INTRODUCTION

Species distribution models (SDMs) for geographical range prediction (Segurado & Araújo 2004; Guisan & Thuiller 2005; Elith *et al*. 2006) assume that species' occurrence is determined by an immediate response of individuals to environmental variation (equilibrium of species' distribution in relation to climate, *sensu* Araújo & Pearson 2005). This is expected only under unlimited dispersal (or if dispersal is at least as fast as the changes in environmental conditions) and very high extinction rates outside the limits of species' environmental 'envelope'.

However, non-equilibrium will arise under different ecological and evolutionary scenarios, so that SDMs may produce biased estimates. First, it

could appear by failures in colonization of suitable areas due to recent environmental changes (e.g. habitat destruction, abrupt climatic shifts or physical barriers) or will appear in the initial stages of species' invasion ('colonization-lag' non-equilibrium, or CNE, hereafter). In this case, although distribution is actually determined by the environment, generating strong range cohesion (*sensu* Rahbek *et al*. 2007), a mismatch between the actual and potential distributions is expected due to historical time lag. Second, complex colonization–extinction dynamics within the species' environmental envelopes, generated by local processes as, for instance, biotic interactions or metapopulation dynamics, will appear as random noise in geographical space. We call this a demographic non-equilibrium (DNE) scenario, which is expected to disrupt range cohesion.

Some recent studies showed that incorporating spatial predictors into SDMs improves model fit (Segurado *et al*. 2006; Bahn & McGill 2007; Dormann 2007). However, as the two scenarios of non-equilibrium described above will generate different spatial structures (highly structured in CNE and not structured in DNE), it is still necessary to test the performance of autocorrelation models under these scenarios and find how they can be conceptually linked to non-equilibrium dynamics (Araújo & Guisan 2006; Heikkinen *et al*. 2006). Here we used simulation models, based on cellular automata, to evaluate how a spatially explicit SDM performs under these non-equilibrium scenarios.

2. MATERIAL AND METHODS

The use of simulated data is an interesting approach to evaluate SDMs because some important species-range properties affecting modelling efficiency can be controlled for (Hirzel *et al*. 2001; Austin *et al*. 2006; Meynard & Quinn 2007). This is usually done by creating a hypothetical species whose occurrences are found within a pre-determined 'bioclimatic envelope', and sampled occurrences are then used to evaluate SDM performance by comparison with a known range. As we were interested in dynamic scenarios, we generated non-equilibrium ranges by a spatially explicit simulation of colonization–extinction mechanisms using cellular automata models.

All simulations were based on the premise that species distribution is affected by a simple 'suitability' measure, established by the combination of unimodal responses to environmental variables (Meynard & Quinn 2007). This suitability measure was defined for each of the 2545 cells (0.24 decimal degrees cell size) covering the geographical area of the cerrado biome (figure 1), based on four climatic variables (mean annual temperature and its seasonality and mean annual precipitation and its seasonality, from the WORLDCLIM database; available at http://www.worldclim.org) and two topographic variables (altitude and slope, from the Hydro-1K global digital elevation model). The cerrado realm was used here as a computational facility only.

Range expansion processes were simulated based on local colonization and extinction constrained by local suitability (see electronic supplementary material for details). Under the CNE scenario, the range expansion was based on two simple rules: (i) a species automatically colonizes a cell 'i' if there is any neighbouring cell 'j' successfully colonized at time $t-1$ and (ii) extinction probability is linearly and negatively related to the suitability. A DNE scenario was simulated by adding stochastic colonization–extinction dynamics to the range expansion model. Thus, we allowed for a suitability-independent persistence probability that increased linearly with the proportion of neighbouring cells successfully colonized at time $t-1$.

We modelled species' distribution at 15 time steps (cycles), before all possible suitable areas were occupied by the population in both simulation models, using the maximum entropy principle implemented in the program Maxent v. 3.4 (Phillips *et al*. 2006). At each step, 100 occurrence points were randomly sampled and modelled in Maxent based on the six

modelled distributions
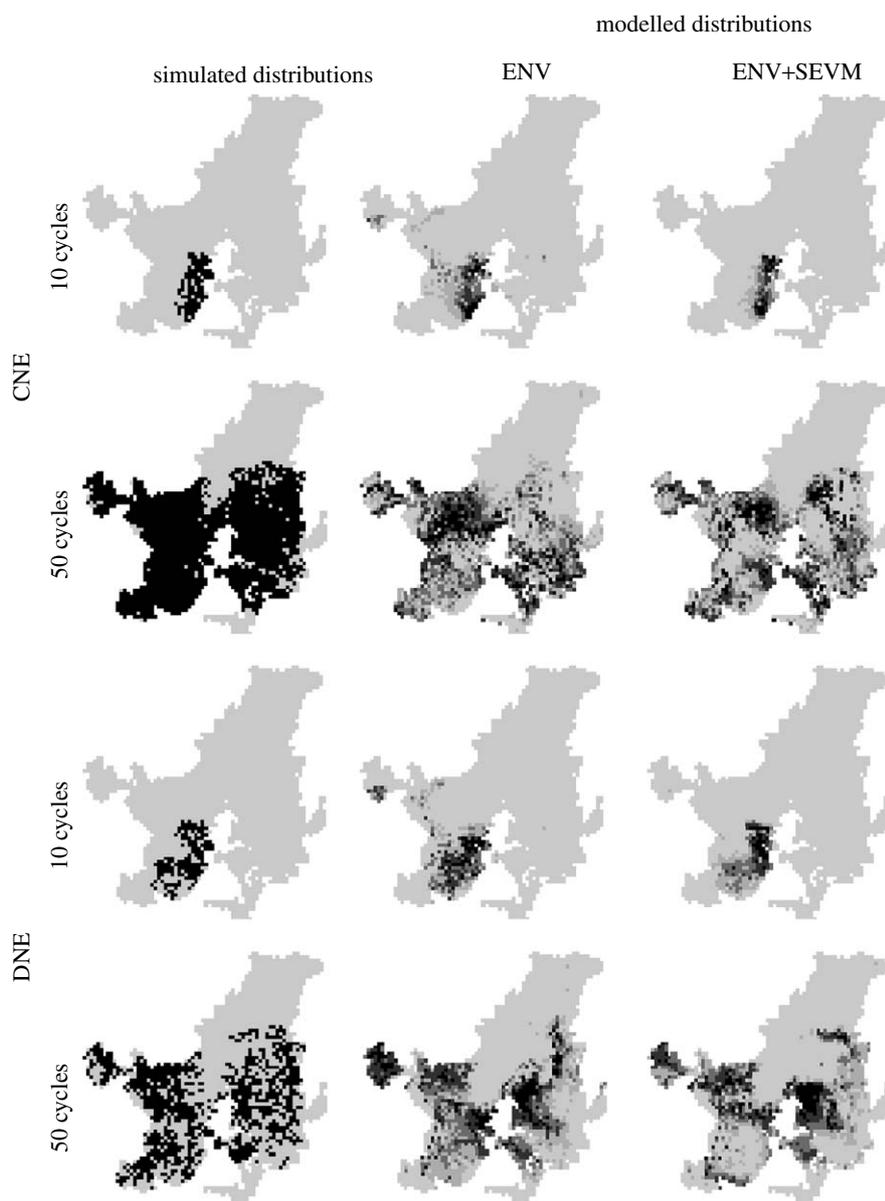
simulated distributions   ENV   ENV+SEVM

Figure 1. Simulated range distribution under two scenarios of non-equilibrium (CNE and DNE) at two selected time steps and their predicted distribution obtained by modelling environment alone (ENV) or environment coupled with spatial eigenvectors (ENV+SEVM).

environmental variables previously described. Model evaluation was done using Cohen's kappa (κ; Allouche *et al*. 2006) obtained after converting probabilities of occurrence to presence–absence data. Cut-off thresholds were established using receiving operator characteristic curves (Liu *et al*. 2005). We then added the first five eigenvectors extracted from a truncated double-centred geographical distance matrix as additional predictors, coupling then MAXENT with spatial eigenvector mapping (see Diniz-Filho & Bini (2005), Griffith & Peres-Neto (2006) and Dormann *et al*. (2007), for reviews). These eigenvectors are orthogonal spatial predictors that capture, at different scales, the geometry of the studied area and were obtained in spatial analysis in macroecology software (Rangel *et al*. 2006).

Spatial autocorrelation in model residuals (i.e. observed occurrence–probability of occurrence given by MAXENT at each cell) was investigated using Moran's *I* coefficients (Dormann *et al*. 2007). We used an analysis of covariance (ANCOVA) to verify whether the gain in κ values ($\Delta\kappa$) after adding spatial predictors was influenced by the type of scenario simulated (CNE versus DNE). A decrease in κ values is expected with the increase of range size due to the loss of statistical power and reduction in prevalence (Allouche *et al*. 2006; Jiménez-Valverde & Lobo 2007), since sample size for MAXENT analysis was held constant. Thus, to account for this relationship and make $\Delta\kappa$ comparable, geographical range size was allowed as a covariate in the ANCOVA.

3. RESULTS AND DISCUSSION

The analyses revealed that, in the initial phases of range expansion, adding spatial variables always provided better fit than using environmental data alone in MAXENT (figure 1; figure S1, see also animations in the electronic supplementary material). Under the DNE scenario, models have lower κ values than in their corresponding simulations for the CNE scenario (figure 2a) up to 30 time steps. However, ranges expanded continuously in the second scenario, whereas in the first there was a tendency to stabilization below the maximum expected by suitability (figure 2b). The relationship between gain in κ values after adding spatial predictors ($\Delta\kappa$ values) and range size was independent of the scenario, as the interaction between this factor and range size was not significant ($F_{1,26}=1.05$; $p=0.3144$). After accounting for the effect of range size ($F_{1,27}=58.5$; $p<0.001$), a significant effect of scenario was detected ($F_{1,27}=9.14$;

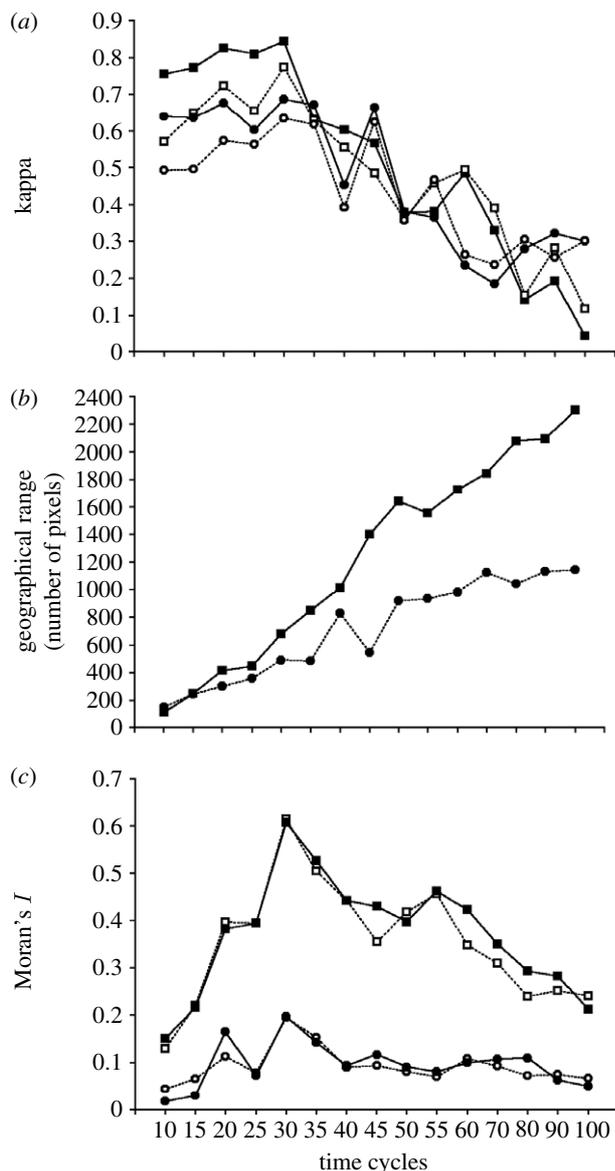Figure 2. (*a*) Kappa statistics (open squares, CNE-ENV; filled squares, CNE-ENV-SEVM; open circles, DNE-ENV; filled circles, DNE-ENV-SEVM), (*b*) range size (squares, CNE; circles, DNE) and (*c*) residual autocorrelation (Moran's *I*) in the first distance class for MAXENT based on environment and spatial eigenvectors models under CNE and DNE simulations across time cycles (open squares, CNE-ENV; filled squares, CNE-ENV-SEVM; open circles, DNE-ENV; filled circles, DNE-ENV-SEVM).

$p < 0.01$) and the adjusted mean value of $\Delta\kappa$ was actually 10 times higher in the CNE (0.06) than in DNE (0.005; see figure S2 in the electronic supplementary material).

Under CNE, using environmental variables alone overestimates the range in the initial phases of the range expansion (figure 1). This occurs because in these initial phases the occurrences are sampled within a restricted part of the range, so there is a systematic bias in sampling environmental suitability values. By including spatial predictors, a better fit was obtained because these additional predictors forced range cohesion independently of the spatial distribution of the environmental suitability. Under the CNE scenario, spatial autocorrelation in the residuals was higher than in the DNE scenario, due to the higher levels of range

cohesion within a more concentrated part of the potential range defined by suitability (figure 2*c*). On the other hand, the low levels of spatial autocorrelation in residuals under DNE shows that suitability is enough to ensure accurate predictions and, consequently, this explains why spatial models tend to be ineffective to improve fit in this case (figure 2*a*).

It is well known that biotic interactions and stochastic colonization processes also determine species' range (Heikkinen *et al.* 2006; Araújo & Luoto 2007; Soberon 2007). Spatial eigenvector mapping and other spatial autocorrelation techniques can account for these processes only if they are spatially structured. Our analyses reveal that adding spatial components can be a promising approach to modelling CNE processes, such as, for instance, those occurring under fast climate change allowing species' range expansion towards new suitable areas. However, they are ineffective under DNE, in which departures from bioclimatic envelopes are caused by local processes related to biotic interactions or meta-population structure within species' ranges. This is coherent with theoretical expectations based on the origins of autocorrelation in biogeographical data (Diniz-Filho *et al.* 2003). So, despite the uncertainty associated with particular SDM techniques (Thuiller 2003, 2004; Araújo & New 2007) and recent criticisms of the limited transferability of MAXENT (Peterson *et al.* 2007; but see Phillips 2008), our main conclusions must hold in general.

Although further studies are necessary to show how these spatial predictors can be coupled with projected environmental changes, spatial eigenvector mapping is particularly suitable for this task as it allows representing spatial relationships at different spatial scales. Also, they can be easily introduced as new predictors in any SDM, with the advantage of not being intrinsically related to observed species' distribution, as it occurs with autologistic terms (Dormann 2007). This is in line with recent suggestions that it is necessary to expand SDMs to incorporate other more complex dynamic scenarios in a spatially explicit context.

Allouche, O., Tsoar, A. & Kadmon, R. 2006 Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistics (TSS). *J. Appl. Ecol.* **43**, 1223–1232. (doi:10.1111/j.1365-2664.2006.01214.x)

Araújo, M. B. & Guisan, A. 2006 Five (or so) challenges for species distribution modelling. *J. Biogeogr.* **33**, 1677–1688. (doi:10.1111/j.1365-2699.2006.01584.x)

Araújo, M. B. & Luoto, M. 2007 The importance of biotic interactions for modelling species distributions under climate change. *Glob. Ecol. Biogeogr.* **16**, 743–753. (doi:10.1111/j.1466-8238.2007.00359.x)

Araújo, M. B. & New, M. 2007 Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42–47. (doi:10.1016/j.tree.2006.09.010)

Araújo, M. B. & Pearson, R. G. 2005 Equilibrium of species' distributions with climate. *Ecography* **28**, 693–695. (doi:10.1111/j.2005.0906-7590.04253.x)

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. (doi:10.1016/j.ecolmodel.2006.05.023)

Bahn, V. & McGill, B. J. 2007 Can niche-based distribution models outperform spatial interpolation. *Glob. Ecol. Biogeogr.* **16**, 733–742. (doi:10.1111/j.1466-8238.2007.00331.x)

Diniz-Filho, J. A. F. & Bini, L. M. 2005 Modelling geographical patterns in species richness using eigen-vector-based spatial filters. *Glob. Ecol. Biogeogr.* **14**, 177–185. (doi:10.1111/j.1466-822X.2005.00147.x)

Diniz-Filho, J. A. F., Bini, L. M. & Hawkins, B. A. 2003 Spatial autocorrelation and red herrings in geographical ecology. *Glob. Ecol. Biogeogr.* **12**, 53–64. (doi:10.1046/j.1466-822X.2003.00322.x)

Dormann, C. F. 2007 Assessing the validity of autologistic regression. *Ecol. Model.* **207**, 234–242. (doi:10.1016/j.ecolmodel.2007.05.002)

Dormann, C. F. *et al.* 2007 Methods to account for spatial autocorrelation in the analysis of distributional species data: a review. *Ecography* **30**, 609–628. (doi:10.1111/j.2007.0906-7590.05171.x)

Elith, J. *et al.* 2006 Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151. (doi:10.1111/j.2006.0906-7590.04596.x)

Griffith, D. A. & Peres-Neto, P. 2006 Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**, 2603–2613. (doi:10.1890/0012-9658(2006)87[2603:SMIETF]2.0.CO;2)

Guisan, A. & Thuiller, W. 2005 Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009. (doi:10.1111/j.1461-0248.2005.00792.x)

Heikkinen, R. K., Luoto, M., Araújo, M. B., Virkkala, R., Thuiller, W. & Sykes, M. T. 2006 Methods and uncertainties in bioclimatic envelope modelling under climate change. *Prog. Phys. Geogr.* **30**, 751–777. (doi:10.1177/0309133306071957)

Hirzel, A. H., Helfer, V. & Metral, F. 2001 Assessing habitat-suitability models with a virtual species. *Ecol. Model.* **145**, 111–121. (doi:10.1016/S0304-3800(01)00396-9)

Jiménez-Valverde, A. & Lobo, J. M. 2007 Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecol.* **31**, 361–369. (doi:10.1016/j.actao.2007.02.001)

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. (doi:10.1111/j.0906-7590.2005.03957.x)

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. (doi:10.1111/j.1365-2699.2007.01720.x)

Peterson, A. T., Papes, M. & Eaton, M. 2007 Transferability and model evaluation in ecological niche modeling: a comparison of GARP and MAXENT. *Ecography* **30**, 550–560.

Phillips, S. J. 2008 Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson *et al.* (2007). *Ecography* **31**, 272–278. (doi:10.1111/j.0906-7590.2008.5378.x)

Phillips, S. J., Anderson, R. P. & Schapire, R. E. 2006 Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231–259. (doi:10.1016/j.ecolmodel.2005.03.026)

Rahbek, C., Gotelli, N. J., Colwell, R. K., Entsminger, G. L., Rangel, T. F. L. V. B. & Graves, G. R. 2007 Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proc. R. Soc. B* **274**, 165–174. (doi:10.1098/rspb.2006.3700)

Rangel, T. F. L. V. B., Diniz-Filho, J. A. F. & Bini, L. M. 2006 Towards an integrated computational tool for spatial analysis in macroecology and biogeography. *Glob. Ecol. Biogeogr.* **15**, 321–327. (doi:10.1111/j.1466-822X.2006.00237.x)

Segurado, P. & Araújo, M. B. 2004 An evaluation of methods for modelling species distributions. *J. Biogeogr.* **31**, 1555–1568. (doi:10.1111/j.1365-2699.2004.01076.x)

Segurado, P., Araújo, M. B. & Kunin, W. E. 2006 Consequences of spatial autocorrelation for niche-based models. *J. Appl. Ecol.* **43**, 433–444. (doi:10.1111/j.1365-2664.2006.01162.x)

Soberon, J. 2007 Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* **10**, 1115–1123. (doi:10.1111/j.1461-0248.2007.01107.x)

Thuiller, W. 2003 BIOMOD: optimising predictions of species distributions and projecting potential future shifts under global change. *Glob. Change Biol.* **9**, 1353–1362. (doi:10.1046/j.1365-2486.2003.00666.x)

Thuiller, W. 2004 Patterns and uncertainties of species' range shifts under climate change. *Glob. Change Biol.* **10**, 2020–2027. (doi:10.1111/j.1365-2486.2004.00859.x)