Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2014) **23**, 811–820

Global Ecology and Biogeography
A Journal of Macroecology

MACROECOLOGICAL METHODS

Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation

Kévin Le Rest[1]*, David Pinaud[1], Pascal Monestiez[1,2,3], Joël Chadoeuf[3] and Vincent Bretagnolle[1]

[1]*Centre d'Études Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France,* [2]*INRA (USC 1339), CEBC-CNRS, 79360 Beauvoir-sur-Niort, France,* [3]*Unité Biostatistique et Processus Spatiaux (BioSP), INRA Provence-Alpes-Côte d'Azur, Domaine Saint-Paul Site, Agroparc, 84914 Avignon Cedex 9, France*

*Correspondence: Kévin Le Rest, Agripop, Centre d'Études Biologiques de Chizé, CEBC-CNRS UPR 1934, Beauvoir-sur-Niort, 79360, France.
E-mail: lerest.k@gmail.com

ABSTRACT

Aim Processes and variables measured in ecology are almost always spatially autocorrelated, potentially leading to the choice of overly complex models when performing variable selection. One way to solve this problem is to account for residual spatial autocorrelation (RSA) for each subset of variables considered and then use a classical model selection criterion such as the Akaike information criterion (AIC). However, this method can be laborious and it raises other concerns such as which spatial model to use or how to compare different spatial models. To improve the accuracy of variable selection in ecology, this study evaluates an alternative method based on a spatial cross-validation procedure. Such a procedure is usually used for model evaluation but can also provide interesting outcomes for variable selection in the presence of spatial autocorrelation.

Innovation We propose to use a special case of spatial cross-validation, spatial leave-one-out (SLOO), giving a criterion equivalent to the AIC in the absence of spatial autocorrelation. SLOO only computes non-spatial models and uses a threshold distance (equal to the range of RSA) to keep each point left out spatially independent from the others. We first provide some simulations to evaluate how SLOO performs compared with AIC. We then assess the robustness of SLOO on a large-scale dataset. R software codes are provided for generalized linear models.

Main conclusions The AIC was relevant for variable selection in the presence of RSA if the independent variables considered were not spatially autocorrelated. It otherwise failed because highly spatially autocorrelated variables were more often selected than others. Conversely, SLOO had similar performances whether the variables were themselves spatially autocorrelated or not. It was particularly useful when the range of RSA was small, which is a common property of spatial tools. SLOO appears to be a promising solution for selecting relevant variables from most ecological spatial datasets.

Keywords
AIC, common buzzard *Buteo buteo*, GLM, residual spatial autocorrelation, simulations, spatial cross-validation.

INTRODUCTION

Ecological processes *in natura* are inherently spatial, either for environmental or intrinsic biological reasons (Legendre & Fortin, 1989; Legendre, 1993; Koenig, 1999; Keitt *et al.*, 2002). Data collected in the field are thus usually spatially autocorrelated. Spatial autocorrelation can alter the statistical independence of residuals in regression models, leading to bias such as falsified tests or likelihoods (Lennon, 2000; Bahn *et al.*, 2006; Hoeting *et al.*, 2006; Dormann, 2007; but see Diniz-Filho *et al.*, 2003). Statistical methods able to capture the residual spatial autocorrelation (RSA), so-called 'spatial models', are required to correct for those biases (see Lichstein *et al.*, 2002; Fortin & Dale, 2005; Griffith & Peres-Neto, 2006; Dormann *et al.*, 2007; Betts *et al.*, 2009; Beale *et al.*, 2010; Saas & Gosselin, 2014). While such methods were shown to be efficient for the

© 2014 John Wiley & Sons Ltd

DOI: 10.1111/geb.12161
http://wileyonlinelibrary.com/journal/geb

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

estimation of model parameters from spatial datasets, they are mainly applied after the process of variable selection. This begs the question of the validity of variable selection in the presence of RSA.

Model selection has gained a wide audience in ecology (Johnson & Omland, 2004), with the main aim of selecting pertinent variables by comparing several models with different subsets of variables and choosing which ones are most likely to explain the observed pattern in relation to the studied process. A few metrics have been proposed to help in this process, e.g. Mallows' C_p (Mallows, 1973), the Akaike information criterion (AIC; Akaike, 1973) or the Bayesian information criterion (BIC; Schwarz, 1978). These selection criteria usually reflect a balance between the data fit and model complexity (George, 2000). The principal difference between them comes from the penalty accorded to the model complexity. For instance, Mallows' C_p and the AIC are almost equivalent and correspond to a penalty of 2 whereas the BIC corresponds to a penalty of $\log(n)$, where n is the number of independent observations (George, 2000). BIC is asymptotically consistent since it will select the true model as $n \to \infty$ (see Stone, 1979; George, 2000). An implicit assumption, however, is the existence of a 'true model' in the set of candidate ones (Stone, 1979; Shao, 1997; George, 2000). In biological sciences this assumption is unrealistic because the number of variables affecting the processes can be very high, if not infinite (see Burnham & Anderson, 2002). It will thus be better to allow the dimension of the true model to increase with n (Stone, 1979), which is a fundamental property of Mallows' C_p and the AIC.

The presence of RSA invalidates the use of classical model selection criteria such as Mallows' C_p, the AIC or the BIC since they are based on the overall likelihood assuming independent residuals (Cordy & Griffith, 1993; Cassemiro *et al.*, 2007). However, in practice these criteria were still used without accounting for RSA (see, e.g., Kühn *et al.*, 2009). This may lead to the selection of overly complex models having a much larger number of variables than necessary (Hoeting *et al.*, 2006; Cassemiro *et al.*, 2007; Diniz-Filho *et al.*, 2008). As acknowledged by Dormann *et al.* (2007), variable selection in the presence of RSA has received surprisingly little interest so far in the literature. Identifying the relevant variables in the presence of RSA thus remains challenging. The classical strategy for variable selection in the presence of RSA consists of first accounting for RSA in all candidate models and then to compare them with a classical model selection criterion. The computed criterion is this time valid since RSA has been removed. For instance, Hoeting *et al.* (2006) underlined the need to account for RSA when using the AIC for variable selection in a geostatistical modelling framework, as did Diniz-Filho *et al.* (2008) who compared two methods for accounting for RSA. This approach, however, has three main drawbacks: first, models accounting for RSA need a much longer computation time, making model selection very difficult when the number of variables is large; second, the variables that are finally selected may depend on the method used to account for RSA (see Diniz-Filho *et al.*, 2008); and third, most 'spatially explicit methods' may lead to a 'spatial confounding' effect between the variables and the spatial term, hiding the importance of some spatially autocorrelated variables (Reich *et al.*, 2006; Betts *et al.*, 2009; Bini *et al.*, 2009; Hodges & Reich, 2010; Paciorek, 2010; Hughes & Haran, 2013). This latter effect is less well known but is of primary interest when one wants to perform variable selection with spatially autocorrelated variables, which probably happens in most real applications.

Yet another method based on a modification of a cross-validation procedure has been frequently used for model evaluation in the presence of RSA and should also be used for variable selection in this context. Cross-validation usually consists in splitting the initial dataset in two subsets (see Arlot & Celisse, 2010, for an overview), one is used to estimate model parameters (the training set) and the other one is used to evaluate the predictive power of the model (the validation set). A critical prerequisite is that the training and validation sets should be independent (Arlot & Celisse, 2010), at least under the model being evaluated. Otherwise, the difference between the observation and the prediction may be unreliable (Altman, 1990). In a spatial context, most observations are related to each other, so training and validation sets are rarely independent, which highly reduces the power of cross-validation to evaluate a model. An intuitive way to solve this problem consists in splitting the spatial data into several non-overlapping geographical areas that are used as training and validation sets, a technique often referred to as spatial cross-validation (see Chung & Fabbri, 2003; Brenning, 2005; Pinkerton *et al.*, 2010; Russ & Brenning, 2010; Hijmans, 2012; Bahn & McGill, 2013). It is also necessary that the distance between the training and the validation areas is greater than the range of RSA (i.e. the distance at which a pair of observations are independent) of the model evaluated in order to guarantee full independence (Brenning, 2005; Russ & Brenning, 2010). Unfortunately this minimal distance between the training and validation sets is almost always ignored (see, e.g., Chung & Fabbri, 2003; Pinkerton *et al.*, 2010; Russ & Brenning, 2010; Bahn & McGill, 2013). Spatial cross-validation is actually a spatial version of delete-d cross-validation (Geisser, 1975), where d is the number of observations in the validation set. If there is no true model, which is expected in ecological applications, delete-d cross-validation is only useful for variable selection when $d = 1$, i.e. when a simple leave-one-out (LOO) cross-validation (Allen, 1974; Stone, 1974) is used (see Shao, 1997). The current form of spatial cross-validation considering $d >> 1$ should thus not be useful for variable selection in this context. The special case of $d = 1$ would, however, provide a useful criterion since LOO cross-validation is known to be asymptotically equivalent to the AIC (Stone, 1977).

In this paper we thus propose to evaluate the performance of the spatial LOO (SLOO) cross-validation for variable selection in the presence of RSA. The selection criterion is computed by applying LOO in a spatial context, i.e. by using a threshold distance between the training and the validation sets that removes some data in order to eliminate the bias due to RSA. This approach has recently been used by Le Rest *et al.* (2013), though these authors did not provide either a suitable calculus of the selection criterion or an evaluation of its performance. We first give a full description of the method and the way to

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

compute the selection criterion. Then, we use a simulation approach to evaluate how SLOO performs compared with the AIC in selecting a continuous variable that affects the studied process while avoiding another one that does not affect the process. We quantify in particular the relative effects of: (1) the RSA, (2) the threshold distance used to calculate SLOO, and (3) the spatial autocorrelation in the explanatory variables. In 'Application to a real case study' we use SLOO with different threshold distances on a real data set composed of 28 environmental variables suspected to influence the abundance of a diurnal raptor species, the common buzzard *Buteo buteo* (Linnaeus, 1758). This case study illustrates the utilization of SLOO and its robustness for the selection of variables in a species distribution model. Finally, Appendix S1 in the Supporting Information provides computing codes and an example based on the simulations showing how to calculate SLOO with R software from generalized linear models (GLMs).

THE SPATIAL LEAVE-ONE-OUT METHOD

The SLOO method relies in practice on four steps. The first step removes one observation from the initial dataset (the grey cross in Fig. 1). The second step removes all the observations that are spatially correlated with this removed observation, i.e. removes all data inside a buffer of a radius equal to the range of the RSA of the model considered (see for example the grey buffer in Fig. 1). All remaining observations (black points in Fig. 1) constitute the training set and are used in a GLM framework to estimate the parameters. A prediction (step three) is then made at the location of the removed observation (validation set, grey point in Fig. 1) using the estimated parameters of the GLM. The fourth step calculates a score between the observed value and the predicted one. This procedure is repeated for every single observation of the dataset, which allows calculation of an overall criterion of fit. Note that LOO is a special case of SLOO when the threshold distance used is null and SLOO is thus asymptotically equivalent to the AIC in absence of RSA (Stone, 1977), allowing a direct comparison between these two selection criteria.

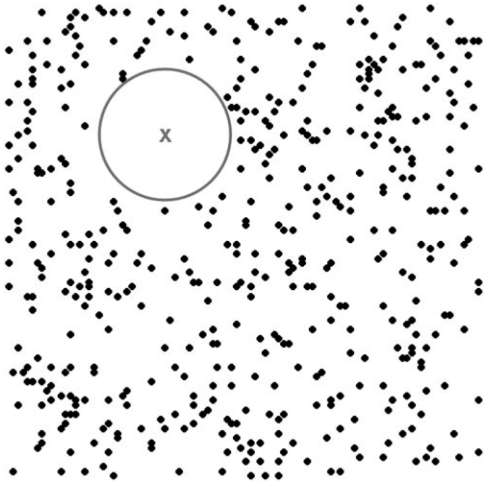

Figure 1 One example of spatial leave-one-out on a grid of 100 pixels × 100 pixels having 500 observations. Here the threshold distance is set arbitrarily to 15 pixels (radius of the grey buffer). The grey cross is the point left out, i.e. the validation set, and the black points are the training set.

The criterion of the SLOO (equation 1) is based here on likelihood instead of the classical sum of squares of errors, because it is more adapted for non-normally distributed response variables (see Knafl & Grey, 2007, for details of the likelihood versus least square cross-validation) and is therefore more suitable for GLMs. In practice, we compute the probability *P* for a discrete response variable (or the density probability for a continuous one) of the left out observed value y_i according to the predicted one $\hat{y}_i$ by using the training set. This is achieved by using the theoretical distribution of the model (normal, binomial, Poisson, etc. . . .). The sum of the logarithm of these probabilities leads to an overall cross-validated log-likelihood for the model, which is the selection criterion to be maximized.

$$SLOO_{\log \text{Lik}} = \sum_{i=1}^{n} \log[P(y_i|\hat{y}_i)] \tag{1}$$

All simulations and analysis were performed using R version 2.13.0 (R Core Team, 2013; http://www.R-project.org). Full details on how to calculate this criterion with R and an example can be found in Appendix S1.

SIMULATIONS

We conducted the first simulation on a 100 pixel × 100 pixel regular grid approximating a continuous field and iterated the following process 10,000 times.

1. Generate three Gaussian random fields (GRFs) with a spherical spatial structure with mean equal to zero, variance (sill) equal to one, no nugget effect and a range chosen randomly between 1 and 100 pixels.

2. Generate the response variable *Y* such as:

$$Y = \text{GRF1} + \beta\,\text{GRF2}. \tag{2}$$

GRF1 was considered as unavailable (unknown process) and was used to generate RSA from its spatial properties. GRF2 played the role of an available and influential variable, and the parameter β reflected its actual importance. β was taken from *N*(0,1), which allowed us to scan a wide range of values avoiding values that were too high. High β (> 2) were irrelevant since they always led to select the influential variable (GRF2) in the model whatever the selection criterion used. GRF3 did not affect the response *Y* in equation 2 and can be considered as an available but non-influential variable. A random sample of 500 observations from the 100 × 100 grid was considered as the available dataset.

3. Run two variable selection procedures (one using the AIC and one using SLOO) with GRF2 and GRF3 being the candidate variables of the model. Note that SLOO was computed by using the range of GRF1 as threshold distance.

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

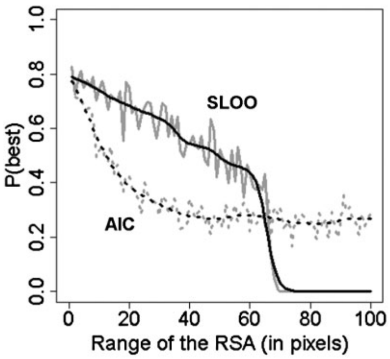

Figure 2 Probability (P) of selecting the best model depending on the range of the residual spatial autocorrelation (RSA), by using two selection criteria: the Akaike information criterion (AIC, dotted lines) and the spatial-leave-one-out (SLOO, solid lines). Black lines are cubic smoothing splines and grey lines consider the range of RSA as a factor (measure of the variability).

4. Record which variables (GRF2 and/or GRF3) were selected for each selection criterion used, i.e. either by minimizing the AIC or by maximizing $\mathrm{SLOO}_{\mathrm{logLik}}$.

Note that the 'true model', i.e. the one holding the two influential variables (GRF1 and GRF2) and avoiding the non-influential one (GRF3), could never be selected since GRF1 was considered as unavailable; thus the term 'best model' was used to qualify the model holding the influential variable (GRF2) and avoiding the non-influential one (GRF3).

Binomial regression models were used to represent graphically either the probability of selecting the best model, the probability of selecting the influential variable (GRF2) or the probability of selecting the non-influential one (GRF3), by considering alternatively the AIC or SLOO and depending on varying levels of RSA. The range of RSA was first used as a factor having 100 modalities (between 1 and 100 pixels) and then smoothed using cubic smoothing splines in order to provide an easier graphical representation.

Effect of the residual spatial autocorrelation

Figure 2 shows that the probability of selecting the best model, i.e. one holding the influential variable (GRF2) and avoiding the non-influential one (GRF3), was always higher using SLOO than the AIC except when the range of RSA was higher than 60 pixels (i.e. it had a width of 60% of the grid), a value at which most of the training set was removed (see 'The spatial leave-one-out method' and Fig. 1). Indeed in our simulated area, the farthest distance between two locations was about 140 pixels (the diagonal of the grid), which explained why the capacities of SLOO decreased with threshold distance between 60 and 70 pixels. SLOO could thus not be used when the RSA was higher than 70. We also found that the probability of selecting the best model decreased as RSA increased, considering either the AIC or SLOO (Fig. 2).

Using the AIC, the probability of selecting the influential variable (GRF2 in Fig. 3) was always high and actually slightly increased with the range of RSA. But on the other hand, the probability of selecting the non-influential one (GRF3 in Fig. 3) dramatically increased with the range of RSA. These results were expected since AIC is known to select overcomplex models in presence of RSA (Hoeting *et al.*, 2006; Cassemiro *et al.*, 2007; Diniz-Filho *et al.*, 2008). Conversely, using SLOO, the probability of selecting the influential variable (GRF2 in Fig. 3) decreased with increasing RSA and the probability of selecting the non-influential one (GRF3 in Fig. 3) slightly increased as RSA increased (up to a certain limit, see comments above) but remained rather low. However, it was not possible to determine if these effects were due to RSA because the range of RSA was also the threshold distance used in the SLOO, which caused a decrease in the number of observations in the training set (see 'The spatial leave-one-out method' and Fig. 1), also decreasing the statistical power of the SLOO. The effect of the threshold distance was thus investigated with another simulation.

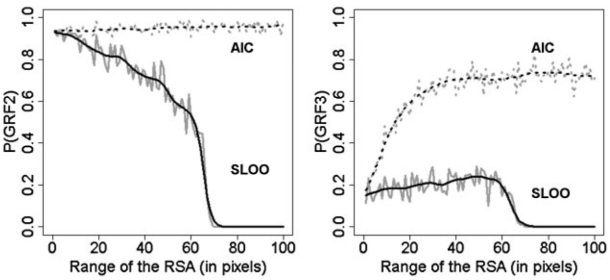

Figure 3 Probability (P) of selecting the influential variable (GRF2) and the non-influential one (GRF3) depending on the range of the residual spatial autocorrelation (RSA), by using two selection criteria: the Akaike information criterion (AIC, dotted lines) and the spatial leave-one-out (SLOO, solid lines). Black lines are cubic smoothing splines and grey lines consider the range of RSA as a factor (measure of the variability).

Effect of the threshold distance used in the absence of RSA

The first simulation procedure (see 'Effect of the residual spatial autocorrelation') was modified in order to separately study the threshold distance: here there was no RSA (i.e., GRF1 had no spatial structure), and varying threshold distances were used for SLOO, chosen randomly between 1 and 100 pixels (as for the range of GRF1 in the first simulation).

The probability of selecting the influential variable in the absence of RSA using SLOO was very close to AIC performances whatever the threshold distance used (see GRF2 in Fig. 4), except when the threshold distance was higher than half of the extent of the study area (see 'Effect of the residual spatial autocorrelation' for an explanation). Moreover the probability of selecting the non-influential variable using SLOO (see GRF3 in Fig. 4) was only slightly increased by the threshold distance used. Overall the threshold distance used in SLOO only slightly affected the probability of selecting the variables, and was thus not the cause of the important decrease in the probability of selecting the influential variable when increasing the range of RSA (GRF2 in Fig. 3). This latter result may be explained by pseudo-replication caused by RSA, leading naturally to a loss of

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

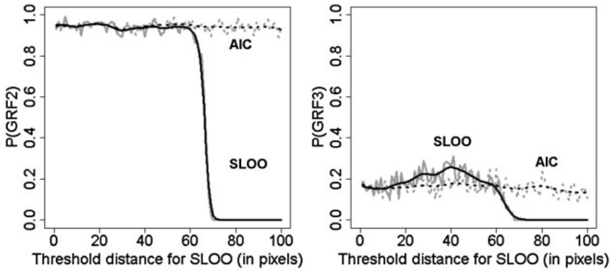

Figure 4 Probability (P) of selecting the influential variable (GRF2) and the non-influential one (GRF3) in the absence of residual spatial autocorrelation but according to the threshold distance used for spatial leave-one-out (SLOO, solid lines). Results for the Akaike information criterion (AIC, dotted lines) are given for an easier visual comparison. Black lines are cubic smoothing splines and grey lines consider the threshold distance as a factor (measure of the variability).

power by decreasing the true number of degrees of freedom (see Legendre, 1993).

Effect of the spatial autocorrelation of the variables

The first simulation (see 'Effect of the residual spatial autocorrelation') also showed that the probability of selecting the variables depended on their own spatial autocorrelation. In the presence of RSA, the probability of selecting the non-influential variable using the AIC also increased with its own spatial autocorrelation (Fig. 5). Lennon (2000) found a similar result by considering correlations and levels of significance on explanatory variables. Thus in the presence of RSA, both the amount of RSA and the amount of spatial autocorrelation of the explanatory variables could affect the probability of selecting the variables when using the AIC. SLOO, conversely, showed less sensitivity to RSA; in particular, the probability of selecting the non-influential variable was not affected by its own spatial autocorrelation (see Fig. 5).

Effect of sample size and the number of explanatory variables

We also analysed the effect of sample size on the initial simulation set (from 100 to 10,000). Ten thousand observations led to a dramatic increase in the probability of selecting the non-influential variable using the AIC, which sharply increased the difference between the AIC and SLOO in selecting the best model (see Fig. S2(c) in Appendix S2); conversely, reducing the number of observations (to 100) led to a reduction in the difference between the AIC and SLOO (see Fig. S2(a) in Appendix S2). This could be explained by the fact that observations were chosen randomly on the grid and were thus far apart for small sample size, which reduced the impact of RSA.

Including higher numbers of explanatory variables (10 influential variables and 10 non-influential ones) did not affect the previous results (compare Fig. S3 in Appendix S3 with Fig. 3). This was expected because the variables were independent.

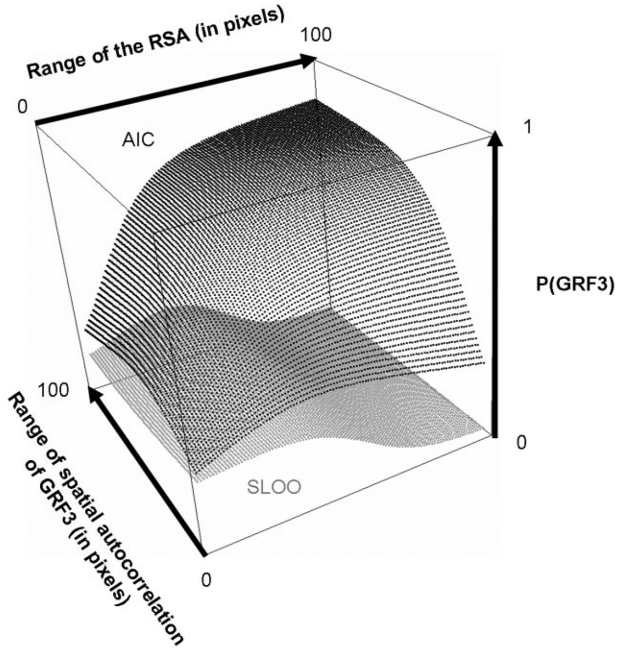

Figure 5 Probability (P) of selecting the non-influential variable (GRF3) depending on its own range of spatial autocorrelation and the range of residual spatial autocorrelation (RSA) by considering two selection criteria: the Akaike information criterion (AIC, in black) and the spatial leave-one-out (SLOO, in grey). The surface plots are obtained from cubic smoothing splines accounting for each dimension and also accounting for the interaction between them.

APPLICATION TO A REAL CASE STUDY

We applied the AIC and SLOO to a dataset from a French national survey of breeding diurnal raptors (Thiollay & Bretagnolle, 2004; Le Rest *et al.*, 2013). Our aim was to identify the environmental variables affecting abundance of the most abundant raptor that breeds in France, the common buzzard *B. buteo.* We used 1206 sampled quadrats of 5 km × 5 km (Fig. 6) and 28 environmental variables suspected of influencing raptor abundance (19 climatic and 9 land-use variables). The climatic variables came from the Bioclim dataset (Hijmans *et al.*, 2005; http://www.worldclim.org/bioclim) and the land-use variables came from the Corine land-cover dataset (http://www.eea.europa.eu). A principal component analysis (PCA) was performed separately on each environmental dataset (climate and land use) because high correlations occurred between initial environmental variables. These principal components were then used in place of the initial environmental variables for analysis. The principal components were labelled as follows: 'ClimDim.x' denoted the xth principal component from the climate dataset and 'ClcDim.x' was used in the same way for the land-cover dataset.

Variable selection was performed by assuming a Poisson distribution in a GLM framework. An automated forward step-by-step algorithm was used in order to reduce the computation time. We first used the AIC without accounting for RSA and

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

then, in order to access the robustness of the SLOO for this dataset, we tested several threshold distances for SLOO, from 0 to 630 km (respectively, the minimum and the maximum distance at which SLOO could be used) every 15-km step (the minimal distance between actual observations). In each case, only the best model, either minimizing AIC or maximizing $\mathrm{SLOO}_{\mathrm{logLik}}$, was retained for simplicity.

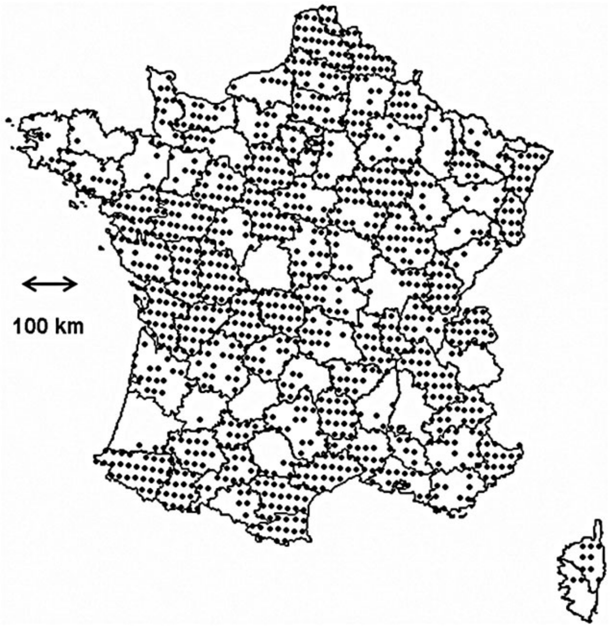

Figure 6 Map of the 1206 sampling quadrats of 5 km × 5 km over France (projection: Lambert azimuthal equal area, ETRS89, EPSG3035). The minimum and maximum distances between observations are, respectively, 15 and 1200 km.

Use of the AIC without accounting for RSA led to the selection of 23 variables in the model (Fig. 7). All but one of these variables was also selected by using LOO (SLOO 0 in Fig. 7), i.e. without a threshold distance. The variable that was not selected by LOO only reduced the AIC by 0.14 units, which meant that these two models were almost equivalent. This was expected because of the asymptotic equivalence between these two criteria (Stone, 1977). Increasing the threshold distance used for the SLOO then led to the selection of fewer variables in the model until reaching 15 variables when using a threshold distance of 45 km (SLOO 45 in Fig. 7). The expected range of RSA was given by the residuals of the full model since it gave the spatial autocorrelation that could not be accounted for by the available variables. It was between 40 and 50 km (see Fig. 8), emphasizing the fact that SLOO led to the selection of more variables in the model when the threshold distance used was lower than the range of RSA. Conversely when the threshold distance reached and exceeded the range of RSA (i.e. from 45 to 255 km in Fig. 7), SLOO selected a stable set of variables, with only very marginal differences (differences of $\mathrm{SLOO}_{\mathrm{logLik}} < 1$). This was in line with the fact that SLOO accounted for RSA when the threshold distance was equal to or higher than the range of RSA. Above a threshold distance of 255 km, variables selected in the model became unstable and their number decreased dramatically for threshold distances above 300 km (i.e. one-quarter of the extent of the studied area), which suggested that the SLOO was not efficient, with too large a threshold distance.

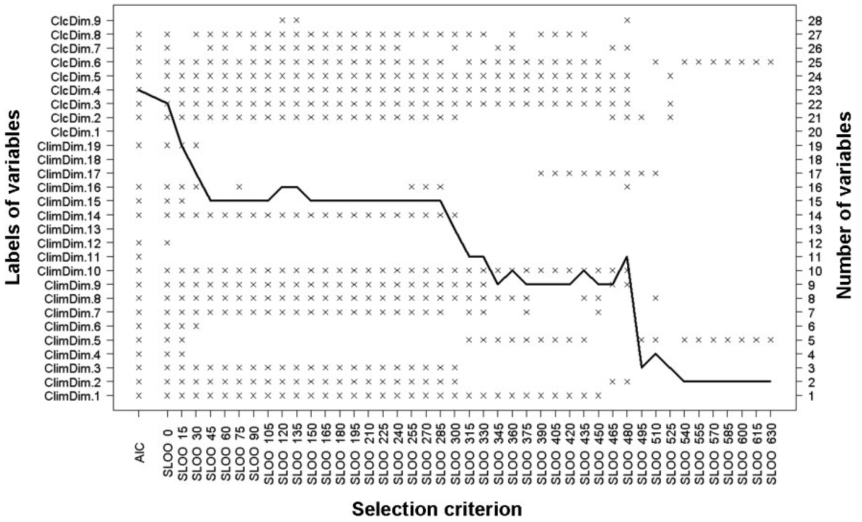

Figure 7 List and number of variables selected in the model by using two selection criteria: the Akaike information criterion (AIC) and the spatial leave-one-out used with different threshold distances (SLOO xxx). The abbreviation SLOO xxx denotes the use of SLOO with a threshold distance of xxx km. Labels ClimDim.x and ClcDim.x denote the xth principal component axis resulting from PCAs done on each type of environmental variables [either climatic (ClimDim) or land cover (ClcDim)], which highly reduces collinearity problems between initial variables. The black line represents the number of variables selected in the best model depending on the selection criterion considered and the marks identify the labels of the variables selected.

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

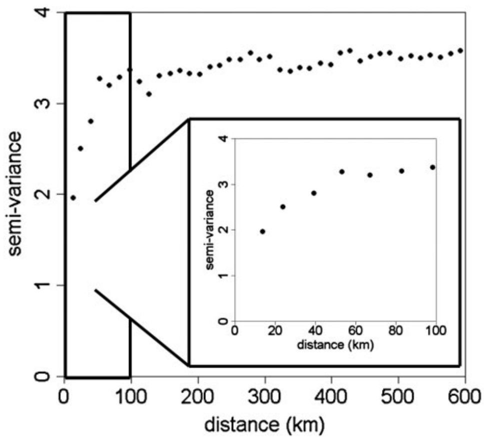

Figure 8 Variogram of the (deviance) residuals of the full model (including all the 28 environmental variables) to explain the abundance of the common buzzard *Buteo buteo*.

The common buzzard is present over the entire territory of France (Thiollay & Bretagnolle, 2004). This species is thus particularly adapted over all climatic constraints of that country. That is why we expected that climatic variables should play a rather small role in explaining the abundance of this species. However, almost all principal components of climatic variables were selected by using the AIC on this dataset, which could suggest the opposite. These climatic variables were also highly spatially autocorrelated and we showed in the simulations that the spatially autocorrelated variables have more chance of being selected by using the AIC in the presence of RSA. The fact that several variables were selected by using the AIC but not selected by using the SLOO with a correct threshold distance should thus be evidence of spatial autocorrelation present in both the residuals and the explanatory variables.

DISCUSSION

In the simulations, using the AIC for variable selection in the presence of RSA led to the selection of unnecessary variables in the regression models. These results were consistent with the conclusions of Diniz-Filho *et al.* (2008). However, this only occurred when the explanatory variables considered were themselves spatially autocorrelated (see Fig. 5). This could be explained by the fact that two random variables are more likely to be correlated (based on the absolute value of their correlation coefficient) when they are both spatially autocorrelated (Liebhold & Sharov, 1998). The correlation between the explanatory variables and the residuals of the regression model was thus inflated in the presence of spatial autocorrelation. The highly spatially autocorrelated variables were then selected more often (see Fig. 5). Conversely, SLOO had similar performances whether the variables were themselves spatially autocorrelated or not (see Fig. 5), providing a great alternative to the AIC for variable selection in the presence of spatial autocorrelation. However, SLOO became less efficient when the threshold distance increased (see Fig 4). This phenomenon resulted from the fact that increasing the threshold distance reduced the number of observations in the training set. When the training set had only a few observations, the estimated parameters were quite unstable between samples and $\mathrm{SLOO}_{\mathrm{logLik}}$ was improved by chance. For the same reasons, SLOO could not be used when the threshold distance exceeded half of the extent of the studied area (there were no observations in the training set).

The results of the case study were highly concordant with the simulations despite the important theoretical constraints that the simulations are never entirely verified with a real dataset, e.g. random sampling in space, stationarity and isotropy. Even if the truth remains unknown for the real dataset, we found evidence that use of the AIC led to unnecessary climatic variables being kept to explain the abundance of the common buzzard (see Fig. 7). It was no coincidence that these variables were also highly spatially autocorrelated. The spurious inclusion of meaningless variables in a model may lead to misguided statistical inference (Johnson & Omland, 2004). This is particularly significant when variables are used to outline ecological processes, for example when they are used to predict the spatial response of species to climate or land-use changes. Using the AIC for variable selection in this case study thus reduced the ability of the data collected to bring relevant ecological information about species. Conversely, SLOO appeared useful for variable selection as soon as the threshold distance exceeded the range of RSA. It was, however, not relevant when the threshold distance was lower than the range of RSA since many unnecessary variables were still selected in the model. Moreover, SLOO became unstable when the threshold distance exceeded one-quarter of the extent of the studied area (about 250 km here, see Fig. 6).

The modification of LOO by removing the non-independent data between the training and validation sets was initially proposed in non-spatial settings (see Chu & Marron, 1991; Burman *et al.*, 1994). It has already been mentioned that removing too much data may have an impact on the effectiveness of the expected prediction error and a limit of one-quarter has also been invoked by Burman *et al.* (1994). SLOO thus appears to be a safe technique for variable selection when the range of RSA does not exceed one-quarter of the extent of the studied area. This limit is not so restrictive since spatial tools (such as the variogram estimation) become less recommended when the range of RSA exceeds one-third of the extent of the studied area, and unreliable when it exceeds one-half (Rossi *et al.*, 1992; Dungan *et al.*, 2002). It is easy to use since it computes GLMs and only needs the range of RSA as additional spatial information (used as threshold distance). A prerequisite is, however, to correctly estimate the range of RSA. We propose using the range of RSA on the residuals of the full model, i.e. the model including all available variables, because it gives the RSA that cannot be accounted for by the available variables. Note that this strategy may underestimate the true range of RSA by including unnecessary spatially autocorrelated variables in the model. Caution must also be taken in establishing the threshold distance used with SLOO, and one must keep in mind that if SLOO appears robust when estimating the range of RSA at an upper limit, it may not be robust against underestimation (see Fig. 7). The performance of SLOO may also

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

depend on how the space has been sampled. All spatial tools are affected by irregular sampling and we do not expect there to be more concerns with SLOO than with other methods. Further investigation of the performance of SLOO on irregular spatial datasets remains necessary to confirm this claim.

In most situations, SLOO can be used for variable selection in the presence of RSA instead of computing all candidate models in a spatially explicit framework. However, even if it has a real advantage in terms of computation time it does not address the problem of correctly modelling the RSA. This can be seen as both an advantage and an inconvenience: an advantage because it avoids the choice between the many spatially explicit methods that are available, which may give different results (see Diniz-Filho *et al*., 2008), but an inconvenience because it prevents us understanding the unknown ecological processes that generated the RSA. Finding and describing spatial patterns of the RSA is often critical. Indeed, it reflects the spatial patterns not accounted for by available variables and is thus highly valuable, especially for predictive purposes when one could use it to interpolate at unsampled locations. SLOO is thus only the first step in the statistical analysis. Once the variables have been selected, it remains to use a spatially explicit framework to correctly model the RSA (e.g. by introducing a spatial term that is orthogonal to the variables considered; see Reich *et al*., 2006; Hughes & Haran, 2013) and make correct inference from the dataset.

ACKNOWLEDGEMENTS

We thank the 'Agripop' team of the CEBC for their helpful comments and in particular Luca Borger for proofreading the manuscript. We also gratefully thank Creagh Breuner for the English revision. Particular thanks are expressed to the associate editor Alberto Jiménez-Valverde and two anonymous referees for their helpful comments. Finally, we thank the 'Région Poitou-Charentes' and the 'Conseil Général des Deux-Sèvres' for funding K.LeR.'s PhD grant.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* (ed. by B.N. Petrov and F. Csaki), pp. 267–281. Akademiai Kiado, Budapest.

Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Altman, N.S. (1990) Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749–759.

Arlot, S. & Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.

Bahn, V. & McGill, B.J. (2013) Testing the predictive performance of predictive models. *Oikos*, **122**, 321–331.

Bahn, V., O'Connor, R.J. & Krohn, W.B. (2006) Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography*, **29**, 835–844.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246–264.

Betts, M.G., Ganio, L.M., Huso, M.P., Som, N.A., Huettmann, F., Bowman, J. & Wintle, B.A. (2009) Comment on 'Methods to account for spatial autocorrelation in the analysis of species distributional data: a review'. *Ecography*, **32**, 374–378.

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B. *et al*. (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography*, **32**, 193–204.

Brenning, A. (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, **5**, 853–862.

Burman, P., Chow, E. & Nolan, D. (1994) A cross-validatory method for dependent data. *Biometrika*, **81**, 351–358.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer-Verlag, New York.

Cassemiro, F.A.S., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B. & Bini, L.M. (2007) Spatial autocorrelation, model selection and hypothesis testing in geographical ecology: implications for testing metabolic theory in New World amphibians. *Neotropical Biology and Conservation*, **2**, 119–126.

Chu, C.-K. & Marron, J.S. (1991) Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics*, **19**, 1906–1918.

Chung, C.J.F. & Fabbri, A.G. (2003) Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards*, **30**, 451–472.

Cordy, C.B. & Griffith, D.A. (1993) Efficiency of least squares estimators in the presence of spatial autocorrelation. *Communications in Statistics – Simulation and Computation*, **22**, 1161–1179.

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

Diniz-Filho, J.A.F., Rangel, T.F.L.V. & Bini, L.M. (2008) Model selection and information theory in geographical ecology. *Global Ecology and Biogeography*, **17**, 479–488.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, 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.

Dungan, J.L., Perry, J.N., Dale, M.R.T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M. & Rosenberg, M. (2002) A balanced view of scale in spatial statistical analysis. *Ecography*, **25**, 626–640.

Fortin, M.-J. & Dale, M.R.T. (2005) *Spatial analysis: a guide for ecologists*. Cambridge University Press, Cambridge.

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.

George, E.I. (2000) The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Griffith, D.A. & Peres-Neto, P.R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, **87**, 2603–2613.

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., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hodges, J.S. & Reich, B.J. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. *American Statistician*, **64**, 325–334.

Hoeting, J.A., Davis, R.A., Merton, A.A. & Thompson, S.E. (2006) Model selection for geostatistical models. *Ecological Applications*, **16**, 87–98.

Hughes, J. & Haran, M. (2013) Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 139–159.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25**, 616–625.

Knafl, G.J. & Grey, M. (2007) Factor analysis model evaluation through likelihood cross-validation. *Statistical Methods for Medical Research*, **16**, 77–102.

Koenig, W.D. (1999) Spatial autocorrelation of ecological phenomena. *Trends in Ecology and Evolution*, **14**, 22–26.

Kühn, I., Nobis, M.P. & Durka, W. (2009) Combining spatial and phylogenetic eigenvector filtering in trait analysis. *Global Ecology and Biogeography*, **18**, 745–758.

Le Rest, K., Pinaud, D. & Bretagnolle, V. (2013) Accounting for spatial autocorrelation from model selection to statistical inference: application to a national survey of a diurnal raptor. *Ecological Informatics*, **14**, 17–24.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm. *Ecology*, **74**, 1659–1673.

Legendre, P. & Fortin, M.-J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–138.

Lennon, J.J. (2000) Red-shifts and red herrings in geographical ecology. *Ecography*, **23**, 101–113.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Liebhold, A.M. & Sharov, A.A. (1998) Testing for correlation in the presence of spatial autocorrelation in insect count data. *Population and community ecology for insect management and conservation* (ed. by J. Baumgartner, P. Brandmayr and B.F.J. Manly), pp. 11–117. Balkema, Rotterdam.

Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Paciorek, C.J. (2010) The importance of scale for spatial confounding bias and precision of spatial regression estimators. *Statistical Science*, **25**, 107–125.

Pinkerton, M.H., Smith, A.N.H., Raymond, B., Hosie, G.W., Sharp, B., Leathwick, J.R. & Bradford-Grieve, J.M. (2010) Spatial and seasonal distribution of adult *Oithona similis* in the Southern Ocean: predictions using boosted regression trees. *Deep-Sea Research I*, **57**, 469–485.

R Core Team (2013) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ (accessed 25 September 2013).

Reich, B.J., Hodges, J.S. & Zadnik, V. (2006) Effect of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**, 1197–1206.

Rossi, R.E., Mulla, D.J., Journel, A.G. & Franz, E.H. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, **62**, 277–314.

Russ, G. & Brenning, A. (2010) Data mining in precision agriculture: management of spatial information. *Computational intelligence for knowledge-based systems design* (ed. by E. Hüllermeier, R. Kruse and F. Hoffmann), pp. 350–359. Springer-Verlag, Berlin.

Saas, Y. & Gosselin, F. (2014) Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. *Ecography*, doi: 10.1111/j.1600-0587.2013.00279.x.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shao, J. (1997) An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221–264.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**, 111–147.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 44–47.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**, 276–278.

Thiollay, J.M. & Bretagnolle, V. (2004) *Rapaces nicheurs de France: distribution, effectifs et conservation*. Delachaux et Niestlé, Paris.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Appendix S1 Computing SLOO with R.
Appendix S2 Extending simulations for different amount of data.

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix S3 Extending simulations for a higher number of variables.

BIOSKETCH

Kévin Le Rest is interested in the application of statistics in all fields of ecology. He has just finished his PhD on statistical methods for modelling the distribution and the abundance of populations at large scales. Its topic mainly concerned the evaluation and the development of methods dealing with spatial autocorrelation and overdispersion in spatial count data analysis. A particular emphasis concerned the variable selection step.

Editor: Alberto Jiménez-Valverde

14668238, 2014, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/geb.12161 by King Abdullah University Of Science & Technology, Wiley Online Library on [30/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License