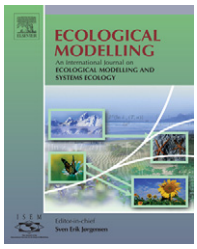
available at www.sciencedirect.com

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.

© 2007 Elsevier B.V. All rights reserved.

* *Corresponding author*. Tel.: +49 441 798 3608; fax: +49 441 798 5659.
E-mail address: barbara.strauss@uni-oldenburg.de (B. Strauss).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
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

Table 1 – Overview of model transfers

Training data	Test data	# of models	Transfer
Bremen 2003 (157)	Bremen 2004 (149)	30	Temporal
Bremen 2003 (157)	Berlin 2004 (89)	10	Spatial, temporal
Berlin 2004 (89)	Bremen 2003 (157)	10	Spatial, temporal
Berlin 2004 (89)	Bremen 2004 (149)	10	Spatial

Sample size in brackets.

in 2003 and 2004). These study areas are located at a distance of 300 km. In Berlin (52°30′ N, 13°28′ E, mean temperature 9.7 °C, mean annual precipitation 560 mm), 89 plots had been set up in a random stratified way, in Bremen (53°05′ N, 8°44′ E, mean temperature 8.8 °C, mean annual precipitation 694 mm), 157 plots. For each species with a prevalence ≥ 10%, models had been built using logistic regression (i.e. generalized linear models (GLMs) with a logistic link) and model averaging (Burnham, 2002; Gibson et al., 2004b). Only monotonic and univariate relationships were considered. Several 'good' models for a species had been weighted and averaged. Each model entering the model averaging process consisted of one to four environmental variables. This process resulted in averaged models for 28 species in Berlin and 30 in Bremen. Details on the model building process can be found in Strauss and Biedermann (2006). For 10 species, models were available from both study areas.

Environmental variables covered four main driving factors: vegetation structure (e.g. several height and density measures and litter cover), landscape context (proportions of different brownfield types within different radii around the plots), soil parameters (e.g. pH, available water capacity, soil nutrients), and site age (for details, see Strauss and Biedermann, 2006). Note that environmental variables approximately covered the same ranges of values in both study areas. However, the distribution of values within the total range differed between Bremen and Berlin.

We applied the habitat models to different test data (Table 1). To test transferability in time, the Bremen models were used on test data from the same plots, recorded in the following year. Transferability in space we tested for the 10 species that had models for both study areas. Bremen models were applied to Berlin data, and vice versa. The transfers from the Bremen 2003 models to the Berlin 2004 data and from the Berlin 2004 models to the Bremen 2003 data represented transfers in both time and space. Such transfers might be expected to lead to poorer models than transfer in time only. All calculations were performed using Splus 6.1.

2.2. *Assessing model discrimination*

We assessed model discrimination by threshold-dependent (*c*-index) and threshold-independent measures. A chance model has a *c*-index of 0.5 (Hanley and McNeil, 1982). With small datasets and/or few observations, confidence limits grow large (McPherson et al., 2004). We therefore performed a randomisation test (Manly, 2001) to test if species occurrences were associated with significantly higher predicted probabilities of occurrence. The model's predicted probabilities for the data were randomly distributed over the sites and the *c*-index was calculated. This procedure was repeated 10,000 times to produce a null (or chance) distribution with a median of 0.5. If a model's *c*-index exceeded the 95%-percentile of this chance distribution, we considered it to be significantly ($p \leq 0.05$) different from chance.

We applied two methods to dichotomise predictions. First, we used P_{Kappa} (threshold that maximizes Cohen's Kappa) of the original models (Liu et al., 2005). Second, we assigned presence to the plots with the highest predicted occurrence probabilities. The number of plots that was assigned presence we chose to be the same as the observed number of presences (prevalence based proportion of highest probabilities = pbp). For a species with a prevalence of 30%, the 30% of plots with the highest predicted occurrence probabilities were assigned presence. Since models with high discriminatory power assign the highest occurrence probabilities to occupied sites, we expected this method to correctly classify a substantial proportion of plots. The quality of dichotomised predictions we assessed with four measures of agreement: sensitivity, specificity, CCR, and Cohen's Kappa (Fielding and Bell, 1997). To illustrate 'chance', we generated a chance distribution for each of these measures: for each species, given its prevalence and the number of plots, we randomly distributed the observations over the plots and calculated the measures of agreement. This we repeated 10,000 times. The resulting chance distributions have a median corresponding to the prevalence (for sensitivity), 1 − prevalence (for specificity), [prevalence × # of presences] + [(1 − prevalence) × # of absences] (for CCR) (Fielding and Bell, 1997), and approximately 0 (for Kappa). The 95%-percentiles depend on prevalence and sample size. We considered the model to perform better ($p \leq 0.05$) than chance with the respective threshold if all measures of agreement exceeded the 95%-percentile of their chance distribution.

2.3. *Assessing model calibration*

For every model transfer, we calculated a calibration curve as described by Pearce and Ferrier (2000b). It relates the logit-transformed model predictions ($\ln[\pi_i/(1-\pi_i)]$) to the observed presences/absences by means of logistic regression. In case of a perfectly calibrated model, the resulting regression line has an intercept of zero and a slope of one (Miller et al., 1991). Transforming logits to probabilities results in curved logistic lines (Fig. 1). Significant deviations from perfect calibration we tested with likelihood ratio tests (Pearce and Ferrier, 2000b; Miller et al., 1991). Deviations of the intercept from 0 indicate bias, with intercepts < 0 resulting in predictions that are too high, and with intercepts > 0 giving too low predictions. Slopes > 1 result in predictions that

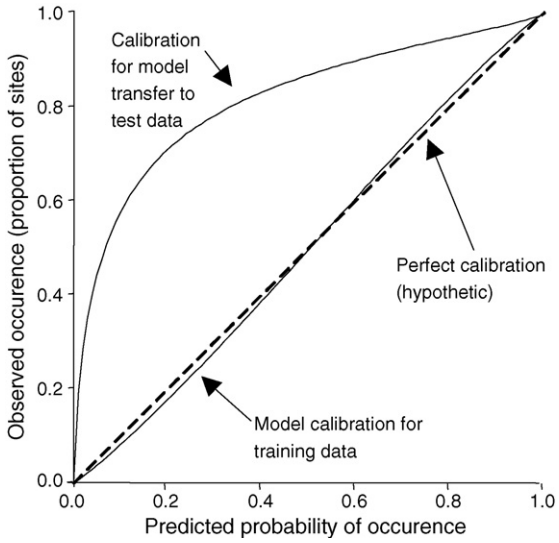


Fig. 1 – Calibration curves, resulting from relating logit-transformed model predictions to observed occurrences by logistic regression. The example shows the Bremen model for *Chorthippus mollis* and its transfer to Berlin data. After transfer, significant bias is obvious: consistent underestimate of occurrence probabilities, due to an increase in prevalence from 39% (Bremen) to 81% (Berlin).

are too extreme, i.e. too low for probabilities < 0.5 and too high for probabilities > 0.5, indicating overfitting. The reverse occurs for slopes between 0 and 1. If slopes are < 0, the overall trend of predictions is wrong with unoccupied sites having the highest predicted occurrence probabilities. Note that with slopes significantly different from 1, the intercept merely describes the bias for $p = 0.5$ (Vaughan and Ormerod, 2005).

2.4. *Qualitative comparison of models*

For the 10 species that had models for both study areas, we qualitatively compared these models. Model averaging, which we had used for model building, considers a number of models for each species and does not eliminate significant variables or models like, e.g. stepwise procedures. It also allows to assess the weight of each variable within a species' averaged model (Burnham, 2002). Thus, via the qualitative model comparison, we could check if the same variables were important in both regions. Moreover, we compared the functional form of the relationships (Altman and Royston, 2000).

3. Results

3.1. *Transfer in time*

The detailed results for models transfers in time (transfer of Bremen models 2003 to Bremen data 2004) are shown in Fig. 2. Numbers below the species name give prevalences: in the case of *Aphrodes makarovi*, 28% in the test data and 15% in the training data. The first black dot gives the *c*-index of the model transfer: 0.57 for *A. makarovi*. This does not exceed the 95%-percentile of the null distribution generated by a randomisation test. The span between the 50% and the 95%-percentile of this null distribution is indicated by the solid black line. The not successful transfer in terms of *c*-index is indicated by the minus on top of the species column. For the training data, the *c*-index was 0.88, shown as an open circle. This far exceeded the 95%-percentile of the null distribution, the distance is shown as a dashed line. The next pair of symbols represents Kappa for test and training data. Again, the open circle represents the training data. In this case of a threshold dependent measure, the black dot represents Kappa for pbp, the 'x' for P_{Kappa} of the original model. The + and − on top indicate that for P_{Kappa}, in *A. makarovi*, Kappa was higher than chance (+, upper symbol), for pbp not higher than chance (−, lower symbol). The next pairs of symbols represent sensitivity, specificity and CCR in the same way.

Results of all species are summarized in Tables 2 and 3. For 28 out of 30 species, the *c*-index of the model transfer was significantly better than chance. Most models assigned highest occurrence probabilities to occupied test sites, with the exceptions of *A. makarovi* and *Macrosteles cristatus*. *c*-Index values mostly decreased with model transfer (Table 3). Median of this decrease (for the transfers with significant *c*-index) was −0.11 with a maximum of −0.3 and a minimum of +0.05. Applying a threshold caused difficulties. With the models' original P_{Kappa}-thresholds, dichotomised predictions exceeded chance for only four models. Pbp performed better, 19 species models exceeded chance (these species' names are printed in bold in Fig. 2). For the models that could be successfully transferred using pbp, Kappa decreased considerably (median of difference: −0.19).

Most models partly lost their calibration when transferred in time (Table 2). About half of the models showed significant spread (slope of calibration curve differs from 1) with new data. However, for models that could be successfully transferred using pbp, only 6 out of 19 exhibited significant spread. For the two species where the *c*-index was not significant, the slope of the calibration curve was < 0. This indicates that the overall trend of probabilities was wrong with high predicted occurrence probabilities where observed probabilities were low and vice versa. The other slopes different from 1 were between 0 and 1, indicating overfitting. Most models showed significant bias (intercept different from 0). This could mostly be traced back to differences in prevalence. If prevalence decreased with respect to the training data, the intercept was < 0, resulting in consistent overestimation of occurrence probabilities. Where prevalences were similar, intercepts were not different from 0 (e.g. *Oedipoda caerulescens*, *Neophilaenus minor*). For only three species (*O. caerulescens*, *N. minor*, *Rhopalopyx vitripennis*), models were well calibrated to the test data.

3.2. *Transfer in space*

Details of model transfers in space are given in Fig. 3. Transfers in space worked well for 9 out of 10 species regarding the *c*-index (Table 2). Transfer of the model for *Doratura homophyla* failed for all test data and models. Overall, transfer in space worked better from Bremen to Berlin than vice versa: *c*-index decrease was minor for the transfer of Bremen models to Berlin data (median −0.04), but larger for the transfer

Table 2 – Results of model transfers: discrimination and calibration with test data

	Bremen '03 → Bremen '04			Bremen '03 → Berlin '04			Berlin '04 → Bremen '03			Berlin '04 → Bremen '04		
	discr.	cal.		discr.	cal.		discr.	cal.		discr.	cal.	
Athysanus argentarius	0.82			0.88	++	++	0.67	+	+	0.67	+	+
Chorthippus mollis	0.72	+	+	0.86	+	+	0.63	+		0.69	+	
Cicadula quadrinotata	0.76	+		0.85	+	+	0.85	++	+	0.72		+
Doratura homophyla	0.82	+	+	0.36		+	0.44		+	0.57		+
Euscelis incisus	0.78	+	+	0.73	+	+	0.77	+	+	0.78	++	++
Macrosteles quadripunctulatus	0.91	+	+	0.90	+		0.76	+	+	0.85	+	+
Metrioptera roeseli	0.79	+	+	0.86	+	+	0.72	+		0.70	++	+
Oedipoda caerulescens	0.86	+	++	0.72	+		0.70			0.61		
Ophiola decumana	0.86	+	+	0.79	++	++	0.66	+		0.77	+	+
Psammotettix confinis	0.75	++	+	0.77	+	+	0.85	+		0.82	+	
Aphrodes makarovi	0.57		+									
Arocephalus longiceps	0.67											
Arthaldeus pascuellus	0.74	++										
Chorthippus biguttulus	0.68		+									
Cicadella viridis	0.60											
Cixius nervosus	0.86	+	+									
Elymana sulphurella	0.78	+										
Jassargus pseudocellaris	0.96	+										
Javesella pellucida	0.69											
Macropsis prasina	0.77											
Macrosteles cristatus	0.50		+									
Macrosteles ossiannilssoni	0.76	+	+									
Macrosteles sexnotatus	0.68		+									
Myrmeleotettix maculatus	0.77	+										
Neophilaenus minor	0.76	++	++									
Philaenus spumarius	0.65		+									
Psammotettix excisus	0.94											
Psammotettix nodosus	0.76	+										
Rhopalopyx vitripennis	0.85	++	++									
Ribautodelphax collina	0.72	+	+									
# of successful transfers	28/4/19			9/2/9			9/1/8			9/3/6		

Discrimination ('discr.'): threshold independent c-index (first column, non-significant values marked grey), threshold dependent measures with P_{Kappa} (second column), and with pbp (third column). + indicates that all four criteria (sensitivity, specificity, CCR, Kappa) exceed chance. Calibration ('cal.'): intercept (bias, first column) and slope (spread, second column), + indicating no significant deviation from 0 (intercept) and 1 (slope).

Table 3 – Overview of model discrimination for training data ('Train.') and test data ('Test'), and difference ('Diff.') of model performance between training and test data

Transfer	c-index				Kappa (p_{crit} = pbp)			
	#	Train.	Test	Diff.	#	Train.	Test	Diff.
Bremen'03 → Bremen'04								
Med.	28	0.89	0.76	−0.11	19	0.64	0.41	−0.19
Min.		0.81	0.60	−0.30		0.46	0.25	−0.42
Max.		0.98	0.96	0.05		0.69	0.58	−0.07
Bremen'03 → Berlin								
Med.	9	0.89	0.85	−0.04	9	0.64	0.41	−0.25
Min.		0.82	0.72	−0.19		0.46	0.24	−0.42
Max.		0.91	0.90	0.03		0.68	0.57	0.08
Berlin → Bremen'03								
Med.	9	0.93	0.72	−0.17	8	0.68	0.34	−0.33
Min.		0.83	0.63	−0.31		0.58	0.15	−0.66
Max.		0.94	0.85	−0.06		0.81	0.60	−0.07
Berlin → Bremen'04								
Med.	9	0.93	0.72	−0.22	6	0.71	0.34	−0.33
Min.		0.83	0.61	−0.31		0.58	0.18	−0.63
Max.		0.94	0.85	−0.06		0.81	0.46	−0.18

Median, minimum and maximum values for c-index and Kappa (threshold = pbp). Only successfully transferred models are presented (# = number of models).

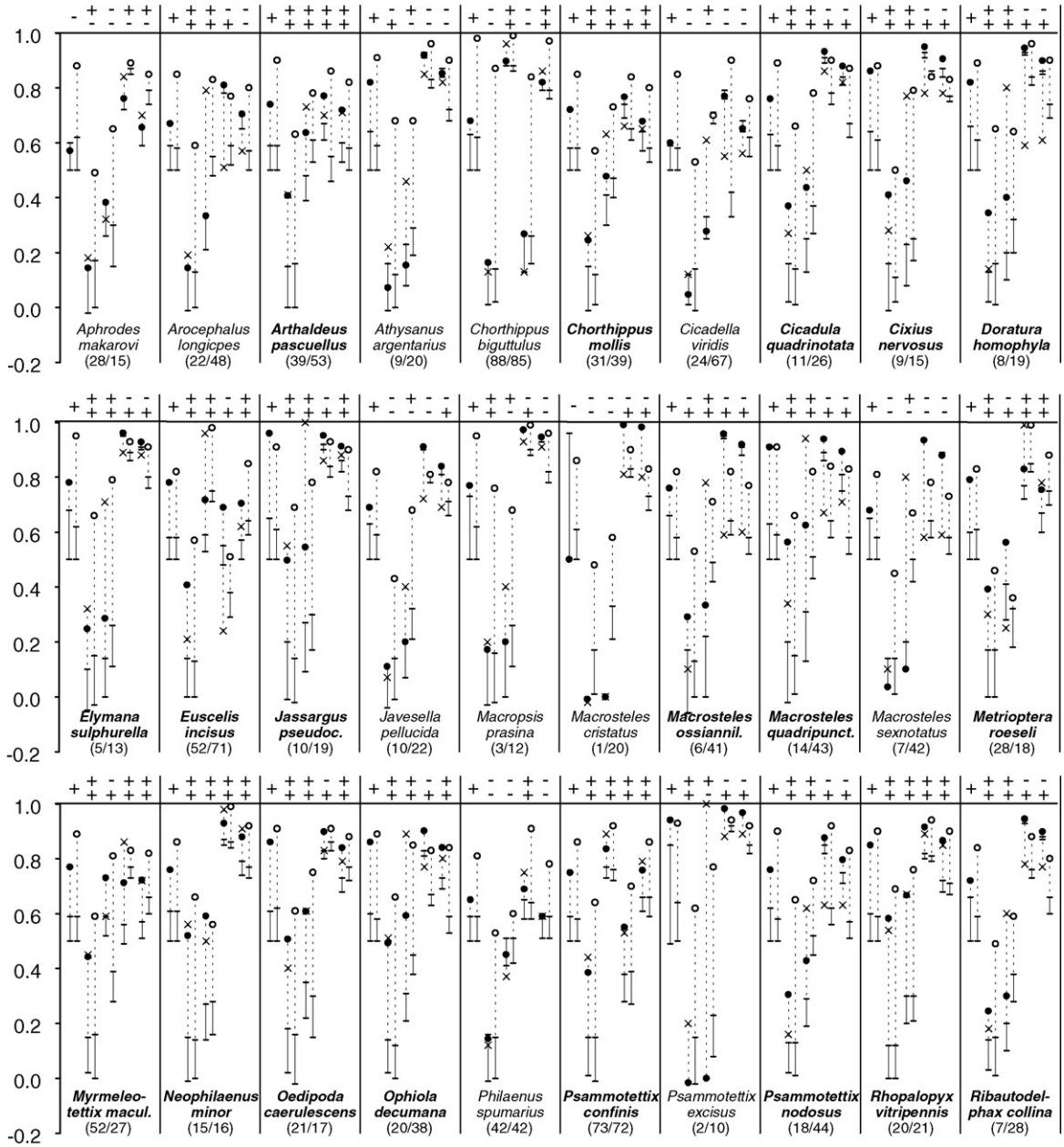


Fig. 2 – Model transfer in time. Discriminatory ability assessed by the threshold independent *c*-index, and threshold dependent Kappa, sensitivity, specificity and CCR. Measures for model transfer (x for P_{Kappa}, black dots for pbp), and for original models (open circles). Chance distributions (50–95%-percentiles) for each measure are indicated by black bars. +/− on top of each measure indicate whether the model transfer is better than chance predictions. The upper row represents P_{Kappa}, the lower pbp. Numbers under the species names give prevalences for test/training data. See text for further explanations.

of Berlin models to Bremen 2003 data (−0.17) and Bremen 2004 data (−0.22) (Table 3). For the Bremen'03 → Berlin'04 transfer, all species with significant *c*-index also reached significant dichotomised predictions with pbp as a threshold, even though Kappa decrease was considerable (median difference −0.25). For the Berlin'04 → Bremen'03 transfer, 8 species reached significant 0/1 predictions, for Berlin'04 → Bremen'04, 6 species. P_{Kappa} of the original models performed poorly as a threshold, even though for one species (*Athysanus argentarius*) it performed better than pbp.

Like discrimination, calibration of the Bremen → Berlin transfer was better than vice versa. In the first case, six models showed no significant spread, in the latter one model (2003 data) and three models (2004 data). *D. homophyla* calibration curves had slopes < 1 in two cases. This enhances that models for this species could not be transferred in space, which had already been indicated by the lack of discriminatory power. Most models showed bias with the test data, reflecting differences in prevalence. *C. mollis* is an example for the resulting consistent underestimate of occurrence

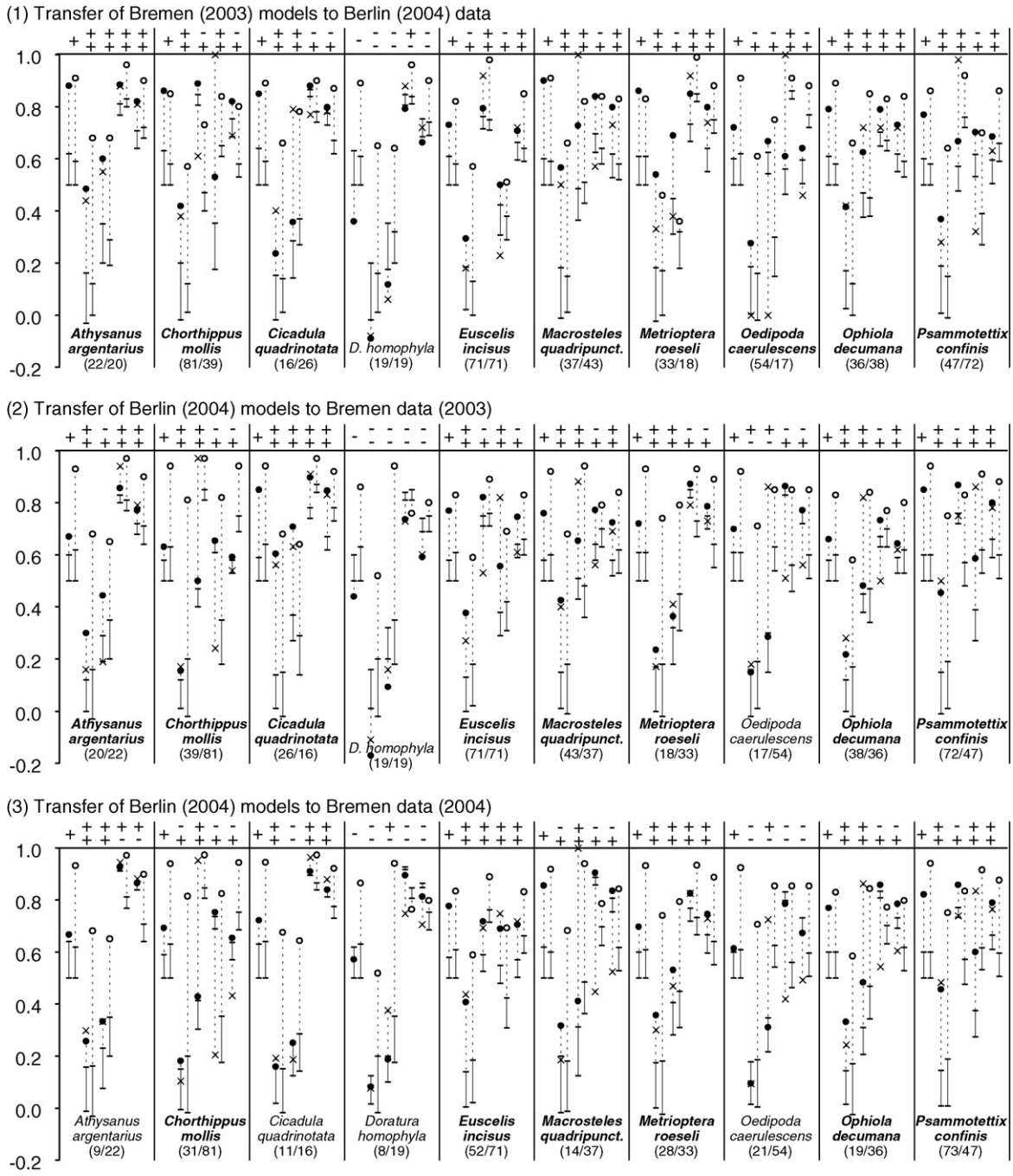


Fig. 3 – Model transfer in space. For details, see Fig. 2.

probabilities in case of prevalence increase (Fig. 1). Only two models were overall well calibrated to the test data for the Bremen'03 → Berlin'04 transfer (*Athysanus argenarius, Ophiola decumana*), none for Berlin'04 → Bremen'03, and one model for Berlin'04 → Bremen'04 (*Euscelis incisus*).

3.3. *Qualitative comparison of models*

The qualitative comparison of Berlin and Bremen models revealed that, regarding the main driving factors, only vegetation structure appeared in all model pairs (Fig. 4). Age was in both models for three species, landscape context for five and soil parameters for four. Since soil parameters in the models differed between the two study area models for all species, comparison was not possible for these parameters. Investigation of the shape of response curves showed no case of opposite influence of one parameter on a species in either study area. In some cases however (e.g. vegetation height for *Cicadula quadrinotata*), one relationship was unimodal and the other monotonic. Overall, vegetation parameters largely agreed between models, even though variable weights usually differed. In contrast to this, for landscape context, mostly different parameters entered into the models.

	A. argent.		*C. mollis*		*S. quadrin.*		*D. homoph.*		*E. incisus*		*M. quadrip.*		*M. roeseli*		*O. caerulesc.*		*O. decum.*		*P. confinis*	
	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br	Be	Br
site age	**0**	**0.01** +	**0**	**22** u	**15** +	**0.2** u	**0**	**0**	**0**	**0**	**0.3** -	**0.02** -	**6** u	**0**	**0.002** -	**0**	**0**	**20** -	**0.3** u	**38** u
vegetation structure	**36**	**69**	**61**	**49**	**85**	**71**	**50**	**30**	**100**	**77**	**65**	**36**	**51**	**76**	**54**	**4**	**100**	**30**	**98**	**39**
density	0.09 +	5 +	22 u	25 u	16 u/+	12 +	0	0	0	6 u	24 -	5 -	8 u/+	31 +	54 u/-	4 u/-	0	11 u/-	29 u/-	37 u
height	2 u/+	13 u/+	13 u	0.01 u	14 u	17 +	50 u	2 u	50 +	28 u	3 -	0	21 u/+	5 +	0.05 u/-	9E-07 u	0	0.02 u	21 u/-	1 u
host plants	2 +	0.5 +	0	0	38 +	29 +	0	0	50 +	33 +	0	0	0	0	0	0	0	0	0	0
moss cover	0	25 +	0	0	0	0	0	28 +	0	0.51 u	8 -	9 -	3.2 +	0	0	0	100 -	17 -	0	0
litter cover	30 +	8 +	0	0	0	12 +	0	0	0	0	30 -	21 -	14 +	14 +	0.01 -	0.14 u	0	1 -	0.09 u	0.008 u
bare soil	1 -	17 -	0	24 -	18 -	0.8 -	0	0	0	10 u	0.03 +	0.08 u	4.1 -	26 -	0.003 u/-	0	0	0.02 u	0	0
landscape context	**32**	**31**	**30**	**29**	**0**	**29**	**0**	**26**	**0**	**23**	**35**	**37**	**25**	**0**	**45**	**54**	**0**	**23**	**2E-04**	**24**
open brownfields	0.08 -	0.1 -	30 -	0.03 -	0	0	0	3 -	0	0	0.01 +	0.006 u	20 -	0	0	5 -	0	0	0	0.004 -
grassy brownfields	0	30 +	0	4 u	0	21 u	0	0	0	23 u	2 -	37 u/-	4 +	0	0	0	0	0	0	24 +
herbaceous brownfields	32 u/+/-	0.04 -	0	0	0	0	0	0	0	0	33 -/+	0.09 +	2 -	0	44 -	5 -/+	0	12 +	2E-04 u	0
moist/wet brownfields	0	0	0	25	0	7	0	23	0	0	0	0	0	0	0	44	0	11	0	0
soil parameters	**32**	**0**	**9**	**0**	**0**	**0**	**50**	**44**	**0**	**0**	**0.02**	**27**	**18**	**24**	**1**	**42**	**0**	**27**	**2**	**0**

Fig. 4 – Qualitative comparison of models based on Berlin (Be) vs. Bremen (Br) data. Main driving factors in bold. Numbers indicate variable weight [%], symbols illustrate the functional form of the relationship between species and variable: (u) unimodal (bell shaped); (–) monotonic, decreasing; (+) monotonic, increasing. If a species showed different reactions to one variable complex (e.g. vegetation density: + for density at the 0–5 cm height, u for density at the 25–50 cm height), all are listed. Black frames indicate a variable or variable complex to be present in both models, grey background indicates presence in only one model. No details are given for soil parameters because no species reacted to the same soil parameters in both models.

4. Discussion

4.1. *Model discrimination – c-index*

It seems that the majority of habitat models tested in this study are general in the sense that they successfully rank test sites from suitable to unsuitable. Models for both taxa (grasshoppers and leafhoppers) could be transferred equally well. However, the limited number of grasshoppers (five species) does not allow to detect possible differences between the taxa. Fifty-seven of 60 transfers achieved a *c*-index exceeding chance. Forty-five transfers (75%) reached *c*-values $\geq$ 0.70. This is considered as 'good' discrimination by Hosmer and Lemeshow (2000), and Randin et al. (2006) require a *c*-index of 0.7 for model transferability. In comparison to our results, Bulluck et al. (2006) found only 56% of their breeding bird models to reach *c*-indexes $\geq$ 0.70 when transferred to new data in time or space. Randin et al. (2006) achieved sufficient spatial transferability for less than half of their models for alpine plant species. This suggests our models to be robust and general, indicating that model averaging might lead to more stable models than stepwise procedures. The modelling technique showed an influence on model transferability in previous studies (Randin et al., 2006; Araujo and Guisan, 2006). However, a clearly superior method leading to transferable models has not yet been identified.

Despite the overall good transferability, model transfer mostly went along with a loss of accuracy. This loss was not necessarily larger for transfer in space than for transfer in time, since models could be transferred better from Bremen to Berlin than from Bremen 2003 to Bremen 2004 regarding the *c*-index. Bulluck et al. (2006) found in their study as well that some transfers in space worked better than those in time. Possibly, the Bremen 2004 test data were unusual, being affected by the exceptionally hot and dry summer of 2003. This assumption is supported by the fact that Berlin 2004 models could be transferred more successfully to Bremen 2003 than to Bremen 2004 data. Thus, transfer in space and time did not lead to poorer results than transfer in space or time only. Jensen et al. (2005), who extensively tested model transfer in time for the blue crab *Callinectes sapidus*, found that some years showed unique habitat relationships that were not well predicted by models from the other years. The comparison between the years showed that there can be enormous differences in species prevalence, particularly in dynamic habitats like brownfields. Predictive habitat models are generally static (Bulluck et al., 2006; Guisan

and Zimmermann, 2000), i.e. not explicitly considering population dynamics or dispersal. Even though modelling species in disequilibrium using such static models is problematic (Gibson et al., 2004a), and theories based on equilibrium might be inadequate for urban communities (Rebele, 1994), models from 1 year were mostly valid in the next and in another region.

A closer look at which species models transferred worse than others reveals that low *c*-values might be associated with eurytopic species. None of the species considered eurytopic (*A. makarovi*, *Chorthippus biguttulus*, *Javesella pellucida* and *Philaenus spumarius*) (Detzel, 1998; Nickel, 2003) reached a *c*-index ≥ 0.7. It seems reasonable that habitat generalists do not exhibit strong species–habitat relationships. If strong relationships are found for such species, they might be an artifact within a particular dataset. In fact, none of these species, though present, had shown significant relationships in the Berlin dataset. Investigations on the relation between species properties (biological traits) and model transferability might be able to reveal more general patterns (Randin et al., 2006).

Model transfer from Bremen to Berlin worked better than vice versa. Randin et al. (2006) suggest such asymmetrical transferability to be caused by differences in the width of environmental ranges or in species abundances. In our case, however, environmental ranges had comparable widths, and abundance differences did not seem to produce particularly asymmetric values for the *c*-index. Thus, it seems likely that the Bremen models, based on a larger dataset of 147 plots, were more general than the Berlin ones (based on 89 plots). Harrell et al. (1984) found that smaller training samples had an apparent higher quality, but a large loss in quality when applied to test data. The opposite was true for large training samples. McPherson et al. (2004) obtained best models for very large sample sizes (300–500), Pearce and Ferrier (2000a) recommend sample sizes of > 250. In this light, the Berlin dataset in particular might have been too small to build general models. Considering these results it seems desirable to base models on large datasets. On the other hand, test datasets as well require a certain size, Vaughan and Ormerod (2005) suggest 200. This is particularly important for rare species, since otherwise the *c*-index cannot be calculated reliably (McPherson et al., 2004). Pearce et al. (2001) required sufficient evaluation data to have at least nine species records. Since sample sizes for labor intensive field data are usually restricted by logistic constraints, available money and manpower, it will be difficult to follow these recommendations in practice. Our results indicate that also sample sizes of 150 lead to general models, even though larger samples might allow even better results.

An interesting finding was that those species that made models in both study areas could be transferred better in time than the others. The fact that they exhibited statistically strong relationships to the measured environmental factors in both regions might indicate that they show stable, general relationships to these factors.

4.2. *Model calibration*

Models applied to new data hardly ever showed good calibration. Considerable bias was fully expected, since prevalences between training and test data differed on a large scale (Pearce and Ferrier, 2000b; Vaughan and Ormerod, 2005). Bremen models applied to Berlin showed less spread than the other way round. Since spread indicates overfitting, this is another hint that Berlin models might have had a stronger tendency to be overfitted to their small dataset.

The consequence of these findings is that predicted occurrence probabilities cannot be used in a quantitative way since they do not express the true probability of a site as being occupied. Sites are ranked according to their relative probability of being occupied, thus predictions are ordinal rather than quantitative. They should be displayed as ranked categories to avoid quantitative interpretation (Vaughan and Ormerod, 2005). If poor calibration of a model is due to a subset of plots for which the model can not be transferred well, such plots can be identified and restrictions placed on the model's use (Miller et al., 1991). In general, intercept and slope of the calibration curve can be used to adjust model predictions (Steyerberg et al., 2003). Such fine-tuning leads to a better model adjustment to the local circumstances of the test data, but not necessarily to a more general model. Thus, this method should be used with caution (Miller et al., 1991).

4.3. *Dichotomising predictions*

Converting occurrence probabilities into presences/absences raises the problem of finding an optimal threshold. When applying models to new environmental data, this problem cannot be overcome without information of the species' prevalence. We clearly showed that using the training data's threshold was doomed to failure since models were mostly poorly calibrated to new data. This resulted in consistent over- or underestimation of occurring probabilities, making the original threshold useless. In some studies, a new optimal threshold was calculated for the test data (e.g. Eyre et al., 2005; Schröder and Richter, 1999/2000). This allows assessment of model transfer to the test data. However, it does not give any hint on what threshold should be used with new data, where true species presence/absence is not known, but to be predicted with the model. If it is possible to gain information on the species prevalences within the area where a model is to be applied, pbp is a promising alternative to defining thresholds, at least for the majority of species. Before relying on it, this should probably be verified with more than one set of test data. If information on prevalence is not accessible, there does not seem to be much point in dichotomising occurrence probabilities since misclassifications are likely.

4.4. *Qualitative comparison*

Comparing the models for Bremen and Berlin allowed a deeper insight into the question why model transfers might succeed or fail. Parameters contained in both models are likely to have a stable relationship to the species' presence and can probably be generalised. Parameters that were in only one of the models might have an indirect influence. The relationship to the underlying direct variable might not be the same in other datasets (Vaughan and Ormerod, 2003). This seems to be the case with all soil parameters. Even though they might have had considerable influence in one model, the same param-

eters never went in the model for the other study area. In the case of *Doratura homophyla*, this led to a model that could be transferred in time (Bremen) but not in space. Landscape context also showed large differences between models. Fisher et al. (2005) note that research has rarely been undertaken to test the assumption that the response to landscape structure from one area can be extrapolated to another. Purtauf et al. (2005) believe that there is a high risk of artificial correlations in hierarchical multi-scale landscape analyses when ecological data are related to the landscape context. Thus, it might well be that species–landscape context relationships are region specific. In our study this might stem from the fact that distribution of herbaceous and grassy brownfields differed considerably between the study areas, and moist to wet brownfields were not present in Berlin. Overall, it seems likely that the vegetation structure variables have a more direct relationship to species occurrences than the other variables (Strauss and Biedermann, 2006). Thus, their influence was comparable between the study areas. This highlights that for generalizations on the species–environment relationship as well as for model transfer, models basing on direct parameters are more suitable.

The overall trend that there were more variables in the Bremen models was probably caused by the larger dataset. With more data, variables more easily exceeded the significance level. Therefore, differences between models did not necessarily result in poor transferability and might mainly be due to statistical reasons during the model building process. Reactions to vegetation structure were considerably similar in both cities. Since vegetation structure was the most important driving factor, this probably enabled the good overall transferability. Some species exhibited responses that seemed to be relocated (monotonic to unimodal and vice versa) between the study areas. Even though this could be caused by incomplete stratification not covering the whole gradient (Vaughan and Ormerod, 2003), it seems more likely that: (1) there was a true difference in species reactions between the oceanic Bremen and the more continental Berlin or (2) differences were due to differences in data distributions. The ranges of values (minimum/maximum values) were comparable for all variables between the study areas, but data distribution within the range often differed, in particular for the landscape context variables.

In this context, when applying models in nature conservation practice, one has to keep in mind that models that proofed to be transferable are only valid for data ranges present in the test data until the model is tested under different conditions (Vaughan and Ormerod, 2003). Therefore, test sites should be carefully selected, representing the full range of environmental conditions present in the training data. This requirement was met within our study, where all plots had been chosen in a random-stratified design, covering the whole gradient of urban brownfield stages.

5. Conclusions

The vast majority of models tested in this study turned out to be transferable to new data from different years and different regions. However, some models could not be transferred at all in time or space. This implies that generality always needs to be tested if inference about general relationships is to be drawn or models are to be applied on independent data. Both temporal and spatial transferability should be tested, since single years may exhibit unusual relationships. Certain factors seem to enhance model generality: (1) Large sets of training data. (2) Strong influence of direct variables within models. (3) Species are not eurytopic. (4) Species show significant relationships to environmental variables in more than one study area/dataset. It is likely that these general findings also hold for other taxa.

Model accuracy usually decreases with model transfer. Thus, models that do not fit their training data well should not be transferred. On the other hand, well fitting models do not necessarily transfer well. In most cases, model transfer leads to poor calibration. Predicted occurrence probabilities can therefore not be used quantitatively and should not be presented as such, but as ordinal information on habitat quality.

Dichotomisation of predictions should be avoided without information about species' prevalences. With prevalence information available, the prevalence based proportion of highest probabilities (pbp) allows classification with reasonable accuracy.

Acknowledgements

This study was conducted as part of the TEMPO-project and was financially supported by the German Ministry of Education and Research (BMBF, grant 01LM0210). We thank Ute Schadek for providing soil and plant composition data and Nora Lange for providing leafhopper data and a habitat type map for Berlin.

REFERENCES

Altman, D.G., Royston, P., 2000. What do we mean by validating a prognostic model? Stat. Med. 19 (4), 453–473.

Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33 (10), 1677–1688.

Bulluck, L., Fleishman, E., Betrus, C., Blair, R., 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. Global Ecol. Biogeogr. 15 (1), 27–38.

Burnham, K., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second ed. Springer, New York.

Detzel, P., 1998. Die Heuschrecken Baden-Württembergs. Ulmer, Stuttgart.

Eyre, M.D., Rushton, S.P., Luff, M.L., Telfer, M.G., 2005. Investigating the relationships between the distribution of British ground beetle species (Coleoptera, Carabidae) and temperature, precipitation and altitude. J. Biogeogr. 32 (6), 973–983.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence–absence models. Environ. Conserv. 24 (1), 38–49.

Fisher, J.T., Boutin, S., Hannon, S.J., 2005. The protean relationship between boreal forest landscape structure and red squirrel distribution at multiple spatial scales. Landsc. Ecol. 20 (1), 73–82.

Fleishman, E., Mac Nally, R., Fay, J.P., 2003. Validation tests of predictive models of butterfly occurrence based on environmental variables. Conserv. Biol. 17 (3), 806–817.
Gibson, L.A., Wilson, B.A., Cahill, D.M., Hill, J., 2004a. Modelling habitat suitability of the swamp antechinus (*Antechinus minimus maritimus*) in the coastal heathlands of southern Victoria, Australia. Biol. Conserv. 117 (2), 143–150.
Gibson, L.A., Wilson, B.A., Cahill, D.M., Hill, J., 2004b. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. J. Appl. Ecol. 41 (2), 213–223.
Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135, 147–186.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.
Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., 1984. Regression modeling strategies for improved prognostic prediction. Stat. Med. 3 (2), 143–152.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second ed. Wiley, New York.
Jensen, O.P., Seppelt, R., Miller, T.J., Bauer, L.J., 2005. Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. Mar. Ecol. Progr. Ser. 299, 239–255.
Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28 (3), 385–393.
Manly, B.F., 2001. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman and Hall, London.
McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41 (5), 811–823.
Miller, M.E., Hui, S.L., Tierney, W.M., 1991. Validation techniques for logistic-regression models. Stat. Med. 10 (8), 1213–1226.
Nickel, H., 2003. The Leafhoppers and Planthoppers of Germany (Hemiptera, Auchenorrhyncha). Patterns and Strategies in a Highly Diverse Group of Phytophagous Insects. Series Faunistica 28. Pensoft, Sofia.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. Trans. Am. Fisheries Soc. 131 (2), 329–336.
Pearce, J., Ferrier, S., 2000a. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecol. Modell. 128 (2–3), 127–147.
Pearce, J., Ferrier, S., 2000b. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 224–245.
Pearce, J., Ferrier, S., Scotts, D., 2001. An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. J. Environ. Manage. 62 (2), 171–184.
Purtauf, T., Thies, C., Ekschmitt, K., Wolters, V., Dauber, J., 2005. Scaling properties of multivariate landscape structure. Ecol. Indicators 5 (4), 295–304.
Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33 (10), 1689–1703.
Rebele, F., 1994. Urban ecology and special features of urban ecosystems. Global Ecol. Biogeogr. Letters 4 (6), 173–187.
Schröder, B., Richter, O., 1999/2000. Are habitat models transferable in space and time? Zeitschrift für Ökologie und Naturschutz 8, 195–205.
Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobbee, D.E., Moons, K.G.M., 2003. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J. Clin. Epidemiol. 56 (5), 441–447.
Strauss, B., Biedermann, R., 2006. Urban brownfields as temporary habitats: driving forces for the diversity of phytophagous insects. Ecography 29, 928–940.
Vaughan, I.P., Ormerod, S.J., 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conserv. Biol. 17 (6), 1601–1611.
Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution models. J. Appl. Ecol. 42 (4), 720–730.