Ecological Informatics 14 (2013) 17–24

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

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

Accounting for spatial autocorrelation from model selection to statistical inference: Application to a national survey of a diurnal raptor

Kévin Le Rest *, David Pinaud, Vincent Bretagnolle

Centre d'Etudes Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France

ARTICLE INFO

Article history:
Received 26 January 2012
Accepted 30 November 2012
Available online 12 December 2012

Keywords:
Generalized Linear Models
Spatial cross-validation
Population size
Residual spatial autocorrelation
Spatial filtering
Species distribution

ABSTRACT

Planning actions for species conservation involves working at both an ecologically meaningful spatial scale and a scale suitable for implementing management or conservation plans. Animal populations and conservation policies often operate across wide areas. Large-extent spatial datasets are thus often used, but their analyses rarely deal with problems inherent to spatial datasets such as residual spatial autocorrelation, which can bias or even reverse results. Here we propose a procedure for analysing a large-scale count dataset integrating residual spatial autocorrelation in a Generalized Linear Model framework by combining and extending previously published methods. The first step concerns the selection of the environmental variables by a modified cross-validation procedure allowing for residual spatial autocorrelation. Then the second step consists in evaluating the spatial effect of the model using a spatial filtering approach based on the variogram parameters. We apply this method to the Black kite (*Milvus migrans*) to estimate the distribution and population size of this species in France. We found some divergence in estimated population size between spatial and non spatial models, as well as in the distribution map. We also found that the uncertainty of the model was underestimated by the residual spatial autocorrelation. Our analysis confirms previous results, that residual spatial autocorrelation should be always accounted for, especially in conservation where false results may lead to poor management decisions.

1. Introduction

Animal populations and conservation policies often operate across wide areas. Large-extent spatial datasets (Scheiner et al., 2000) can therefore be extremely valuable to determine population parameters for conservation purposes, e.g. the geographical distribution of species, its population size or trends. However, the statistical analyses used often ignore issues that may bias conclusions. In particular, they rarely deal with inference problems inherent from spatial datasets such as residual spatial autocorrelation (hereafter RSA), which may actually reverse observed patterns (Kühn, 2007).

Spatial autocorrelation arises when the measure of a variable of interest in multiple sample units are not independent of each other (Griffith, 1987), which often occurs in ecological data. Such spatial patterns are usually explained by environmental features (e.g. climatic variables or habitat structure) that are themselves spatially structured. Therefore, including all environmental variables that are spatially structured may be sufficient to remove RSA of a regression model (Diniz-Filho et al., 2003). However, it is often impossible to measure all spatially structured variables: for instance, variables accounting for social behaviour or for the availability of food resources, are very difficult to measure and often miss in the dataset. In such cases, the inclusion of all available variables does not fully remove RSA and the important assumption of independence of residuals is violated (see Dormann et al., 2007). It is well known that this problem mostly affects the uncertainty of statistical models (Legendre, 1993; Legendre et al., 2002), i.e. the confidence interval around the regression coefficients, which is commonly measured by the standard error. A positive RSA, i.e. closer locations having more similar residual values than others, tends to underestimate the true standard errors of parameters, which lead to an over-precise estimation of the regression coefficients. In turn this can lead to an erroneously low p-value, wrong R^2 and wrong likelihood (Legendre, 1993; Legendre et al., 2002; Lennon, 2000).

RSA raises two main concerns. The first relates to model selection, since classical criterion such as the Akaïke information criterion (hereafter AIC) are biased in the presence of RSA (see Cassemiro et al., 2007; Diniz-Filho et al., 2008; Hoeting et al., 2006). The most common strategy

Abbreviations: AIC, Akaïke Information Criterion; GLM, Generalized Linear Model; PCA, Principal Component Analysis; RMSEP, Root Mean Squared Error of Prediction; RSA, Residual Spatial Autocorrelation.

* Corresponding author. Tel.: +33 5 49 09 35 13; fax: +33 5cbu 49 09 65 26.
E-mail addresses: lerest.k@gmail.com (K. Le Rest), pinaud@cebc.cnrs.fr (D. Pinaud), breta@cebc.cnrs.fr (V. Bretagnolle).

1574-9541/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecoinf.2012.11.008

to overcome this problem involves correcting first the RSA by considering a spatially explicit model and then, using a classical criterion such as AIC. However, accounting for RSA for all biologically pertinent candidate models can be extremely time consuming, especially if the number of candidate models is high (see Craig et al., 2007). As a consequence, AIC is often used without accounting for RSA (see for example Kühn et al., 2009). Kissling and Carl (2008) proposed several strategies to choose the spatial structure that should be added to the model in order to correct for RSA, but they did not provide solutions for the selection of variables. The second concern relates to the model estimation since model parameters are not estimated correctly (Dormann, 2007; Keitt et al., 2002; Kühn, 2007). To overcome this problem, some tools were made available for Generalized Linear Models (hereafter GLMs) (see Carl and Kühn, 2010; Dormann et al., 2007). Among these, the spatial filtering techniques are recognized as one of the most efficient, both practically and theoretically (Diniz-Filho et al., 2009; Dormann et al., 2007). Spatial filtering consists in using a weighted distance matrix to address the issue of RSA, by adding several spatial filters (eigenvectors) to a GLM (see Diniz-Filho and Bini, 2005; Dray et al., 2006; Getis and Griffith, 2002; Griffith, 2000). However, there is evidence that the choice of the weight matrix highly influences the set of spatial filters and thus the model (Patuelli et al., 2006). In addition, although there are several possibilities for defining the weight matrix (see Getis and Aldstadt, 2004; Tiefelsdorf et al., 1999), it remains mainly based on basic functions of the distance (binary, linear, quadratic) which may not always satisfy the complexity of the residual spatial structure underlined in the ecological processes.

In this paper, our aim is to provide a guideline for analysing spatial datasets integrating RSA within a GLM, by extending different methods within the same framework. As a first step, we deal with model selection, by using a cross-validation approach. In order to overcome the problem of RSA in the selection step, we use a threshold distance between the training and the validation sets to ensure that they are fully independent. The second step consists in accounting for the RSA of the selected model. We use a spatial filtering technique, where the weighted matrix has been modified in order to directly use the shape of the variogram to calculate the eigenvectors. We then apply this approach on a real case study and compare results of the spatial and non spatial models. As a practical example, we used a French national dataset collated for the Black kite (*Milvus migrans*), a diurnal raptor. A particular emphasis was given to the estimation of species distribution and its population size, which are major issues in management and conservation plans.

2. Material and methods

2.1. Survey and datasets

A national survey aiming to estimate the distribution and population size of all diurnal raptors was undertaken between 2000 and 2002, with around 1600 volunteers. For this study, we used a subset of the available data, consisting in 683 sampling units in France (see Fig. 1) with known searching effort. Sampling protocol consisted in counting the number of breeding pairs of diurnal raptors on 25-km^2 quadrats (5×5 km; see Thiollay and Bretagnolle, 2004 for details). The time spent on each quadrat was recorded by observers. Each quadrat was also described using environmental variables from a climatic dataset (Hijmans et al., 2005, Bioclim, www.worldclim.org/bioclim) and a land cover dataset (CLC: Corine Land Cover, www.eea.europa.eu). The climatic dataset consisted in 19 variables measured between 1960 and 1990, which provided robust estimates of measures such as average temperature, rainfall, temperature variation and rainfall variation at a resolution of approximately 1-km. The land cover dataset had 44 variables giving land use in 2000 on a 1-hectar cell. From these 44 classes, 9 habitat hyper-classes were built from a functional (ecological) point of view for raptors (see Table A1 in Appendix A). The percentage of coverage per 25-km^2 quadrat was calculated for each of these habitat hyper-classes. High correlations occurred between several environmental variables, which cause matrix inversion problems (null determinant). In order to overcome multicollinearity, a Principal Component Analysis (hereafter PCA) was performed separately on each dataset (climate and land use) and principal components were used as environmental variables. The label "ClimDim.x" was used to nominate the xst principal component from the climate dataset and the label "ClcDim.x" was used in the same way for the land cover dataset.

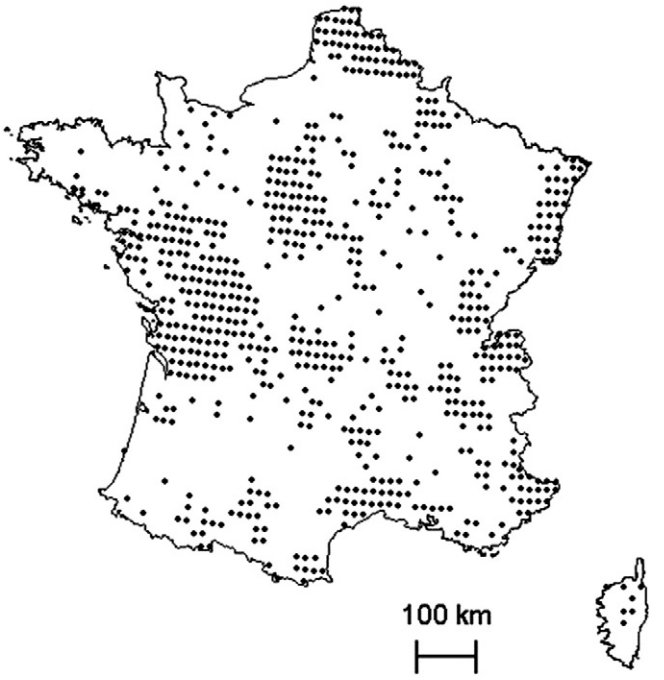

Fig. 1. Map of the 683 locations (25-km^2 quadrats) used for analyses. Each location is represented by a black point.

2.2. Model selection by spatial cross-validation

Model selection consisted in a comparison of candidate models in order to select which predicted best the observed data. As the number of environmental variables k was high (19 climatic and 9 habitat variables), the number of candidate model became oversized (2^k). A stepwise procedure was used to reduce computation time (Efroymson, 1960; Hocking, 1976). The stepwise process was implemented in two steps: first, environmental variables with linear effects were selected and then, quadratic terms and interactions. A Poisson distribution was assumed for the number of breeding pairs per quadrat, considering that there was no additional overdispersion, other than that due to RSA (see Griffith and Haining, 2006; Haining et al., 2009 for details about the relationship between overdispersion and RSA). The time spent per quadrat was included as an offset.

The error of prediction was considered as a selection criterion because the aim of this model was to predict at unsampled points. Error of prediction was calculated by cross-validation (Allen, 1974; Geisser, 1975; Stone, 1974), a widely used technique for model selection and model validation involving many different splittings (see Arlot and Celisse, 2010 for a recent overview of the cross-validation procedures for model selection). Here, leave-one-out cross-validation was used, consisting in deleting one observation (the validation set) and use all the others as training dataset, i.e. to estimate model parameters.

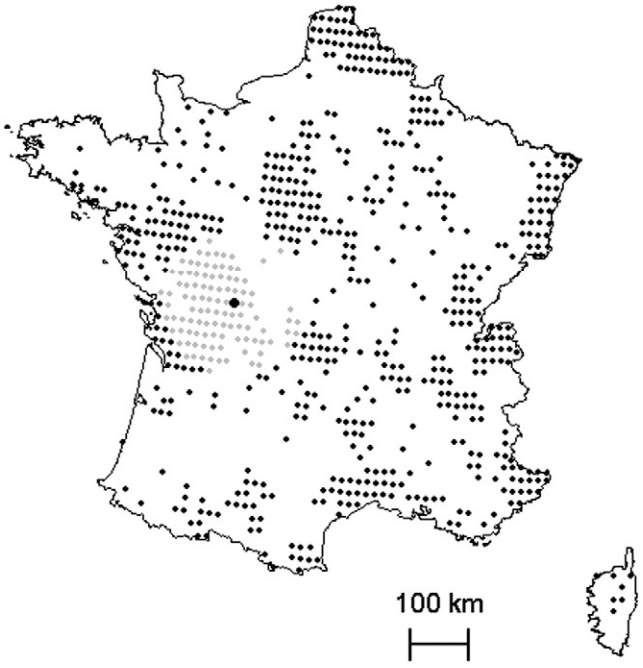

Fig. 2. One example of the modified leave-one-out cross-validation applied in a spatial context with a threshold distance of 125 km. The heavy black point is the point which was left out of the model and where the error of prediction was calculated. Grey points were also excluded due to the residual spatial autocorrelation. Others black points were used in the training data set. This procedure was used iteratively for each observation in order to calculate an overall prediction error.

An overall prediction error can then be calculated using the Root Mean Square Error of Prediction (hereafter RMSEP), see Eq. (1):

$$\mathrm{RMSEP} = \sqrt{E\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]}. \quad (1)$$

In Eq. (1), n is the sample size (number of quadrats), y_i is the deleted observation (the validation set) and $\hat{y}_i$ is the predicted abundance at this location using parameters estimated from all the others (the training set). When using cross-validation, a critical prerequisite is that the training set and the validation set are independent, thus dependent model residuals may bias the error of prediction (Altman, 1990). Several possible alternatives have been proposed to correct the cross-validation procedures (in context of nonparametric regression, see Chu and Marron, 1991; Burman et al., 1994). We extended the "Modified Cross Validation" (Chu and Marron, 1991) to a spatial context by using a threshold distance between the validation and the training dataset, guaranteeing these datasets to be spatially independent. This threshold was chosen as the value of the range of the variogram on the residuals from the model including all covariates (125 km in our case). This represented the spatial autocorrelation that could not be accounted for by our environmental variables (see Fig. 2). Deviance residuals were chosen because Pearson residuals had some extreme values, which could affect the variogram quality (Cressie and Hawkins, 1980). The selected model was labeled non-spatial model because it did not incorporate an explicit spatial component, unlike the model below.

2.3. Accounting for residual spatial autocorrelation

The spatial structure of residuals can be easily evaluated using a correlogram or a variogram, both based on a measure of the covariance between observations according to the distance between them. A variogram was thus estimated using the residuals of the previously selected model, i.e. the non-spatial model (see Fig. 3a). We then used an approach based on spatial filters (see Diniz-Filho and Bini, 2005; Dray et al., 2006; Getis and Griffith, 2002; Griffith, 2000 and see also Griffith, 2002, 2006, for developments with Poisson regressions, i.e. count data) with a modification of the weight matrix. The weight matrix W was defined through the shape of the variogram (as tested in Getis and Aldstadt, 2004) constructed from the deviance residuals of the non-spatial model (see Fig. 3a). As in Getis and Griffith (2002), the diagonal of the matrix W is composed of zeros, which only enables the estimation of the relationship between observations (see also Dray et al., 2006). Other values matched with the Eq. (2):

$$f(d) = 1 - \left[\frac{\gamma(sill, range, d)}{sill}\right]. \quad (2)$$

In Eq. (2), *sill* and *range* are parameters of the variogram, d is the distance between observations and γ is the exponential variogram function. This equation could be interpreted as the degree of connectivity between two observations and was defined between 0, i.e. no connectivity and 1, i.e. maximal connectivity. The nugget effect did not appear in this equation because we expected that $f(d) = 1$ when $d = 0$.

Eigenvectors were then extracted from the $(I\text{-}11'/n)\ W\ (I\text{-}11'/n)$ matrix transformation (see Getis and Griffith, 2002), where n is the

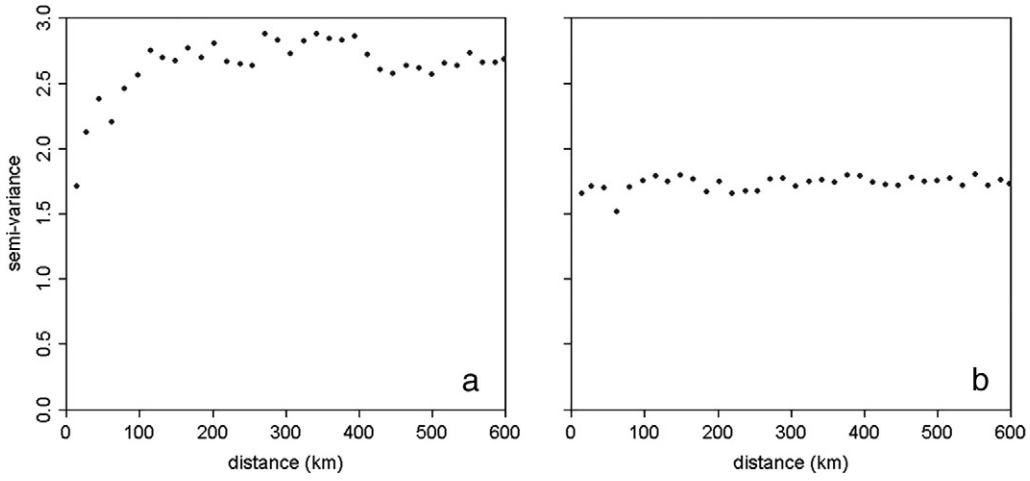

Fig. 3. Variogram of the deviance residuals of the non-spatial model (a) and the spatial model (b).

number of observations, I is the n by n identity matrix, 1 is an n by 1 vector of ones and W is the weight matrix. The Moran's I was also calculated for each eigenvector and only those having positive values were retained. This rule led to selecting eigenvectors having only positive spatial autocorrelation (Griffith, 2003). Kissling and Carl (2008) recommended that the selection of the spatial term be based on a metric of RSA as well as a metric of fit. Therefore, the set of candidate eigenvectors were included linearly in the selected GLM by two stepwise procedures. The first one minimized the RSA, i.e. the sill of the variogram. Relevant eigenvectors were added until sill = 0, which indicated a totally "flat" theoretical variogram. Second, unnecessary eigenvectors were removed by minimizing the AIC. This model was labeled spatial model.

2.4. Distribution and population size

The non spatial and the spatial model were compared with regard to their ability to predict the distribution and the population size of Black kites. Predictions were made over a grid of France constituted by 22,500 cells of 25 km^2 using a partial regression (see Legendre and Legendre, 1998) on environmental variables (climatic and habitat) excluding eigenvectors. Thus the non spatial and the spatial model relied, for prediction, on the same environmental variables. However in the non spatial model, RSA was not taken into account whereas it was in the spatial model. Population size and its confidence interval were calculated by running 10,000 Monte Carlo simulations of the sum of abundance predicted at the 22,500 cells, i.e., the total abundance expected. At each simulation, the value of each parameter was set randomly, by considering a normal distribution with the mean equal to the regression coefficient, and the standard deviation equal to the standard error. This process allowed the full range of model parameters to be considered.

All analyses were performed using the R software version 2.13.0 (R Development Core Team, 2011).

3. Results

The value of RMSEP for the null model, i.e. the model only including a constant and the offset, was 3.30. The value of RMSEP for the selected model (Table 1) was 3.00, with a total of 11 parameters associated to environmental features (7 linear terms, 3 quadratics terms and 1 interaction). Thus selected environmental variables reduced the RMSEP by about 9%. The RSA was fully removed by adding 42 eigenvectors to the non-spatial model (see Fig. 3). As a consequence, the estimation of the model parameters from the non-spatial and the spatial model was not the same (see Table 1).

Distribution maps built from the non spatial and spatial model were also different, with a lower predicted abundance of Black kites in western of France using the spatial model compared to the non-spatial model (see Fig. 4). Estimation of population size was also different using the non-spatial and the spatial model (see Fig. 5). The average population size prediction using the non-spatial model was 36,122 (95% confidence interval 28,780–45,683) breeding pairs of Black kite in France, whereas using the spatial model it was 32,133 pairs (21,426–47,072).

4. Discussion

Differences in coefficient estimation were found between the spatial or the non-spatial model (see Table 1), with some coefficients even changing sign (see Clim.Dim2 and ClimDim.6). The latter reflects a possible inverse fit of the data when the RSA was not accounted for, a pattern previously observed (see Kühn, 2007). Other coefficients showed moderate to rather large reduction of their statistical effect (see coefficients and standard errors in Table 1).

The selection of eigenvectors by a two step procedure minimizing both RSA and AIC (a strategy first suggested by Kissling and Carl, 2008) seems promising. The first step (minimizing the RSA), led to select 49 eigenvectors, while the second step (minimizing AIC) allowed the removal of 7 eigenvectors without impacting on the RSA. We also recommend checking for collinearity between the selected environmental variables and the selected eigenvectors, which could artificially increase the standard error of the regression parameters (see Freckleton, 2002). Here the correlation matrix between the environmental variables showed only a slight correlation (correlation coefficient never above 0.5, see Table C1 in the Appendix C). These correlations were lower than the critical value of 0.7 proposed by Dormann et al. (2012).

The PCA approach was used to overcome a problem of multicollinearity between the environmental variables. However, there is a cost in using PCA, since it raises some difficulties in understanding model coefficients from an ecological point of view. In order to check that selected PCA axes are biologically relevant, one must interpret these axes. In our case, the axe ClcDim.1 represented mainly a natural gradient (see Fig. B1 in Appendix B) where positive values indicated a high percentage of forest in the quadrat and negative values indicated a high percentage of intensive farming (i.e. no forest). The ClcDim.1 effect in the model resulted from a linear effect (−0.32, see Table 1) as well as a quadratic effect (−0.46, see Table 1), the latter indicating that this effect was stronger in quadrats dominated by forest. ClcDim.4 mostly represented wetlands (see Fig. B1 in Appendix B) where positive values

Table 1
The coefficients (Coef), standard errors (StdE), t-value (t) and p-value (p) for the non spatial and spatial models.

	Non-spatial model				Spatial model			
Label	Coef	StdE	t	p	Coef	StdE	t	p
Intercept	−3.48	0.06	−54.66	$<1.10^{-324}$	−4.10	0.12	−34.95	$<1.10^{-267}$
ClimDim.2	0.12	0.05	2.59	0.010	−0.68	0.11	−6.35	$<1.10^{-09}$
ClimDim.3	1.24	0.08	14.84	$<1.10^{-50}$	1.91	0.13	14.70	$<1.10^{-49}$
ClimDim.3^2	−0.69	0.07	−10.21	$<1.10^{-24}$	−0.67	0.11	−6.17	$<1.10^{-09}$
ClimDim.6	0.05	0.04	1.45	0.148	−0.26	0.04	−6.18	$<1.10^{-09}$
ClimDim.6^2	0.12	0.03	4.43	$<1.10^{-05}$	−0.06	0.04	−1.60	0.111
ClimDim.11	0.31	0.04	7.53	$<1.10^{-13}$	0.02	0.06	0.38	0.705
ClcDim.1	−0.48	0.05	−10.05	$<1.10^{-24}$	−0.32	0.06	−5.16	$<1.10^{-06}$
ClcDim.1^2	−0.59	0.05	−11.58	$<1.10^{-31}$	−0.46	0.06	−7.37	$<1.10^{-12}$
ClcDim.4	0.19	0.05	4.05	$<1.10^{-04}$	0.24	0.06	4.13	$<1.10^{-04}$
ClcDim.6	0.15	0.04	4.16	$<1.10^{-04}$	0.11	0.04	3.04	0.002
ClcDim.1:ClimDim.6	−0.15	0.04	−3.62	$<1.10^{-03}$	−0.27	0.05	−5.34	$<1.10^{-07}$

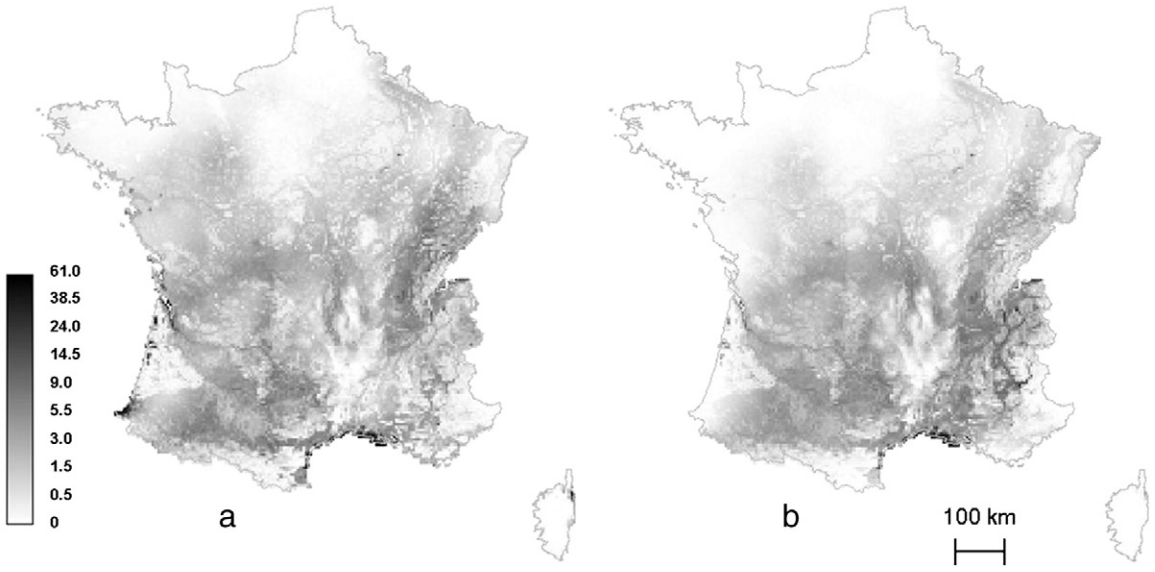

Fig. 4. Predicted distribution of the Black kite (in number of pairs per 25 km^2) using the non-spatial model (a) and the spatial model (b).

indicated a high percentage of wetlands. The coefficient of ClcDim.4 was 0.24 (see Table 1), underlying a strong preference of the Black kite for wetland habitats. These two aspects of the landscape fit well with previous knowledge for this species: Black kites avoid large forests, and prefer anthropic environments (agricultural lands) and wetlands (Thiollay and Bretagnolle, 2004). Therefore, despite interpretation problems, we believe that in our case, the advantages to use PCA outweighed the disadvantages, since our aim was first to provide unbiased parameter estimation in order to predict population distribution and size.

The selection of these PCA axes was done using a leave-one-out cross-validation, which is known to be asymptotically equivalent to AIC (Stone, 1977), but allows dealing with RSA using a threshold distance between the subsets. However, as cross-validation criterion, we used the RMSEP as it is used in linear models, i.e. without accounting for the fact that we used a Poisson distribution. This means that there is equal penalization between an observed abundance of 0 and a predicted one of 1, and an observed abundance of 50 and a predicted one of 51, whereas mean equal variance on a Poisson distribution, i.e. high count have more variation. Thus high errors of prediction have more probability to occur on high counts. A better or refined RMSEP should be found, for example using a logarithmic scale, which would select better predictors than those found here.

Moreover, we did not account for overdispersion in our analysis, whereas the analysis of count data often presents overdispersion. This choice was made because there is a strong link between RSA and overdispersion since RSA generates overdispersion (see Griffith and Haining, 2006; Haining et al., 2009). Here we were interested into the effects of RSA, and therefore we have fixed overdispersion, considering that there was no additional overdispersion than the one due to RSA. We have however checked the overdispersion in the final spatial model, which turned to be about 3.32 units using the Pearson Chi Square statistic (the Poisson distribution fixes it to 1). For comparison, the non spatial model had an overdispersion of 9.99 units, which clearly shows that accounting for RSA also reduces overdispersion. However, since some overdispersion remains in the spatial model, further improvement seems necessary, e.g. by considering a Quasi-Poisson distribution.

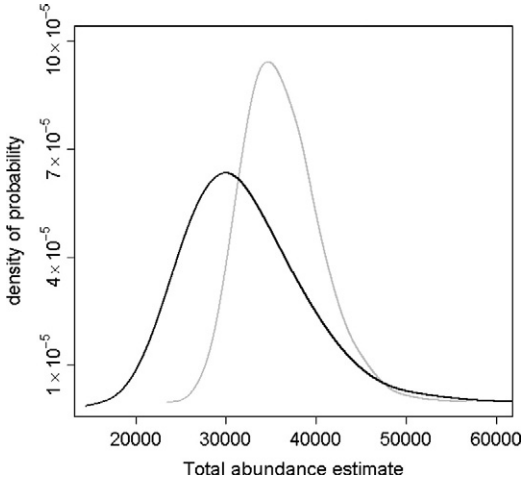

Fig. 5. Predicted population size and confidence interval using the non-spatial model (in grey) and the spatial model (in black). The two curves are obtained by dividing the 10,000 Monte Carlo simulations in 100 breaks points and calculating the density of probability for each break point. A cubic smoothing spline was then used (with 15 equivalent degree of freedom) in order to make the figure clearer.

5. Conclusions

Predicted distribution and population size of the Black kite between the non-spatial and the spatial model were similar, but there were also substantial differences. The spatial model predicted lower abundance in western France compared to the non spatial model (see Fig. 4), and hence a lower population size than the spatial model (see Fig. 5). The model uncertainty was also larger for the spatial model than for the non-spatial model (see Table 1), which was expected, but may strongly impact model predictions (see Fig. 5). Thus in addition to give misleading distribution maps of species, RSA also gave a false feeling of precise predictions, which a priori may suggest that this model shows a better fit of the data. In terms of conservation applications, a poorly predicted map of abundance may have serious consequences: in our case, not accounting for RSA would lead to the interpretation that western and eastern of France are

equal important and suitable breeding habitat for Black kites, whereas in fact eastern France is the main area of breeding for this species.

Acknowledgements

Particular thanks are deserved to an anonymous referee for his very helpful comments (including missed references) and suggestions which greatly improved the paper. We also acknowledge the Editor of this proceeding (Dr Mark O'Connell) for his additional comments. We also thank all voluntaries who have carried out the field survey. We particularly thank Jean Sériot who coordinated the national survey "Rapaces nicheurs de France", and Fabienne David who is now coordinating the French monitoring raptor scheme. We thank Arzhela Hemery and Maxime Passerault who have participated at the preparation of the dataset used here and particularly for extraction of satellite data. We thank people from "Biostatistics and spatial processes" team (INRA Avignon), and particularly Pascal Monestiez, Joël Chadoeuf and Rachid Senoussi, for their advices and comments about spatial statistics. We also thank Patrick Duncan and Samantha Patrick for the English revision and their very interesting comments. Finally we thank the *Région Poitou-Charentes* and the *Département des Deux-Sèvres* for funding PhD grant.

Appendix A

Table A1
The nine habitat hyper-classes used in our analyses and the Corine Land Cover initial classification. One row corresponds to one initial Corine Land Cover class, which is defined by three distinct labels.

44 corine land cover nomenclatures			9 habitat hyper-classes
Label 1	Label 2	Label 3	
Artificial surfaces	Urban fabric	Continuous urban fabric	Anthropic areas
Artificial surfaces	Urban fabric	Discontinuous urban fabric	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Industrial or commercial units	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Road and rail networks and associated land	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Port areas	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Airports	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Mineral extraction sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Dump sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Construction sites	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Green urban areas	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Sport and leisure facilities	Anthropic areas
Agricultural areas	Arable land	Non-irrigated arable land	Intensive agriculture
Agricultural areas	Arable land	Permanently irrigated land	Intensive agriculture
Agricultural areas	Arable land	Rice fields	Intensive agriculture
Agricultural areas	Permanent crops	Vineyards	Permanent agriculture
Agricultural areas	Permanent crops	Fruit trees and berry plantations	Permanent agriculture
Agricultural areas	Permanent crops	Olive groves	Permanent agriculture
Agricultural areas	Pastures	Pastures	Extensive farming
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Natural grasslands	Extensive farming
Agricultural areas	Heterogeneous agricultural areas	Annual crops associated with permanent crops	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Complex cultivation patterns	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Agro-forestry areas	Forest areas
Forest and semi-natural areas	Forests	Broad-leaved forests	Forest areas
Forest and semi-natural areas	Forests	Coniferous forest	Forest areas
Forest and semi-natural areas	Forests	Mixed forest	Forest areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Moors and heathland	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Sclerophyllous vegetation	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Transitional woodland-shrub	Transitional areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Beaches, dunes, sands	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Bare rocks	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Sparsely vegetated areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Burnt areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Glaciers and perpetual snow	Open areas
Wetlands	Inland wetlands	Inland marshes	Wetlands
Wetlands	Inland wetlands	Peat bogs	Wetlands
Wetlands	Maritime wetlands	Salt marshes	Wetlands
Wetlands	Maritime wetlands	Salines	Wetlands
Wetlands	Maritime wetlands	Intertidal flats	Wetlands
Water bodies	Inland waters	Water courses	Wetlands
Water bodies	Inland waters	Water bodies	Wetlands
Water bodies	Maritime waters	Coastal lagoons	Wetlands
Water bodies	Maritime waters	Estuaries	Wetlands
Water bodies	Maritime waters	Sea and oceans	Not used

Appendix B

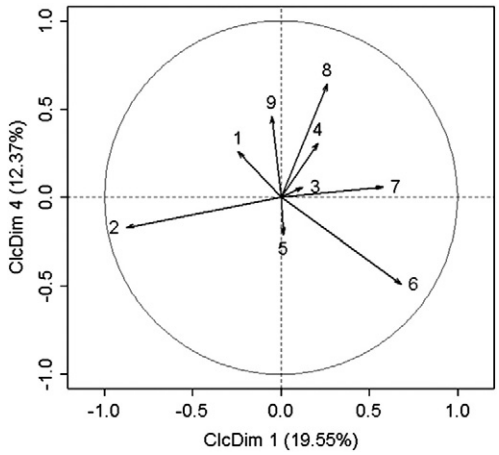

Fig. B1. Correlation circle of ClcDim.1 and ClcDim.4 principal components. 1: Anthropic areas. 2: Intensive agriculture. 3: Permanent agriculture. 4: Extensive farming. 5: Heterogeneous agriculture. 6: Forest areas. 7: Transitional areas. 8: Open areas. 9: Wetlands.

Appendix C

Table C1
Correlation matrix between environmental variables (columns) and eigenvectors (rows).

Eigenvectors	ClimDim.2	ClimDim.3	ClimDim.6	ClimDim.11	ClcDim.1	ClcDim.4	ClcDim.6
E1	0.50	0.27	−0.14	−0.14	−0.15	−0.03	0.10
E3	0.18	−0.26	−0.06	0.17	−0.04	0.18	−0.05
E6	0.30	−0.08	−0.02	−0.04	−0.01	0.03	0.09
E7	−0.07	0.04	−0.21	−0.12	0.11	−0.01	0.03
E11	0.14	−0.01	−0.11	0.03	0.09	0.01	−0.02
E13	0.12	−0.15	−0.02	−0.10	0.00	0.05	0.00
E15	−0.09	0.17	−0.14	−0.10	0.09	−0.12	−0.04
E17	0.05	−0.08	−0.10	−0.09	−0.04	0.02	0.09
E23	0.01	0.05	0.19	−0.11	−0.04	0.01	−0.01
E27	0.13	−0.02	0.01	0.11	0.11	0.08	0.09
E30	−0.14	−0.01	−0.19	−0.02	0.03	0.05	−0.02
E32	0.04	0.06	−0.13	0.06	−0.07	0.07	−0.03
E33	−0.07	−0.04	−0.01	0.01	−0.04	0.07	0.04
E34	0.05	−0.09	−0.01	0.11	0.01	0.02	0.01
E35	−0.03	0.06	−0.09	−0.06	−0.06	−0.05	0.00
E37	0.18	−0.05	−0.03	−0.10	−0.01	0.09	−0.01
E40	0.07	0.08	0.03	0.00	0.08	0.03	0.07
E43	0.09	−0.01	−0.09	0.09	0.06	0.07	0.01
E56	−0.06	0.03	−0.02	0.04	−0.04	0.01	−0.02
E58	−0.07	0.06	0.03	−0.09	0.04	−0.18	0.01
E65	−0.03	0.02	0.04	0.05	−0.08	−0.08	0.03
E70	0.01	−0.02	−0.03	−0.05	−0.06	0.07	−0.04
E78	0.02	−0.03	−0.02	0.00	0.01	0.00	0.01
E82	−0.01	−0.02	−0.06	−0.01	0.02	0.06	0.04
E87	0.01	0.03	−0.02	0.00	−0.04	−0.05	0.05
E90	−0.01	0.00	0.00	0.02	0.04	−0.01	0.03
E93	0.01	0.01	−0.03	−0.04	0.01	−0.02	−0.02
E96	0.04	0.00	0.01	0.00	0.01	0.03	−0.05
E103	0.02	0.00	0.03	0.00	0.06	−0.01	−0.02
E107	0.00	0.00	0.01	0.01	−0.03	0.03	−0.02
E108	0.00	0.03	0.02	0.00	−0.04	−0.02	0.01
E111	0.02	0.01	−0.01	−0.10	0.03	0.01	0.03
E112	-0.01	0.00	0.00	−0.02	0.01	−0.03	0.06
E115	0.00	0.01	−0.01	0.00	−0.01	0.01	0.02
E121	0.02	0.02	0.01	−0.02	0.06	0.02	−0.06
E123	0.00	0.02	0.02	−0.03	0.01	−0.05	−0.01
E129	−0.02	0.00	−0.01	−0.04	−0.03	−0.04	−0.03
E134	−0.01	−0.01	−0.02	0.03	0.01	0.00	−0.05
E135	0.02	0.01	0.00	−0.07	0.02	−0.01	0.00
E141	0.02	0.02	0.02	−0.01	0.03	−0.08	−0.01
E142	0.03	−0.01	0.03	0.00	−0.01	0.01	0.04
E143	0.05	0.03	0.04	−0.05	−0.02	−0.01	0.03

References

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 Association 85, 749–759.

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

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

Carl, G., Kühn, I., 2010. A wavelet-based extension of Generalized Linear Models to remove the effect of spatial autocorrelation. Geographical Analysis 42, 323–337.

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. The Annals of Statistics 19, 1906–1918.

Craig, M.H., Sharp, B.L., Mabaso, M.L.H., Kleinschmidt, I., 2007. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. International Journal of Health Geographics 6, 44.

Cressie, N., Hawkins, D., 1980. Robust estimation of the variogram: I. Mathematical Geology 12, 115–125.

Diniz-Filho, J.A.F., Bini, L.M., 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography 14, 177–185.

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.B., Bini, L.M., 2008. Model selection and information theory in geographical ecology. Global Ecology and Biogeography 17, 479–488.

Diniz-Filho, J.A.F., Nabout, J.C., Telles, M.P.C., Soares, T.N., Rangel, T.F.L.V.B., 2009. A review of techniques for spatial modeling in geographical, conservation and landscape genetics. Genetics and Molecular Biology 32 (2), 203–211.

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.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J.R., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Launtenbach, S., 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 35, 001–020.

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196, 483–493.

Efroymson, 1960. Multiple regression analysis. In: Ralston, A., Wilf, H.S. (Eds.), Mathematical Methods for Digital Computers. Wiley.

Freckleton, R.P., 2002. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. Journal of Animal Ecology 71, 542–545.

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

Getis, A., Aldstadt, J., 2004. Constructing the spatial weights matrix using a local statistic. Geographical Analysis 36, 90–104.

Getis, A., Griffith, D.A., 2002. Comparative spatial filtering in regression analysis. Geographical Analysis 34, 130–140.

Griffith, D.A., 1987. Spatial Autocorrelation: A Primer. Association of American Geographers.

Griffith, D.A., 2000. A linear regression solution to the spatial autocorrelation problem. Journal of Geographical Systems 2, 141–156.

Griffith, D.A., 2002. A spatial filtering specification for the auto-Poisson model. Statistics & Probability Letters 58, 245–251.

Griffith, D.A., 2003. Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization. Springer-Verlag, Berlin Heidelberg.

Griffith, D.A., 2006. Assessing spatial dependence in count data: winsorized and spatial filter specification alternatives to the auto-Poisson model. Geographical Analysis 38, 160–179.

Griffith, D.A., Haining, R., 2006. Beyond mule kicks: the Poisson distribution in geographical analysis. Geographical Analysis 38, 123–139.

Haining, R., Law, J., Griffith, D.A., 2009. Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics & Data Analysis 53, 2923–2937.

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.

Hocking, R.R., 1976. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 32, 1–49.

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

Keitt, T.H., Ottar, N., Bjørnstad, O.N., Dixon, P.M., Citron-Pousty, S., 2002. Accounting for spatial pattern when modeling organismenvironment interactions. Ecography 25, 616–625.

Kissling, W.D., Carl, G., 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecology and Biogeography 17, 59–71.

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Diversity and Distributions 13, 66–69.

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

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

Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, 601–615.

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

Patuelli, R., Griffith, D.A., Tiefelsdorf, M., Nijkamp, P., 2006. The use of spatial filtering techniques: the spatial and space–time structure of German unemployment data. Tinbergen Institute Discussion Papers 06-049/3.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/.

Scheiner, S.M., Cox, S.B., Willig, M.R., Mittelbach, G.G., Osenberg, C., Kaspari, M., 2000. Species richness, species-area curves and Simpson's paradox. Evolutionary Ecology Research 2, 791–802.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B (Methodological) 36 (2), 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 (1), 44–47.

Thiollay, J.M., Bretagnolle, V., 2004. Rapaces nicheurs de France: Distribution, Effectifs et Conservation. Delachaux et Niestlé, Paris.

Tiefelsdorf, M., Griffith, D.A., Boots, B., 1999. A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A 31 (1), 165–180.