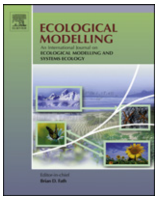
Ecological Modelling 221 (2010) 1995–2002

Contents lists available at ScienceDirect

Ecological Modelling

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



Review

Ecological relevance of performance criteria for species distribution models

Ans M. Mouton[a,b,c,*], Bernard De Baets[b], Peter L.M. Goethals[a]

[a] Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium
[b] KERMIT: Research Unit "Knowledge-based Systems", Ghent University, Coupure links 653, B-9000 Ghent, Belgium
[c] Research Institute for Nature and Forest, Kliniekstraat 25, 1070 Brussels, Belgium

ARTICLE INFO

Article history:
Received 5 November 2009
Received in revised form 23 April 2010
Accepted 27 April 2010
Available online 27 May 2010

Keywords:
Prevalence
Omission/commission errors
CCI
Kappa
TSS

ABSTRACT

Species distribution models have often been developed based on ecological data. To develop reliable data-driven models, however, a sound model training and evaluation procedures are needed. A crucial step in these procedures is the assessment of the model performance, with as key component the applied performance criterion. Therefore, we reviewed seven performance criteria commonly applied in presence–absence modelling (the correctly classified instances, Kappa, sensitivity, specificity, the normalised mutual information statistic, the true skill statistic and the odds ratio) and analysed their application in both the model training and evaluation process. Although estimates of predictive performance have been used widely to assess final model quality, a systematic overview was missing because most analyses of performance criteria have been empirical and only focused on specific aspects of the performance criteria. This paper provides such an overview showing that different performance criteria evaluate a model differently and that this difference may be explained by the dependency of these criteria on the prevalence of the validation set. We showed theoretically that these prevalence effects only occur if the data are inseparable by an *n*-dimensional hyperplane, *n* being the number of input variables. Given this inseparability, different performance criteria focus on different aspects of model performance during model training, such as sensitivity, specificity or predictive accuracy. These findings have important consequences for ecological modelling because ecological data are mostly inseparable due to data noise and the complexity of the studied system. Consequently, it should be very clear which aspect of the model performance is evaluated, and models should be evaluated consistently, that is, independent of, or taking into account, species prevalence. The practical implications of these findings are clear. They provide further insight into the evaluation of ecological presence/absence models and attempt to assist modellers in their choice of suitable performance criteria.

Contents

1. Introduction

In past decades, species distribution models have increasingly received attention due to their wide management applications in the context of biogeography, conservation biology and climate

[*] Corresponding author at: Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium. Tel.: +32 9 264 39 96; fax: +32 9 264 41 99.
E-mail address: Ans.Mouton@UGent.be (A.M. Mouton).

0304-3800/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2010.04.017

change studies (Guisan and Theurillat, 2000; Guisan and Thuiller, 2005; Araújo and Rahbek, 2006; Kerr et al., 2007). Concurrent with numerous tests, new applications and reviews (Guisan and Zimmermann, 2000; Austin, 2007; Meynard and Quinn, 2007), there has been careful attention to the algorithms and methods used, including comparisons of the relative performance of different methods (Hirzel et al., 2001; Elith et al., 2006; Heikkinen et al., 2007; Meynard and Quinn, 2007; Peterson et al., 2007).

A crucial step in the model comparison procedure is the assessment of the model performance (Fielding and Bell, 1997; Manel et al., 1999; McPherson and Jetz, 2007). Most authors refer to this step as model evaluation (Boyce et al., 2002; Anderson et al., 2003; Barry and Elith, 2006; Guisan et al., 2007), hereby situating the model evaluation procedure at the end of the modelling process. However, model performance is also assessed during the model development process to compare trained models and select the best performing models (Hastie et al., 2001; Van Broekhoven et al., 2006; Mouton et al., 2009b). To avoid misleading terminology in this paper, the term model *evaluation* will refer to performance assessment of the final model, whereas *training* performance assessment will refer to model performance assessment during model training.

The key component of model performance assessment is the performance criterion applied to quantify model performance. Since Fielding and Bell (1997) reviewed the performance criteria most commonly applied in conservation presence/absence methods, the performance assessment of models developed from presence–absence data has been a recurrent focus (Pearce and Ferrier, 2000; Manel et al., 2001; Nielsen et al., 2005; Vaughan and Ormerod, 2005; Allouche et al., 2006). Estimates of predictive performance have been widely applied to assess final model quality, especially in papers that compare the relative performance of different methods (Hirzel et al., 2001; Elith et al., 2006; Heikkinen et al., 2007; Meynard and Quinn, 2007; Peterson et al., 2007). Manel et al. (2001) reviewed a sample ($n = 87$) of published ecological literature between 1989 and 1999, which revealed that many users of presence–absence models made no evaluation at all, even in leading ecological journals.

We aim to review the performance criteria most commonly applied in presence–absence modelling and to analyse their application in both the model training and the model evaluation process. We review the role of performance criteria in both processes and analyse the role of performance criteria in model training theoretically. Although we focus on presence–absence models, we will also discuss some concepts and problems that occur across ecological modelling in general. Finally, we give recommendations on the application of different performance criteria on model evaluation and training.

2. Performance criteria in presence-absence modelling

The key component of the model training and validation procedures is the performance criterion which evaluates the model performance. Performance criteria can deal with either continuous or discrete model outputs, or with both. If a model generates discrete predictions, these outputs can be summarised in a confusion matrix (Fielding and Bell, 1997; Manel et al., 2001) which compares the model predictions to the observations. Several performance criteria have been derived from this confusion matrix and are listed in Table 1, including overall predictive accuracy or the percentage of correctly classified instances (CCI; Buckland and Elston, 1993; Fielding and Bell, 1997), sensitivity, specificity, the normalised mutual information statistic (*NMI*; Forbes, 1995), the odds ratio (Fielding and Bell, 1997), Kappa (Cohen, 1960) and the true skill statistic (TSS; Allouche et al., 2006). An optimal odds ratio may reach positive infinity, whereas all other criteria are optimal at their maximum, one.

Table 1
Measures of predictive accuracy calculated from the confusion matrix. The percentage of correctly classified instances (*CCI*) is the rate of correctly classified cells. Sensitivity (*Sn*) is the probability that the model will correctly classify a presence. Specificity (*Sp*) is the probability that the model will correctly classify an absence. *NMI* quantifies the information included in the model predictions compared to that included in the observations. The *Kappa* statistic and *TSS* normalise the overall accuracy by the accuracy that might have occurred by chance alone. The odds ratio is the ratio of correctly assigned cased to the incorrectly assigned cases. In all formulae $n = a + b + c + d$.

Measure	Formula
CCI	$\frac{a+d}{n}$
Sn	$\frac{a}{a+c}$
Sp	$\frac{d}{b+d}$
NMI	$1-\frac{-a\ln(a)-b\ln(b)-c\ln(c)-d\ln(d)+(a+b)\ln(a+b)+(c+d)\ln(c+d)}{n\ln(n)+(a+c)\ln(a+c)+(b+d)\ln(b+d)}$
Kappa	$\frac{((a+d)/n)-((a+b)(a+c)+(c+d)(d+b)/n^2)}{1-((a+b)(a+c)+(c+d)(d+b)/n^2)}$
Odds ratio	$\frac{ad}{cb}$
TSS	$Sn+Sp-1$

The most popular measure for the accuracy of presence–absence predictions is Cohen's *Kappa* (Manel et al., 2001; Loiselle et al., 2003; Petit et al., 2003; Berg et al., 2004; Parra et al., 2004; Pearson et al., 2004; Rouget et al., 2004; Segurado and Araújo, 2004; Allouche et al., 2006). This measure allows an assessment of the extent to which models correctly predict occurrence at rates that are better than chance expectation (Fielding and Bell, 1997). However, *Kappa* has been criticised based on statistical theoretical grounds (Sprott, 2000). Several authors argued that *Kappa* may be less appropriate for model evaluation due to its dependence on the proportion of sites in the training dataset at which a species was recorded as present, hereafter defined as prevalence (Fielding and Bell, 1997; Manel et al., 2001; Allouche et al., 2006). Three performance criteria were proposed to avoid this problem because they were assumed to be independent of prevalence (Manel et al., 2001; Allouche et al., 2006): the *NMI* (Forbes, 1995; Fielding and Bell, 1997), *TSS* (Allouche et al., 2006) and *AUC* (Fielding and Bell, 1997; Manel et al., 2001).

All performance criteria which were developed to evaluate discrete model predictions can also handle continuous predictions since the latter can be discretised using threshold values.

Although numerous threshold selection methods were suggested (Manel et al., 1999, 2001; Liu et al., 2005; Jiménez-Valverde and Lobo, 2007), the choice of an appropriate threshold often remains difficult and arbitrary (Fielding and Bell, 1997; Manel et al., 1999, 2001; Liu et al., 2005). Moreover, selection of a threshold often depends on the conservationist's preferences and can significantly affect reserve selection for conservation planning (Liu et al., 2005; Wilson et al., 2005). Therefore, some criteria, such as the average deviation (*AD*; Van Broekhoven et al., 2007), do not require this arbitrary threshold to process continuous data.

The receiver operator characteristic (ROC; Fielding and Bell, 1997) approach is another alternative method for assessing the accuracy of probabilistic output models. The area under the ROC curve (*AUC*) is often used as a single threshold-independent measure for model performance (Manel et al., 2001; Thuiller et al., 2003, 2005; McPherson et al., 2004). *AUC* was shown to be independent of the prevalence, and is an effective measure of discriminatory ability for probabilistic models (Vaughan and Ormerod, 2005). Consequently, some authors considered *AUC* to be the current best practice for assessing model success for presence/absence data (Pearce and Ferrier, 2000; Thuiller et al., 2003; Rushton et al., 2004; Austin, 2007). However, the *AUC* approach cannot be applied to dichotomous presence–absence model outputs (Allouche et al., 2006). Moreover, it could be shown that models with the same

Table 2
The number of validation criteria used in the model evaluation process. A sample of 385 papers on species distribution modelling listed in the Web of Science was evaluated, of which 67% used data for model training or evaluation. The numbers in the table are percentages of this group ($n = 257$). The papers which applied 0 performance criteria did not evaluate model performance.

	Number of performance criteria applied in model evaluation						Total
	0	1	2	3	4	5	
1998–2002	2	18	61	8	8	3	15
2003–2007	3	15	64	12	4	2	85
Total	3	15	64	11	5	2	100

Taxonomic groups most frequently involved in presence–absence models were trees and other angiosperms (31%), birds (26%), invertebrates (21%), mammals (17%), fish (9%), reptiles (4%) and amphibians (2%). Less frequent applications were related to bacteria, plankton and fungi. Model and optimisation techniques included regression techniques such as GLM or GAM (84%), classification trees (17%), artificial neural networks (14%), genetic algorithms (13%) and other methods (12%).

or very similar *AUC* values may predict very different distribution patterns. Finally, Maggini et al. (2006) found that the *AUC* is systematically lower at extreme prevalence values (prevalence <0.05 or >0.70). The *AUC* appears to be independent of prevalence only in its middle range (Maggini et al., 2006; McPherson and Jetz, 2007). Consequently, reliance on *AUC* as a sufficient test of model success needs to be re-examined (Termansen et al., 2006; Austin, 2007; Lobo et al., 2008).

Several authors suggested some desirable properties of accuracy statistics for the assessment of species distribution model performance (Forbes, 1995; Fielding and Bell, 1997; Vaughan and Ormerod, 2005). For Vaughan and Ormerod (2005), the most important is generality, defined as the ability to compare accuracy meaningfully between the same model in different applications or between models developed for different species or with different training and test data. As a consequence of such definition, from the evaluation perspective, a suitable performance criterion should be independent of the prevalence of the data to which the criterion is applied. This limitation of generality attempts to avoid that two identical models which are evaluated on two different datasets, would show different model performance. Vaughan and Ormerod (2005) suggest that, if performance criteria values are affected by prevalence, the criteria could also be corrected for this prevalence to assure generality.

Another desirable property of performance criteria is the ability to weigh a consistent under- or overestimation of the species prevalence, which is also referred to as omission or commission error (Rondinini et al., 2006) or as false-negative or false-positive error (Loiselle et al., 2003), respectively. Several authors have addressed this issue in ecological modelling and emphasised the significant negative correlation between both errors (Fielding and Bell, 1997; Anderson et al., 2003; Rondinini et al., 2006; Lobo et al., 2010). Consequently, performance criteria should allow species distribution modellers to choose between both errors. Knowledge on the data used to develop the model may substantially influence this choice because these data contain either more commission or omission errors (Loiselle et al., 2003; Wilson et al., 2005; Austin, 2007). Other desirable characteristics of performance criteria that are directly or indirectly related to the generality or to the ability to distinguish between omission and commission errors are listed in Table 8.

3. Application of performance criteria for model evaluation

The comparison of model performance is based on the assessment of the performance of the final model which is obtained after model training. Most authors refer to this step as model evaluation, although the terms model testing and model validation are also being used (Anderson et al., 2003; Vaughan and Ormerod, 2005). Both latter terms are less appropriate to designate model performance assessment because they may overlap with other steps in the modelling process. In neural network applications for instance, model testing is applied to stop the supervised learning procedure and to avoid overfitting of the training data, which occurs when idiosyncrasies in the training set are modelled in addition to the underlying species-environment relationship (Lek and Guégan, 1999). Model validation implies the quantification of the model performance on an independent dataset. Ideally, this dataset should be completely independent from the data used to train or calibrate the model, e.g. collected on other areas (Fielding and Bell, 1997; Hastie et al., 2001).

In recent years, model evaluation has increasingly received attention in species distribution modelling. The aforementioned sample of ecological literature between 1989 and 1999 (Manel et al., 2001), revealed that only 52% of the modellers ($n = 87$) evaluated model performance. In this paper, we evaluated a sample of ecological literature on presence-absence or presence-only models listed in the Web of Science ($n = 385$). This not only indicated that the number of papers on species distribution modelling increased significantly, but also that 97% of the model users evaluates model performance (Table 2). Araújo et al. (2005) found similar results for a sample of species-climate envelope models under climate change between 1995 and 2004 ($n = 29$), of which 93% was evaluated. Although 82% of the modellers applied two or more criteria, the percentage of papers applying at least two performance criteria has only increased slowly. Table 3 summarizes the performance criteria used in the evaluated sample of ecological literature ($n = 385$), if any. Almost all papers use *CCI*, but the application of *Kappa* and *AUC* is increasing, with *AUC* being the second most applied performance criterion for model evaluation, after *CCI*. In recent years, some papers introduced other performance criteria into species distribution modelling, such as the odds ratio, the *TSS* and the *NMI*

Table 3
The performance criteria used for model evaluation in the same sample of papers as in Table 2 ($n = 257$). *CCI* = percentage correctly classified instances; *CCI* only = papers which only applied *CCI* for model evaluation; *Kappa* = Cohen's *Kappa*; *AUC* = area under the curve; *Sn* = sensitivity; *Sp* = specificity. The percentages do not cumulate to 100% because some papers used more than one performance criterion.

	Performance criteria applied in model evaluation						*Sp*	Other
	No criterion	*CCI*	*CCI* only	*Kappa*	*AUC*	*Sn*		
1998–2002	2	98	18	26	53	16	16	1
2003–2007	3	97	15	33	61	8	7	3
Total	3	97	15	32	60	9	8	2

Table 4
The different scenarios, their corresponding elements of the confusion matrices and the values of the 7 different performance criteria for the 4 scenarios, assuming that 0 ln(0) = 0. To calculate the odds ratio, a continuity correction was performed by adding 0.5 to each of the cells in the confusion matrix (Forbes, 1995; Vaughan and Ormerod, 2005). These assumptions have no effect on the characteristics of the presented criteria.

Scenario	Element of the confusion matrix			Criterion							
	a	*b*	*c*	*d*	*CCI*	*Sn*	*Sp*	*NMI*	*Kappa*	*TSS*	Odds ratio
1	2	0	0	2	1	1	1	0	1	1	25
2	0	2	2	0	0	0	0	0	−1	−1	0
3	2	1	0	1	0.75	1	0.5	0.23	0.5	0.5	5
4	1	0	1	2	0.75	0.5	1	0.23	0.5	0.5	5

statistic. Current practice in species distribution modelling is to apply at least two different performance criteria for model evaluation. However, this approach does not avoid that misleadingly high performance criteria values could be obtained if strongly correlated criteria are being chosen.

To assess whether performance criteria for model evaluation are able to distinguish between omission and commission errors, 4 different scenarios were defined, of which the confusion matrix is given in Table 4. In the first scenario, all instances are classified correctly, whereas in the second scenario, all instances are classified erroneously. The third scenario is characterised by slight overprediction, whereas in the fourth scenario, the model is underpredicting the observed data. Previous research demonstrated that overprediction does not necessarily imply a model error, in contrast to underprediction (Mouton et al., 2008, 2009a). In line with this assumption, an ecologically relevant performance criterion would classify scenario 1 as the best scenario and scenario 2 as the worst, while scenario 3 may be considered ecologically more sound than scenario 4 due to the false-negative predictions in the latter scenario. Table 4 shows the values of seven different performance criteria (*CCI*; Cohen's *Kappa*, *TSS*; *NMI*, *Sn*, *Sp* and odds ratio) for the 4 scenarios. The performance criteria *CCI*, *NMI*, *Kappa*, *TSS*, and odds ratio do not distinguish between scenarios 3 and 4, whereas *Sn* and *Sp* allow differentiation between these two scenarios. However, *Sn* and *Sp* do not distinguish between scenario 1, and scenarios 3 and 4, respectively. Consequently, no performance criterion clearly distinguishes between the four different scenarios.

4. Application of performance criteria for model training

Although numerous papers on performance criteria assessment focus on the application of these criteria in model evaluation (Fielding and Bell, 1997; Manel et al., 2001; Vaughan and Ormerod, 2005; Allouche et al., 2006), few authors describe the impact of performance criteria on model training. From the aforementioned group of papers which used data in the modelling process, 99% applied a training performance criterion which assesses the predictive accuracy of the model, such as *CCI*, Pearson's correlation coefficient, or the root mean squared error. The application of these criteria for model training is based on the assumption that, provided the "true" model is nested within the model specification, the model estimated using predictive accuracy techniques will converge to the true model as the sample size increases (Welsh, 1996). However, the true model is rarely nested within the model specification due to various reasons (Tyre et al., 2001; Barry and Elith, 2006). Moreover, it could be shown theoretically that none of the criteria assessing predictive accuracy distinguishes between omission and commission errors. Less than 1% of the evaluated papers applied different performance criteria for model training and compared the results.

The impact of the dependency between the performance criterion and the prevalence differs between model training and model evaluation. During the first process, the prevalence is constant because only one training dataset is used, whereas during model evaluation, prevalence may vary due to the use of different evaluation datasets. Despite the constant prevalence during model training, effects of species prevalence on model training results have been reported (Hirzel et al., 2001; Syphard and Franklin, 2009). We suggest that the impact of the species prevalence on the model training results is related to the performance criterion used for model training.

The purpose of this section is to analyse the impact of the performance criteria used for model training on the final model which is obtained after training. This impact is based on the two aforementioned characteristics of performance criteria: the dependency on prevalence and the distinction between omission and commission errors. These characteristics are assumed to be independent of each other in model evaluation, whereas the following analyses show that both characteristics are related in model training. This section attempts to analyse this relation between prevalence dependency and omission–commission distinction of performance criteria during the model training process.

Three of the most commonly used performance criteria are analysed theoretically: *CCI*, *Kappa* and *TSS*. More specifically, the effect of an increase of the true-positive predictions (a) in the confusion matrix (Δa) on the performance criterion value is assessed. We assumed that the ecological data are linearly inseparable. Consequently, the change of the model parameters which results in a change of the true-positive predictions Δa, leads to a decrease of the true-negative predictions, $-\Delta d$. If the data would be linearly separable, Δd would equal zero. Since prevalence remains constant during model training ($a + c = a^* + c^*$), the confusion matrix will change as described in Table 5.

The impact of prevalence on the effect of an adjustment of the model parameters which results in Δa will be discussed. The prevalence P is described as:

$$P = \frac{a + c}{N} \qquad (1)$$

4.1. Analysis of CCI

It can be shown that the change in *CCI*, ΔCCI, which results from a change Δa as described in the confusion matrix in Table 5 is equal to:

$$\Delta CCI = \frac{\Delta a - \Delta d}{N} \qquad (2)$$

Table 5
The confusion matrix after a change of the model parameters which results in an increase Δa of the true-positive predictions a. The table cross-tabulates observed values against predicted values: true-positives, a; false-positives, b; false-negatives, c; true-negative values, d.

	Observed	
	Present	Absent
Predicted		
Present	$a + \Delta a$	$b + \Delta d$
Absent	$c - \Delta a$	$d - \Delta d$

If ΔCCI exceeds zero, the adjustment of the model parameters will result in a better model and the optimisation algorithm will continue with this adjusted model. At low prevalences, Δa tends to be considerably smaller than Δd, whereas at high prevalences, almost all changes of a model parameter will result in a Δa which is greater than Δd. Consequently, the likelihood that ΔCCI will be greater than zero if the prevalence is high is higher than the likelihood if the prevalence is low. This will result in a high number of present model predictions at high prevalences and a high number of absent predictions at low prevalences. At high prevalences, optimisation based on *CCI* will thus lead to high overprediction errors, whereas at low prevalences, optimisation based on *CCI* will result in high underprediction errors.

4.2. Analysis of Kappa

To assess the effect of a change Δa as described in the confusion matrix in Table 5 on *Kappa*, $\Delta Kappa$ can be described as a function of the prevalence *P*, the sensitivity *Sn* and the specificity *Sp*:

$$\Delta Kappa = \frac{2(Sn+\Delta Sn+Sp+\Delta Sp-1)P(1-P)}{1+2(Sn+\Delta Sn+Sp+\Delta Sp-1)P(1-P)-P(Sn+\Delta Sn)-(1-P)(Sp+\Delta Sp)} - \frac{2(Sn+Sp-1)P(1-P)}{1+2(Sn+Sp-1)P(1-P)-PSn-(1-P)Sp} \quad (3)$$

with

$$\Delta Sn = \frac{\Delta a}{a+c} \quad (4)$$

and

$$\Delta Sp = \frac{\Delta d}{b+d} \quad (5)$$

(2) can be rewritten as:

$$\Delta Kappa = P(1-P)\left(\frac{2(TSS)+2(\Delta Sn+\Delta Sp)}{U} - \frac{2TSS}{V}\right) \quad (6)$$

with

$$\begin{aligned} U &= 1+2(Sn+\Delta Sn+Sp+\Delta Sp-1)P(1-P)-P(Sn+\Delta Sn)-(1-P)(Sp+\Delta Sp) \\ &= 1+2(TSS+\Delta Sn+\Delta Sp)P(1-P)-P(Sn+\Delta Sn)-(1-P)(Sp+\Delta Sp) \end{aligned} \quad (7)$$

and

$$\begin{aligned} V &= 1+2(Sn+Sp-1)P(1-P)-PSn-(1-P)Sp \\ &= 1+2TSSP(1-P)-PSn-(1-P)Sp \end{aligned} \quad (8)$$

It can be further shown that

$$\Delta Kappa = \frac{2}{UVN^2}(\Delta a((1-P)b+Pd)-\Delta d((1-P)a+Pc)) \quad (9)$$

It can be shown that *U* and *V* are strictly positive and thus (4) shows that

$$\Delta Kappa > 0 \Leftrightarrow \frac{\Delta a}{((1-P)a+Pc)} > \frac{\Delta d}{((1-P)b+Pd)}. \quad (10)$$

If $\Delta Kappa$ exceeds zero, the adjustment of the model parameters will result in a better model and the optimisation algorithm will continue with this adjusted model. Whether a change in the confusion matrix Δa will lead to a positive $\Delta Kappa$, depends on the prevalence *P* and on the proportions in the confusion matrix. For instance, if *P* exceeds 0.5 and $c \ll d$, a small change Δa will lead to a positive $\Delta Kappa$, even at high Δd values. This relation between $\Delta Kappa$, *P* and the proportions in the confusion matrix is shown in Table 6.

Table 6
The effect of the prevalence *P* and the proportions in the confusion matrix on the likelihood that $\Delta Kappa$ will exceed zero at a fixed increase of the true-positive predictions, Δa. 'High' indicates that this relative likelihood is high, while 'low' indicates that this likelihood is low.

Prevalence	$a < b$	$a > b$	$c < d$	$c > d$
$P < 0.5$	Low	High	Low	High
$P = 0.5$	See 'Analysis of *TSS*'			
$P > 0.5$	High	Low	High	Low

4.3. Analysis of TSS

The change in *TSS*, ΔTSS, resulting from a change Δa as described in the confusion matrix is equal to:

$$\Delta TSS = \frac{\Delta a}{a+c} - \frac{\Delta d}{b+d} \quad (11)$$

Consequently,

$$\Delta TSS > 0 \Leftrightarrow \Delta a > \frac{\Delta d(a+c)}{b+d} \quad (12)$$

If ΔTSS exceeds zero, the adjustment of the model parameters will result in a better model and the optimisation algorithm will continue with this adjusted model. If prevalence is greater than 0.5, the difference $\Delta a - \Delta d$ should be greater than when prevalence is smaller than 0.5. In the latter case, ΔTSS could even exceed zero for some situations in which $\Delta a < \Delta d$. Consequently, a change Δa in the confusion matrix will more easily lead to adjusted model parameters if prevalence <0.5 and overprediction (*b*) will be stimulated in this situation. At prevalences higher than 0.5, a change Δa in the confusion matrix will only result in a change of the model parameters if Δa substantially exceeds Δd. The optimisation algorithm will thus stimulate underprediction (*c*) (Table 7). Notice that *TSS* is a special case of *Kappa*, when *P* equals 0.5 (Allouche et al., 2006). Furthermore, the stimulation of either underprediction or overprediction is more complex with *Kappa* than with *TSS*, which is reflected by the more complex denominators in Eq. (10) compared to Eq. (12).

Table 7 shows that the result of model optimisation based on *TSS* depends on the prevalence of the training data set. This emphasises the difference between model training and evaluation. Although several authors proved that *TSS* is independent of prevalence when it is used for model evaluation (Allouche et al., 2006), this performance criterion clearly depends on prevalence when it is applied in model training. Specifically, model training based on *TSS* attempts to compensate for the prevalence of the training data: if this prevalence is high, underprediction is stimulated, whereas low prevalences correspond to the stimulation of overprediction. This issue is of key importance in ecological modelling studies because it shows the effect of the performance criterion used for model training on the resulting model, and thus on the decisions supported by this model. At the start of the model development process, modellers should clearly define the goals of the model and then choose a performance criterion which reflects these model purposes.

Table 7
The relation between ΔTSS, Δa and Δd as a function of prevalence. The possible relation between Δa and Δd given is the relation at which ΔTSS could exceed 0.

Prevalence	$\frac{a+c}{b+d}$	Possible relations between Δa and Δd such that $\Delta TSS > 0$	Stimulation of
$P > 0.5$	>1	$\Delta a \gg \Delta d$	Underprediction
$P = 0.5$	=1	$\Delta a > \Delta d$	–
$P < 0.5$	<1	$\Delta a < \Delta d$ or $\Delta a > \Delta d$	Overprediction

Table 8
Characteristics of the most frequently applied performance criteria for model training and evaluation. *NMI* = normalised mutual information statistic; *TSS* = true skill statistic; *AUC* = area under the curve; *CCI* = correctly classified instances; *Sn* = sensitivity; *Sp* = specificity; v = the characteristic fully applies to the performance criterion; (–) = the characteristic does not apply to the performance criterion; (?) = the characteristic may apply to the performance criterion; n.a. = not applicable.

Performance criterion	*Kappa*	*NMI*	Odds ratio	*TSS*	*AUC*	*CCI*	*Sn, Sp*
Characteristic							
Quantifies the extent to which models correctly predict occurrence better than chance expectation	v	v	v	v	v	–	–
Depends on prevalence	v	?	?	?	?	v	v
Takes into account the complete information included in the confusion matrix	v	v	v	v	n.a.	–	–
Does distinguish between omission and commission errors	–	–	–	–	–	–	–
Compensates for extreme prevalence values when applied on model training	v	–	–	–	–	–	–
Requires discretisation of model predictions by applying threshold values	v	v	v	v	–	v	v
Allows zero values in the confusion matrix	v	–	–[a]	v	n.a.	v	v[b]
Is proportional (the same performance is found if all elements of the confusion matrix are divided by the same constant)	v	–	–	v	n.a.	v	v
Can be applied for model training	v	–	v	v	v	v	v

[a] Cannot be applied directly when both the number of false-positive predictions and false-negative predictions is zero; adding a constant value to each element of the confusion matrix changes the relative value of the odds ratio.

[b] Cannot be applied if the prevalence of the evaluation or training set is 0.

5. Towards adaptive model evaluation

Our analysis of model evaluation showed that different performance criteria evaluate a model (or its resulting confusion matrix) differently. Several authors attributed this difference to the relation between the performance criteria and the prevalence of the validation set (Manel et al., 1999, 2001; McPherson et al., 2004; Allouche et al., 2006). Specifically, if models derived from different datasets are being compared, the prevalence of these datasets may affect the value of the performance criteria and consequently influence the results of the comparison. Similar problems rise when the performances of a model on a training set and a validation set with different prevalences are compared (Allouche et al., 2006). Some authors suggested that this problem would be avoided if validation sets would be collected such that prevalence would be around 50% (Lantz and Nebenzahl, 1996; Hoehler, 2000; McPherson et al., 2004). However, various authors agree that this recommendation is of questionable practicability in species distribution modelling, particularly for rare species for which a small number of presence data is available (Maclure and Willett, 1987; Mackenzie and Royle, 2005; Allouche et al., 2006). Moreover, an appropriate performance criterion is meant to be a tool for communication (Maclure and Willett, 1987). Consequently, it should be very clear which aspect of the model performance is evaluated, and models should be evaluated consistently, that is, independent of, or taking into account, species prevalence. Given the questionable value of *Kappa* for distribution modelling (McPherson et al., 2004; Allouche et al., 2006), Vaughan and Ormerod (2005) agreed that measures other than *Kappa* may thus be preferable to evaluate model predictions.

The presented results highlight the relative importance that the performance criteria give to omission and commission errors, as a possible explanation for the differing evaluation scores among performance criteria for the same model. Theoretical analysis revealed that performance criteria may value a perfect model equally, but yet focus on very different aspects of model performance. An example is the assessment of model discriminatory ability: although all performance criteria will attain their optimum for a model with excellent discrimination, not all criteria adequately quantify the discriminatory ability of a model (Vaughan and Ormerod, 2005). Consequently, model developers should carefully choose an appropriate performance criterion for model evaluation which corresponds to the ecological objectives of the optimised model. The ecological literature has recognised these problems and the ROC technique in particular has received considerable attention (Elith et al., 2006; Meynard and Quinn, 2007).

Concerning the model training process, we showed that the effect of the performance criterion on the optimal model depends on the separability of the training data. Training sets which are linearly separable could lead to perfect predictions for all performance criteria, as shown in scenario 1. Jimenez-Valverde et al. (2009) even showed that prevalence effects may be limited if data separability is stimulated by avoiding false absences or non-explanatory variables. However, in ecological case studies, an increase of the true-positive predictions *a* often results in a decrease of the true-negative predictions *d* and vice versa because the training data are rarely linearly separable. Our analysis indicated that the balance between *a* and *d* is affected by the performance criterion which was used to train the model. Like for model evaluation, model developers should thus also carefully choose an appropriate training performance criterion which reflects the ecological model purpose (Segurado and Araújo, 2004). To assist modellers in this choice, Table 8 provides an overview of the most important characteristics and restrictions of the most frequently applied performance criteria.

We showed theoretically that training data prevalence may significantly affect the over- or underprediction rate of a model. Several authors agree that the relative importance of omission and commission errors may vary among applications (Loiselle et al., 2003; Vaughan and Ormerod, 2005; Wilson et al., 2005; Prates-Clark et al., 2008; Vaclavik and Meentemeyer, 2009). Applying a general performance criterion in model training ignores these subtle but significant differences between different applications. Consequently, an appropriate performance criterion should allow modellers to implement this relative importance in the model training process, for example by including a parameter which can be adjusted to the specific situation (Mouton et al., 2008, 2009a,c).

Although an optimal parameter value could be found by applying sensitivity analysis, a more important problem with these flexible performance criteria could be the difficulty of identifying which models are better (Vaughan and Ormerod, 2005). Differences in species' dispersal patterns and associated gene flow may lead to subtle variations in habitat preferences of some species due to local adaptations (Holt, 2003; McPherson and Jetz, 2007). Even in the absence of genetically driven differences in habitat use, species could express different realised niches (Hutchinson, 1957) as a result of spatial variation in predators, competitors or other biotic factors (Hutchinson, 1957; Osborne and Suarez-Seoane, 2002; Holt, 2003; Peterson and Holt, 2003; Hernandez et al., 2006; McPherson and Jetz, 2007). Consequently, conservationists should consider not only statistical but also ecological factors that are actually affecting the reliability of a given distribution model (Jiménez-Valverde and Lobo, 2006; McPherson and Jetz, 2007; Chefaoui and Lobo, 2008).

Given the complexity of the modelled ecological relations, the most robust modelling approaches are likely to be those in which care is taken to match the model with knowledge in ecology. These models should be constrained to be congruent with ecological knowledge, with successive improvement in model performance that is driven by increasing knowledge of the ecology of the system (Barry and Elith, 2006; Elith and Leathwick, 2009). Uncertainty in model predictions can thus be viewed from two perspectives: uncertainty as an obstacle that needs to be reduced or removed, or uncertainty as a fact of life (Barry and Elith, 2006). The first approach attempts to change the model structure by seeking more powerful modelling techniques or to improve the data by collecting more samples, removing errors or selecting variables, whereas the second approach tries to understand, characterise and analyse uncertainty by sensitivity analysis, explorations of error or the application of decision strategies that aim to be robust to likely errors (Burgman et al., 2001). Both perspectives are valid and not necessarily mutually exclusive (Barry and Elith, 2006). Finally, selection of the best model will always depend on the preferences of modellers since no model can excel on all aspects of model performance. This could be shown theoretically and is referred to as the “no free lunch theorem” (Wolpert and Macready, 1997). Consequently, meta-modelling techniques may be applied to overcome the shortcomings of different algorithms and performance criteria. By providing further insight into the behaviour of performance criteria, this paper may contribute to the development of such meta-modelling approaches.

6. Conclusions

Performance criteria are the key element of the presence/absence model evaluation process and assess the performance of both the final model and the model during training. Although numerous studies on species distribution modelling focus on the role of performance criteria for evaluation of the final model, few authors have addressed the effect of these criteria on model training. We provide a theoretical analysis of the impact of the performance criteria applied for model training on the final model. The results show that, like criteria for evaluation of the final model, the appropriate performance criteria for model training should be chosen carefully and that this choice is dictated by the end-use of the model. For both model training and evaluation, we suggest that prevalence-independent measures should be preferred, and that at least some of these measures should allow modellers to distinguish between omission and commission errors. The practical implications of this paper are clear. It provides further insight in the evaluation of ecological presence–absence models and attempts to assist modellers in their choice of suitable performance criteria. As such, it may be an important step towards more reliable species distribution models.

Acknowledgement

Ans Mouton was a recipient of a Ph.D. grant financed by the Special Research Fund of Ghent University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2010.04.017.

References

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232.

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species-climate impact models under climate change. Global Change Biol. 11, 1504–1513.

Araújo, M.B., Rahbek, C., 2006. How does climate change affect biodiversity? Science 313, 1396–1397.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423.

Berg, A., Gardenfors, U., von Proschwitz, T., 2004. Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden—importance of environmental data quality and model complexity. Ecography 27, 83–93.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300.

Buckland, S.T., Elston, D.A., 1993. Empirical-models for the spatial-distribution of wildlife. J. Appl. Ecol. 30, 478–495.

Burgman, M.A., Breininger, D.R., Duncan, B.W., Ferson, S., 2001. Setting reliability bounds on habitat suitability indices. Ecol. Appl. 11, 70–78.

Chefaoui, R.M., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecol. Model. 210, 478–486.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.

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, 38–49.

Forbes, A.D., 1995. Classification-algorithm evaluation—5 performance-measures based on confusion matrices. J. Clin. Monitor. 11, 189–206.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F., 2007. Sensitivity of predictive species distribution models to change in grain size. Divers. Distrib. 13, 332–340.

Guisan, A., Theurillat, J.P., 2000. Equilibrium modeling of alpine plant distribution: how far can we go? Phytocoenologia 30, 353–384.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 536 pp.

Heikkinen, R.K., Luoto, M., Kuussaari, M., Toivonen, T., 2007. Modelling the spatial distribution of a threatened butterfly: impacts of scale and statistical technique. Landsc. Urban Plann. 79, 347–357.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Hoehler, F.K., 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J. Clin. Epidemiol. 53, 499–503.

Holt, R.D., 2003. On the evolutionary ecology of species' ranges. Evol. Ecol. Res. 5, 159–178.

Hutchinson, G.E., 1957. Population studies – animal ecology and demography – concluding remarks. Cold Spring Harb. Symp. Quant. Biol. 22, 415–427.

Jiménez-Valverde, A., Lobo, J.M., 2006. The ghost of unbalanced species distribution data in geographical model predictions. Divers. Distrib. 12, 521–524.

Jiménez-Valverde, A., Lobo, J.M., 2007. Threshold criteria for conversion of probability of species presence to either-or presence–absence. Acta Oecol. 31, 361–369.

Jimenez-Valverde, A., Lobo, J.M., Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecol. 10, 196–205.

Kerr, J.T., Kharouba, H.M., Currie, D.J., 2007. The macroecological contribution to global change solutions. Science 316, 1581–1584.

Lantz, C.A., Nebenzahl, E., 1996. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. J. Clin. Epidemiol. 49, 431–434.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Liu, C.R., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.

Lobo, J.M., Jimenez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33, 103–114.

Lobo, J.M., Jimenez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. 17, 145–151.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conserv. Biol. 17, 1591–1600.

Mackenzie, D.I., Royle, J.A., 2005. Designing occupancy studies: general advice and allocating survey effort. J. Appl. Ecol. 42, 1105–1114.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the Kappa-Statistic. Am. J. Epidemiol. 126, 161–169.

Maggini, R., Lehmann, A., Zimmermann, N.E., Guisan, A., 2006. Improving generalized regression analysis for the spatial prediction of forest communities. J. Biogeogr. 33, 1729–1749.

Manel, S., Dias, J.M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120, 337–347.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.

McPherson, J.M., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography 30, 135–151.

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, 811–823.

Meynard, C.N., Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J. Biogeogr. 34, 1455–1469.

Mouton, A., De Baets, B., Van Broekhoven, E., Goethals, P.L.M., 2009a. Prevalence-adjusted optimisation of fuzzy models for species distribution. Ecol. Model. 220, 1776–1786.

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2009b. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. Environ. Modell. Softw. 24, 982–993.

Mouton, A.M., Jowett, I., Goethals, P.L.M., De Baets, B., 2009c. Prevalence-adjusted optimisation of fuzzy habitat suitability models for aquatic invertebrate and fish species in New Zealand. Ecol. Inform. 4, 215–225.

Mouton, A.M., Schneider, M., Peter, A., Holzer, G., Müller, R., Goethals, P.L.M., De Pauw, N., 2008. Optimisation of a fuzzy habitat model for spawning European grayling (*Thymallus thymallus* L.) in the Aare river (Thun. Switzerland). Ecol. Model. 215, 122–132.

Nielsen, S.E., Johnson, C.J., Heard, D.C., Boyce, M.S., 2005. Can models of presence–absence be used to scale abundance?—two case studies considering extremes in life history. Ecography 28, 197–208.

Osborne, P.E., Suarez-Seoane, S., 2002. Should data be partitioned spatially before building large-scale distribution models? Ecol. Model. 157, 249–259.

Parra, J.L., Graham, C.C., Freile, J.F., 2004. Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography 27, 350–360.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245.

Pearson, R.G., Dawson, T.P., Liu, C., 2004. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography 27, 285–298.

Peterson, A.T., Holt, R.D., 2003. Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. Ecol. Lett. 6, 774–782.

Peterson, A.T., Papes, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550–560.

Petit, S., Chamberlain, D., Haysom, K., Pywell, R., Vickery, J., Warman, L., Allen, D., Firbank, L., 2003. Knowledge-based models for predicting species occurrence in arable conditions. Ecography 26, 626–640.

Prates-Clark, C.D., Saatchi, S.S., Agosti, D., 2008. Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data. Ecol. Model. 211, 309–323.

Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H., Possingham, H.P., 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecol. Lett. 9, 1136–1145.

Rouget, M., Richardson, D.M., Nel, J.L., Le Maitre, D.C., Egoh, B., Mgidi, T., 2004. Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. Divers. Distrib. 10, 475–484.

Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200.

Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.

Sprott, D.A., 2000. Statistical Inference in Science. Springer, 245 pp.

Syphard, A.D., Franklin, J., 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. Ecography 32, 907–918.

Termansen, M., McClean, C.J., Preston, C.D., 2006. The use of genetic algorithms and Bayesian classification to model species distributions. Ecol. Model. 192, 410–424.

Thuiller, W., Araújo, M.B., Lavorel, S., 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. J. Veg. Sci. 14, 669–680.

Thuiller, W., Lavorel, S., Araújo, M.B., 2005. Niche properties and geographical extent as predictors of species sensitivity to climate change. Global Ecol. Biogeogr. 14, 347–357.

Tyre, A.J., Possingham, H.P., Lindenmayer, D.B., 2001. Inferring process from pattern: can territory occupancy provide information about life history parameters? Ecol. Appl. 11, 1722–1737.

Vaclavik, T., Meentemeyer, R.K., 2009. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecol. Model. 220, 3248–3258.

Van Broekhoven, E., Adriaenssens, V., De Baets, B., 2007. Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: an ecological case study. Int. J. Approx. Reason. 44, 65–90.

Van Broekhoven, E., Adriaenssens, V., De Baets, B., Verdonschot, P.F.M., 2006. Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. Ecol. Model. 198, 71–84.

Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution models. J. Appl. Ecol. 42, 720–730.

Welsh, A.H., 1996. Aspects of Statistical Inference. John Wiley & Sons, New York, USA, 452 pp.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biol. Conserv. 122, 99–112.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82.