Application of consensus theory to formalize expert evaluations of plant species distribution models[i]

Maarten van Zonneveld[1,2], Nora Castañeda[3,4], Xavier Scheldeman[4], Jacob van Etten[1], Patrick Van Damme[2,5,6]

van Zonneveld, M. (corresponding author, m.vanzonneveld@cgiar.org) & **van Etten, J.** (j.vanetten@cgiar.org): Bioversity International, Costa Rica Office, c/o CATIE, 7170 Turrialba, Costa Rica
van Zonneveld, M. & Van Damme, P. (Patrick.VanDamme@ugent.be): Ghent University, Faculty of Bioscience Engineering, Coupure links 653, 9000 Gent, Belgium
Castañeda, N. (n.p.castaneda@cgiar.org) & **Scheldeman, X.** (xschelde@gmail.com): Bioversity International, Regional Office for the Americas, PO Box 6713, Cali, Colombia
Castañeda, N.: Centro Internacional de Agricultura Tropical (CIAT), PO Box 6713, Cali, Colombia.
Van Damme, P.: World Agroforestry Centre (ICRAF), Global Research Programme 1, PO Box 30677, Nairobi, 00100, Kenya.
Van Damme, P.: Institute of Tropics and Subtropics, Czech University of Life Sciences. Prague, Kamycka 129, Prague 6, Suchdol, 165 21, Czech Republic.

Abstract

Aim: Application of Environmental Envelope Modelling (EEM) for conservation planning requires careful validation. Opinions of experts who have worked with species of interest in the field can be a valuable and independent information source to validate EEM because of their first-hand experience with species occurrence and absence. However, their use in model validation is limited because of the subjectivity of their feedback. In this study we present a method on the basis of cultural consensus theory to formalize expert model evaluations.

Methods: We developed for five tree species, distribution models with nine different variable combinations and Maxent EEM software. Species specialists validated the generated distribution maps through an online Google Earth interface with the scores from *Invalid* to *Excellent*. Experts were also asked about the commission and omission errors of the distribution models they evaluated. We weighted expert scores according to consensus theory. These values were used to get to a final average expert score for each of the produced distribution models. The consensus-

---

[i] This is the accepted authors' manuscript of the final and published article: van Zonneveld, M., et al. (2013). Application of consensus theory to formalize expert evaluations of plant species distribution models. Applied Vegetation Science http://dx.doi.org/ 10.1111/avsc.12081.

.

weighed expert scores were compared with un-weighed scores and correlated to four conventional model performance parameters after cross-validation with test data: Area Under Curve (AUC), maximum Kappa, commission error and omission error.

Results: The median consensus-weighed expert score of all species-variable combinations was close to *Fair*. In general, experts that reached more consensus with peers were more positive about the EEM outcomes compared to those that had more opposite judgements. Both consensus-weighed and un-weighed scores were significantly correlated to corresponding AUC, maximum Kappa and commission error values, but not to omission errors. More than half of the experts indicated that the distribution model they considered best, included areas where the species is known to be absent. One third also indicated areas of species presence that were omitted by the model.

Conclusions: Our results indicate that experts are fairly positive about EEM outcomes but its application for conservation actions remains limited according to them. Methods to formalize expert knowledge allow a wider use of this information in model validation and improvement, and they complement conventional validation methods of presence-only modelling. Online GIS and survey applications facilitate the consultation of experts.

Keywords: Cultural consensus theory; Environmental Envelope Modelling; Expert opinion; Google Earth; *In situ* conservation; MAPFORGEN; Model validation; Online survey; Species distribution modelling

Nomenclature: Germplasm Resources Information Network (GRIN) taxonomy

Running head: Expert evaluation of species distribution models

Introduction

A good understanding of the actual distribution of any plant species is one of the key parameters allowing evaluation of its conservation status and the formulation of effective conservation strategies. However, for most plant species, only a limited amount of data on their distribution is available (Nic Lughadha et al. 2005; Newton and Oldfield 2008). This is particularly true for regions that harbour high levels of plant diversity, including tropical and subtropical zones in Africa, Asia, Latin America and the Caribbean (Nic Lughadha et al. 2005).

Environmental Envelope modelling (EEM) can be used to develop predictive models that make inferences about species´ geographic distributions (Araújo and Peterson 2012). EEM is therefore considered a useful tool to overcome the lack of complete distribution data (Guarino et al. 2002). This kind of modelling technique defines a species' ecological niche to predict areas of potential species occurrence. This is done on the basis of environmental data obtained for occurrence sites where a species has been observed and from sites where it is absent. Because absence points are difficult to obtain, often randomly generated background points are used as an alternative to discriminate less suitable environments from more suitable environments in areas where the species has been observed (Pearce and Boyce 2006). Presence points can be derived from georeferenced herbarium specimens, genebank accessions and/or vegetation/plant species inventories. The latter are made increasingly available online by herbaria and gene banks through portals like the Global Biodiversity Information Facility (GBIF) (www.gbif.org). One of the advantages of EEM is that no prior knowledge on the ecophysiology or reproductive biology of plant species is needed to develop a model (Guisan and Zimmerman 2000). This allows a systematic approach to predict distributions and assess the conservation status for large species numbers.

EEM is therefore now widely applied in ecological and biodiversity conservation studies (Araújo and Peterson 2012). Yet, application of this tool in conservation planning should be critically evaluated. To it is, the algorithm chosen to model species distribution from actual observation data influences the outcomes. This may lead to modelled distributions that deviate significantly from reality (Loiselle et al. 2003). An additional challenge comes from the fact that the modelled distribution ranges are influenced by the environmental variables included and/or omitted in the model. An adequate selection of determinant variables for any species distribution can thus improve the model significantly (Austin 2007).

The results of EEM presence-only modelling have therefore been extensively cross-validated with test data consisting of presence and pseudo-absence points using statistical parameters like maximum Kappa and/or Area Under Curve (AUC) (e.g. Loiselle et al. 2003; Elith et al. 2006; Hernandez et al. 2006). Nevertheless, because of the lack of confirmed points of species absence, it remains difficult to provide a good estimate of the commission error; this is the extent to which models predict occurrence in areas where the species is actually absent (Anderson et al. 2003; Rupprecht et al. 2011). In addition, sampling bias can result in dependence between presence points to train a distribution model and the presence points to test model performance (Dorman et al. 2007). This may lead to high rates of model performance whereas a model may actually omit many not-yet-sampled areas of species occurrence (Hijmans 2012).

Opinions of experts, like foresters, ecologists, botanists and park managers are another key information source that can be used to validate and fine-tune the outcomes of EEM because of their experience with specific species in the field (Thuiller 2003; Beauvais et al. 2006). In addition to species presence, they also can provide valuable information about the extent to which models predict species absence in areas where species don't occur naturally. They may also be a good source to validate model performance in under-sampled areas. Expert feedback also provides insight on how relevant potential users consider distribution modelling to be for their field activities on *in situ* conservation, seed collection and inventories of specific species, to name just a few potential uses. This fits in a wider discussion about the applicability of species distribution mapping and EEM for conservation and sustainable use of biodiversity (Knight et al. 2008; Araújo and Peterson 2012; Guisan et al. 2013).

Increased computer capacity and internet availability during the last decade have allowed the development and widespread application of many new, powerful EEM tools to predict species distribution (e.g. Elith et al. 2006; Thuiller et al. 2009). At the same time, this has allowed the development of online mapping tools, like ArcGIS Server, Google Earth and GeoWiki, which make it possible to remotely consult specialists including botanists, ecologists or park managers located in different parts of a country and in the world, and consider their opinion in distribution model validation and improvement. In many cases only a few experts are available for specific plant species. In these cases, these tools make it possible to connect them online and ask for their feedback in a systematic and efficient way.

Expert-based judgements are often not applied or reported in evaluating EEM because they are considered to be subjective. Measurements of model performance that use presence and pseudo-absence points, like AUC or maximum Kappa, are, despite their limitations, preferred in EEM studies. This is because of their formal nature which allows repeatability and comparability between different studies.

It is possible, though, to analyse expert-based opinions in a more objective way. Romney et al. (1986) developed an approach to formalize informant knowledge on the basis of cultural consensus theory. The consensus model estimates the probability that an informant provides correct answers in function of the concordance of her/his answers with overall group consensus. This technique has been applied in social and ethnobotanical sciences to weigh informant responses (e.g. Weller and Mann 1997; van Etten 2006; Benz et al. 2007).

In this study, we present an approach on how to formalize expert evaluation applying consensus theory to examine the relevance of distribution models for species' conservation assessment and planning. First, the rate of expert agreement indicates how reliable the expert evaluations are to select the best distribution model. Secondly, consensus theory allows identifying for each expert how trustworthy his/her answers are compared to other specialists. This information can be used to weigh the opinions of different experts in average scores for model evaluation.

In EEM, expert knowledge has been used to identify critical environmental variables and species environmental ranges in the case of small sample sizes (Barry and Elith 2006) or to identify areas for crop suitability (Ecocrop 2007). It is also being incorporated in the development of distribution models (Bierman et al. 2010). But we found only a few references that reported the use of experts for model evaluation (Anderson et al. 2003; Ramírez-Villegas et al. 2010). To our knowledge, this is the first time an approach is presented to formalize expert knowledge for the validation of EEM outcomes.

For five socio-economically important tree species native to Latin America and the Caribbean, we present distribution models run in Maxent with nine different environmental variable combinations. Species specialists evaluated model outcomes through an online survey in Google docs with a dynamic Google Earth interface. We compare expert judgements, with and without applying consensus theory, with four commonly used measures of model performance after cross-validation with test data; maximum Kappa, Area Under Curve (AUC), and commission and omission error. We further examine the patterns of variable selection and model appreciation by experts with and without applying consensus theory.

Methods

Species

The five tree species we tested here are *Annona cherimola* Mill. (cherimoya), *Bactris gasipaes* Kunth (peach palm), *Bertholletia excelsa* Bonpl. (Brazil nut), *Cedrela odorata* L. (Spanish cedar) and *Nothofagus nervosa* Phil. (raulí; synonym for *N. alpina* (Poepp. & Endl.) Oerst.). These species were prioritized by LAFORGEN, the Latin American Forest Genetic Resources Network of scientists and practitioners, and have been selected in a project named MAPFORGEN (www.mapforgen.org)). This project aims at evaluating the conservation status of 100 socio-economically important woody species native to Latin America and the Caribbean. As part of this analysis, the species distribution ranges are modelled. The five selected species occur in different ecological and geographical zones in Latin America and the Caribbean, and their distribution has been studied relatively well compared to other MAPFORGEN species.

EEM

We applied a presence-only EEM approach using the Maxent program (Phillips et al. 2006). This is a widely used EEM tool of which the algorithm is reported as predicting species distribution well, in comparison to other modelling software (Elith et al. 2006; Hernandez et al. 2006). It is already used by several environmental agencies (Elith et al. 2011).

We obtained presence points coming from herbaria and genebank passport data for the five selected species through GBIF. This dataset was complemented with presence points provided by several members of LAFORGEN: Corporación para el Desarrollo de los Recursos Naturales (CEDERENA) Ecuador; World Agroforestry Centre (ICRAF) Peru; Instituto Forestal (INFOR) Chile; and Instituto Nacional de Tecnología Agropecuaria (INTA) Argentina. We only considered points within the native distribution ranges defined according to the Germplasm Resources Information Network (GRIN) of the United States Department of Agriculture, Agricultural Research Service, National Genetic Resources Program (USDA, ARS, NGRP) (http://www.ars-grin.gov/). The timber species *C. odorata*, and *N. nervosa*, and non-timber species *B. excelsa* occur in general only in natural populations. The fruit species *A. cherimola* and multi-use palm species *B. gasipaes* are in phases of incipient or semi-domestication (National Research Council 1989; Clement et al. 2010). These species rely partly on conservation through use on farms and presence records include observations from natural populations, and trees maintained *circa situm* in backyards, home gardens and smallholder farms.

We checked the presence points for inconsistencies between the recorded coordinates and the reported highest-level administrative unit in a country (e.g. departments or states), after Scheldeman and van Zonneveld (2010). Inconsistent points were removed. In addition, we used a Mahalanobis distance analysis to identify points in atypical climates ($0.025 < p > 0.975$) as they are probably errors (Chapman 2005). Distances between points were calculated with values of 19 bioclimatic variables as defined by Busby (1991) representing different interannual bioclimatic conditions important for a plant's establishment and survival. Climate data were derived, for each species presence point, from the 30 seconds resolution Worldclim dataset (Hijmans et al. 2005).

Each of the nine models that we developed in Maxent, used as input a different environmental variable combination from the 19 bioclimatic variables, one soil-type classification map and a categorical ecological zone map (Table 1). Climatic variables are important factors to explain geographic patterns of species diversity and distribution at large spatial scales (Pearson and Dawson 2003; Field et al. 2008). Soils play an important role in shaping plant distribution and diversity at smaller spatial scales (Willis and Whittaker 2002; Pearson and Dawson 2003). Data on ecological zones help to further define species distribution areas.

The 19 bioclimatic variables can be clustered in different groups of highly correlated variables of mean annual values and intra-annual fluctuations of temperature and precipitation (van Zonneveld et al. 2009). We therefore also selected a core set of four bioclimatic variables that represent precipitation and temperature mean annual values and seasonality. This set of variables consisted of annual mean temperature (C°), annual precipitation (mm/y), temperature seasonality (standard deviation of monthly temperature x 100) and precipitation seasonality (variation coefficient of monthly precipitation). The map of soil units was derived from the SOTERLAC database (Batjes 2005) and followed FAO´s classification of soil units (FAO 1988). The map of ecological zones was derived from FAO's terrestrial ecological zone classification (FRA 2001).

We used Maxent default settings and applied the 10 percentile training presence threshold to restrict potential distribution areas. The latter is one of the threshold values provided by Maxent and limits the modelled areas of occurrence to a distribution range in which 90% of the presence points are located inside the modelled area while 10% of the presence points are outside the modelled areas of occurrence. Background points were taken from the whole study area that comprises Latin America and the Caribbean (maximum longitude in decimal degrees = -32.375, minimum longitude = -121.125, maximum latitude = 34.5833, minimum latitude = -55.9583). From the modelled areas, we excluded intensive agricultural areas, bare lands and urban areas as delineated by the Global Land Cover 2000 Project (Fritz et al. 2003). We anticipate that our tree species don't occur in these land use types because they have low forest cover and no natural vegetation.

Online expert evaluation survey

For each species, we developed an online survey in Spanish (see Appendix 1). Hyper Text Markup Language (HTML) code and Cascading Style Sheets (CSS) were used to develop a web page to present a questionnaire for each of the five species. The script of the Webpages can be adapted for own use. Within the web page of each species respectively, the nine modelled distribution maps were presented in Keyhole Markup Language (KML) format in an Application Programming Interface (API) of Google Earth. An embeddable form hosted in Google Docs was included in the web pages to store the evaluation scores provided by experts. For each species, we sent an invitation with a link to the online survey to: 1) LAFORGEN members who had indicated research interest in conservation and use of the respective species (many of them are actively involved in such research); and 2) researchers who we found to have studied these species, following a literature review of genetic and ecological studies for each respective species. In total, 99 persons were invited to participate. The survey took place from 10 August 2009 to 29 September 2009.

In the Google Earth interface, each respective species expert could select and view the modelled distribution derived from each of the nine variable combinations to evaluate them visually. Experts were asked to concentrate on the areas they knew best under the assumption that the variable combinations would predict species occurrence with the same quality across the whole distribution range. We asked them to indicate their geographic area of expertise (e.g. country and/or departments or provinces). Distribution maps were presented on a scale from low (yellow) to high probability (red) of species occurrence. Experts did not receive information about the environmental datasets that had been used to generate each model. Specialists could zoom to the geographical distribution area of their expertise (we recommended a minimum eye height of about 25 miles (~ 40 km), whereas they could choose one of five scores to rate the modelled distributions: 1 (invalid), 2 (bad), 3 (fair), 4 (good) and 5 (excellent). Maps in Appendix 2 show the concordance and disagreement of the species distribution models under the different variable combinations in their predictions of species occurrence.

Commission and omission errors according to experts

Distribution models used in conservation planning should ideally have a low commission error to minimize the costs for implementing conservation measures to protect species (Araújo and Peterson 2012). Over-prediction resulting in a high rate of commission errors can occur because migration limitations to species movement are not taken into account in EEM. For example, past and current barriers can substantially restrict real distributions compared to their potential distributions (Svenning and Skov 2007). On the other hand, for the discovery of new populations it is important that models have a low omission error (Araújo and Peterson 2012). Accessing new populations is important for germplasm collecting and to improve *in situ* conservation of species´ genepools. Omission errors may occur because of sampling bias resulting from over-sampling in areas which are easy accessible, like areas close to roads. Sampling is much more difficult in more remote areas with potentially new species populations, which remain under-represented and may consequently be under-predicted in EEM (Hijmans 2012).

Therefore, we also asked each expert if the model that he or she had selected as producing a distribution most similar to the species distribution in their area of expertise contained commission and/or omission errors. We further asked the reasons for commission error; whether model prediction in areas of species absence was due to human-mediated species extinction and/or because these areas were outside the native distribution range.

Application of consensus theory to formalize expert evaluation

The consensus model assumes that each informant has a probability to provide the correct answers to questions on which a researcher doesn´t have the right answers prior to inquiry (Romney et al. 1986). In our case, we didn't know how the different distribution models are related to the real species distributions. The model further assumes that respondents make their observations within the same cultural context (Romney et al. 1986). In our case, we tapped into a community of biological scientists. We assume that this community consists of one cultural group, although our experts come from different biological disciplines and were maybe trained with other conceptual backgrounds. A third postulation is that informants' answers are independent from each other (Romney et al. 1986). We consulted each expert individually about their opinion on the produced models.

The consensus model estimates the level of accuracy of an informant´s response to questions by its concordance to the answers of the other informants in a group. The levels of accuracy or competence rates (D) are calculated for each informant, they are between 0 and 1, and can be used to weigh each informant's response in the final analyses. Indeed, the results from several case studies support consensus theory confirming that within a cultural group, informants whose answers are closer to consensus also have more correct answers compared to persons whose answers are more opposed to consensus (see Romney et al. 1986). The former persons tend also to be more consistent in their answers when they are being asked again after a certain period.

In our study, we used the rate of agreement between species experts as a way to estimate the reliability of the overall expert model evaluation and model selection for a specific species. Secondly, we used the expert competence rates to weigh average expert scores per species-variable combination. In the remaining text of this paper, we will refer to these scores as consensus-weighed expert scores. Similarly, un-weighed expert scores were calculated, but without taking into account competence values.

We will examine how consensus-weighing influences (1) best model selection according to experts; (2) quality of the distribution models according to experts; (3) expert score correlation with Maximum Kappa and AUC, and commission and omission errors; and (4) commission and omission errors according to experts. The steps involved to calculate competence values were written with the basic functions included in R (R Development Core Team 2010).

The first step in consensus model calculation is the development of a matrix with the proportions of agreement in answers between paired experts. Originally, Romney et al. (1986) developed this matrix on the basis of the rates of matches between 0 and 1 in answers on true/false or multiple choice questions (Romney et al. 1986). Later this has been extended to covariance matrices (Weller & Mann 1997). In our case, each species expert provided a rank score from 1 to 5 for nine different models. Instead of rate of matches or covariance, we then calculated the proportion of agreement between respondents as Spearman correlation coefficients. The main difference between correlation coefficients and rates of matches is that correlation coefficients can also be negative when two experts systematically disagree, and thus range from -1 to 1.

The second step is correction of matches for guessing (Romney et al. 1986). In our case, the chances that two respondents return the same series of scores by simply guessing are practically zero. However, to avoid singular computations in further analysis of the correlation matrix, we subtracted 0.0001 from pairwise correlation coefficients.

We then carried out a maximum-likelihood factor analysis on the correlation coefficient matrices. This was only done with one factor, as indicated by Romney et al. (1986). The amount of variance explained in this first factor reflects the rate of consensus between experts (Weller &

Mann 1997). We used this as an indicator of the rate of expert agreement on model performance and best model selection.

The results from the maximum-likelihood factor analysis were also used to obtain for each expert its competence rate on the basis of his/her concordance with group consensus. Expert scores can only be weighed with zeros or positive competence rates ($0 \leq D \leq 1$). In our case, though, an expert could receive negative competence rates when he or she rated consistently opposite to consensus scores. In these cases, values were converted to zero, i.e. the lowest competence value that can be contributed to weigh expert scores.

We use the terminology of competence rates following Romney et al. (1986) to estimate expert agreement and accuracy in model validation. By no means, these rates refer to the overall professional skills and knowledge of our experts.

Selection and relevance of variable combinations

We carried out a non-parametric ANOVA test (Friedman) to test if the model outcomes of one or more of the nine variable combinations were consistently more appreciated by the experts of the five different species compared to the results with the other combinations. We also examined if there were differences in variable combination appreciation between consensus-weighed and un-weighed expert scores.

Correlation of expert-based judgement with conventional model performance parameters

We compared consensus-weighed and un-weighed expert scores with the corresponding values of four commonly used parameters in the validation of EEM outcomes: AUC, maximum Kappa, and commission and omission error values from cross-validation. Correlations were calculated as Pearson´s coefficient. Kappa measures the proportion of agreement between the test data and the modelled areas of species occurrence and absence (Fielding and Bell 1997). In presence-only modelling, AUC is the likelihood that a randomly selected presence point from test data is located at a site with a higher probability of species occurrence than that of a pseudo-absence point, i.e. a randomly selected point in the study area (Philips et al. 2006). Commission errors were calculated as the percentage of false positives in the test data, yielding a predicted distribution area of where the species in reality is absent (Araújo and Peterson 2012). In a similar way, omission errors were calculated as the percentage of false negatives in the test data.

To calculate these four parameters, we trained every distribution model in Maxent with 75% of randomly selected presence points and each model was cross-validated with test data in DIVA-GIS. Test data consisted of 25% of the remaining observation data and pseudo-absence points (five times the number of presence points), randomly generated in the geographic bounding box of the test data. Pseudo-absence points were restricted to this bounding box to reduce the number of such points that are located far away from the known, observed distribution range. This type of points may inflate AUC and maximum Kappa values, and reduce artificially commission errors derived from cross-validation (Lobo et al. 2008).

Finally, we tested with homogeneity χ^2 tests if application of consensus theory changes the rate of commission and omission errors according to experts.

Results

Expert evaluation

Of the 99 persons we invited to participate in the validation exercise, 45 responded. This yielded on average almost nine experts per species. Experts came from 13 countries and were affiliated with universities, herbaria, and international, national, regional or non-governmental agricultural and environmental research institutions. One *B. excelsa* expert and one *C. odorata* expert were excluded from the analysis because they considered model outcomes under all variable combinations as being invalid. Although this helps us understanding how relevant these models are for some experts in general, it does not give us information to discriminate between the variable combinations.

N. nervosa experts reached the highest consensus between each other compared to experts for the other species. Therefore the variance explained by the first axis of the factor analysis and average competence value was highest for their expert score correlation matrix (Table 2). For the other four species considerably less variance was explained by the first factor axis and the average competence values for these species experts were also lower (Table 2).

In the case of *C. odorata*, we received sufficient expert response to look at species agreement among experts in two different geographic areas: 1) Mexico and Central America (n=6); and 2) South America (n=7). The variance explained by the first factor was 0.36 for both expert groups. We compared this value for both groups with a normal distribution of 1000 bootstraps of respectively six and seven randomly drawn experts without replacement ($\mu = 0.37$ sd = 0.049; $\mu = 0.35$ sd = 0.047). In either case there is a high probability to find randomly better values of consensus than 0.36 ($p = 0.56$; $p = 0.45$).

Quality and selection of distribution models

The median of consensus-weighed expert scores over all 45 species-variable combinations was 2.91, in other words near to *Fair* according to the qualitative scores initially defined. These scores were significantly higher than the corresponding un-weighed scores (Figure 1; Wilcoxon paired test, $p = 0.049$). The median of un-weighed scores was 2.71.

On average across all species, variable combination 8 -which included the 19 bioclimatic variables plus the soil and ecological zone layer- resulted in the best distribution models according to un-weighed expert scores (Figure 2; Friedman, df = 8, $\chi^2 = 16.37$, $p = 0.04$). However, according to consensus-weighed expert scores, no variable combination resulted in consistently better or worse models when taken into account all five species (Figure 2; Friedman, df = 8, consensus-weighed average expert scores: $\chi^2 = 14.05$, $p = 0.08$).

The ranges between maximum and minimum consensus-weighed expert scores of the nine variable combinations per species were much higher compared to un-weighed scores (Appendix 3; Friedman, df = 4, $\chi^2 = 37.44$, $p < 0.001$). These wider ranges made it easier to select the best variable combination per species compared to un-weighed scores (Figure 3; Appendix 3).

For specific species, some variable combinations performed particularly well according to our consensus-weighed expert scores. The best *A. cherimola* and *N. nervosa* predictive models were close to the qualitative score *Good* (Figure 3; Appendix 3; respectively score 3.90 with variable combination 4 –which included the 19 bioclimatic variables plus the soil layer- and 3.82 with variable combination 7 that consisted of the four bioclimatic variables plus the ecological zone layer. In the case of *B. excelsa*, the score of the best model was even between *Good* and *Excellent* (Figure 3; Appendix 3; score 4.30 with variable combination 2 that consisted only of the four bioclimatic variables).

Correlation with model performance parameters

Both consensus-weighed and un-weighed specialist judgements correlated significantly to corresponding AUC, maximum Kappa and commission error when all 45 species-variable combinations were taken together (Table 3). Correlation between these parameters and un-weighed expert scores were similar to the correlation with un-weighed judgements of species specialists (Table 3). Expert opinions did not correlate significantly with omission error (Table 3). Almost all correlations with commission and omission errors were negative. This would be because expert appreciation and rate of these errors are inversely related.

The best variable combinations according to the conventional parameters were different from the best model choice according to the experts independently if they were consensus-weighed or not. According to the AUC, maximum Kappa and commission error values, variable combination 4 –which included the 19 bioclimatic variables plus the soil layer- resulted in the best distribution models (Appendix 3; Friedman AUC, df = 8, $\chi^2 = 25.63$, $p < 0.01$; Friedman Kappa, df = 8, $\chi^2 = 20.98$, $p < 0.01$; Friedman commission error, df = 8, $\chi^2 = 28.59$, $p < 0.0001$). The lowest omission errors were observed in variable combination 3 that consisted of the 19 bioclimatic variables plus the ecological zone layer (Appendix 3; Friedman, df = 8, $\chi^2 = 15.73$, $p = 0.046$).

Considering each species individually, consensus-weighing only improved for *B. excelsa* the correlations between specialist judgments and the model performance parameters (Table 3). In the case of *A. cherimola*, we found highly significant correlations between the specialist evaluations and AUC, maximum Kappa and commission error (Table 3). Similar results were obtained with consensus-weighed and un-weighed expert scores (Table 3). No clear correlations were observed for *N. nervosa* and *C. odorata* (Table 3). Correlation between *B. gasipaes* expert scores and the model performance parameters worsened much when expert scores were consensus-weighed (Table 3).

Commission and omission error according to experts

Averaged per species, 54 % of the preferred models had a commission error according to consensus-weighed expert judgment (Table 4). Forty-three percent of our species specialists indicated these were areas outside the native distribution range. Twenty-two percent indicated that this was due to human disturbance like selective extraction. Thirty-five percent did not specify the reason for species absence in predicted areas of occurrence (Table 4). For each species on average, 31 % of the experts indicated areas of species occurrence that were not predicted in his/her preferred model (omission) (Table 4). We did not obtain significant differences between the commission errors according to consensus-weighed and un-weighed expert scores; either did we for omission errors. Only a significant difference was observed between consensus-weighed and un-weighed values when we asked for the reasons of commission error (Homogeneity, df = 2, $\chi^2 = 10.80$, $p = 0.004$). The reason for this was that experts with higher competency values tended not to clarify the reasons of the commission error (Table 4).

Discussion

In this paper, we present an approach using consensus theory to formalize expert knowledge to validate the outcomes of EEM. The average consensus-weighted score per species-variable combination was higher than the average un-weighted model score. This suggests that experts who have more favourable opinions about models reach more easily consensus between one another, whereas more sceptic experts appear to diverge from consensus. Our results suggest that application of the consensus model could thus be a way to filter out sceptical "mavericks" in the validation of models by experts.

The variation explained in the first axis of the factor analysis shows overall agreement between experts of a specific species. We propose to use this measure to indicate how reliable consensus-weighed expert evaluations of distribution models. Using this criterion, the expert selection of the best distribution model for *N. nervosa* would seem to be trustworthy because of the high degree of expert consensus on the quality of the modelling outcomes. However, the best model chosen by *N. nervosa* experts didn't coincide with the best model choice according to conventional model performance parameters. Similar discordance between expert evaluation and model performance after cross-validation with presence-pseudo-absence data was observed in other studies as well (Anderson et al. 2003). We suggest that in the case of high agreement between experts –as in the case of *N. nervosa*- their opinion should be considered seriously in the validation and selection of distribution models. In other cases, when experts more disagree, conventional parameters such as AUC, Kappa, commission and omission error could be the lead parameters for model evaluation and selection.

N. nervosa occurs in the South American temperate rainforests that occurs only in Chile and Argentina. Consensus was much lower for the other four species that have a more extensive distribution range and cover three or more countries. Opposite opinions may arise between experts belonging to geographical zones with differences in species´ niches. There also could be a sampling bias towards specific geographic zones. To improve EEM outcomes, distribution models could be developed for different geographic zones and accordingly be evaluated separately by expert groups from these different geographic zones. In the software Flora map, for example, it is possible to model species distributions for separate sub clusters of presence points located in different climate zones (Jones et al. 2002). An extra advantage of this additional step is that a possible sampling bias for a specific climate zone would be reduced because records from different zones are modelled independently. This requires further research and expansion of expert validation exercises.

Low expert agreement may also simply occur because of poor model quality in every geographic zone due to sampling bias across the whole distribution range and the use of suboptimal sets of environmental layers. We examined how *C. odorata* experts from two different areas - i.e. Central America and Mexico, and South America - agreed with other species specialists from their own area. Yet, for both regions, expert agreement was not higher than that of a randomly group of experts drawn from both regions.

All consulted experts had a scientific biological background. Yet their disciplines may differ and they could view the distribution models from different perspectives depending on to the culture of their discipline. It requires further research to understand whether expert agreement could be improved when specialists are consulted in separate groups according to their disciplines.

Experts and conventional parameters did not coincide in their best model choice. So, even though there is a significant relation between three of the four conventional parameters and expert evaluation results, there are several discrepancies. Of the four conventional parameters, only the omission error values did not correlate significantly with expert scores. We don ´t know why they don´t correspond, but calculation of omission errors may have been affected by spatial sorting bias, this occurs when test data is located nearby model training data. Experts are less restricted by sampling bias and may consider as well under and not-sampled areas when they estimate omission errors.

We didn't find strong evidence that consensus-weighing improves the correlation between expert scores and AUC, maximum Kappa, commission and omission errors. Only for one of the five species, i.e. *B. excelsa*, correlations between expert scores and conventional parameters clearly improved when these were consensus-weighed. Interestingly, this was also the species with the highest expert scores after consensus-weighing. This model also had the lowest commission and omission errors after cross-validation (Appendix 3). On the other hand, we also found a significant decrease in correlation for *B. gasipaes* when expert scores were weighed. As it is, we only had very few *B. gasipaes* experts (n=5) compared to other species (n≥7). The low number of experts in combination with relatively low degree of consensus may explain why the consensus model didn't perform well for *B. gasipaes* in reference to the conventional parameters.

The number of informants necessary to receive confident results depends on the consensus between the consulted informants (Romney et al. 1986). The more likely informants agree with each other, the lower number of experts is required in model validation. On the basis of responses on true-false questions, Romney et al. (1986) estimated for different competence rates, the minimum number of informants that is necessary to get accurate responses. Sometimes only a few experts are required; in their evaluation of true-false responses up to four informants with competence rates above 0.7 provided accurate responses (Romney et al. 1986). This makes this type of validation exercises also potentially relevant for plant species for which only a low number of specialists exists. So, the challenge becomes to estimate the competence rates of the invited experts before the validation exercise. For species such as *N. nervosa* that have a restricted distribution in a specific ecosystem, seven informants seem to be more than sufficient as they reached in our study an average competence rate of 0.7. For other, more widely distributed species the number of experts may need to be higher to get confident results.

According to our consensus-weighed expert scores average, model quality was towards *Fair*, whereas the best model choices per species, yielded an average between *Fair* and *Good*. This indicates that these models are considered useful by our experts although their applicability remains limited in their opinion. In part this may be explained because Maxent generated considerable commission errors, predicting areas of occurrence where the species is absent. This affects EEM application for reserve design because areas may be included where the species is actually absent, which results in non-efficient investment in conservation (Araújo and Peterson 2012) For each species on average, more than half of our experts indicated that the model they considered best-performing, had a commission error and included areas where the species is absent. One third of the experts also indicated that areas of actual species presence were omitted by the model of their preference. The low omission percentage compared to the rate of commission error suggests that these models are more appropriate for new population discovery and germplasm collecting than for reserve design.

Scale also affects applicability of the modelled distributions (Guisan and Thuiller 2005). Maxent and other EEM software can predict the full distributions of a species and therefore is useful to assess species´ conservation status across their whole distribution ranges. However it doesn´t give that much precision about which interventions should be carried out at a local scale. Many experts tend to work at this scale and are only familiar with a part of the species distribution range which they know in detail. On such a local scale, modelled distributions tend to be less accurate than any expert's knowledge of real field situation. Two experts rated all nine potential distribution maps as invalid. This suggests that the modelled distributions were inaccurate, and thus not useful, at the local scale with which they were familiar. It is thus recommended to indicate to which scale distribution maps are accurate (Hurlbert and Jetz 2007; Lobo et al. 2008). On the other hand, EEM should also meet the needs for potential users. Therefore, modellers provide users more and more explanations how to apply Maxent and other EEM algorithms for ecological studies and biodiversity conservation (Elith et al. 2011; Araújo and Peterson 2012; Guisan et al. 2013).

According to our experts, 43 % of the commission errors in their preferred models, were predictions outside the species distribution range. Inclusion of spatial constraints in EEM may help reduce these over-predictions (Blach-Overgaard et al. 2010). According to the experts, 21 % of the identified commission errors in their preferred models comes from the fact that species had become locally extinct due to selective extraction and forest degradation. It is a challenge to take these factors into account in EEM and requires a combined analysis with a threat assessment.

We only asked experts if they observed commission and/or omission errors or not. In further studies, more details can be asked about the nature and extent of these errors. However, a balance should be sought between depth of questioning and the ease for experts to respond.

No variable combination performed consistently better for all five species compared to other variable combinations according to consensus-weighed expert scores. This means that at this moment, we cannot recommend a particularly outstanding variable combination to model the distribution of other economically important tree species in the Americas. It can be anticipated that the quality of some environmental layers would require improvement whereas a more optimal variable combination with additional environmental layers could further improve EEM results. Perhaps experts from the different test species will be able to reach higher consensus and also agree an overarching best model on the basis of an improved set of environmental layers.

An important limitation in EEM is the lack of high resolution soil maps. Soil properties are known to be important factors for shaping the distribution of plant species (Coudun et al. 2006). But currently only low-resolution soil maps are available at the regional level in Latin America and the Caribbean. The SOTERLAC soil map we used is still coarse compared to the interpolated bioclimatic layers that we used. Initiatives are underway to develop higher-resolution soil maps (Sanchez et al. 2009). Among other environmental variables that could improve model outcomes are solar radiation (Austin 2007) and normalized difference vegetation index (NDVI) (Prates-Clark et al. 2008).

EEM has been developed in ecology to understand the relationships between wild species and their environment, and normally is not applied to predict the distributions of semi-domesticated and/or cultivated species, as was done with peach palm and cherimoya in this study. Nevertheless, the technique has also been used to model the distribution of tropical fruit species and locally important crops (e.g. Miller and Knouft 2006; Scheldeman et al. 2007; Solano and Feria 2007). Many of these species are grown in traditional low-input production systems and/or maintained in semi-natural habitats (e.g. Clement et al. 2010; Scheldeman et al. 2003). This suggests that they are adapted to specific environmental conditions and are not intensively managed. Even so, the ecological range in which cultivated plant species are grown can be expected to be wider compared to the environmental ranges in which wild species populations occur. These plants were domesticated to adapt to different types of growing conditions and management practices, where they can grow well with less competition of other plants (Miller and Knouft 2006; van Zonneveld et al. 2009). As distribution of semi-domesticated and/or cultivated plants is determined by both cultural and environmental factors, it would be interesting to study whether distribution modelling of these species can improve when cultural variables are included. It remains a challenge to find good quality data to develop geospatial layers of cultural variables like localities of archaeological plant remains, historical human routes and linguistic diversity (e.g. Pearsall 1992; Levis et al. 2012; Gorenflo et al. 2012).

It is clear that the results of EEM can also be improved by using better presence point quality and quantity (Anderson et al. 2003; Feely and Silman 2011). Despite the fact that data points are increasingly shared by genebanks and herbaria through online portals like GBIF, for many plant species only few presence points are available. Sampling gaps and sorting bias are especially a problem when EEM is used to better understand species-environmental relationships (Elith et al. 2011). Therefore there is an urge for more data collection in the field (Feely and Silman 2011). Yet field collection is expensive.

At the same time, incomplete sampling is the main reason to use EEM in the case of predicting other areas where a species occurs, on the basis of initial knowledge on its distribution. This is the principal use of EEM for *in situ* conservation planning and targeted collecting for herbaria and germplasm samples (Guarino et al. 2002). In this study, we view EEM from this perspective.

Another, less-costly approach to improve the knowledge about species distributions is combining existing information obtained from experts with the results of EEM. This can be done e.g. by combining modelled species distributions with distribution range maps drawn by experts (Graham and Hijmans 2006) or correcting them based on existing descriptive literature (Rámirez-Villegas et al. 2010). This could be done in much more detail, when species specialists are directly involved in identifying the extent of species distributions and in the revision of presence point data. Especially relevant is local knowledge on species occurrence from under-sampled areas which are difficult to access for field inventories and germplasm collecting because of logistic and administrative constraints. Equally important, species specialists can also provide absence points (Tognelli et al. 2009). Both types of information enrich the understanding of species distributions and help to improve EEM results as well.

Active involvement of existing national and international networks of foresters, taxonomists, ecologists, and/or nature conservationists could increase the number of participants in validation exercises. Amongst others, such networks are often established to facilitate sharing information. Indeed, several studies indicate that local experts, including amateurs, are willing to share information on species occurrence. The clearest examples are with bird watching and reporting (Silvertown 2009), but there also cases where weed or other plant species are monitored (Silvertown 2009; Bradley and Marvin 2011). Such knowledge could be relevant to enrich inventory programs that aim to minimize sampling biases (Feely and Silman 2011). Expert feedback could also be used to iteratively improve EEM to better predict species geographic distribution ranges and better understand species-environmental relationships.

Finally, online GIS and survey applications and the involvement of networks can facilitate the development of methods to carry out this type of consultation for large numbers of species, to interact in a time-effective way with many experts and present the generated species distribution maps for evaluation in an attractive and user-friendly way.

Conclusions

Experts were fairly positive about the distribution model outcomes. This is encouraging although the applicability of EEM for conservation planning remains limited according to expert opinion. To get better results, EEM will require several improvements like the inclusion of better environmental layers.

We obtained several interesting results about expert agreement, model appreciation and correlation of expert scores with conventional parameters. This confirms the potential of expert knowledge and the use of consensus theory for model validation. At the same time, we observed for several species low expert agreement and substantial discrepancies between expert scores and conventional parameters. We suggest that expert judgements should be considered seriously in model selection and evaluation when species specialists reach high consensus. In addition, consensus theory allows to increase the weight of the most knowledgeable experts in final model validation and to filter out sceptical "mavericks". In the case of low expert agreement, however, conventional parameters may remain the leading reference to measure model performance. Low expert agreement may be a result of overall poor model quality or geographically differences in model performance and expert knowledge domains. Further research should be carried out to better understand the possible occurrence of these zones and how to form geographically separate expert groups.

Acknowledgements

We thank all experts who evaluated the modelled species distribution ranges and made this study possible. Judy Loo provided valuable comments to an early version of the manuscript. This study has been carried out in the framework of the project "Strengthening Regional Collaboration in Conservation and Sustainable Use of Forest Genetic Resources in Latin America and Sub-Saharan Africa" financed by the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) Spain. Maarten van Zonneveld is thankful to the CGIAR Research Programme on Forests, Trees and Agroforestry for financial support.

References

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

Araújo, M.B. & Peterson, A.T. 2012. Uses and misuses of bioclimatic envelope modelling. Ecology 93:1527-1539.

Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200: 1-19.

Barry, S. & Elith, J. 2006 Error and uncertainty in habitat models. Journal of Applied Ecology 43: 413–423.

Batjes, N.H. 2005. SOTER-based soil parameter estimates for Latin America and the Caribbean (ver. 1.0). Report 2005/02. ISRIC – World Soil Information, Wageningen.

Beauvais, G.P., Keinath, D.A., Hernandez, P., Master, L. & Thurston, R. 2006. Element Distribution Modeling: A Primer (Version 2). Natureserve, Arlington, Virginia.

Benz, B., Perales, H., Brush, S. 2007. Tzeltal and Tzotzil Farmer Knowledge and Maize Diversity in Chiapas, Mexico. Current Anthropology 48: 289-300.

Bierman, S.M., Butler, A., Marion, G., Kühn, I. 2010 Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. Ecography 33: 451-460.

Blach-Overgaard, A., Svenning, J.-C., Dransfield, J., Greve, M. & Balslev, H. 2010. Determinants of palm species distributions across Africa: the relative roles of climate, non-climate environmental factors, and spatial constraints. Ecography 33: 380 – 391.

Bradley, B.A. & Marvin D.C. 2011. Using expert knowledge to satisfy data needs: mapping invasive plant distributions in The Western United States. Western North American Naturalist 71: 302–315

Busby, J.R. 1991. BIOCLIM a bioclimatic analysis and prediction system. In: Margules, C.R. and Austin, M.P. (eds.) Nature conservation: cost effective biological surveys and data analysis, pp. 64-68 CSIRO, Canberra.

Chapman, A.D. 2005. Principles and methods of data cleaning – Primary species and species occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Clement, C.R., de Cristo-Araújo, M., Coppens d'Eeckenbrugge, G., Alves Pereira, A., Picanço, D. 2010. Origin and domestication of native Amazonian crops. Diversity 2: 73-106.

Coudun, C., Gégout, J.-C., Piedallu, C. & Rameau, J.-C. 2006. Soil nutritional factors improve models of plant species distribution: an illustration with Acer campestre (L.) in France. Journal of Biogeography 33: 1750–1763.

Dormann, C.F., McPherson J.M., Araújo M.B., Bivand, R., Bolliger, J., Carl G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr F.M., Wilson R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609-628.

Ecocrop (2007) www.ecocrop.fao.org. Accessed 5 January 2013

Elith, J., Graham, C.H., Anderson, R.P., Dudík, 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.Mc., Townsend, C., Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, 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., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43-57.

FAO (Food and Agriculture Organization of the United Nations). 1988. Soil Map of the World. Revised Legend. Reprinted with corrections. World Soil Resources Report 60. FAO, Rome.

Feely, K.J. & Silman, M.R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. Diversity and Distributions doi: 10.1111/j.1472-4642.2011.00813.x.

Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guégan, J.-F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M. & Turner, J.R.G. 2008. Spatial species-richness gradients across scales: a meta-analysis. Journal of Biogeography 36: 132-147.

Fielding, A.H. & Bell, J.F.1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38–49.

FRA (Forest Resources Assessment Programme). 2001. Global ecological zoning for the global forest resources assessment 2000. Final Report. Working Paper 56. FAO. Rome.

Fritz, S., Bartholomé, E., Belward, A., Hartley, A., Stibig, H.-J., Eva, H., Mayaux, P., Bartalev, S., Latifovic, R., Kolmert, S., Roy, P.S., Agrawal, S., Bingfang, W., Wenting, X., Ledwith, M., Pekel, J.-F., Giri, C., Mücher, S., de Badts, E., Tateishi, R., Champeaux, J.-L. and Defourny, P. 2003. Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version). EUR 20849 EN. European Commission, Joint Research Centre.

Gorenflo, L.J., Romaine, S., Mittermeier, R.A. & Walker-Painemill, K. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. Proceedings of the National Academy of Science 109: 8032–8037.

Graham, C.H. & Hijmans, R.J. 2006. A comparison of methods for mapping species ranges and species richness. Global Ecology and Biogeography 15: 578–587.

Guarino, L., Jarvis, A., Hijmans, R.J. & Maxted, N. 2002. Geographic Information Systems (GIS) and the Conservation and Use of Plant Genetic Resources. In: Engels, J.M.M., Ramanatha Rao, V., Brown, A.H.D. & Jackson, M.T. (eds.) Managing plant genetic diversity, pp. 387-404. International Plant Genetic Resources Institute (IPGRI) Rome, Italy.

Guisan, A. & Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8: 993-1009.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P. & Buckley, Y.M. 2013. Predicting species distributions for conservation decisions. Ecology Letters DOI: 10.1111/ele.12189

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 modelling methods. Ecography 29: 773-785.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978.

Hijmans, R.J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93: 679-688.

Hurlbert, A.H. & Jetz, W. 2007. Species richness, hotspots, and the scale dependence o range maps in ecology and conservation. Proceedings of the National Academy of Science 104: 13384-13389.

Jones, P.G., Guarino, L. & Jarvis A. 2002. Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap. Plant Genetic Resources Newsletter 130: 1-6,

Knight, A.T., Cowling, R.M., Rouget, M., Balmford, A., Lombard, A.T. & Campbell, B.M. 2008. Knowing but not doing: selecting priority conservation areas and the research–implementation gap. Conservation Biology 22: 610-617.

Levis, C., Souza, P.F.d., Schietti, J., Emilio, T., Pinto, J.L.P.d.V., Clement, C.R. & Costa, F.R.C. 2012. Historical human footprint on modern tree species composition in the Purus-Madeira interfluve, central Amazonia. PLoS ONE 7: e48559.

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.

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. Conservation Biology 17: 1591–1600.

Miller, A.J. & Knouft, J.H. 2006. GIS-based characterization of the geographic distributions of wild and cultivated populations of the mesoamerican fruit tree *Spondias purpurea* (Anacardiaceae). American Journal of Botany 93: 1757–1767.

Nic Lughadha, E., Baillie, J., Barthlott, W., Brummitt, N.A., Cheek, M.R., Farjon, A., Govaerts, R., Hardwick, K.A., Hilton-Taylor, C., Meagher, T.R., Moat, J., Mutke, J., Paton, A.J., Pleasants, L.J., Savolainen, V., Schatz, G.E., Smith, P., Turner, I., Wyse-Jackson, P. & Crane, P.R. 2005. Measuring the fate of plant diversity: towards a foundation for future monitoring and opportunities for urgent action. Philosophical Transitions of the Royal Society B 360: 359–372.

Newton, A.C. & Oldfield, S. 2008. Red Listing the world's tree species: a review of recent progress. Endangered Species Research: 6: 137–147.

Pearce, J.L. & Boyce, M.S. 2006. Modelling distribution and abundance with presence-only data. Journal of Ecology 43: 405-412.

Pearsall, D.M. 1992. The origins of plant cultivation in South America. In: Cowan CW, Watson PJ, editors. The origins of agriculture: an international perspective. Smithsonian Series in Archaeological Inquiry. pp. 173-205.

Pearson, R.G. & Dawson, T.P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12: 361–371.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. 2006. Maximum entropy modelling of species geographic distributions. Ecological Modelling 190: 231-259.

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

Ramírez-Villegas, J., Khoury, C., Jarvis, A., DeBouck, D.G. & Guarino, L. 2010. A gap analysis methodology for collecting crop genepools: a case study with *Phaseolus* beans. PLoS One 5: e13497 doi:10.1371/journal.pone.0013497.

R Development Core Team. 2010. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org. Accessed 26 November 2012.

Romney, A.K., Weller, S.C., Batchelder, W.H. 1986. Culture as consensus: a theory of culture and informant accuracy. American Anthropologist 88: 313-338.

Rupprecht, F., Oldeland, J. & Finckh, M. 2011. Modelling potential distribution of the threatened tree species *Juniperus oxycedrus*: how to evaluate the predictions of different modelling approaches? Journal of Vegetation Science 22: 647–659.

National Research Council. 1989. Lost crops of the Incas: little-known plants of the Andes with promise for worldwide cultivation. National Academy Press, Washington DC, U.S.A.,415 p. (1989) Cherimoya. In: Lost crops of the Incas: Little-known plants of the Andes with promise for worldwide cultivation, National Academy Press, Washington, DC, pp. 228-239.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B,, McKenzie, N.J., Mendonça-Santos, M.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A. & Zhang, G.-L. 2009. Digital Soil Map of the World. Science 325: 680-681.

Scheldeman, X., Van Damme, P., Ureña Alvarez, J.V, & Romero Motoche, J.P. 2003. Horticultural potential of Andean fruit crops exploring their centre of origin. Acta Horticulturae 598: 97-102.

Scheldeman, X., Willemen, L., Coppens, D., Eeckenbrugge, G., Romeijn-Peeters, E., Restrepo, M.T., Romero Motoche, J., Jiménez, D., Lobo, M., Medina, C.I., Reyes, C., Rodríguez, D., Ocampo, J.A., Van Damme, P. & Goetgebeur, P. 2007. Distribution, diversity and environmental adaptation of highland papayas (*Vasconcellea* spp.) in tropical and subtropical America. Biodiversity and Conservation 16: 1867-1884.

Scheldeman. X. & van Zonneveld, M. 2010. Training manual on spatial analysis of plant diversity and distribution. Bioversity International, Rome, Italy.

Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology and Evolution 24: 467–471.

Solano, E. & Feria, T.P. 2007. Ecological niche modeling and geographic distribution of the genus *Polianthes* L. (Agavaceae) in Mexico: using niche modeling to improve assessments of risk status. Biodiversity and Conservation 16: 1885–1900.

Svenning, J.-C. & Skov, F. 2007. Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? Ecology Letters 10: 453–460.

Thuiller, W. 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9: 1353–1362.

Thuiller, W., Lafourcade, B., Engler, B. Araújo, M.B. 2009. BIOMOD: a platform for ensemble forecasting of species distributions. Ecography 32: 369-373.

Tognelli, M.F., Roig-Juñent, S.A., Marvaldi, A.E., Flores, G.E. & Lobo, J.M. 2009. An evaluation of methods for modelling distribution of Patagonian insects. Revista Chilena Historia Natural 82: 347-360.

van Etten, 2006. Changes in farmers' knowledge of maize diversity in highland Guatemala, 1927/37-2004 Journal of Ethnobiology and Ethnomedicine 2. doi:10.1186/1746-4269-2-12.

van Zonneveld, M., Jarvis, A., Dvorak, W., Lema, G. & Leibing, C. 2009. Climate change impact predictions on *Pinus patula* and *Pinus tecunumanii* populations in Mexico and Central America. Forest Ecology and Management 257: 1566–1576.

Weller, S.C. & Mann N.C. 1997. Assessing rater performance without a “gold standard” using consensus theory.

Willis, K.J. & Whittaker, R.J. 2002. Species diversity - scale matters. Science 295: 1245-1248.

Table 1. The nine different variable combinations that experts validated to develop species distribution models with Maxent.

Variable combination	19 bioclimatic variables*	4 bioclimatic variables**	Soil units***	Ecological zones****
1	X			
2		X		
3	X		x	
4	X			x
5			x	x
6		X	x	
7		X		x
8	X		x	x
9		X	x	x

*see www.worldclim.org for more details about the 19 bioclimatic variables

** annual mean temperature, annual precipitation, temperature seasonality, precipitation seasonality

*** Layer derived from SOTERLAC database (Batjes 2005) following FAO's soil classification (FAO 1988)

**** FAO 's terrestrial ecological zone classification (FRA 2001)

Table 2. Agreement on the evaluation of Maxent distribution modelling with nine different variable combinations. Rates of concordance are indicated by the variance explained in the first axis of a maximum likelihood factor analysis and the average competence value (D) according to consensus theory.

Species	Number of experts	Variance explained in first factor	Average competence value (D)
A. cherimola	9	0.29	0.31
B. gasipaes	5	0.39	0.45
B. excels	9	0.39	0.36
C. odorata	13	0.29	0.25
N. nervosa	7	0.59	0.70

Table 3. Pearson correlation coefficients between average expert scores of specific variable combinations and corresponding model performance parameters. Correlations are provided for both consensus-weighed and un-weighed scores.

	Consensus-weighed expert scores			
	AUC	max Kappa	Commission error	Omission error
All experts (n = 43)	0.30*	0.37*	-0.33*	-0.07
A. cherimola (n = 9)	0.90**	0.83**	-0.90**	-0.40
B. gasipaes (n = 5)	0.19	0.11	-0.11	0.09
B. excelsa (n = 9)	0.87**	0.73*	-0.52	-0.43
C. odorata (n = 13)	0.23	-0.03	-0.22	0
N. nervosa (n = 7)	0.20	0.58	-0.21	0.06

	Un-weighed expert scores			
	AUC	max Kappa	Commission Error	Omission error
All experts (n = 43)	0.29*	0.39**	-0.31*	-0.05
A. cherimola (n = 9)	0.85**	0.86**	-0.83**	-0.38
B. gasipaes (n = 5)	0.54	0.42	-0.37	-0.18
B. excelsa (n = 9)	0.76*	0.65	-0.25	-0.23
C. odorata (n = 13)	0.13	0.06	-0.14	-0.08
N. nervosa (n = 7)	0.27	0.59	-0.32	0.11

*p < 0.05, **p < 0.01

Table 4. Expert feedback per species (%) with respect to inclusion of areas where the species is absent (commission) in the model which they selected as best-fitting.

Weighed expert scores				Un-weighed expert scores		
Species	No	Yes	No answer	No	Yes	No answer
A. cherimola	3.98	40.91	55.11	11.11	33.33	55.56
B. excelsa	0.00	67.62	32.38	0	66.67	33.33
B. gasipaes	0.00	31.24	68.76	0	60	40
C. odorata	15.62	70.33	14.05	23.08	69.23	7.69
N. nervosa	19.18	61.09	19.73	14.29	71.43	14.29
Mean	7.76	54.24	38.00	9.69	60.13	30.17

Expert feedback per species (%) about reasons for species absence in predicted areas of occurrence in the model which they selected as best-fitting.

Consensus-weighed expert scores				Un-weighed expert scores		
Species	Human disturbance	Outside distribution range	No answer	Human disturbance	Outside distribution range	No answer
A. cherimola	0.00	47.59	52.41	0	66.67	33.33
B. excelsa	14.92	61.28	23.79	33.33	66.67	16.67
B. gasipaes	0.00	0.00	100.00	0	66.67	33.33
C. odorata	68.57	31.43	0.00	55.56	44.44	0
N. nervosa	24.04	75.96	0.00	20	80	0
Mean	21.51	43.25	35.24	20.11	63.22	16.67

Expert feedback per species (%) with respect to exclusion of areas where the species is present (omission) in the model which they selected as best-fitting.

Consensus-weighed expert scores				Un-weighed expert scores		
Species	No	Yes	No answer	No	Yes	No answer
A. cherimola	40.91	3.98	55.11	22.22	22.22	55.56
B. excelsa	36.28	53.40	10.33	44.44	44.44	11.11
B. gasipaes	0.00	27.42	72.58	0.00	60.00	40.00
C. odorata	56.68	21.22	22.10	53.85	38.46	7.69
N. nervosa	30.75	49.52	19.73	42.86	42.86	14.29
Mean	32.92	31.11	35.97	32.67	41.60	25.73

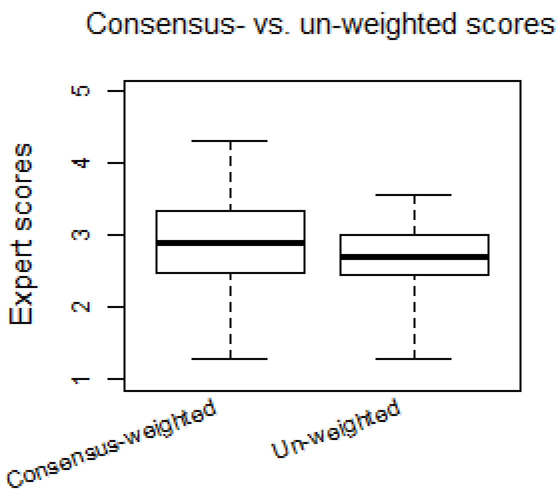


Figure 1. Boxplots of the consensus-weighed and un-weighed expert scores of the produced distribution models for all 45 species-variable combinations.

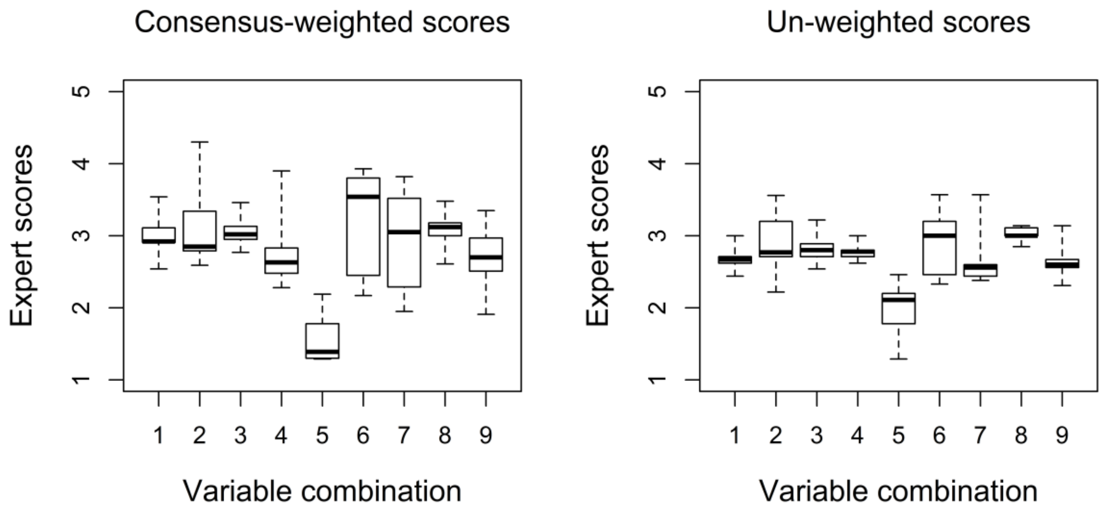


Figure 2. These boxplots show per variable combination the consensus-weighed and un-weighed expert scores of the five test tree species *Annona cherimola*, *Bactris gasipaes*, *Bertholletia excelsa*, *Cedrela odorata* and *Nothofagus nervosa*.

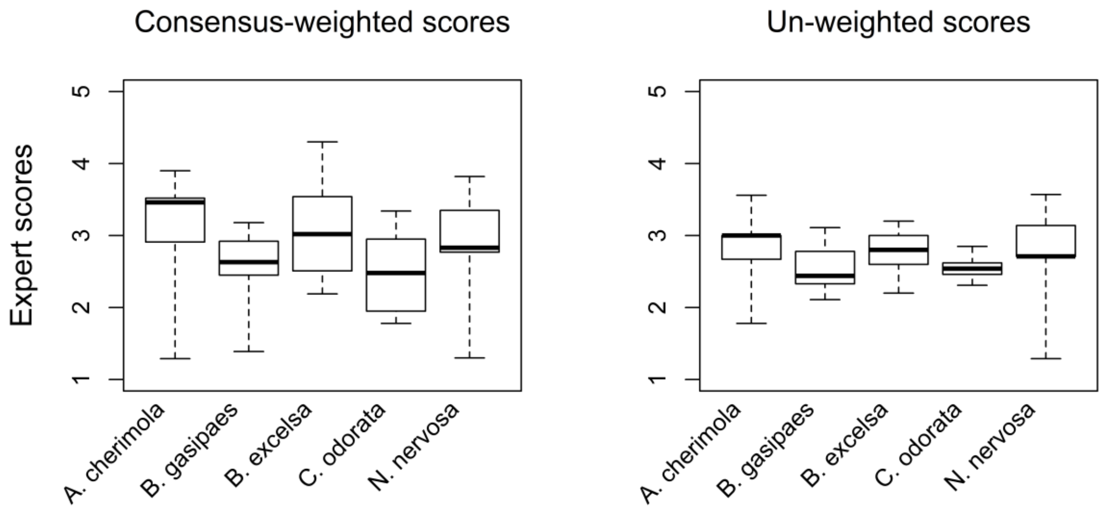
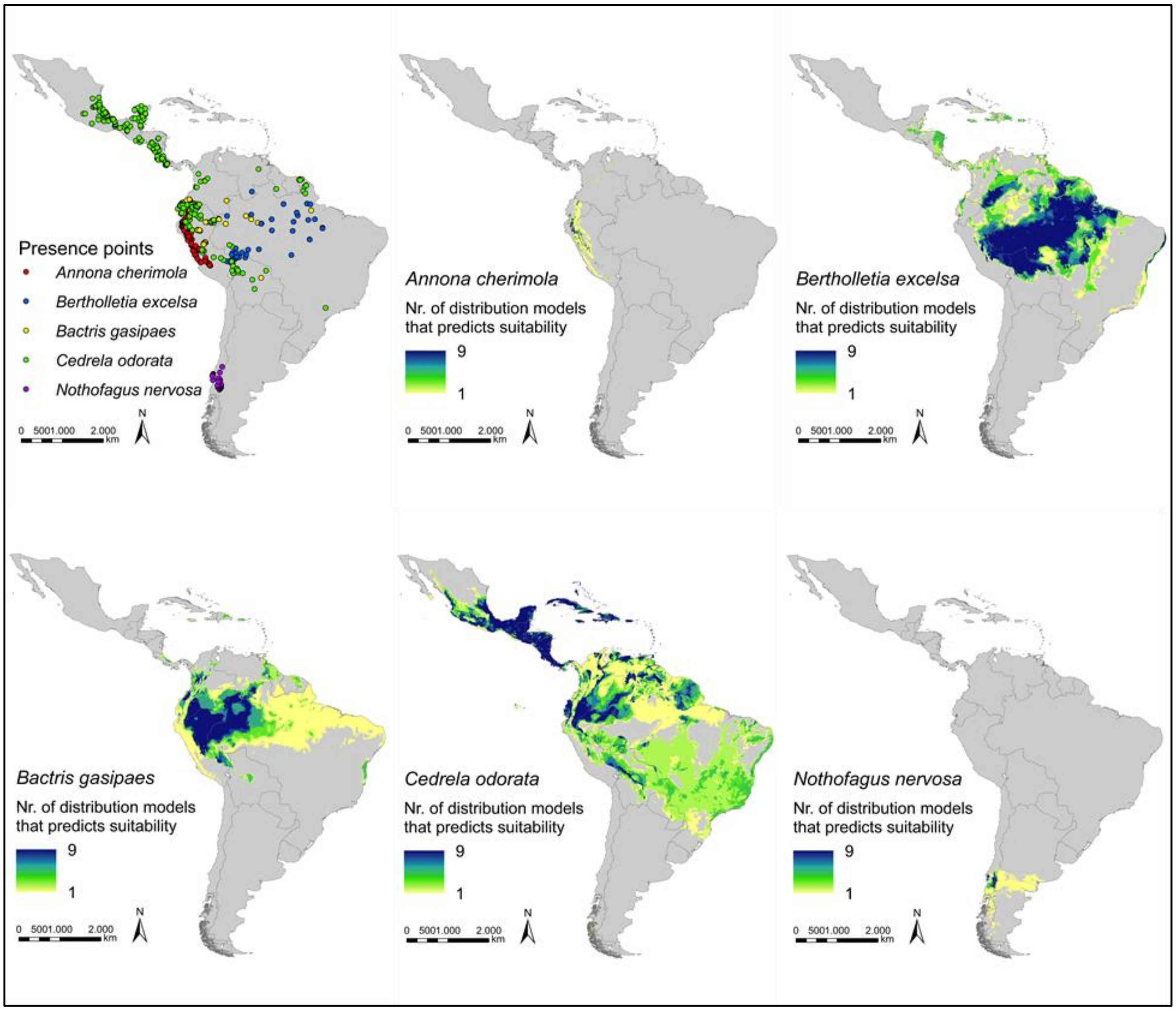


Figure 3. These boxplots show per species the consensus-weighed and un-weighed expert scores for each of the nine variable combinations. These combinations consist of a subset of the 19 bioclimatic variables of Worldclim, a soil layer from SOTERLAC and ecological zone layer from FAO.

Appendix 1. URL links to the Spanish questionnaire to validate plant species distribution models

Annona cherimola: http://gisweb.ciat.cgiar.org/mapforgen/ann_che.html
Bactris gasipaes: http://gisweb.ciat.cgiar.org/mapforgen/bac_gas.html
Bertholletia excelsa: http://gisweb.ciat.cgiar.org/mapforgen/ber_exc.html
Cedrela odorata: http://gisweb.ciat.cgiar.org/mapforgen/ced_odo.html
Nothofagus nervosa: http://gisweb.ciat.cgiar.org/mapforgen/not_ner.html

Appendix 2. Observed and modelled distributions of the five test species. The first map shows the presence points used in the environmental envelope modelling in Maxent. The other five maps show for each species the concordance in species occurrence prediction of the generated distribution models using the nine different variable combinations. The 10 percentile training presence threshold was used to restrict potential distribution areas. Maps were edited in Arc map.



Appendix 3. Consensus-weighed and un-weighed expert scores and the values of four conventional measurements of model performance through cross-validation (AUC maximum Kappa, commission error and omission error) for Maxent Environmental Envelope Modelling (EEM) outcomes for each of the five test species and nine variable combinations.

Consensus-weighed expert scores

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	2.91	2.85	3.46	**3.90**	1.29	3.54	3.52	3.48	2.97	3.46	2.61
B. gasipaes	2.92	2.59	3.13	2.63	1.39	2.45	2.29	**3.18**	2.70	2.63	1.79
B. excelsa	3.54	**4.30**	3.02	2.28	2.19	3.93	3.05	3.00	2.51	3.02	1.69
C. odorata	3.11	**3.34**	2.95	2.48	1.78	2.17	1.95	2.61	1.91	2.48	1.55
N. nervosa	2.54	2.79	2.77	2.83	1.30	3.80	**3.82**	3.12	3.35	2.83	2.52
Median score	2.92	2.85	3.02	2.63	1.39	**3.54**	3.05	3.12	2.70		
Range max-min score	1.00	1.71	0.69	1.62	0.90	1.76	1.87	0.87	1.44		

Un-weighed expert scores

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	2.67	**3.56**	3.22	3.00	1.78	3.00	2.56	3.00	2.67	3.00	1.78
B. gasipaes	2.44	2.22	2.89	2.78	2.11	2.33	2.44	**3.11**	2.56	2.44	1.00
B. excelsa	3.00	**3.20**	2.80	2.80	2.20	**3.20**	2.60	3.00	2.60	2.80	1.00
C. odorata	2.62	2.77	2.54	2.62	2.46	2.46	2.38	**2.85**	2.31	2.54	0.54
N. nervosa	2.71	2.71	2.71	2.71	1.29	**3.57**	**3.57**	3.14	3.14	2.71	2.29
Median score	2.67	2.77	2.80	2.78	2.11	**3.00**	2.56	**3.00**	2.60		
Range max-min score	0.56	1.34	0.68	0.38	1.17	1.24	1.19	0.29	0.83		

Area Under Curve (AUC) of cross-validated models

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	0.963	**0.983**	0.967	0.978	0.891	0.976	0.975	0.976	0.965	0.975	0.085
B. gasipaes	0.844	0.779	0.857	**0.875**	0.601	0.738	0.758	0.87	0.786	0.758	0.269
B. excelsa	0.844	0.801	0.832	**0.889**	0.683	0.822	0.784	0.881	0.84	0.822	0.198
C. odorata	0.887	0.796	0.883	**0.901**	0.792	0.816	0.851	0.877	0.858	0.851	0.085
N. nervosa	0.84	**0.889**	0.849	0.84	0.721	0.786	0.791	0.784	0.786	0.786	0.07
Median score	0.844	0.801	0.857	**0.889**	0.721	0.816	0.791	0.877	0.84		
Range max-min score	0.123	0.204	0.135	0.138	0.29	0.238	0.217	0.192	0.179		

Maximum values are in bold and underlined

Appendix 3. Continuation

Maximum Kappa of cross-validated models

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	0.836	**0.917**	0.828	0.867	0.623	0.867	0.861	0.871	0.842	0.861	0.248
B. gasipaes	0.609	0.487	0.574	**0.6**	0.27	0.436	0.539	0.617	0.583	0.539	0.347
B. excelsa	0.653	0.547	0.627	**0.693**	0.427	0.599	0.667	0.653	0.68	0.653	0.253
C. odorata	0.593	0.463	0.607	**0.684**	0.489	0.509	0.596	0.642	0.605	0.596	0.153
N. nervosa	0.600	**0.694**	0.635	0.663	0.400	0.682	0.529	0.565	0.682	0.565	0.282
Median score	0.609	0.547	0.627	**0.684**	0.427	0.599	0.596	0.642	0.68		
Range max-min score	0.243	0.454	0.254	0.267	0.353	0.431	0.332	0.306	0.259		

Commission error (%) of cross-validated models

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	7.63	3.97	7.42	4.25	25.37	5.12	5.93	**3.80**	6.25	5.93	21.57
B. gasipaes	39.39	34.78	34.78	**33.33**	40.25	41.48	45.45	35.90	42.68	41.48	9.56
B. excelsa	**39.13**	**39.13**	45.31	39.66	47.37	42.27	51.13	39.66	44.00	44.00	11.47
C. odorata	23.14	36.56	28.34	**21.63**	38.85	35.75	28.77	26.39	27.95	28.77	12.46
N. nervosa	**16.67**	25.00	19.75	20.35	39.02	29.11	31.19	24.42	26.58	29.11	14.61
Median score	23.14	34.78	28.34	**21.63**	39.02	35.75	31.19	26.39	27.95		
Range max-min score	31.76	35.16	37.89	35.41	22.00	37.15	45.20	35.86	37.75		

Omission error (%) of cross-validated models

Variable combination	1	2	3	4	5	6	7	8	9	Median score	Range max- min score
A. cherimola	18.2	14.48	**12.74**	15.48	16.02	13.13	13.34	15.18	14.76	14.76	2.89
B. gasipaes	23.08	27.17	**14.49**	18.75	28.17	33.45	38.46	20.27	34.25	33.45	18.19
B. excelsa	**14.29**	**14.29**	22.73	14.71	29.41	18.25	58.82	14.71	20	20.00	44.11
C. odorata	20.50	21.35	**9.65**	13.89	19.08	20.48	16.20	12.98	16.87	16.87	7.49
N. nervosa	25.51	27.78	22.47	**13.87**	21.28	30.34	16.39	23.81	20.99	21.28	13.94
Median score	20.50	21.35	**14.49**	14.71	21.28	20.48	16.39	15.18	20.00		
Range max-min score	11.22	13.49	13.08	4.88	13.39	20.32	45.48	10.83	19.49		

Maximum Kappa values with a bold and underlined font are the maximum values for a specific species and variable combination. In the case of commission and omission errors, minimum values are in bold and underlined font.