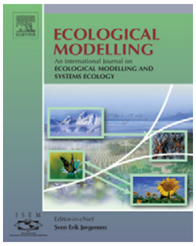
ECOLOGICAL MODELLING 204 (2007) 104–114

available at www.sciencedirect.com

ScienceDirect

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



Evaluating temporal and spatial generality: How valid are species–habitat relationship models?

B. Strauss[*], *R. Biedermann*

Institute for Biology and Environmental Sciences, University of Oldenburg,
26111 Oldenburg, Germany

ARTICLE INFO

Article history:
Received 11 May 2006
Received in revised form
12 December 2006
Accepted 22 December 2006
Published on line 6 February 2007

Keywords:
Habitat model
Model transfer
Model discrimination
Model calibration
Classification threshold
Leafhoppers
Grasshoppers

ABSTRACT

Prior to making general inferences or predictions from habitat models, their generalizability requires thorough assessment. However, systematic testing of model generality is often claimed, but rarely done. We used existing models for phytophagous insects (grasshoppers and leafhoppers) from a study on urban brownfields. Data for model building had been collected in two major cities of Northern Germany, Berlin and Bremen. We transferred these models to test data from another year (Bremen, 30 model transfers), and to test data from different geographic regions (transfer from Berlin to Bremen and vice versa, 30 model transfers). We evaluated discriminatory ability as well as model calibration for the test data. Most transfers (28 in time, 27 in space) were successful, i.e. occupied sites within the test data were assigned higher occurrence probabilities than unoccupied sites, the threshold independent *c*-index for the test data exceeded chance. Our results indicated that models built on the larger dataset (147 plots, Bremen) were more general than the ones basing on the smaller dataset (89 plots, Berlin).

The overall good transferability had three important drawbacks: (1) models were mostly not well calibrated to the test data, thus predicted occurrence probabilities may not be used as absolute values, but as ordinal ranks. (2) Model fit to the test data often decreased considerably compared to the training data. (3) Dichotomising occurrence probabilities to presence/absence predictions required prior information about species prevalence. Assigning presences to the sites with the highest predicted occurrence probabilities, with the number of presences corresponding to the prevalence, proofed to be a comparatively simple and reliable way of dichotomising predictions. Still, it only allowed predictions exceeding chance for 19 model transfers in time and 23 transfers in space, and required information about species' prevalences.

We qualitatively compared pairs of models for 10 species, with one model basing on the Bremen data, one on the Berlin data. Both models had been built with the same modeling technique. Vegetation structure variables were largely comparable between models. It seemed that they were more directly related to species' occurrences and thus more general than landscape context variables and soil parameters.

---

[*] *Corresponding author*. Tel.: +49 441 798 3608; fax: +49 441 798 5659.
E-mail address: barbara.strauss@uni-oldenburg.de (B. Strauss).

doi:10.1016/j.ecolmodel.2006.12.027

1. Introduction

Habitat models, also called species distribution models or species–habitat relationship models, quantify species–habitat relationships. Habitat models see increasing use in ecology and conservation biology (Guisan and Zimmermann, 2000; Vaughan and Ormerod, 2005). The availability of satellite data and remote sensing techniques enables predictions of species occurrences over large areas. A commonly ignored drawback is that models based on data from one study year or site ('training data') may lose most of their predictive power when applied to data from other years or different geographic regions (Bulluck et al., 2006). Such failure might stem from overfitting of the model to its training data as well as from different conditions experienced in new data (Vaughan and Ormerod, 2005). Even though it is often claimed that prior to their application, the validity of models beyond their training data needs to be tested with independent test data (Pearce and Ferrier, 2000b; Vaughan and Ormerod, 2005; Araujo and Guisan, 2006), this is rarely done. Internal validation (e.g. bootstrapping) enables unbiased estimates of model performance for the training data, but it cannot assess a model's generalizability, i.e. its capacity to predict a species' distribution with new data from different regions or different years (Altman and Royston, 2000; Vaughan and Ormerod, 2005; Randin et al., 2006). Vaughan and Ormerod (2003) propose that independent test data, collected from a geographically discrete region, are the only valid test. Still, few studies systematically investigate the generalizability of models (but see Bulluck et al., 2006; Fleishman et al., 2003; Jensen et al., 2005; Randin et al., 2006). It is common to split one data set in training and test data to evaluate a model's performance and generalizability (e.g. Eyre et al., 2005). However, the significance of such tests may not exceed what could be achieved with internal validation as well. The generalizability of habitat models needs to be evaluated with respect to two aspects: (1) discrimination, and (2) calibration (Pearce and Ferrier, 2000b).

Discriminatory power of a model is the capacity to distinguish occupied from unoccupied sites (Pearce and Ferrier, 2000b). It can be evaluated by several threshold dependent and threshold independent measures. Threshold dependent measures require dichotomisation of a model's quantitative output (probabilities of occurrence) into presences and absences (Fielding and Bell, 1997). The choice of the threshold largely determines the result. Sensitivity (the model's ability to correctly predict presences), specificity (ability to correctly predict absences), and the overall correct classification rate (CCR) are easy to interpret. However, they can be highly misleading if chance is not considered. For instance, a model for a rare species can achieve high correct classification if all sites are predicted as absences (Olden et al., 2002). Such a model is of limited use for ecological applications. In general, prevalences different from 0.5 allow high chance predictions. Thus, when using threshold dependent measures, it is necessary to assess if a model's predictions are better than what could be achieved by chance alone (Vaughan and Ormerod, 2005).

Despite these threshold related problems, a common goal in ecological applications is to produce presence/absence predictions, making the choice of a threshold unavoidable. During model building, a threshold may be chosen based on the data (Fielding and Bell, 1997). If a model is applied to new environmental data, where nothing is known about a species' presence or absence, this way of finding an optimal threshold is not possible. Applying the 'training threshold' to new data might be risky, in particular if prevalences differ between the training data and the area where the model is to be applied.

The selection of one particular threshold tests accuracy under only one scenario and thus limits the capacity to describe generalizability (Pearce and Ferrier, 2000b). Threshold independent, non-parametric correlation coefficients like the *c*-index (equivalent to the AUC and the Wilcoxon statistic) overcome this problem by making direct use of the occurrence probabilities (Vaughan and Ormerod, 2005). They compare the mean rank of occurrence probabilities for occupied sites with those of unoccupied sites. The *c*-index represents the probability that the model assigns a higher probability of occurrence to a randomly chosen occupied site than to a randomly chosen unoccupied one (Hanley and McNeil, 1982).

Model calibration addresses the numerical accuracy of predictions, i.e. if each predicted probability is an accurate estimate of the likelihood of detecting a species at a given site (Pearce and Ferrier, 2000b). Calibration can be split up into two measurable components: bias and spread. Consistent over- or underestimation (bias) typically results when a species' prevalence differs from the training data (Pearce and Ferrier, 2000b). Probabilities that are too extreme (spread), i.e. too low at unoccupied sites and too high at occupied ones, indicate overfitting (Vaughan and Ormerod, 2005). Even if a model successfully discriminates new data, calibration might be poor (Vaughan and Ormerod, 2005). This becomes a problem if maps with probabilities of occurrence are produced, where, for example, an estimated probability of 0.9 represents an actual probability of only 0.6.

In this paper, we transfer habitat models for phytophagous insects in time (data from 2 years) and space (data from different geographic regions). With these model transfers, we address the following questions:

1. Can species models from 1 year and region be used to predict species occurrence in another year and/or different geographic region, namely:
   - Are sites correctly ranked from unsuitable to suitable?
   - Is it possible to apply a threshold that successfully separates occupied from unoccupied sites?
   - Are transferred models well calibrated, allowing quantitative predictions of occurrence probabilities?
2. Do data from different regions lead to similar models, if the same modeling techniques are applied?

2. Methods

2.1. Habitat models, training data and test data

For this paper, we used existing habitat models for grasshoppers and leafhoppers (Orthoptera and Hemiptera: Auchenorrhyncha) in urban brownfields (Strauss and Biedermann, 2006). Models were available from two study areas in Northern Germany, Berlin (sampled in 2004) and Bremen (sampled