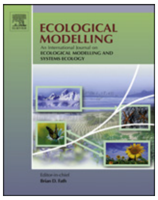
Contents lists available at ScienceDirect

Ecological Modelling

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



Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models

Aitor Gastón*, Juan I. García-Viñas

EGOGESFOR Research Group, Universidad Politécnica de Madrid, Escuela de Ingeniería Forestal y del Medio Natural, Ciudad Universitaria s/n, 28040, Madrid, Spain

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form 7 April 2011
Accepted 17 April 2011
Available online 12 May 2011

Keywords:
Species distribution models
Regularisation
Generalised linear models
Calibration

ABSTRACT

An important aspect of species distribution modelling is the choice of the modelling method because a suboptimal method may have poor predictive performance. Previous comparisons have found that novel methods, such as Maxent models, outperform well-established modelling methods, such as the standard logistic regression. These comparisons used training samples with small numbers of occurrences per estimated model parameter, and this limited sample size may have caused poorer predictive performance due to overfitting. Our hypothesis is that Maxent models would outperform a standard logistic regression because Maxent models avoid overfitting by using regularisation techniques and a standard logistic regression does not. Regularisation can be applied to logistic regression models using penalised maximum likelihood estimation. This estimation procedure shrinks the regression coefficients towards zero, causing biased predictions if applied to the training sample but improving the accuracy of new predictions. We used Maxent and logistic regression (standard and penalised) to analyse presence/pseudo-absence data for 13 tree species and evaluated the predictive performance (discrimination) using presence–absence data. The penalised logistic regression outperformed standard logistic regression and equalled the performance of Maxent. The penalised logistic regression may be considered one of the best methods to develop species distribution models trained with presence/pseudo-absence data, as it is comparable to Maxent. Our results encourage further use of the penalised logistic regression for species distribution modelling, especially in those cases in which a complex model must be fitted to a sample with a limited size.

1. Introduction

Predictive modelling of species distributions has become an increasingly important tool for ecological research and management in the last three decades (see Guisan and Thuiller, 2005 for a review). Three major components make up species distribution models: an ecological model, a data model, and a statistical model (Austin, 2002). An important aspect of the statistical model is the choice of the modelling method because a suboptimal choice may cause poor predictive performance. The ecological modelling community has shown significant interest in the effect of the modelling method on the predictive ability of species distribution models (e.g., Muñoz and Felicísimo, 2004; Segurado and Araújo, 2004)

A comparative study of the predictive performance of different modelling techniques developed by a working group at the National Center for Ecological Analysis and Synthesis (NCEAS) is the most comprehensive assessment to date (Elith et al., 2006). The study compared the performance of 16 modelling techniques using data from six regions and 226 species. They modelled species distributions using presence-only or presence/pseudo-absence data and evaluated the predictive performance using presence–absence data. Their results showed that novel methods, such as maximum entropy models (Maxent), outperform well-established modelling methods, such as logistic regressions (fitted either using generalised linear models (GLM) or generalised additive models (GAM)). Other studies also found better predictive performance for the Maxent compared to the logistic regression (Gibson et al., 2007; Elith and Graham, 2009; Roura-Pascual et al., 2009; Tognelli et al., 2009; Marini et al., 2010).

Further research using the NCEAS dataset showed that the differences between the Maxent and logistic regression models decrease inversely to sample size (Wisz et al., 2008). These results suggest that Maxent models are less sensitive to overfitting and consequently outperform logistic regressions when analysing small samples. The NCEAS working group used an average of 7.5 occurrences per estimated parameter in logistic regression models, a ratio under the recommended minimum of 10 (Harrell, 2001). Other model comparisons involving Maxent and GLM used occurrence/parameter ratios below 10 (Roura-Pascual et al., 2009; Marini et al., 2010) or slightly over (Gibson et al., 2007; Elith and Graham,

* Corresponding author. Tel.: +34 913367669; fax: +34 915446025.
E-mail address: aitor.gaston@upm.es (A. Gastón).

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.04.015

2009). In such small sample size scenarios, regularisation techniques may help avoid the performance problems caused by overfitting (Steyerberg et al., 2000). Unlike in Maxent models, the previous model comparisons did not use regularisation in the logistic regression models, which could be the cause of the observed difference in the predictive performance.

A way of applying regularisation to logistic regression models is using penalised maximum likelihood estimation (Harrell, 2001). The penalised regression outperformed an alternative regularisation technique called *Lasso* (Tibshirani, 1994) with small sample sizes in a comparison of regularisation methods applied to species distribution models (Reineking and Schröder, 2006).

In the penalised logistic regression, we maximise the penalised log likelihood (PML):

$$\mathrm{PML} = \log L - 0.5\lambda \sum (s_i\beta_i)^2$$

where L is the usual likelihood function, λ is a penalty factor, β_i are the estimated regression coefficients and s_i are the scale factors to make $s_i\ \beta_i$ unitless. This estimation procedure shrinks the regression coefficients towards zero, causing biased predictions if applied to the training sample but improving the accuracy of new predictions. Penalisation reduces the effective number of estimated parameters and, therefore, helps avoid performance problems caused by overfitting (Harrell, 2001).

The objectives of this study were to compare the penalised logistic regression with Maxent (one of the best methods in the NCEAS comparison) and to analyse the factors that may explain the differences in predictive performance. Our hypothesis was that Maxent would outperform the logistic regression in the NCEAS comparison because the Maxent included regularisation techniques and the logistic regression did not. If this is true, then the penalised logistic regression and Maxent should have similar predictive performance values. An alternative hypothesis could be that generative methods (like Maxent) have better predictive performance than discriminative methods (like the logistic regression) when the sample size is small (Phillips and Dudík, 2008). If this is true, then Maxent should outperform the penalised logistic regression. We attempted to test these hypotheses by comparing the predictive performance of Maxent and the penalised logistic regression models for varying numbers of tree species occurrence records in Spain.

2. Materials and methods

2.1. Experimental framework

We fitted species distribution models for 13 tree species using presence/pseudo-absence data and evaluated the predictive performance using presence–absence data. We varied the number of occurrences in the training datasets from 10 to 1280 to assess the effect of sample size on model performance. Four modelling strategies were tested (see below for a detailed description): standard logistic regression (full models and stepwise selection of predictors), penalised logistic regression, and Maxent with default settings.

2.2. Species occurrence data

We used the Spanish National Forest Inventory (NFI) dataset to generate training and evaluation data. NFI comprises a systematic grid with 91,889 plots, each of which is 0.2 ha in size. Two different approaches were used to split the NFI dataset into training and evaluation datasets. The first approach consisted in a random split of the NFI dataset into two subsets of equal number of plots for model training and evaluation. To reduce the effect of spatial autocorrelation between training and evaluation datasets a spatial split approach was additionally used. For each species, we split the NFI dataset along the meridian that leaves one half of the species occurrences in each side (West or East). The half with the higher number of plots was used as training sample.

For both approaches, we varied the number of occurrences from 10 to 1280 (by considering non-nested subsets of the full training dataset). Ten thousand plots were randomly drawn from the whole training dataset (including plots with presence records) and used as pseudo-absences. The pseudo-absences were the same for every model run in the random split approach and varied between species in the spatial split approach.

We modelled the distribution of tree species from the Pinaceae and Fagaceae families native to continental Spain, excluding species with fewer than 1300 occurrences in the training dataset to allow the same sample size range for every species considered. A total of 13 species were used: *Castanea sativa* Miller, *Fagus sylvatica* L., *Pinus halepensis* Miller, *Pinus nigra* Arnold, *Pinus pinea* L., *Pinus pinaster* Aiton, *Pinus sylvestris* L., *Quercus faginea* Lam., *Quercus ilex* L., *Quercus humilis* Miller, *Quercus pyrenaica* Willd., *Quercus robur* L., and *Quercus suber* L.

2.3. Environmental predictors

Maxent's default settings are optimised for models with 11–13 environmental predictors (Phillips and Dudík, 2008); therefore, we selected 11 environmental predictors for our comparison.

We derived climatic data grids by applying the models for climatic estimation developed by Sánchez Palomares et al. (1999) to the STRM 3-arc-second (≈90 m) elevation dataset (Farr et al., 2007). The climatic estimation models interpolate monthly climate data from weather stations using latitude, longitude, and elevation as independent variables. We used a set of 10 climatic predictors commonly considered in tree species autoecology in Spain (Alonso Ponce et al., 2010): mean summer rainfall, mean annual rainfall, mean summer temperature, mean annual temperature, mean of maximum temperatures of the warmest month, mean of minimum temperatures of the coldest month, dry season length, mean annual potential evapotranspiration, mean annual water surplus, and mean annual water deficit.

Some of the climatic variables are highly collinear (the correlation was greater than 0.8 in 12 out of 45 pairs of variables) and a variable reduction may be advisable. Nevertheless, we kept all the variables to reproduce conditions for which Maxent default settings were optimised (i.e., 11–13 collinear environmental predictors, see Phillips and Dudík, 2008; Elith et al., 2006, Table 3). Collinearity can cause inflated standard errors of the regression coefficients, but does not affect predictions made on new data that have the same degree of collinearity as the training data, as long as extreme extrapolation is not attempted (Harrell, 2001). Our study focused on model predictions and the degree of collinearity of the training and evaluation datasets were almost the same (i.e., the values of the correlation matrix did not differ more than 0.019 between the training and evaluation datasets).

The distribution of calcareous parent materials is a useful predictor of plant species distribution in our study area (Gastón et al., 2009). We used the European Soil Database (Van Liedekerke et al., 2006) to allocate each plot to a parent material class (calcareous or siliceous).

2.4. Modelling strategies

We modelled species distributions using Maxent (Phillips et al., 2006) version 3.3.2 with the default settings. Maxent's default settings are a set of model parameters obtained as a result of a tuning approach using the NCEAS dataset (Elith et al., 2006). This approach focused on tuning the regularisation parameters and the choice of