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 x$^{st}$ principal component from the climate dataset and the label “ClcDim.x” was used in the same way for the land cover dataset.

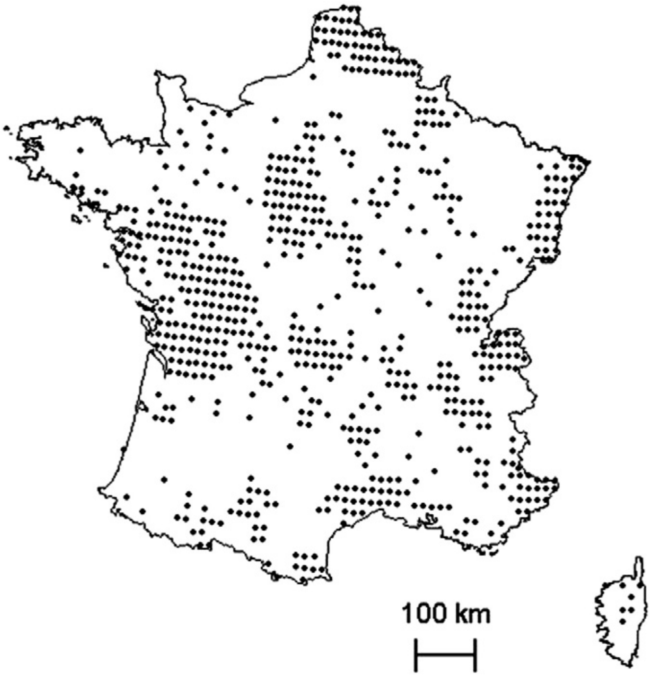

**Fig. 1.** Map of the 683 locations (25-km$^2$ quadrats) used for analyses. Each location is represented by a black point.

### 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.