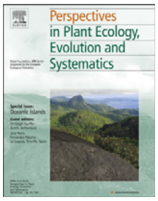
Perspectives in Plant Ecology, Evolution and Systematics 12 (2010) 219–225

Contents lists available at ScienceDirect

Perspectives in Plant Ecology, Evolution and Systematics

journal homepage: www.elsevier.de/ppees



Research article

Predictive performance of plant species distribution models depends on species traits

Jan Hanspach*, Ingolf Kühn, Sven Pompe[1], Stefan Klotz

UFZ, Helmholtz Centre for Environmental Research—UFZ, Department of Community Ecology, Theodor-Lieser-Str. 4, 06120 Halle, Germany

ARTICLE INFO

Article history:
Received 19 June 2009
Received in revised form
22 March 2010
Accepted 6 April 2010

Keywords:
Atlas Florae Europaeae
Cross-validation
GLM
AUC
Kappa

ABSTRACT

Predictive species distribution models are standard tools in ecological research and are used to address a variety of applied and conservation related issues. When making temporal or spatial predictions, uncertainty is inevitable and prediction errors may depend not only on data quality and the modelling algorithm used, but on species characteristics. Here, we applied a standard distribution modelling technique (generalized linear models) using European plant species distribution data and climatic parameters. Predictive performance was calculated using AUC, (Cohen's) Kappa and true skill statistic (TSS), that were subsequently correlated with biological and life-history traits. After accounting for phylogenetic dependence among species, model performance was poorest for species having a short life span and occurring in human disturbed habitats. Our results clearly indicate that the performance of distribution models can be dependent on functional traits and provide further evidence that a species' ecology is likely to affect the ability of models to predict its distribution. Biased and less reliable predictions could misguide policy decisions and the management and conservation of our natural heritage.

Introduction

Species distribution models (SDMs) are widely used tools in ecological research (Guisan and Zimmermann, 2000). Given the increasing importance of SDMs in ecological risk assessment, an evaluation of their predictive capacity is necessary (Barry and Elith, 2006). Within temporal or spatial predictions, model uncertainty is inevitable and thus prediction error needs to be assessed thoroughly in order to derive reasonable interpretation of the model results (Heikkinen et al., 2006).

Fielding and Bell (1997) distinguished between 'algorithmic' and 'biotic' prediction errors. Algorithmic errors have been widely studied (e.g. Segurado and Aráujo, 2004; Pearson et al., 2006; Meynard and Quinn, 2007) and are an artefact of the data-collection process or stem from limitations of the modelling approach. In contrast, biotic errors occur when ecological parameters are omitted from the modelling framework and lead to an inaccurate description of the species' distribution. SDMs are based on the assumption that species are in equilibrium with climate, i.e. they occur in all climatically suitable areas whilst being absent from all unsuitable ones (Guisan and Zimmermann, 2000). However, this assumption is violated by the influence of biotic interactions (Brown et al., 1996; Pearson and Dawson, 2003) and dispersal limitations (Svenning et al., 2008). Further, SDMs assume that species show no intraspecific variability regarding their niche but this is not always true, for example large-ranging species show adaptations to local conditions (Stockwell and Peterson, 2002). Distribution models commonly apply climatic predictors but other biologically relevant parameters such as land use, geological and pedological properties, groundwater influence and biotic interactions were, until recently, rarely considered mostly due to limited data availability. This has been addressed in recent studies and applications are promising (Heikkinen et al., 2007; Luoto et al., 2007; Pompe et al., 2008; Rickebusch et al., 2008; Schweiger et al., 2008).

Besides modelling algorithm and environmental parameters, predictive performance can be associated with species' traits. Geographical attributes of ranges (McPherson et al., 2004; Luoto et al., 2005) and climatic niche position (Kadmon et al., 2003; Hernandez et al., 2006) have been identified as correlates of predictive performance. The importance of range size for predictive performance is often studied although results are inconclusive. While some studies report that models of species with a high prevalence perform better compared to species with low prevalence (birds: Kadmon et al., 2003, artificial species: Meynard and Quinn, 2007), others found opposite results (birds: Stockwell and Peterson, 2002, butterflies: Luoto et al., 2005, various animal species: Hernandez et al., 2006). Furthermore, the influence of prevalence can be a statistical artefact because of the systematic dependence of modelling algorithms and accuracy

---

* Corresponding author. Tel.: +49 345 558 5307; fax: +49 345 558 5329.
E-mail address: Jan.Hanspach@ufz.de (J. Hanspach).

[1] Current address: Max-Planck-Institute for Biogeochemistry, P.O. Box 10 01 64, 07701 Jena, Germany.

1433-8319/$ - see front matter © 2010 Rübel Foundation, ETH Zürich. Published by Elsevier GmbH. All rights reserved.
doi:10.1016/j.ppees.2010.04.002

measures on prevalence (McPherson et al., 2004; Allouche et al., 2006). Results concerning climatic niches are less ambiguous and models of species with clumped distributions (Luoto et al., 2005) and a narrow climatic niche are generally more accurate (Kadmon et al., 2003; Hernandez et al., 2006) than models of species with scattered distributions and wider climatic tolerance.

Since range and niche characteristics within these studies were derived from the data used for modelling, they strongly resemble data characteristics within the study area and do not necessarily depict true species characteristics. Independent data on species' traits have only been correlated with prediction errors for birds (Seoane et al., 2005; Carrascal et al., 2006; McPherson and Jetz, 2007), butterflies (Pöyry et al., 2008) and trees (Guisan et al., 2007). A low model accuracy was reported for fast growing, early successional tree species (Guisan et al., 2007) and locally scarce and inconspicuous bird species (Seoane et al., 2005). In contrast, Carrascal et al. (2006) found poor models for ubiquitous species, as well as species with very variable distributions, i.e. expanding/shrinking with climatic fluctuation. Distribution models for butterflies revealed a better predictive performance for large, more easily detectable, and less mobile species that are less influenced by stochasticity. Furthermore, the habitat in which species predominantly occur also influences the predictive power of distribution models (Pöyry et al., 2008).

It is evident, therefore, that the knowledge of the relationship between plant species traits and the predictive performance of SDMs is poor and requires further examination (Guisan et al., 2007). In our study, we use a standard distribution modelling technique and a widely used species distribution dataset to test the relationship between model performance and plant species traits. We hypothesise that the following traits will affect the quality/predictive performance of plant species distribution models:

- Life span/life form: The absence of short-lived species in a mapping unit may result from true climatic unsuitability but may also be caused by the failure to detect the species or by the absence of natural habitat in the mapping unit (e.g. following human disturbance) (Guisan and Thuiller, 2005). The distribution and occurrence of long-lived and conspicuous species are more likely to be recorded (Seoane et al., 2005). As modelling accuracy improves with data quality, we hypothesise that SDMs for these species will have a greater predictive power.
- Ecological strategy type (after Grime, 1979)/habitat dependence: Species ranges are not only restricted by environmental parameters but by biotic interactions (Gaston, 2003). We hypothesise that a slow growth rate and a good competitive ability promotes filling of the climatic niche and a longer persistence in the landscape and thus improves the predictive performance of SDMs (Guisan et al., 2007); or vice versa: if the occurrence of a species is prevented by competition with another species and the SDM does not include the interaction, it will tend to overestimate the species distribution (McPherson and Jetz, 2007). We also hypothesise that species with a lower competitive ability are associated with specific habitat types (e.g. extreme habitats, disturbances, pioneer species). The absence of such species may not indicate unsuitability and SDMs will therefore have a lower predictive quality.
- Dispersal type: It is assumed that trees (Svenning and Skov, 2004) and forest herbs (Svenning et al., 2008) in Europe have not yet reached their post-glacial equilibrium with climate. Species with low dispersal ability may not have filled their climatic niches and thus the predictive performance of these SDMs is reduced.
- Pollination type: Relative frequencies of pollination types are dependent on specific climatic and non-climatic factors (Kühn et al., 2006). Insect pollination is strongly dependent on land use, topology and geology (Kühn et al., 2006), wind pollination is facilitated by open vegetation (Culley et al., 2002) and a moderate wind speed (Whitehead, 1983) and selfing is regarded to be favoured when the environment is variable (e.g. due to disturbances), climatic conditions are poor or mates are absent (Baker, 1955). Dependence of a pollination type on non-climatic parameters or climatic parameters commonly not included in SDMs (e.g. wind speed, disturbance) would probably reduce predictive performance of SDMs for species having this pollination type.
- Niche width/habitat tolerance: Beside the prime importance of growth rate, Guisan et al. (2007) reported a low predictive performance for generalist trees, i.e. trees with wider elevation ranges. We hypothesise that this holds for other plant species with a low degree of specialization. The lack of ecological contrast among occupied and unoccupied mapping units should constrain a statistical description of the species' distribution.

Methods

Data

Species distribution data of vascular plants were obtained from the Atlas Florae Europaeae database (AFE) maintained by the Botanical Museum, University of Helsinki, at a resolution of 50 km × 50 km. Species traits were derived from BiolFlor (Klotz et al., 2002), a database of biological and ecological traits for Central European plant species and from a dataset on dispersal type (Frank and Klotz, 1990). We extracted all AFE species with available trait information. We used the following traits to address our hypotheses: dispersal type, life span, life form, pollination type, strategy type, number of vegetation units a species is affiliated to and hemerobic level (see Table 1 for details). Although the BiolFlor database covers Central Europe only, we associated the data with models covering the whole of Europe, as the chosen traits are generally stable and show low intraspecific variability. Hence, this spatial mismatch should not influence the results. The AFE database covers approximately 20% of the European flora but does not provide distribution data on some species rich herb families such as Asteraceae, Poaceae, Cyperaceae and Fabaceae. Preliminary tests for regions with known distributions of the full flora revealed that our modelled species are generally adequate to represent the trait compositions of the whole flora (Hanspach et al., unpublished). Species with less than 50 presences or absences in the AFE database were excluded to allow for reliable modelling (Kadmon et al., 2003). Data on recent climate (1961–1990) were taken from Mitchell et al. (2004) and were aggregated from the original resolution of 10 min × 10 min onto the 50 km × 50 km resolution of the AFE data. We derived a set of 17 standard climatic variables (see Appendix Table 1).

Modelling of species distribution

We used generalized linear models (GLMs) with a binomial error distribution to model the distribution of 638 plant species. Recently, authors have suggested that several methods be combined in an ensemble to minimize algorithmic errors (Araújo and New, 2007; Marmion et al., 2009). GLMs are, however, among the methods with a good performance (Elith et al., 2006). They are more robust to overfitting and have a better transferability or