Ecological Modelling 299 (2015) 51–63

Contents lists available at ScienceDirect

Ecological Modelling

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

Modeling spatial patterns of rare species using eigenfunction-based spatial filters: An example of modified delta model for zero-inflated data

Yan Li [a,b,*], Yan Jiao [a]

[a] Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0321, USA
[b] Department of Veterinary Medicine, Wildlife Health Center, University of California, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 5 June 2014
Received in revised form 3 November 2014
Accepted 5 December 2014

Keywords:
Spatial pattern
Rare species
Eigenfunction
Spatial filter
Delta model
Zero-inflated data

ABSTRACT

Data of rare species usually contain a high percentage of zero observations due to their low abundance. Such data are generally referred as zero-inflated data. Modeling spatial patterns in such data has been challenging, especially when large datasets are involved and intensive computing are required. The eigenfunction-based spatial filtering provides a flexible tool that allows the existing modeling approaches that can handle zero-inflated data such as the delta model to be applied in the presence of spatial dependence. With a real dataset, the longline seabird bycatch data, the present study demonstrated a modification of delta model with the spatial filters to investigate spatial patterns in zero-inflated data for rare species. We explored a total of 108 spatial weighting matrices, and modified the delta model by incorporating the spatial filters generated from the best spatial weighting matrix. We applied the five-fold cross-validation to compare performance of the modified delta model with other three candidate models based on the mean absolute error and the mean bias. The three candidate models included the baseline model without spatial dependence considered, the trend-surface generalized additive model and the random areal effect model. The delta model modified with spatial filters showed superior performance over the other three candidate models in the seabird bycatch example. With the seabird bycatch example, we illustrated a modification of delta model with the eigenfunction-based spatial filters to investigate spatial patterns. This study provides an alternative to incorporate spatial dependence in the existing approaches for modeling spatial patterns in zero-inflated data for rare species.

1. Introduction

Data of rare species are usually characterized by a high percentage of zero observations due to their low abundance. Such data are generally referred as zero-inflated data. The excess zeros may invalidate the normality assumption that we commonly use in ecological data analyses and may cause computational problems (Cunningham and Lindenmayer, 2005). Ignoring a considerable proportion of zeros or merging multiple records would likely result in a loss of information that reflects the spatial or temporal distribution characteristics of the species. Ignoring the spatial patterns that are likely to exist may also cause bias in estimation because the assumption of independence among observations is violated. Therefore, analysis of such data requires specialized statistical technics, and it is quite challenging to incorporate spatial patterns in the analysis. Because rare species are frequently of ecological, conservation or management interest, development of appropriate models for such data may provide valuable information for management and conservation of rare species, e.g., to help identify hotspots of bycatch events or provide guidance for habitat restoration.

Several modeling approaches have been developed to deal with zero-inflated data. In these approaches, the non-zero data and zero data are handled either in two separate sub-models (e.g., the delta model or hurdle model: Pennington, 1983; Lo et al., 1992; Fletcher et al., 2005 and the zero-inflated model: Welsh et al., 1996; Hall, 2000; Minami et al., 2007), or simultaneously in a single model (e.g., the Tweedie distribution model: Tweedie, 1984 and the classification tree model: Li and Jiao, 2011b). Given these existing modeling approaches, we are left with the question of how to incorporate spatial dependence in these approaches.

* Corresponding author currently at: University of California, Department of Veterinary Medicine, Wildlife Health Center, Davis, CA 95616, USA.
Tel.: +1 540 8088373.
E-mail addresses: linli@ucdavis.edu, yanli9010@me.com (Y. Li).

http://dx.doi.org/10.1016/j.ecolmodel.2014.12.005

An early example of incorporating spatial patterns in the traditional regression model is the trend-surface analysis. The trend-surface analysis models the spatial patterns using polynomial terms of geographic coordinates (Legendre, 1990; Borcard et al., 2011). This method is devised to capture broad-scale spatial patterns with simple shapes like planes, saddles or parabolas, but insufficient to capture fine-scale structures that may require too many parameters to estimate (Borcard and Legendre, 2002). Another major problem with the trend-surface analysis is the correlation inherent in those polynomial terms, which may cause bias in parameter estimation and statistical testing; although this problem can be minimized by orthogonalizing the polynomial terms, difficulties may arise for interpretation and prediction.

The eigenfunction-based spatial filtering provides an alternative. The principal coordinates of neighbor matrices (PCNM) proposed by Borcard and Legendre (2002) is one of the commonly used eigenfunction-based spatial filtering methods. The Moran's Eigenvector Maps (MEM) is a generalization of PCNM by Dray et al. (2006). To apply the eigenfunction-based spatial filtering, a spatial weighting matrix is constructed based on the relationship between each pair of locations: whether they are neighbors and how strong their dependence is given that they are neighbors. The spatial dependence between two locations could be assumed to be a function of their Euclidean distance. The eigenvalues and their associated eigenvectors of a centered spatial weighting matrix constitute the eigenfunction. These eigenvectors (called the spatial filters hereafter) are believed to describe all the possible spatial patterns among sampling locations, where the spatial filters associated with larger eigenvalues represent broader-scale variation and those associated with smaller eigenvalues represent finer-scale variation. The spatial filters can further be used as explanatory variables in the traditional regression model. The eigenfunction-based spatial filtering has been used with success in several ecological applications (Borcard and Legendre, 2002; Diniz-Filho and Bini, 2005; Dray et al., 2006; Griffith and Peres-Neto, 2006; Borcard et al., 2011). Its advantages may include the flexibility in conjunction with the regression model framework, less demand for computation for large datasets and the capability of capturing both large- and fine-scale variations. Therefore, we were motivated to modify existing modeling approaches that handle zero-inflated data with eigenfunction-based spatial filters, and are hoping this modification may provide an alternative to incorporate spatial dependence in the data analysis for rare species.

In the present study, we demonstrated such a modification on the delta model. Delta model, also called hurdle model is developed on the basis of delta distribution to deal excess zeros in data analyses. Delta model and zero-inflated model are similar in that they both model zero and non-zero data in two separate sub-models. Their difference lies in the sub-model that handles zero data, called the probability sub-model, which models all zeros in the delta model while models part of the zeros in the zero-inflated model. Although in this study, we adopted the delta model as an example to demonstrate the modification with spatial filters for zero-inflated data, the methodology can also be applied to other modeling approaches such as the zero-inflated model.

The overall goal of this study was to demonstrate the modification of existing models with eigenfunction-based spatial filters to explore spatial patterns in zero-inflated data for rare species using the example of modified delta model fitted to a real dataset, the longline seabird bycatch observer data. Specifically, we aimed to (1) extract spatial filters from appropriate spatial weighting matrices; (2) modify the delta model by incorporating these spatial filters; (3) compare the performance of the modified delta model with other commonly used modeling approaches.

2. Materials and methods

2.1. Delta model

We applied the delta model as an example to deal with zero-inflated data for rare species. The delta model consists of two sub-models, one sub-model (positive sub-model) to analyze positive data, and the other one (probability sub-model) to estimate the probability of obtaining positive data. Product of the estimates from these two sub-models gives the final estimates:

$$\hat{c} = \hat{d} \times \hat{p}, \tag{1}$$

where $\hat{d}$ is the estimated value when only the positive data are analyzed, $\hat{p}$ is the estimated probability of obtaining positive data, and $\hat{c}$ is the final estimates from the delta model. For the positive sub-model, a generalized linear model was applied where we assumed a certain statistical distribution (e.g., a normal distribution) and a link function (e.g., an identity link) for the positive data:

$$g\left(\hat{d}\right) = \beta_0 + \sum \beta_q X_q, \tag{2}$$

where β_0 is the intercept; β_q is the coefficient for the qth explanatory variable X_q, and $g(.)$ is the link function. In the probability sub-model, the data was converted into the presence/absence data (y) that takes a value of one for those positive observations and a value of zero otherwise. A generalized linear model with an assumption of binomial distribution and a logit link function was used as the probability sub-model:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \alpha_0 + \sum \alpha_q X_q, \tag{3}$$

where α_0 is the intercept; α_q is the coefficient for the qth explanatory variable X_q.

The explanatory variables were selected through a stepwise approach based on the Akaike Information Criterion (AIC) and a chi-square test (Burnham and Anderson, 2002). Model development started with a model only including an intercept. At each step of the stepwise selection, the variable that reduced the AIC value most or showed the most significant effects on the response variable was selected into the model. We repeated this step until no substantial improvement was obtained from including an additional variable. Two-way interactions were not included in the models because they were either insignificant or correlated with main factors.

2.2. Eigenfunction-based spatial filters and the modified delta model

The spatial filters were obtained by calculating the eigenvectors and eigenvalues of the $n \times n$ spatial weighting matrix $W = [w_{ij}]$ after centering, where i and j index the ith and jth locations and n is the total number of locations of the observations (Dray et al., 2006; Griffith and Peres-Neto, 2006). The spatial weighting matrix can be seen as the Hadamard product (element-wise product) of a connectivity matrix $B = [b_{ij}]$ by a weighting function matrix $A = [a_{ij}]$, i.e., $[w_{ij}] = [b_{ij}\ a_{ij}]$ (Dray et al., 2006). Elements of the connectivity matrix B take a value of one for two locations that are neighbors (i.e., connected) and zero otherwise. We constructed five connectivity matrices for comparison, including the distance-based neighborhood based on minimum spanning tree and the one based on semivariogram range, the Delaunay triangulation, the Gabriel graph and the relative neighborhood graph. The last three were based on topology. Details on constructing these five connectivity matrices are provided in Appendix S1, and the technics can refer to Toussaint (1980), Lee and Schachter (1980), and Jaromczyk and Toussaint (1992).