



A new metric to assess the predictive accuracy of multinomial land cover models

Douma, B., Cornwell, W. K., & van Bodegom, P. M.

This article is made publically available in the institutional repository of Wageningen University and Research, under article 25fa of the Dutch Copyright Act, also known as the Amendment Taverne.

Article 25fa states that the author of a short scientific work funded either wholly or partially by Dutch public funds is entitled to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.


For questions regarding the public availability of this article, please contact openscience.library@wur.nl.

Please cite this publication as follows:

Douma, B., Cornwell, W. K., & van Bodegom, P. M. (2017). A new metric to assess the predictive accuracy of multinomial land cover models. *Journal of Biogeography*, 44(6), 1212-1224. <https://doi.org/10.1111/jbi.12983>

METHODOLOGICAL
APPLICATION

A new metric to assess the predictive accuracy of multinomial land cover models

Jacob C. Douma^{1,4*} , William K. Cornwell^{2,4} and Peter M. van Bodegom^{3,4}

¹Wageningen University and Research Centre, Centre for Crop Systems Analysis, 6700 AK Wageningen, The Netherlands, ²Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, UNSW, Sydney, NSW, Australia, ³Leiden University, Institute of Environmental Sciences, 2333 CC Leiden, The Netherlands, ⁴VU University Amsterdam, Institute of Ecological Science, Department of Systems Ecology, 1081 HV Amsterdam, The Netherlands

ABSTRACT

Aim The earth's land cover is often represented by discrete classes, and predicting shifts between these classes is a major goal in the field. One increasingly common approach is to build models that predict land cover classes with probabilities rather than discrete outcomes. Current assessment approaches have drawbacks when applied to these types of models. In this paper we present a new metric, which assesses agreement between model predictions and observations, while correcting for chance agreement.

Location Global.

Methods $\kappa_{\text{multinomial}}$ is the product of two metrics: the first component measures the agreement in the ranks of the predicted and observed classes, the other specifies the certainty of the model in the case of discrete observations. We analysed the behaviour of $\kappa_{\text{multinomial}}$ and two alternative metrics: Cohen's Kappa (κ) and an extension of the area under receiver operating characteristic Curve to multiple classes (mAUC) when applied to multinomial predictions and discrete observations.

Results Using real and synthetic datasets, we show that $\kappa_{\text{multinomial}}$ – in contrast to κ – can distinguish between models that are very far off versus slightly off. In addition, $\kappa_{\text{multinomial}}$ ranks models higher that predict observed classes with an onaverage higher probability. In contrast, mAUC gives the same score to models that are perfectly able to discriminate among classes of outcomes regardless of the certainty with which those classes are predicted.

Main conclusions With $\kappa_{\text{multinomial}}$ we have provided a tool that directly uses the multinomial probabilities for accuracy assessment. $\kappa_{\text{multinomial}}$ may also be applied to cases where model predictions are evaluated against multiple sets of observations, at multiple spatial scales, or compared to reference models. As models develop we assess how well new models perform compared to the real world.

Keywords

cohen's kappa, kappa multinomial, land cover, model predictive accuracy, multinomial models, multiple class AUC, validation

*Correspondence: Jacob C. Douma, Wageningen University and Research Centre, Centre for Crop Systems Analysis, 6700 AK, Wageningen, The Netherlands.
E-mail: bob.douma@wur.nl

INTRODUCTION

Land use and vegetation models are commonly used in earth science and biology to understand the effects of environmental and socio-economic drivers on land cover. For example, they are used to understand and project changes in vegetation type distribution under climate change (Lenihan *et al.*, 2003; Sitch *et al.*, 2008; van Bodegom *et al.*, 2014) and to

predict land use dynamics to support policy and planning (Verburg *et al.*, 2004 and references herein, Le *et al.*, 2010). In these models, land use and vegetation are often represented by discrete classes.

One emerging group of land-use change and vegetation distribution models determine a probability distribution for a set of classes (see for an example Fig. 1; Muller & Zeller, 2002; Hepinstall *et al.*, 2008; Douma *et al.*, 2012b; van

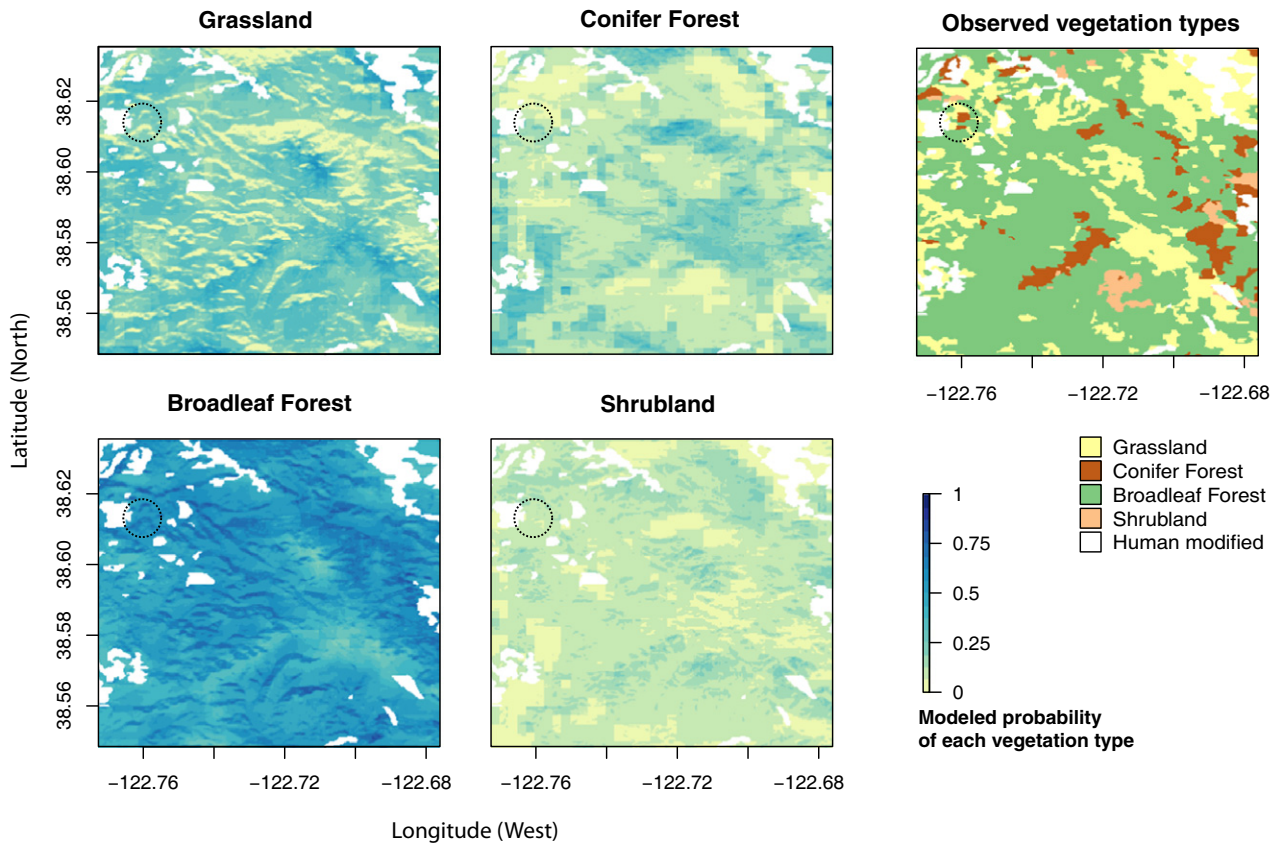


Figure 1 Observed mosaic of vegetation types and predicted probabilities of the four dominant vegetation types (Grassland, Shrubland, Broadleaf forest, and Conifer forest) in the Chaparral in California. The predicted probabilities were fitted simultaneously in a multinomial logistic model, and the probabilities of the four vegetation types in a pixel sum to one. Species composition in this area is partly determined by constant factors (such as topology, soil, and climate), and partly stochastic factors such as fire. The model shows that at some locations multiple vegetation types have an equal chance of occurrence (see dashed circle), while only one vegetation type is observed. For full model specifications we refer to Ackerly *et al.* (2015). [Colour figure can be viewed at wileyonlinelibrary.com]

Bodegom *et al.*, 2014; Ackerly *et al.*, 2015). Such multinomial models estimate the probability of a sample (also called instance, pixel, individual observation or item) being a member of the possible classes with estimated probabilities p_{i1}, \dots, p_{iq} , $p_{ik} \geq 0$ in sample i , and $\sum_{k=1}^q p_{ik} = 1$. The fact that as p_{ik} is less than 1 may reflect either the stochastic nature of possible realizations – whatever the nature of the particular stochastic mechanism(s) in the system may be (Turner *et al.*, 1993), or the inability of the model to represent an important process. The argument for this type of model is that explicitly modelling probabilities is a better representation of both ‘real’ stochasticity and instance-level model uncertainty.

Assessing the agreement of multinomial models with (independent) observations is important to assess and improve model reliability, and to select among competing models. Agreement assessment involves the comparison of the predicted class probabilities to a set of observations y whose class membership is known either with or without uncertainty. However, the probabilistic nature of the new class of models makes traditional methods of model assessment problematic. In multinomial models, the agreement to the data has two components: 1) the degree to which the ranks

of the predicted class frequencies correspond to the observed class frequencies, and 2) the certainty with which those classes are predicted.

The development of metrics for multinomial models and discrete observations is an active field of research (Ferri *et al.*, 2009; Sokolova & Lapalme, 2009; Jurman *et al.*, 2012). The available methods can be classified into three families (Ferri *et al.* (2009): i) metrics based on a threshold and a qualitative understanding of error, ii) metrics based on how well the model ranks the observations, and iii) metrics based on a probabilistic understanding of error, and measuring the difference from the true probability.

The metrics belonging to the first family are most frequently applied and transform class probabilities to success for exactly one out of q outcomes before assessing their agreement with a set of samples (Ferri *et al.*, 2009). The underlying assumption of such a hard classification (also known as crisp classification) is that a class is observed if it exceeds a probability threshold. In this case, the prediction equals the class with the highest predicted probability: $\hat{y}_{ij} = 1$ if $j = \arg \max_k (p_{ik})$ and $\hat{y}_{ij} = 0$ otherwise (Dendoncker *et al.*, 2007; Douma *et al.*, 2012b; van Bodegom *et al.*, 2014). After

transformation, one can calculate the overall accuracy (the fraction correctly classified cells p_0) and a number of other measures (see Dendoncker *et al.*, 2007 and Webb *et al.*, 2008 for applying this measure to multinomial predictions).

A classical and still widely used accuracy metric from this family is the kappa statistic (κ , equation 1; Cohen, 1960). It was originally designed to evaluate the agreement between two classifiers and expresses the observed agreement between those classifiers (p_0) corrected for chance agreement (p_e). p_0 equals the proportion of samples that has the same class attribution by the two classifiers. p_e is the agreement that can be expected after a random allocation of given class sizes. κ has been applied to assess the accuracy of vegetation models (e.g. Monserud & Leemans, 1992). Rescaling ensures that κ reaches one if the model predictions perfectly match the observations, while κ reaches zero if agreement is similar to chance:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

The main reason to use κ is that the scaling against a reference model allows inter-comparison of models from different regions (Gotelli & Graves, 1996). However, this feature has also been a main point of criticism (Foody, 1992; Pontius & Millones, 2011).

The transformation of probabilities to hard classification has critical disadvantages. First of all, the probabilistic nature that is often reality is removed while assessing the accuracy of the model predictions, which eliminates some of the advantages of probabilistic models. Secondly, assessing model agreement on transformed multinomial probabilities very likely affects model agreement because an arbitrary decision is made about which predicted class would be observed. Finally, with transforming probabilities, information is lost about the certainty with which the model predicts a given class.

Well-known metrics from the second family are based on the Area Under the receiver operating characteristic Curve (AUC). AUC has been designed to measure the ability to discriminate between two classes of outcomes. An AUC value of one indicates a perfect ability of the model to discriminate between two classes, and an AUC value of 0.5 indicates similar agreement than what would be expected by chance. AUC has been extended in multiple ways to cases with more than two classes (Hand & Till, 2001; Provost & Domingos, 2001; Ferri *et al.*, 2003). The analogue of the AUC for multinomial predictions is the volume under curve (VOC, Ferri *et al.*, 2003). However, the computation of VOC becomes challenging when a large number of classes is distinguished (dimension equals 2^q – with q classes). A method that is computationally more attractive in case of a high number of classes is by calculating multiple AUCs using a one class versus all other classes approach. Thus in case of q classes, one gets q AUC curves and q AUC values that are weighted according to the occurrence of each class to obtain one AUC value (hereafter referred to as mAUC,

Provost & Domingos, 2001; Fawcett, 2006). A variant of AUC, which is a hybrid between the second and third family, is the probabilistic AUC. Metrics of this kind aim to make the rankings robust to small changes in the predicted probabilities (scored AUC, sAUC, Wu *et al.*, 2007; and probabilistic AUC (pAUC) Ferri *et al.*, 2005). AUC as agreement metric has been criticized when applied to small samples and to predictive distribution models (Lobo *et al.*, 2008; Hanczar *et al.*, 2010). The main criticism of the method when applied to predictive distribution models is also the strength of the method: the discriminating power of the model does not necessarily indicate a good model fit (Lobo *et al.*, 2008).

Agreement metrics that belong to the third family, those based on a probabilistic understanding of error, measure the deviation of observations from the predicted probabilities. Examples include the Brier score and the mean squared error (Brier, 1950; Ferri *et al.*, 2009). A kappa-like metric, hereafter referred to as κ_{group} , was developed by Vanbelle & Albert (2009) to calculate the agreement between a group of classifiers and a single classifier (equation 2). The classification of the group is expressed as a multinomial model with q classes and the classification of the single observer as success for exactly one out of q classes. κ_{group} defines agreement as observed agreement ($p_{0 \text{ multinomial}}$; equation 3) corrected for the difference in maximum agreement that can be obtained with the multinomial model (p_{max}) and chance agreement (p_e). The observed agreement equals the average probability of observed classes, and is also known as the mean probability rate (Ferri *et al.*, 2009). The maximum proportion of agreement was calculated as the average of the most probable classes over all samples (equation 4). Chance agreement is the product of the marginal distributions of the model and the observations (equation 5). Thus, κ_{group} measures the degree to which the most probable class is also observed irrespective of the certainty with which those classes are predicted.

$$\kappa_{\text{group}} = \frac{p_{0 \text{ multinomial}} - p_e}{p_{\text{max}} - p_e} \quad (2)$$

$$p_{0 \text{ multinomial}} = \frac{1}{m} \sum_{i=1}^m \sum_{k=1}^q y_{ik} p_{ik} \quad (3)$$

$$p_{\text{max}} = \frac{1}{m} \sum_{i=1}^m \max_k (p_{ik}) \quad (4)$$

$$p_e = \sum_{k=1}^q y_k p_k \quad \text{with} \quad y_k = \frac{1}{m} \sum_{i=1}^m y_{ik} \quad \text{and} \quad p_k = \frac{1}{m} \sum_{i=1}^m p_{ik} \quad (5)$$

where p_{ik} represents the probability of observing class k in sample i ; y_{ik} represents the observed presence or absence of class k in sample i ; m is the total number of samples; and q is the total number of classes.

κ_{group} reaches one if all observed classes are predicted with the highest probability. Hence, the disadvantage of κ_{group} is that it cannot account for the relative certainty of the model

in its predictions. In model assessment this is a major problem as the observations serve as reference and not the multinomial model. As a consequence one cannot select among multinomial models with $\kappa_{\text{group}} = 1$

From the above it becomes clear that none of the existing metrics is appropriate to apply to multinomial models that are used in the domain of biogeography. Therefore, the aim of this paper is to present a new metric that can be used to assess agreement both between and within multinomial models and that corrects for chance agreement. The proposed metric is a modification of the metric developed by Vanbelle & Albert (2009). In addition, we compare the behaviour of this new metric with the two most commonly applied other measures in biogeography, one from the first and one from the second family.

METHODS

A generic Kappa-like metric to assess accuracy of multinomial models

$\kappa_{\text{multinomial}}$ assesses the agreement between a multinomial model and observations and accounts for agreement obtained by chance when using observations only (equation 6). Although the emphasis of this paper is on the application of $\kappa_{\text{multinomial}}$ to multinomial predictions and observations y with one success for exactly one out of q classes – hereafter referred to as discrete outcomes ($\{y_{ik} \in Z : 0 \leq y_{ik} \leq 1\}$), $\kappa_{\text{multinomial}}$ can also be applied to discrete models with discrete observations and to multinomial predictions and observations y represented by frequencies ($\{y_{ik} \in R : 0 \leq y_{ik} \leq 1, \text{ Table 1}\}$). The first two cases can be considered as special case of the third. Therefore, we first derive the generic equation for $\kappa_{\text{multinomial}}$ and discuss the specific cases afterwards.

$$\begin{aligned} \kappa_{\text{multinomial}} &= \kappa_{\text{prob}} \kappa_{\text{loc}} \\ &= \frac{p_{0 \text{ multinomial}} - p_{e \text{ multinomial}}}{p_{\text{max}} - p_{e \text{ multinomial}}} \frac{p_{\text{max}} - p_{e \text{ multinomial}}}{1 - p_{e \text{ multinomial}}} \end{aligned} \quad (6)$$

$\kappa_{\text{multinomial}}$ can be thought of as having two components, κ_{prob} and κ_{loc} . κ_{prob} measures the degree to which ranks of the predicted class probabilities correspond to the ranks of the observed class frequencies. It thus reaches one if there is a perfect match between the rank orders of the observations and predictions. It reaches zero if the model has similar performance compared to the null model. κ_{loc} , in turn, measures the certainty of the model in the case of discrete observations. For continuous observations, κ_{loc} measures the mean match of the sorted observed and predicted sample frequencies. κ_{loc} equals zero if the performance of the multinomial equals the null model. It equals one if, for each sample, the sorted predictions exactly matches the sorted observations.

For previous generations of models with discrete predictions, $p_{\text{max}} = 1$, and so this decomposition is trivial, but for

multinomial models, κ_{loc} is an interesting descriptor of the relationship between the model and the data: some models make much more ‘certain’ predictions than others – they have a much higher value of $p_{\text{max}} - p_{e \text{ multinomial}}$. This decomposition separates model ‘certainty’ from model ‘correctness’. Understanding the statistical and biological drivers of model ‘certainty’ with the new generations of models is an important challenge.

κ_{loc} and κ_{prob} are calculated from three measures: the observed agreement between model predictions and observations ($p_{0 \text{ multinomial}}$ equation 7), the agreement expected by chance ($p_{e \text{ multinomial}}$) and the maximum possible agreement of the model (p_{max} ; equation 8). $p_{0 \text{ multinomial}}$ measures the agreement between model predictions and observations, and measures the proportion in common between predictions and observations. It is equivalent to the Manhattan distance which assumes no error in the observed data; the safest assumption for model evaluation (Legendre & Legendre, 1998; Warton *et al.*, 2006).

$$p_{0 \text{ multinomial}} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q |y_{ik} - p_{ik}|}{2} \right) \quad (7)$$

When class predictions and observations consist of presence/absence data or consist of probabilities or frequencies, it can be shown that $\sum_{k=1}^q |y_{ik} - p_{ik}|$ is bounded for any i between $[0, 2]$ since for each sample the sum of the probabilities is 1 and the sum of each observation is 1. Hence, this equation can be applied to all three cases. When observations consists of discrete outcomes equation 3 gives similar results.

For evaluating the performance of multinomial models in the domain of biogeography, we define a different reference model compared to what is used in κ_{group} and Cohen’s κ . κ_{group} and Cohen’s κ were originally designed to assess the agreement between classifiers and assumes that classifiers are equally reliable. Hence it computes chance agreement as the product of the marginal distributions (equation 5). However, in case of evaluating model performance, it seems more appropriate to define the reference model as the expected agreement obtained by using the observations only. That would be a fair reference for all candidate multinomial models tested within a study. There is no convenient analytical solution for the generic case of $p_{e \text{ multinomial}}$. However, $p_{e \text{ multinomial}}$ can be calculated as the average agreement of the observed map and a large number of randomized maps. The agreement between the observations and the randomization is calculated with equation 7. Hence, $p_{e \text{ multinomial}}$ and $p_{0 \text{ multinomial}}$ are calculated consistently.

The maximum agreement of a model that can be obtained given the observations (p_{max}) can be generalized to include models fitted to either discrete or continuous observations. For the discrete case: it was calculated by Vanbelle & Albert (2009) as the average of the most probable classes over all samples. For the continuous case, a model reaches maximum agreement if, for each sample, the rank order of the

Table 1 Overview of the cases to which $\kappa_{\text{multinomial}}$ can be used for each case. Cohens κ is added for comparison. For formatting reasons, multinomial is abbreviated to multi. p_{ik} represent the predicted class occurrence or predicted class frequency of sample i and class k , y_{ik} represent the observed class occurrence or the observed class probability of sample i and class k (out of q classes).

Name method	Observations	Predictions	Reference model (p_e)	Observed agreement (p_0)	Maximum agreement of the model (p_{\max})	Model performance, κ
Cohen's κ	Discrete $\{y_{ik} \in Z : 0 \geq y_{ik} \leq 1\}$	Discrete $\{p_{ik} \in Z : 0 \geq p_{ik} \leq 1\}$	$p_e = \sum_{k=1}^q y_{ik} p_{ik}$ with $y_{ik} = \frac{1}{m} \sum_{i=1}^m y_{ik}$ and $p_{ik} = \frac{1}{m} \sum_{i=1}^m p_{ik}$	$p_0 = p_0^{\text{multinomial}}$ $= \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)$ or $\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^q y_{ik} p_{ik}$	NA	$\kappa = \frac{p_0 - p_e}{1 - p_e}$
$\kappa_{\text{multinomial}}$	Discrete $\{y_{ik} \in Z : 0 \geq y_{ik} \leq 1\}$	Discrete $\{p_{ik} \in Z : 0 \geq p_{ik} \leq 1\}$	Randomization of y_{ik} or $p_{e \text{ multi}} = \sum_{k=1}^q y_{ik} y_{ik}$	$p_0 \text{ multi} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)$ or $\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^q y_{ik} p_{ik}$	$p_{\max} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)^*$ or $\frac{1}{m} \sum_{i=1}^m \max_k(p_{ik}) = 1$	$\kappa_{\text{multi}} = \frac{p_0 \text{ multi} - p_{e \text{ multi}}}{1 - p_{e \text{ multi}}}$
$\kappa_{\text{multinomial}}$	Discrete $\{y_{ik} \in Z : 0 \geq y_{ik} \leq 1\}$	Frequencies/ probabilities $\{p_{ik} \in R : 0 \geq p_{ik} \leq 1\}$	Randomization of y_{ik} or $p_{e \text{ multi}} = \sum_{k=1}^q y_{ik} y_{ik}$ or $\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - y_{ik} }{2} \right)$ with $y_{ik} = \frac{1}{m} \sum_{i=1}^m y_{ik}$	$p_0 \text{ multi} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)$ or $\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^q y_{ik} p_{ik}$	$p_{\max} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)^*$ or $\frac{1}{m} \sum_{i=1}^m \max_k(p_{ik})$	$\kappa_{\text{multi}} = \kappa_{\text{loc}}^{\kappa_{\text{prob}}}$ $= \frac{p_{\max} - p_{e \text{ multi}}}{1 - p_{e \text{ multi}}}$ $\frac{p_0 \text{ multi} - p_{e \text{ multi}}}{p_{\max} - p_{e \text{ multi}}}$
$\kappa_{\text{multinomial}}$	Frequencies $\{y_{ik} \in R : 0 \geq y_{ik} \leq 1\}$	Frequencies/ probabilities $\{p_{ik} \in R : 0 \geq p_{ik} \leq 1\}$	Randomization of y_{ik} with $y_{ik} = \frac{1}{m} \sum_{i=1}^m y_{ik}$	$p_0 \text{ multi} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)$	$p_{\max} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q y_{ik} - p_{ik} }{2} \right)^*$	$\kappa_{\text{multi}} = \kappa_{\text{loc}}^{\kappa_{\text{prob}}}$ $= \frac{p_{\max} - p_{e \text{ multi}}}{1 - p_{e \text{ multi}}}$ $\frac{p_0 \text{ multi} - p_{e \text{ multi}}}{p_{\max} - p_{e \text{ multi}}}$

*The ordered probabilities of p_{i1}, \dots, p_{iq} are defined as $p_{i(1)}, \dots, p_{i(q)}$ with $p_{i(q)}$ being the highest class probability, i.e. $p_{i(q)} = \max_k(p_{ik})$. Likewise, the ordered frequencies of y_{i1}, \dots, y_{iq} are defined as $y_{i(q)}, \dots, y_{i(1)}$ with $y_{i(1)}$ being the highest class frequency, i.e. $y_{i(1)} = \max_k(y_{ik})$ and $y_{i(1)} = \min_k(y_{ik})$.

predicted class probabilities exactly matches the rank order of the observed class frequencies. In that case it holds that $y_{i(q)} > y_{i(q-1)} > \dots > y_{i(1)}$ with (q) being the class predicted with highest probability in sample i , i.e. $(q) = \arg \max_k (p_{ik})$

and (1) being the class with lowest predicted class probability, i.e. $(1) = \arg \min_k (p_{ik})$. p_{\max} is calculated as the agreement between the ordered class probabilities and the ordered class frequencies (equation 8).

$$p_{\max} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q |y_{i(k)} - p_{i(k)}|}{2} \right) \quad (8)$$

with p_{i1}, \dots, p_{iq} defined as $p_{i(q)}, \dots, p_{i(1)}$; $p_{i(q)}$ being the highest class probability $p_{i(q)} = \max_k (p_{ik})$, and $p_{i(1)} = \min_k (p_{ik})$. Likewise, the ordered frequencies of y_{i1}, \dots, y_{iq} are defined as $y_{i(q)}, \dots, y_{i(1)}$ with $y_{i(q)}$ being the highest class frequency, i.e. $y_{i(q)} = \max_k (y_{ik})$ and $y_{i(1)} = \min_k (y_{ik})$.

Applying $\kappa_{\text{multinomial}}$ to multinomial models to discrete observations

At present, multinomial models are most frequently used to predict outcomes with one success for exactly one out of q classes – discrete observations. If only one class is observed in a given sample, κ_{prob} measures the degree to which the most probable class is observed above random scaled to the maximum agreement of the model. κ_{loc} measures the upper limit of such a model above random scaled to the difference between maximum and chance agreement. With discrete observations, p_0 multinomial is calculated with equation 7 although equation 3 can be used as well. In addition, equation 8 is a generalization of equation 4. Therefore, κ_{prob} is similar to κ_{group} (sensu Vanbelle & Albert, 2009), except that chance agreement is calculated differently (for reasons explained earlier). Chance agreement can be computed analytically (equation 9). It represents the average of the distance of the samples compared to the average observed occurrence class frequency.

$$p_{\text{e multinomial}} = \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\sum_{k=1}^q |y_{ik} - y_k|}{2} \right) \text{ with } y_k = \frac{1}{m} \sum_{i=1}^m y_{ik} \quad (9)$$

The behaviour of $\kappa_{\text{multinomial}}$ for a representative set of probability models with 5 classes and discrete observations is graphically depicted in Fig. 2.

A special case of multinomial models is when one of the classes in each sample is predicted with a probability of one. If predictions are discrete, the model can only be wrong in the predicting the wrong class (but not in the probabilities assigned to each class). Thus, every discrete model has a p_{\max} of one. Therefore, $\kappa_{\text{multinomial}}$ cannot be decomposed into κ_{loc} and κ_{prob} . $\kappa_{\text{multinomial}}$ is similar to Cohen's κ if the predictions are hard classified (presence/absence) and the marginal totals of the observations and predictions are equal.

Exploring the behaviour of $\kappa_{\text{multinomial}}$ compared to existing methods

We examined the use, behaviour and the relation of $\kappa_{\text{multinomial}}$ to existing metrics in two ways. First, we created a number of contrasting multinomial models and compared their agreement with a dataset consisting of class occurrences with discrete outcomes. Second, we used an existing dataset with vegetation type occurrences for which we predicted the vegetation type occurrence with multinomial models. For each dataset, the agreement was evaluated with $\kappa_{\text{multinomial}}$ and compared with the two most commonly applied metrics: a member from the first family, Cohen's kappa (κ , Cohen, 1960) and the second family, mAUC (Provost & Domingos, 2001).

Evaluation using a synthetic dataset

We created a synthetic dataset consisting of 1000 samples with each sample assigned to one class and we assessed the agreement of 10 sets of pre-defined multinomial models to this dataset. Across the 10 sets, the p_{\max} ranged from high (i.e., one class with high probability of occurrence and hence large differences between the class with the highest probability and those of other classes) to low (Fig. 3, see Appendix S3.1 in Supporting Information). Within each set, class probabilities were varied across samples and classes. This reflects reality as oftentimes class probabilities co-vary with underlying environmental and socio-economic drivers and differ among samples.

Five out of 10 sets were constructed such that for a given p_{\max} (0.18, 0.29, 0.49, 0.65 and 0.75) the ordered class probabilities were partly overlapping across samples. These sets are expected to lead to contrasting behaviour of $\kappa_{\text{multinomial}}$ and κ because κ transforms the most likely class to presence irrespective of the p_{\max} . Two sets of multinomial models were constructed such that the minimum of the class that was predicted with highest probability was higher than the maximum of the class that was predicted with second highest probability (e.g. $\max_i (p_{i(9)}) < \min_i (p_{i(10)})$). Three sets of models were constructed such that the second most probable class is predicted with a probability that is only slightly (0.93) lower to the class with the highest probability. These latter five sets of models may lead to contrasting behaviour of mAUC and $\kappa_{\text{multinomial}}$ because mAUC measures the degree of discrimination between class probabilities, while $\kappa_{\text{multinomial}}$ measures the average probability with which observed classes are predicted.

Within each set, we distinguished 41 multinomial models. In the first model, the observed classes were predicted with highest probability, hereafter referred to as 'perfect agreement' model ($\kappa_{\text{prob}} = 1$). For the remaining 40 multinomial models in each set two types of mispredictions were imposed. In 20 models, the observed class was predicted with second highest probability. In another 20 models, the observed class was predicted with a randomly chosen class

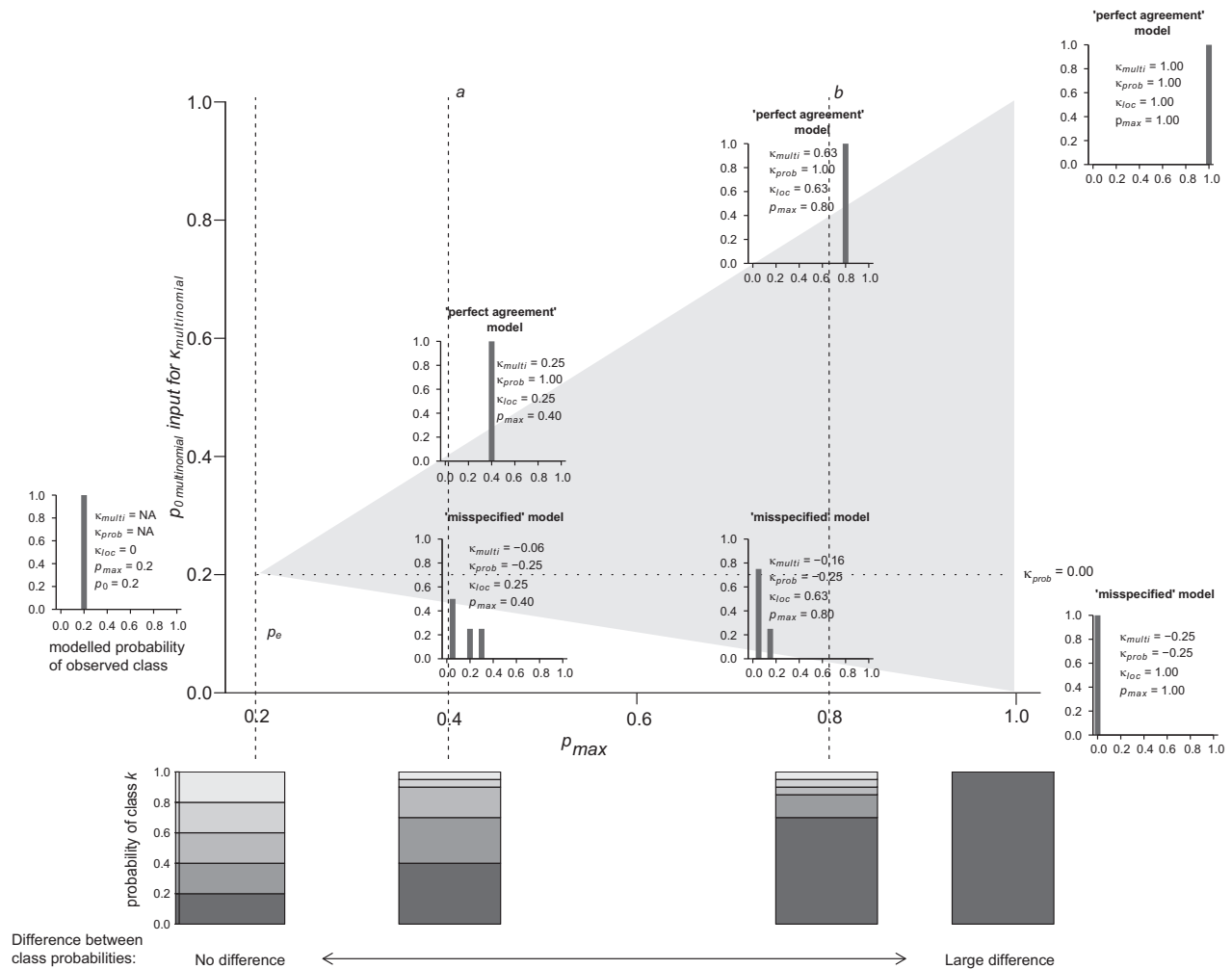


Figure 2 Schematic representation of the relationship between p_{\max} and p_0 multinomial. The shaded grey triangle represents p_0 multinomial of all possible multinomial probability models. The lower part of the graph shows four different possible probabilistic models, with five classes. Based on equal class frequencies it is expected that for 20% of the cells the observed classes is predicted correctly. As the class probabilities differ more from each other, the range in p_0 multinomial increases. For the extreme case of crisp predictions (one class is predicted with a probability of 1), the model can be 100% correct (the probability of observed classes is 1, $\kappa_{\text{multinomial}} = 1$), totally wrong (the probability of observed classes is 0, $\kappa_{\text{multinomial}} = -0.25$) and everything in between. Multinomial models that perform better than random are above the horizontal grey dotted line ($\kappa_{\text{prob}} = 0$). For class probabilities that are more similar to each other (middle two stacked bar plots, dashed black lines a,b), κ_{prob} reaches one if the class that observed is also predicted with highest probability, it reaches lower values when other classes are observed than the one that is predicted with highest probability.

(see Appendix S1 for illustration). Within each type of misprediction, 5% to 100% mispredictions were imposed (20 for each) leading to 10×41 predictive models. For all models, κ , mAUC and κ_{loc} , κ_{prob} and $\kappa_{\text{multinomial}}$ were calculated. The procedure was repeated for a set of predictive models that distinguished five different classes.

Evaluation using observed vegetation type distributions

In a dataset with field observations, the average of the trait values of the plant species in a variety of plant communities was recorded, as well as the vegetation type to which the community belonged. This dataset was used in Douma *et al.* (2012b) to predict the vegetation types based on a number

of plant traits. The plant traits that were used were three expert indicator values for soil acidity (Fa), moisture (Fm), nutrients (Fn) and wood density (SSD). A calibration set was used to derive the relationship between the plant traits and the probabilities of vegetation type occurrence. Model agreement was assessed on an independent set of observations. We compared model agreement to the agreement of a multinomial model distinguishing 38 vegetation types. The 38 vegetation types were a refinement of a classification with 15 different vegetation types, for which we also run the comparison. In addition, we compared model agreement to the agreement of a multinomial model that used three other plant characteristics to predict the 15 vegetation types mentioned earlier. The traits used here were specific leaf area

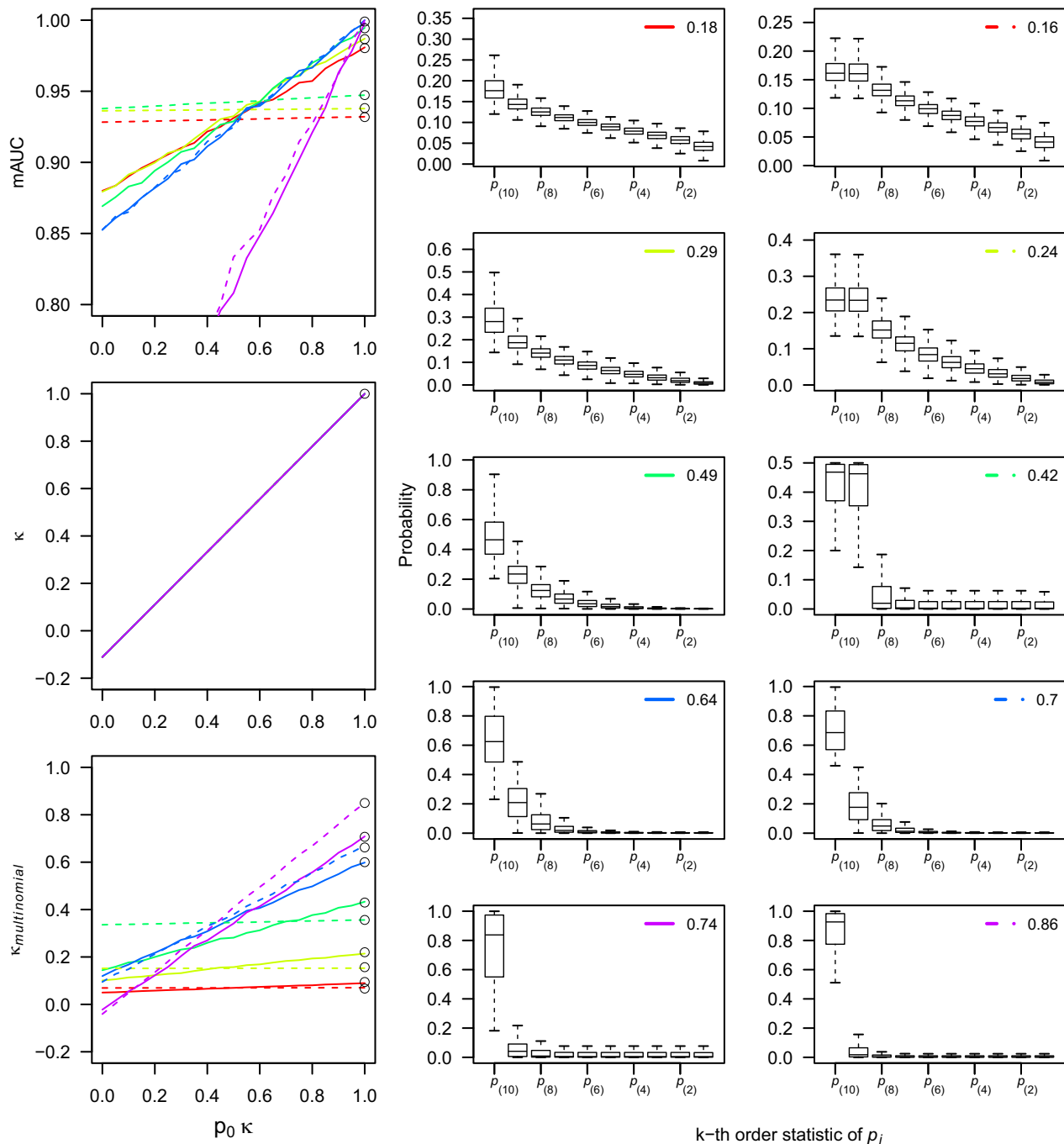


Figure 3 Relationship between the proportion correctly predicted observations (i.e. the proportion of observed classes that was predicted with maximum probability, p_0 in Cohen's Kappa) and three agreement measures: multiple class AUC (mAUC), Cohen's κ and $\kappa_{\text{multinomial}}$ for 10 sets of multinomial models predicting 10 classes, differing in the maximum probability of which classes are predicted (left column). The characteristics of the 10 set of multinomial models are shown in the two right columns, each summarizing the k th order probability over all 1000 samples; $p_{i(10)} = \max_k(p_{ik})$ and $p_{i(1)} = \min_k(p_{ik})$. [Colour figure can be viewed at wileyonlinelibrary.com]

(SLA), average maximum plant height (maxCH) and seed mass (SM). We refer to Douma *et al.* (2012a,b) for details on the plant traits and the technical procedure through which the multinomial probabilities were derived.

$\kappa_{\text{multinomial}}$ has been built into an R package (<https://cran.r-project.org/>). It can be downloaded from https://github.com/bobdouma/kappa_multinomial.git.

RESULTS

Synthetic dataset model agreement behaviour

$\kappa_{\text{multinomial}}$ ranged from 0.09 to 0.85 among the 'perfect agreement' models. The variation in $\kappa_{\text{multinomial}}$ was caused by variation in κ_{loc} , as κ_{prob} was one in all those cases. For a

given value of κ_{loc} , $\kappa_{\text{multinomial}}$ decreases for the ‘non-perfect agreement’ models. The degree to which $\kappa_{\text{multinomial}}$ decreases depends on the certainty with which the classes are predicted. If the model has a high p_{max} and hence a high κ_{loc} than $\kappa_{\text{multinomial}}$ decreases much faster compared to multinomial models with a low p_{max} (compare in Fig. 3: models p_{max} 0.18 versus p_{max} 0.75). Thus, a model with a lower p_{max} may reach a higher $\kappa_{\text{multinomial}}$ than a model with a high p_{max} that does not match the observations.

mAUC also distinguishes among ‘perfect agreement’ models with different p_{max} and decreases when there is a mismatch between the class predicted with highest probability and the class that is observed. mAUC and $\kappa_{\text{multinomial}}$ do, however, differ in their ranking of the multinomial models. In contrast to $\kappa_{\text{multinomial}}$, mAUC cannot differentiate between ‘perfect agreement’ models for which the minimum of the highest class probabilities was higher than the maximum of the second highest class probabilities across samples (i.e. $\min_i(p_{i(10)}) > \max_i(p_{i(9)})$). Thus, the multinomial model with p_{max} of 0.86 and 0.70 both have a mAUC of one. When the probabilities of the most probable class and the second most probable class partly overlap across samples, mAUC decreases. The stronger this overlap, the lower mAUC (compare multinomial models with p_{max} of 0.18 and 0.16). In contrast, among the ‘perfect agreement models’, $\kappa_{\text{multinomial}}$ prefers models based on their p_{max} and ranks the models with a p_{max} of 0.16 lowest among all models.

Cohen’s κ does not distinguish among the 10 different models with different p_{max} , and as hypothesized punishes every mispredictions equally strong. This is particularly undesirable for predictive models that predict two classes with nearly equal probability. If the class that was observed was predicted with second highest probability κ will count this as a full mismatch, while in fact the model is not far off. The behaviour of $\kappa_{\text{multinomial}}$, mAUC and κ did not change when using 5 classes instead of 10 classes (see Appendix S2).

Model agreement behaviour for different vegetation type classifications

In Fig. 4 and Table 2, an overview of the model characteristics and performance metrics are given for the three models (four traits to distinguish 15 vegetation types, three traits to distinguish 15 vegetation types and four traits to distinguish 38 vegetation types; cross tabulation matrices are shown in Appendix 2). The p_{max} had values of 0.85, 0.55, and 0.68 for the three models, respectively. Model agreement was ranked consistently by the three performance metrics. However, κ and mAUC do not reveal how far off the predictive model is from the ‘perfect agreement’ model, while the combination of κ_{loc} and κ_{prob} does. From the bar plots in.

Figure 4 third column, it can be seen that if an observed class was not predicted with highest probability, the second most likely class was observed in most of the cases. The distinctive power of the performance metrics is largest for Cohen’s kappa and smallest for mAUC, the latter hardly

differentiates among the various models. κ scored consistently higher than $\kappa_{\text{multinomial}}$ even though p_e and $p_{e \text{ multinomial}}$ were very similar. This means that mispredictions are punished more heavily in $\kappa_{\text{multinomial}}$ than in κ . This corresponds to the results that were obtained in the synthetic dataset (Fig. 3).

All metrics show that a higher predictive power is obtained when fewer vegetation types are distinguished (each differing in their mean trait values) compared to many classes (each very similar in their mean trait values). In addition, the indicator values (traits derived from expert judgement) were better able to distinguish among vegetation types than the plant traits derived from measurements.

DISCUSSION

The performance of $\kappa_{\text{multinomial}}$

We present a new metric, $\kappa_{\text{multinomial}}$, to assess agreement between a multinomial model and discrete observations while accounting for chance agreement. The metric is independent on how the multinomial probabilities are derived. Probabilities can be derived from multinomial logistic regression models or from Monte Carlo simulations where the average over multiple runs – each with discrete outcomes – are interpreted as probabilities. $\kappa_{\text{multinomial}}$ is applicable to both non-spatial and spatial predictions but it assumes, like Cohen’s Kappa and mAUC that each pixel is independent. The main advantage of $\kappa_{\text{multinomial}}$ is that it partitions model agreement into two components. A first component specifies the correctness of the multinomial model in its prediction of the most likely class (κ_{prob}). Hence, it informs about the degree to which the multinomial model is capable of capturing the (mechanisms that determine) presence and absence of classes. A second component measures the relative certainty with which the multinomial model predicts class occurrences (κ_{loc}). Hence, it informs about the degree to which the system is predictable. κ_{loc} may be used to choose among multinomial models; κ_{prob} to explore the misfit of a given model. κ_{prob} and κ_{loc} are theoretically independent although in the case study in which vegetation types were predicted by a combination of trait values the values of κ_{prob} and κ_{loc} changed consistently over the three models. Hence, these two components may assist researchers to optimize model fit and contrasts to κ and mAUC that measure model agreement with one number.

$\kappa_{\text{multinomial}}$ has two advantages over κ . First, κ seems to overestimate model performance by transforming the probabilities to discrete outcomes. This transformation effectively leads to a p_{max} and a κ_{loc} of one for every multinomial model and thus high κ values. Second, given this transformation κ cannot differentiate between models that differ in the certainty with which classes are predicted (i.e. models that differ in κ_{loc}). In addition, $\kappa_{\text{multinomial}}$ can also be used for hard classified models that are traditionally assessed with Cohen’s kappa (i.e. models with $p_{\text{max}} = 1$). The only difference is the specification of the reference model.

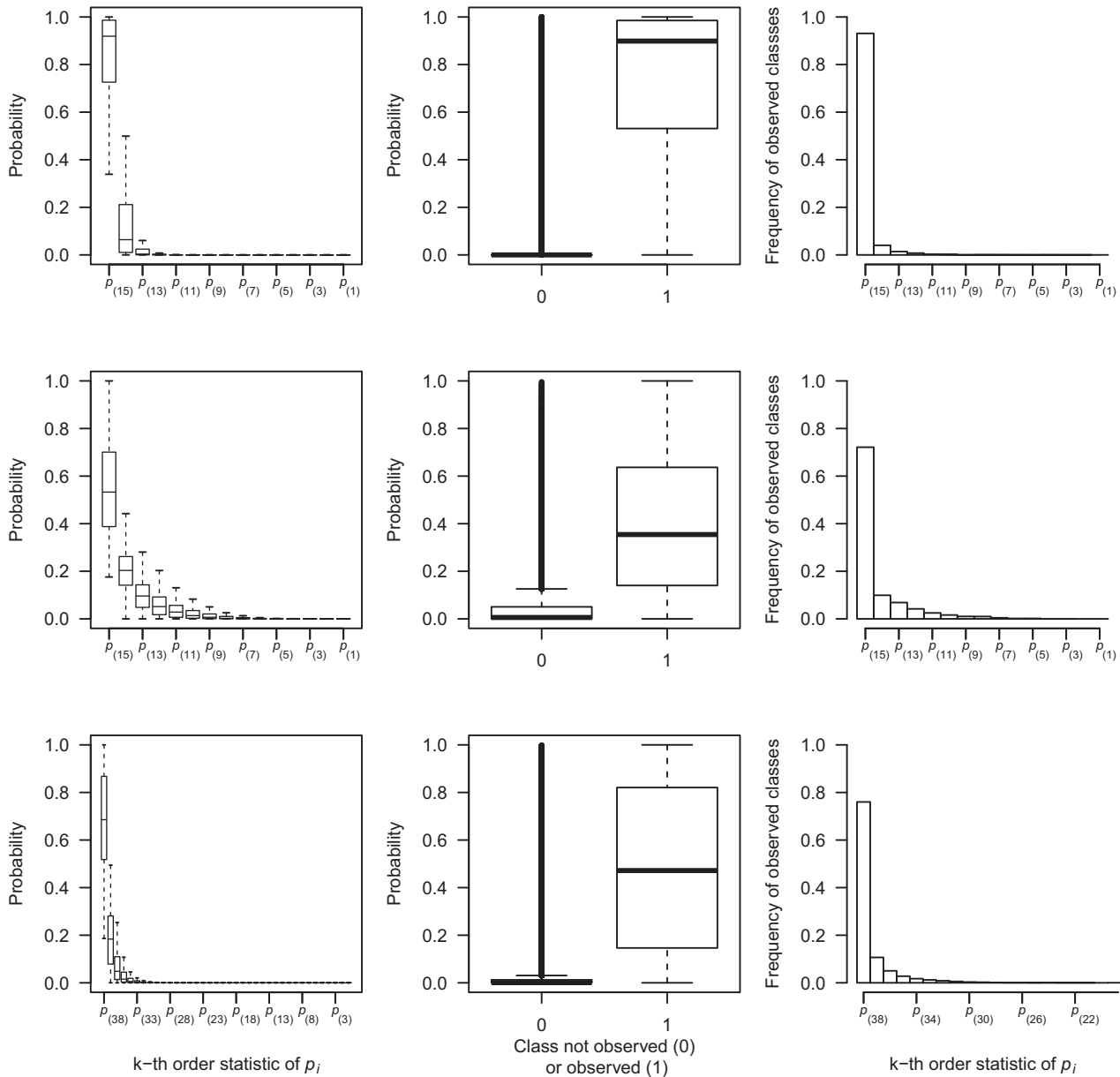


Figure 4 The predictive ability of three predictive models are shown in the rows. The models differ in the maximum probability of which classes are predicted and the number of classes they predict (predicting 15, 15 and 38 vegetation types respectively). First column: distribution of the k th order probable classes over all samples from most probable to least probable class with $p_{i(15)} = \max_k(p_{ik})$ and $p_{i(1)} = \min_k(p_{ik})$ for the three different models. Second column: boxplots of the probabilities with which observed (1) and non-observed (0) classes were predicted. Third column: the frequency with which observed classes were predicted by the i -th order likely class. For the first row, the most probable class was observed in 92% of the cases, and the second most probable class was observed in 6% of the cases. Note that the axis at the bottom right does not display all 38 classes.

$\kappa_{\text{multinomial}}$ differs from mAUC in one important aspect. mAUC cannot discriminate among multinomial models that perfectly discriminate among classes, but that have with different p_{max} . In contrast, $\kappa_{\text{multinomial}}$ does discriminate between such models, as it is based on the average probability with which observed classes are predicted. If the most probable class is always more likely than the second probable class across all samples and the most probable class is observed, mAUC reaches one, irrespective of the certainty with which those classes are predicted. This feature may be desirable

when predicting the right class has important consequences, e.g. for management decisions. However, the condition that the most probable class is consistently more likely than the second probable class over all samples will be hardly met in practice because class probabilities differ across the landscape and co-vary with underlying environmental and socio-economic drivers. This is particularly true in vegetation modelling where a mosaic of vegetation types under homogeneous abiotic conditions may occur (Pacala *et al.*, 1996; Claussen *et al.*, 1999; Scheffer *et al.*, 2001; Rietkerk

Table 2 Coefficients of agreement for κ , mAUC, κ_{loc} , κ_{prob} and $\kappa_{\text{multinomial}}$ for 15 and 38 classes respectively and the plant characteristics used to distinguish the vegetation types.

Number of vegetation types and the plant characteristics	Cohen's kappa (κ)	mAUC	$\kappa_{\text{multinomial}}$
15 (Fa, Fm, Fn, SSD)	$p_0 = 0.78, p_e = 0.08, \kappa = 0.76$	0.98	$p_0 \text{ multinomial} = 0.73, p_e = 0.09, p_{\text{max}} = 0.84, \kappa_{\text{loc}} = 0.85, \kappa_{\text{prob}} = 0.83, \kappa_{\text{multinomial}} = 0.71$
15 (SLA, maxCH, SM)	$p_0 = 0.52, p_e = 0.08, \kappa = 0.48$	0.92	$p_0 \text{ multinomial} = 0.40, p_e = 0.09, p_{\text{max}} = 0.55, \kappa_{\text{loc}} = 0.66, \kappa_{\text{prob}} = 0.51, \kappa_{\text{multinomial}} = 0.34$
38 (Fa, Fm, Fn, SSD)	$p_0 = 0.55, p_e = 0.04, \kappa = 0.53$	0.97	$p_0 \text{ multinomial} = 0.48, p_e = 0.04, p_{\text{max}} = 0.68, \kappa_{\text{loc}} = 0.69, \kappa_{\text{prob}} = 0.67, \kappa_{\text{multinomial}} = 0.46$

et al., 2002), and the occurrence of multiple classes is likely. Therefore, $\kappa_{\text{multinomial}}$ may be better suited to evaluate model agreement in these conditions.

Additional applications of $\kappa_{\text{multinomial}}$

We may extend the application of $\kappa_{\text{multinomial}}$ to multiple other situations. First, $\kappa_{\text{multinomial}}$ can be applied to compare model predictions simultaneously with multiple sources of observations. Different observations on the same sample may disagree in which class is observed because of differences in sampling period, measurement methods and/or decision rules. This is likely to occur in an area that is dynamic in time or in transition between two classes. Averaging over these observations may better represent the possible states that an area can be in and hence more closely match the model predictions. Second, $\kappa_{\text{multinomial}}$ has important applications when assessing the agreement of a multinomial probability model with a reference model that has multinomial probabilities as output. For example, one may directly assess agreement of ground-based derived probabilities with remote sensing derived probabilities. Usually, classification algorithms (tree regression, linear mixture modelling etc.) are used to hard classify spectral information (Xie *et al.*, 2008). However, remote sensing derived probabilities could be used directly in $\kappa_{\text{multinomial}}$. Directly comparing probabilities avoids the rounding error that is introduced by a classification algorithm. Finally, $\kappa_{\text{multinomial}}$ allows assessing the agreement of model predictions and observations at larger spatial scales (Pontius & Cheuk, 2006). Landscape dynamics are to a large extent driven by the spatial and temporal scale of disturbances (Turner *et al.*, 1993). For example, fire outbreaks of relatively large spatial extent and high disturbance intervals will create a mosaic of vegetation types and/or alternative stable states in the landscape (Turner *et al.*, 1993; van Langevelde *et al.*, 2003). If the model fully captures the inherent spatial stochasticity the multinomial model will quickly approach the frequency distribution as derived from the pixels as aggregated to larger scales.

CONCLUSIONS

Research on predicting and understanding spatial and temporal shifts in land cover is increasingly using multinomial

models that model probabilities of outcomes. The argument for this modelling approach is that the uncertainty (both biological and methodological) is preserved in the model output. While probabilistic output is a clearly advantageous in many cases, this new type of output presents a problem in model assessment. Current assessment methods have critical disadvantages when assessing multinomial models. In this paper we presented a new method, $\kappa_{\text{multinomial}}$, that solve these drawbacks.

We show that $\kappa_{\text{multinomial}}$ has several advantages over existing methods such as κ and mAUC. With $\kappa_{\text{multinomial}}$, we have provided a tool that directly uses the multinomial probabilities of model predictions for accuracy assessment. Assessing these models accurately will lead to better and more accurate future generations of these important land cover models.

ACKNOWLEDGEMENTS

This study was carried out in the framework 'Kennis voor Klimaat' theme 6. WKC acknowledges NWO for funding. We thank Rien Aerts and Lieneke Verheijen for critically reading the manuscript, Lia Hemerik and Willem Kruijer for discussion on the performance measures. We also thank A. M. Prasad, R.G. Pontius Jr, J. van Vliet and two anonymous referees for comments that greatly improved earlier versions of this manuscript.

REFERENCES

- Ackerly, D.D., Cornwell, W.K., Weiss, S.B., Flint, L.E. & Flint, A.L. (2015) A Geographic mosaic of climate change impacts on terrestrial vegetation: which areas are most at risk? *PLoS One*, **10**, e0130629.
- van Bodegom, P.M., Douma, J.C. & Verheijen, L.M. (2014) A fully traits-based approach to modeling global vegetation distribution. *Proceedings of the National Academy of Sciences USA*, **111**, 13733–13738.
- Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.
- Claussen, M., Kubatzki, C., Brovkin, V., Ganopolski, A., Hoelzmann, P. & Pachur, H.J. (1999) Simulation of an abrupt change in Saharan vegetation in the mid-Holocene. *Geophysical Research Letters*, **26**, 2037–2040.

- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.
- Dendoncker, N., Rounsevell, M. & Bogaert, P. (2007) Spatial analysis and modelling of land use distributions in Belgium. *Computers Environment and Urban Systems*, **31**, 188–205.
- Douma, J.C., Aerts, R., Witte, J.P.M., Bekker, R.M., Kunzmann, D., Metselaar, K. & van Bodegom, P.M. (2012a) A combination of functionally different plant traits provides a means to quantitatively predict a broad range of species assemblages in NW Europe. *Ecography*, **35**, 364–373.
- Douma, J.C., Witte, J.-P.M., Aerts, R., Bartholomeus, R.P., Ordonez, J.C., Venterink, H.O., Wassen, M.J. & van Bodegom, P.M. (2012b) Towards a functional basis for predicting vegetation patterns; incorporating plant traits in habitat distribution models. *Ecography*, **35**, 294–305.
- Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.
- Ferri, C., Hernández-Orallo, J. & Salido, M. (2003) Volume under the ROC surface for multi-class problems. *Machine learning: ECML 2003*, pp. 108–120. (eds. by N. Lavrac, D. Gamberger, H. Blockeel and L. Todorovski). Springer, Berlin Heidelberg.
- Ferri, C., Flach, P., Hernández-Orallo, J. & Senad, A. (2005) *Modifying ROC curves to incorporate predicted probabilities*. Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning, Bonn, Germany.
- Ferri, C., Hernandez-Orallo, J. & Modroiu, R. (2009) An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, **30**, 27–38.
- Foody, G.M. (1992) On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, **58**, 1459–1460.
- Gotelli, N.J. & Graves, G.R. (1996) *Null models in Ecology*. Smithsonian Institution Press, Washington, D.C.
- Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. & Dougherty, E.R. (2010) Small-sample precision of ROC-related estimates. *Bioinformatics*, **26**, 822–830.
- Hand, D.J. & Till, R.J. (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, **45**, 171–186.
- Hepinstall, J.A., Alberti, M. & Marzluff, J.M. (2008) Predicting land cover change and avian community responses in rapidly urbanizing environments. *Landscape Ecology*, **23**, 1257–1276.
- Jurman, G., Riccadonna, S. & Furlanello, C. (2012) A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE*, **7**, e41882.
- van Langevelde, F., van de Vijver, C., Kumar, L., van de Koppel, J., de Ridder, N., van Andel, J., Skidmore, A.K., Hearne, J.W., Stroosnijder, L., Bond, W.J., Prins, H.H.T. & Rietkerk, M. (2003) Effects of fire and herbivory on the stability of savanna ecosystems. *Ecology*, **84**, 337–350.
- Le, Q.B., Park, S.J. & Vlek, P.L.G. (2010) Land use dynamic simulator (LUDAS): a multi-agent system model for simulating spatio-temporal dynamics of coupled human-landscape system 2. Scenario-based application for impact assessment of land-use policies. *Ecological Informatics*, **5**, 203–221.
- Legendre, P. & Legendre, L. (1998) *Numerical Ecology*, 2nd edn. Elsevier Science BV, Amsterdam.
- Lenihan, J.M., Drapek, R., Bachelet, D. & Neilson, R.P. (2003) Climate change effects on vegetation distribution, carbon and fire in California. *Ecological Applications*, **13**, 1667–1681.
- Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.
- Monserud, R.A. & Leemans, R. (1992) Comparing the global vegetation maps with the Kappa statistic. *Ecological Modelling*, **62**, 275–293.
- Muller, D. & Zeller, M. (2002) Land use dynamics in the central highlands of Vietnam: a spatial model combining village survey data with satellite imagery interpretation. *Agricultural Economics*, **27**, 333–354.
- Pacala, S.W., Canham, C.D., Saponara, J., Silander, J.A., Kobe, R.K. & Ribbens, E. (1996) Forest models defined by field measurements: estimation, error analysis and dynamics. *Ecological Monographs*, **66**, 1–43.
- Pontius, R.G. & Cheuk, M.L. (2006) A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, **20**, 1–30.
- Pontius, R.G. & Millones, M. (2011) Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, **32**, 4407–4429.
- Provost, F. & Domingos, P. (2001) *Well-trained PETs: improving probability estimation trees*. In: CeDER Working Paper #IS-00-04, Stern School of Business, New York University, New York, NY, 10012.
- Rietkerk, M., Boerlijst, M.C., van Langevelde, F., HilleRisLambers, R., van de Koppel, J., Kumar, L., Prins, H.H.T. & de Roos, A.M. (2002) Self-organization of vegetation in arid ecosystems. *The American Naturalist*, **160**, 524–530.
- Scheffer, M., Carpenter, S., Foley, J.A., Folke, C. & Walker, B. (2001) Catastrophic shifts in ecosystems. *Nature*, **413**, 591–596.
- Sitch, S., Huntingford, C., Gedney, N., Levy, P.E., Lomas, M., Piao, S.L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C.D., Prentice, I.C. & Woodward, F.I. (2008) Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five dynamic global vegetation models (DGVMs). *Global Change Biology*, **14**, 2015–2039.

- Sokolova, M. & Lapalme, G. (2009) A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, **45**, 427–437.
- Turner, M.G., Romme, W.H., Gardner, R.H., Oneill, R.V. & Kratz, T.K. (1993) A revised concept of landscape equilibrium - disturbance and stability on scaled landscapes. *Landscape Ecology*, **8**, 213–227.
- Vanbelle, S. & Albert, A. (2009) Agreement between an isolated rater and a group of raters. *Statistica Neerlandica*, **63**, 82–100.
- Verburg, P., Schot, P., Dijst, M. & Veldkamp, A. (2004) Land use change modelling: current practice and research priorities. *GeoJournal*, **61**, 309–324.
- Warton, D.I., Wright, I.J., Falster, D.S. & Westoby, M. (2006) Bivariate line-fitting methods for allometry. *Biological Reviews*, **81**, 259–291.
- Webb, N.F., Hebblewhite, M. & Merrill, E.H. (2008) Statistical methods for identifying wolf kill sites using global positioning system locations. *Journal of Wildlife Management*, **72**, 798–807.
- Wu, S., Flach, P. & Ferri, C. (2007) An improved model selection heuristic for AUC. *Machine Learning: ECML 2007*. pp. 478–489. (eds. by J.N. Kok, J. Koronacki, R.L. de Mantaras, S. Matwin, D. Mladenič and A. Skowron). Springer, Berlin Heidelberg.
- Xie, Y.C., Sha, Z.Y. & Yu, M. (2008) Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology-Uk*, **1**, 9–23.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Example of a misspecified model.

Appendix S2 Systematic comparison of $\kappa_{\text{multinomial}}$, mAUC and κ .

Appendix S3 R-scripts to calculate agreement measures.

BIOSKETCHES

Jacob Douma is a quantitative ecologist. His interests focus on understanding and quantifying the processes that determine plant community assembly, plant–insect interactions, and pest invasion through the development of (statistical) models.

Will Cornwell is a plant ecologist with research interests at the intersection of plant eco-physiology, community ecology, and ecosystem ecology. He is especially interested in using basic ecological tools, especially functional traits, to understand the effects of climate change on terrestrial biodiversity.

Peter van Bodegom is an ecologist, working at the interface between community ecology, macroecology and earth system modelling. With his group, he aims to quantify globally applicable functional relationships between vegetation, soil micro-organisms and their environment by targeted experiments, meta-analyses and process-based modelling. He applies the ecological insights thus derived to better predict human impacts on ecosystem functioning and ecosystem services.

Editor: John Stewart & Jon Sadler.