Computational Statistics and Data Analysis 54 (2010) 1405–1418

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixtures of GAMs for habitat suitability analysis with overdispersed presence/absence data

David R.J. Pleydell [a,*,1,2], Stéphane Chrétien [b]

[a] UMR 6249 Laboratoire Chrono-Environnement, Université de Franche-Comté, Place Leclerc, 25030 Besançon Cedex, France

[b] Laboratoire de Mathématiques, UMR CNRS 6623 et Université de Franche Comté, 16 Route de Gray, 25030 Besançon Cedex, France

ARTICLE INFO

Article history:
Received 18 March 2008
Received in revised form 13 June 2009
Accepted 17 November 2009
Available online 26 November 2009

ABSTRACT

A new approach to species distribution modelling based on unsupervised classification via a finite mixture of GAMs incorporating habitat suitability curves is proposed. A tailored EM algorithm is outlined for computing maximum likelihood estimates. Several submodels incorporating various parameter constraints are explored. Simulation studies confirm that under certain constraints, the habitat suitability curves are recovered with good precision. The method is also applied to a set of real data concerning presence/absence of observable small mammal indices collected on the Tibetan plateau. The resulting classification was found to correspond to species level differences in habitat preference described in the previous ecological work.

1. Introduction

Understanding variations in species distribution has remained one of the key challenges in ecology since its conceptualisation as a discipline (Guisan and Zimmerman, 2000). It is natural that ecologists should seek to model species distribution and early models date from the nineteen twenties (Guisan and Thuiller, 2005). The uses of species distribution models (SDMs) in conservation biology include (Guisan and Thuiller, 2005) quantification of environmental niches for species, testing biogeographical, ecological and evolutionary hypotheses, invasive species monitoring, impact assessment for climatic change, prediction of unsurveyed sites for rare species, management support for species reintroduction and recovery, conservation planning, species assemblage modelling, classification of biogeographic or ecogeographic regions and calibration of ecological distance between patches in metapopulation or gene flow models.

Several techniques have been employed for SDMs including generalised linear models (GLMs), their flexible extensions generalised additive models (GAMs) (Guisan et al., 2002; Greaves et al., 2006; Segurado et al., 2006) and multiple adaptive regression splines Vaniscotte et al. (2009), tree based classification techniques (Franklin, 1998), ordination (Schenková et al., 2001), eco-niche factor analysis (Hirzel et al., 2002), Bayesian approaches (Gelfand et al., 2006), neural networks (Bessa-Gomes and Petrucci-Fonseca, 2003) and support vector machines (Drake et al., 2006). Ecologists have long recognised the bias introduced into SDMs when data are overdispersed with respect to a simple parametric model such as can arise when strong spatial dependence exists between observations, for example Guisan and Thuiller (2005), Barry and Elith (2006) and Segurado et al. (2006), but the proportion of articles published in ecological journals in which these biases are reasonably corrected for remains low. One problem, particularly in the spatial context, has been the lack of available tools for analysing

[*] Corresponding author.
E-mail addresses: pleydell@supagro.inra.fr, d_pleydell@yahoo.com (D.R.J. Pleydell), stephane.chretien@univ-fcomte.fr (S. Chrétien).

[1] Present address: UMR BGPI, Institut National de la Research Agronomique, CIRAD TA 41/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France.

[2] Data and scripts used in the current paper are available for download from https://sites.google.com/site/drjpleydell/.

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.11.016

overdispersed binary or Poisson data. This situation has been slowly changing since the seminal work of Diggle, Tawn and Moyeed (1998) who introduced the geostatistical concept of Gaussian random fields to the GLM literature to account for spatially smooth sources of overdispersion. Since then appropriate tools have become increasingly available: the geoRglm library (Christensen and Ribeiro, 2002) for Bayesian analysis of GLMs with geostatistical priors and the mgcv library for fitting generalised additive mixed models with either geostatistical or spline based random effects using penalised likelihood (Wood, 2006) are just two examples of what is now available for R (R Development Core Team, 2007).

A recent review (with online R code) of available techniques for the estimation of Gaussian random fields within a GLM for spatially dependent Bernoulli data (Paciorek and Ryan, 2005) suggested that the estimation of spatially structured random effects could be reasonable if the underlying spatial structure was simple relative to the sampling density of observation points. However when each curve and bend in a complex hidden surface was sparsely sampled then attempts to estimate the hidden surface proved less successful. The estimation of complicated hidden spatial structure from Bernoulli samples is now recognised to be highly data demanding suggesting that these models might be unreasonable in certain practical situations where logistical constraints limit the quantity of available data. We could ask the question "is it always necessary to estimate continuous spatial random effects plus three or four variogram parameters for binary ecological data sets?" or even "are hidden spatial structures in ecological data sets always smooth?". If the answers to these questions is "no" then perhaps we can simplify and reduce the number of random effects and parameters that we expect to estimate, thereby reducing the demands we place on our data sets. In this paper we attempt to do this using a mixture model approach where the usual single GAM with n continuous random effects might be replaced by say K GAMs. Such a simplification would require a small number of parameters relative to n, especially when further constraints between the mixture components are imposed.

Note that here we do not attempt to explicitly model the sources of overdispersion. The mixture model approach simply provides a general solution to account for various overdispersion sources. According to Robert (1996) mixture components "correspond to particular zones of support of the true distribution" and thus provide local representations of the likelihood function. While these local supports "do not always possess an individual significance or reality for the particular phenomenon modelled", interpretability can be possible in situations such as discrimination or clustering. This is the case for our model and a real data example is Section 5 is found to provide a very natural ecological interpretation.

It is worth noting that the simplification we propose is not necessarily made at the expense of physical interpretation. In a given ecological context a small number of discrete random effects could be a reasonable model for hidden spatial structure or other sources of overdispersion. For example, if the species in question is known to form colonies, one GAM might represent colony formation as a function of habitat suitability under relatively ideal conditions while a second GAM could account for possible absence of colonies in otherwise favourable habitat arising from a complex history of unobserved factors. Similarly, if the observations in question materialised from numerous different processes then a mixture model approach could be expected to outperform its $K = 1$ counterpart. The most pertinent number of random effects K could then be identified using model selection techniques. Herein lies an additional advantage of our approach, our GAM utilises a simple transformation on covariates and so the parameters for our mixture model can be estimated by maximum likelihood (ML). For highly flexible models such as GAMs with splines or random fields ML is known to be prone to over fitting and penalisations are often imposed to compensate. Since we use a mixture of simple GAMs with relatively limited flexibility we can use maximum likelihood directly without penalisation. For model comparison statistics such as Akaike Information Criterion (AIC) (Burnham and Anderson, 2002) are therefore readily available.

In the current paper we implement this proposed model simplification in a habitat suitability identification context. Habitat suitability curves are used to identify non-linear species responses along environmental gradients (see for example Jowett et al. (1991), Roussel et al. (1999) and Mäki-Petäys et al. (2002)). The concept is to identify a curve which transforms a continuous environmental variable to a scale more relevant to the distribution of the species in question thereby giving an index of habitat suitability.

2. A generalised additive model for habitat suitability identification

2.1. Habitat suitability curves in a GAMs framework

Generalised additive models (GAMs) have become popular tools in ecology due to their ability to detect non-linearities. A recent review of GAMs can be found in Wood (2006). The usual approach, when modelling an n length vector $Y = Y_1, \ldots, Y_n$, where Y follows some distribution of the exponential family, is to modify the linear predictor of a generalised linear model (McCullagh and Nelder, 1989) via the inclusion of smooth functions of covariates (Wood, 2006). Here we take the simple case,

$$g(\mu_i) = \beta_0 + \beta_1 \mathcal{H}(x_i), \tag{1}$$

where i provides an index on observations, $\mu_i \equiv E[Y_i]$, $g(\cdot)$ is a link function, β are coefficients and $\mathcal{H}$ is a smooth function of covariate x. Commonly $\mathcal{H}$ is chosen from a class of spline functions such as B-splines, P-splines, thin plate splines etc. (Wood, 2006). Such choices offer highly flexible solutions but the large number of parameters involved requires that practitioners remain cautious to problems of over fitting. Here, we depart from standard practice and adopt a much simpler two-parameter habitat suitability curve based on power functions for modelling $\mathcal{H}$. Our proposed habitat suitability curve (HSC)