18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

# Assessing the accuracy of species distribution models more thoroughly

**Liu, C.[1], M. White[1] and G. Newell[1]**

*[1]Arthur Rylah Institute for Environmental Research, Department of Sustainability & Environment, 123 Brown Street, Heidelberg, Victoria 3084, Australia*
*Email: canran.liu@dse.vic.gov.au*

**Abstract:** Species distribution models (SDMs) are empirical models relating species occurrence to environmental variables based on statistical or other response surfaces. SDMs can be used as a tool to solve some theoretical and applied ecological and environmental problems. The success of their applications depends on the accuracy of the models. In this study we propose an approach to thoroughly assess the accuracy of species distribution models. This includes three aspects:

First is to use several accuracy indices that not only measure model discrimination capability, but also those that measure model reliability. The former is the power of the model that differentiates presences from absences; and the latter refers to the capability of the predicted probabilities to reflect the true probabilities that species occurs in individual locations.

Previous studies have shown that some accuracy measures are sensitive to the prevalence of the test dataset, and that others are not. While all the reliability measures display this sensitivity to prevalence, only do some discriminatory measures fall into the latter group. Many researchers recommend the use of prevalence-insensitive measures in model accuracy assessment. However, using this approach the calibration power of the models cannot be assessed. We argue that calibration measures should also be provided in model accuracy assessments.

The second aspect is to provide confidence intervals associated with the estimates of accuracy indices. Analytical methods, both parametric and nonparametric, have been introduced for constructing the confidence intervals for many accuracy indices. Computer-intensive methods (e.g. bootstrap and jackknife) can also be used to construct confidence intervals that are more attractive than the traditional analytical methods as (1) they have less statistical assumptions; and (2) they are virtually applicable to any accuracy measures.

The third aspect is to provide an assessment of accuracy across a range of test data prevalence, since some accuracy indices are dependant on this quality of the test data. Test data with differing levels of prevalence will provide a range of results for the same accuracy index. Assessing the accuracy at only one level of prevalence will not provide a complete picture of the accuracy of the models. The range of test data prevalence can be set up by researchers according to their knowledge about the target species, or could be taken from the confidence interval of the population prevalence estimated from the sample data if the data can be considered as a random sample of the population.

In this paper, we use an Australian native plant species, Forest Wire-grass (*Tetrarrhena juncea*), as an example to demonstrate our approach to more thoroughly assessing the accuracy of species distribution models. The accuracy of two models, one from a machine learning method (Random Forest, RF) and another from a statistical method (generalized additive model, GAM), were assessed using nine accuracy indices along a range of test data prevalence (i.e. the 95% confidence interval of the population prevalence estimated from the sample data using bootstrap percentile method), and a bootstrap method was used to construct the confidence intervals for the accuracy indices. With this approach, the species distribution models were thoroughly assessed.

***Keywords:*** *species distribution, prediction, accuracy measure, prevalence, confidence interval*

## 1. INTRODUCTION

Species distribution models (SDMs) relate species observations (presence/absence records) and some environmental variables through various modeling techniques, both machine learning and statistical. Species distribution modeling has become an important tool for both theoretical research and environmental applications (Guisan and Zimmermann, 2000). For the former, such modeling exercise can improve our understanding of species-environment relationships; for the latter, SDM predicts species occurrence, which can be used in the assessment of habitat suitability, climate change impact, landscape management impact, and site selection for species reintroduction, etc. However, the success of all these exercises depends on the accuracy of the models. Two aspects are central to examining model accuracy: discrimination capacity and reliability (Pearce and Ferrier, 2000). Discrimination capacity measures model's ability to distinguish between sites where the subject species has been detected (presence sites) and those sites where the species is known to be absent (absence sites). Reliability indicates the correspondence between the predicted probabilities of species presence and the observed proportions of species presences, that is, to which extent the predicted probabilities are accurate. Though the former is generally viewed as more important than the latter (Ash and Shwartz, 1999), the latter can also be important in some situations, e.g. when a map showing the probability of species occurrence is required. A good model should have both good discrimination and good reliability (Pearce and Ferrier, 2000).

Various indices have been introduced to measure the accuracy of the models (some are listed in Table 1). For example, sensitivity, specificity, kappa and the area under the receiver operating characteristic curve (AUC), etc., are the widely used indices measuring the discrimination capacity of models. Brier score (i.e. mean square error, and therefore root mean square error (RMSE)) can be used to measure the reliability of models. Fielding and Bell (1997), Couto (2003), Caruana and Niculescu-Mizil (2004), Liu et al. (2007) provide the detailed explanation and further references for these and other indices.

The effect of prevalence − i.e. the proportion of sites that the species is present within a sample − must also be taken into account in accuracy assessments, since some accuracy indices are sensitive to the prevalence of test dataset. Kappa is particularly susceptible to influences from prevalence (Fielding and Bell, 1997). The positive predictive value (PPV) and negative predictive value (NPV) also depend strongly on prevalence (Riddle and Stratford, 1999; Shapiro, 1999; Bossuyt, 2008). Depending upon the accuracy metric used a variety of conclusions may be drawn from a study where the prevalence of the test data varies. In order to incorporate the test data prevalence within accuracy assessments, we propose the use of a range of 'test data' across the likely range of prevalence to assess the model accuracy. If the test data is randomly sampled from the population, a confidence interval can be estimated for the population prevalence, and the assessment can be taken within this interval.

In the field of species distribution modelling, it is common to provide a value for each accuracy index, however this may be of limited use since sampling variation is not indicated. The uncertainty associated with the estimated accuracy needs to be complemented (Jolliffe, 2007), and confidence intervals serve for this purpose. Various analytical methods, both parametric and nonparametric, have been introduced for constructing the confidence intervals for the accuracy indices (Koopman, 1984; DeLong et al., 1988; Newcombe, 1998; Medina and Zurakowski, 2003; Miao and Gastwirth, 2004; Molodianovitch et al., 2006; Lloyd and Moldovan, 2007; Qin and Hotilovac, 2008). However, many of these are approximations, e.g. derived using delt method, and depend on large sample properties of the statistics used. When sample size is not large, the accuracy of the confidence intervals cannot be guaranteed. In addition, many of these formulae are very complex. As an alternative, computer-intensive methods (e.g. bootstrap and jackknife) have been used for this purpose, e.g. Jolliffe (2007). Bootstrap method provides a general framework for constructing confidence intervals for the accuracy indices of interests. Though there are different methods for constructing bootstrap confidence intervals, it is still difficult to say which method is generally better. The percentile method is a simple and intuitive one, and was used in the study reported here.

In this paper, we take a plant species, Forest Wire-grass, (*Tetrarrhena juncea*), as an example to demonstrate the approach to more thoroughly assessing the accuracy of species distribution models. The accuracy of two models, one from a machine learning method (Random Forest, RF) and another from a statistical method (generalized additive model, GAM), was assessed using nine accuracy indices along a range of test data prevalence, and bootstrap method is used to construct the confidence intervals.

**Table 1**. The nine accuracy measures used in this study, where $n$ is test data size; $n_{11}$, $n_{00}$, $n_{01}$ and $n_{10}$ are the true presences (i.e. positives) and absences (i.e. negatives), and false presences and absences respectively; $n_{1+} = n_{11} + n_{10}$, $n_{0+} = n_{00} + n_{01}$, $n_{+1} = n_{11} + n_{01}$ and $n_{+0} = n_{00} + n_{10}$ are the observed presences and absences, and predicted presences and absences respectively; $o_i$ and $p_i$ ($i = 1,2,\cdots,n$) are the observed species occurrence (1 for presence and 0 for absence) and the predicted probability of species presence at site $i$.

| Index | Definition |
|---|---|
| Overall accuracy | $OA = (n_{11} + n_{00}) / n$ |
| Sensitivity (Recall) | $Se = n_{11} / n_{1+}$ |
| Specificity | $Sp = n_{00} / n_{0+}$ |
| Positive predictive value (Precision) | $PPV = n_{11} / n_{+1}$ |
| Negative predictive value | $NPV = n_{00} / n_{0+}$ |
| True skill statistic | $TSS = Se + Sp - 1$ |
| Kappa | $Kp = (OA - EA)/(1 - EA)$ where $EA = (n_{1+}n_{+1} + n_{0+}n_{+0}) / n^2$ |
| Area under ROC curve | $AUC = \frac{1}{n_{1+}n_{0+}} \sum_{i=1}^{n_{1+}} \sum_{j=1}^{n_{0+}} I(p_{1i}, p_{0j})$ where $I(p_{1i}, p_{0j}) = \begin{cases} 0 & if \quad p_{1i} < p_{0j} \\ 0.5 & if \quad p_{1i} = p_{0j} \\ 1 & if \quad p_{1i} > p_{0j} \end{cases}$ $p_{0i}$ and $p_{1j}$ are the predicted probability of species presence for the absence site i and presence site j. |
| Root mean square error | $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2}$ |

## 2. METHODS

### 2.1 Species and Environmental Data

Forest Wire-grass (*Tetrarrhena juncea* R.Br.) is a tufted perennial grass with branched and scrambling wiry stems. It is common in a range of foothill and montane forest environments in southern and eastern Australia. Species distribution data was extracted from the Victorian Department of Sustainability and Environment's vegetation and plant species database - the Flora Information System (FIS). The FIS is a large repository of vegetation sample plots or quadrats that have been collected from across the Australian State of Victoria, - an area of approximately 22 million hectares. The geographic co-ordinates of all quadrat sites is known with some certainty and as such, many environmental (climatic, radiometric, topographic) and spectral variables from the same locations have been extracted from a 'stack' of data themes stored in a Geographic Information System. From this large range of variables, eleven were considered to be important for modeling species distributions, and were used in this study.

This study aimed to demonstrate an approach to model accuracy assessment by randomly extracting 1000 quadrats from the dataset, among which the target species is present in 164 quadrats, and the other 836 quadrats are considered as absences. We then randomly sampled 70% from the presences and absences respectively to construct a training dataset containing 115 presences and 585 absences respectively, with 49 presences and 251 absences remaining as a test dataset.

### 2.2 Modeling Techniques

Random Forest (RF) and generalized additive model (GAM) were used in this study. While GAM is a statistical model, RF is a machine learning method.

GAMs are semi-parametric extensions of generalized linear models that permit both linear and complex additive shapes or a combination of the two within the same model (Parviainen *et al*. 2008). R Package ***mgcv*** (version 1.3-29) was used in this study. We used a binomial probability distribution for the response and logit function for the link. The degree of smoothness of model terms was estimated as part of model fitting.

RF is an ensemble technique in data mining. It was designed to produce accurate predictions while limiting overfitting of the data (Breiman, 2001). In RF, bootstrap samples are drawn to construct multiple trees, each tree is grown with a randomized subset of predictors, a large number of trees (500 in this study) were grown to maximum size without pruning, and aggregation were produced by averaging the trees (Prasad *et al*., 2006). The R Package randomForest (version 4.5-22) was used to build the model in this study. Exploratory analyses showed that the default values for the parameters worked well for the problems in our study. That is, 500 trees were grown in each forest (i.e. model) and 3 (the closest integer to $\sqrt{11}$ ) environmental variables were randomly chosen at each node to split. However, we used different weights for the two classes — $n_1$ for absence and $n_0$ for presence — to make the total weight balanced for the two classes, where $n_0$ and $n_1$ are the number of training sites for the two classes: absence and presence.

### 2.3 Accuracy Assessment

Nine accuracy indices (Table 1) were used in this study, which included: overall accuracy (OA), sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), true skill statistic (TSS), kappa (Kp), the area under the receiver operating characteristic curve (AUC), and root mean square error (RMSE).

Among the above indices, some are threshold-dependent e.g. OA and Se etc., which need binary result. For each model we used the training data prevalence as the threshold to transform the continuous result into binary one. This threshold-selecting method was shown to be a reliable one (Liu et al., 2005). All the threshold-dependent indices were calculated on the transformed binary result based on the threshold selected in this way.

Since our sample was randomly extracted from a large dataset containing more than 27,000 quadrats that were scattered across the state, we can reasonably assume that the sample was a random one from the population. Thus, the population prevalence can be estimated from such a random sample. Its confidence interval can also be estimated with bootstrap (percentile) method. Each bootstrap sample is completely randomly taken from the original sample. 10,000 bootstrap samples were taken. This interval provided will be taken as the range of prevalence for our assessment. To simplify the process, we made our assessment at three potential levels of prevalence: the estimated prevalence, and the lower and upper limits of the 95% confidence intervals of the prevalence. At each level of prevalence a stratified sampling was adopted to guarantee the prevalence in each bootstrap sample corresponded to the specified prevalence level while the sample size is the same as that of the original sample. At each prevalence level, bootstrap percentile confidence intervals of each index for each model and the difference of each index between the two models were constructed.

## 3. RESULTS

The estimates of the nine accuracy indices for the two models using the original test dataset are shown in Figure 1. It can be seen that the overall accuracy is greater than 75% and AUC is around 0.8, specificity is higher than sensitivity, and NPV is much higher than PPV for both models. The RF model performed better than the GAM model according to OA, Sp, PPV, Kp, AUC and RMSE metrics (since RMSE measures error, the smaller, the better), although they are almost the same according to TSS. The GAM model had better performance than the RF model according to Se and NPV measures.

The population prevalence was estimated as 0.164, and its 95% bootstrap percentile confidence interval is estimated as (0.145, 0.183). The bootstrap estimates and confidence intervals for each index and the difference between the two models for each index at the three levels of prevalence are shown in Figure 2. It can be seen that the confidence intervals are located well above 0 for TSS and Kp, and well above 0.5 for AUC for all the three levels of test data prevalence and the two models. This means that the two models are much better than random predictions, and consequently have predictive ability. Though the confidence intervals for Sp are above 0.7, the confidence interval for Se is only above 0.5. While the confidence intervals

for NPV are above 0.88, those for PPV are below 0.5. This means while most predicted absences are correct, many predicted presences are wrong.

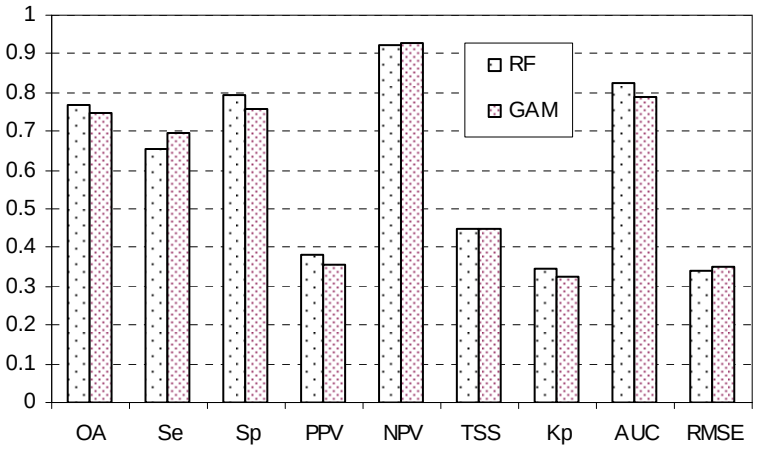


**Figure 1.** Estimates of the nine accuracy indices for the two models (RF and GAM) using the original test dataset.

It can also be seen that for all the indices, the 95% confidence intervals for the two models are overlapping to some degree at each of the three levels of test data prevalence. This indicates that the difference between the two models is not significant. This is also reflected in the 95% confidence intervals of the difference between the two models, which contain 0, except that for AUC at the highest level of prevalence, which is (0.0007, 0.07). This formally tests that the two models are not statistically significantly different at 0.05 significance level according to all the indices at all three prevalence levels. The exception is for AUC at the highest prevalence level where the two models are statistically significantly different at 0.05 significance level. This confirms that some difference cannot be revealed by comparing two separate confidence intervals of an index for two models, which is not a reliable way of testing the difference between them, and directly checking the confidence interval of the difference between the two models is the right way.

A close examination of Figure 3 shows that 0 is very close to the lower limits of the 95% confidence intervals of the difference of AUC between the two models, and to the upper limits of the 95% confidence intervals of the difference of RMSE between the two models for all the three levels of prevalence. However, 0 is not contained in the 90% confidence intervals of the difference between the two models for these two indices (Figure 3). This means that according to these two indices, the two models are statistically significantly different at 0.1 significance level, and RF performs better than GAM with the data provided.

Figure 2 also shows that Se, Sp, TSS and AUC are not dependent on test data prevalence, while PPV, NPV, Kp and RMSE are dependent on this parameter for such a small range (0.145, 0.183) examined in this study. Although there is some indication of OA being dependent on test data prevalence in Figure 2, the trend is not clear because of the small range of prevalence. These results suggest that test data prevalence should be 1) explicitly stated in model accuracy assessment, and 2) examined across a range of prevalence relevant to the accuracy assessment.

## 4. CONCLUSION

In this paper, we have demonstrated an approach to thoroughly assessing the accuracy of species distribution models by modeling a plant species using two modeling methods, RF and GAM. Specifically, this approach includes three aspects. The first is to include several accuracy indices measuring both model discrimination capability and reliability, i.e. in addition to the commonly used accuracy indices measuring model discrimination capability, e.g. OA, Se, Sp, TSS, Kp and AUC, we also employed RMSE to measure model reliability. The second is to provide confidence intervals associated with the estimates for the accuracy measures. Bootstrap method provides a general framework for constructing confidence intervals for all the accuracy measures. The third is to examine accuracy assessments along a range of test data prevalence. This range can be set up by researchers according to their knowledge about the target species, or can be substituted by the confidence interval of the population prevalence estimated from the sample data if the data can be considered as a random sample of the population. Using this approach, species distribution models can be thoroughly and rigorously assessed, and more reliably applied to fields such as conservation biology, landscape planning and environmental impact assessments.

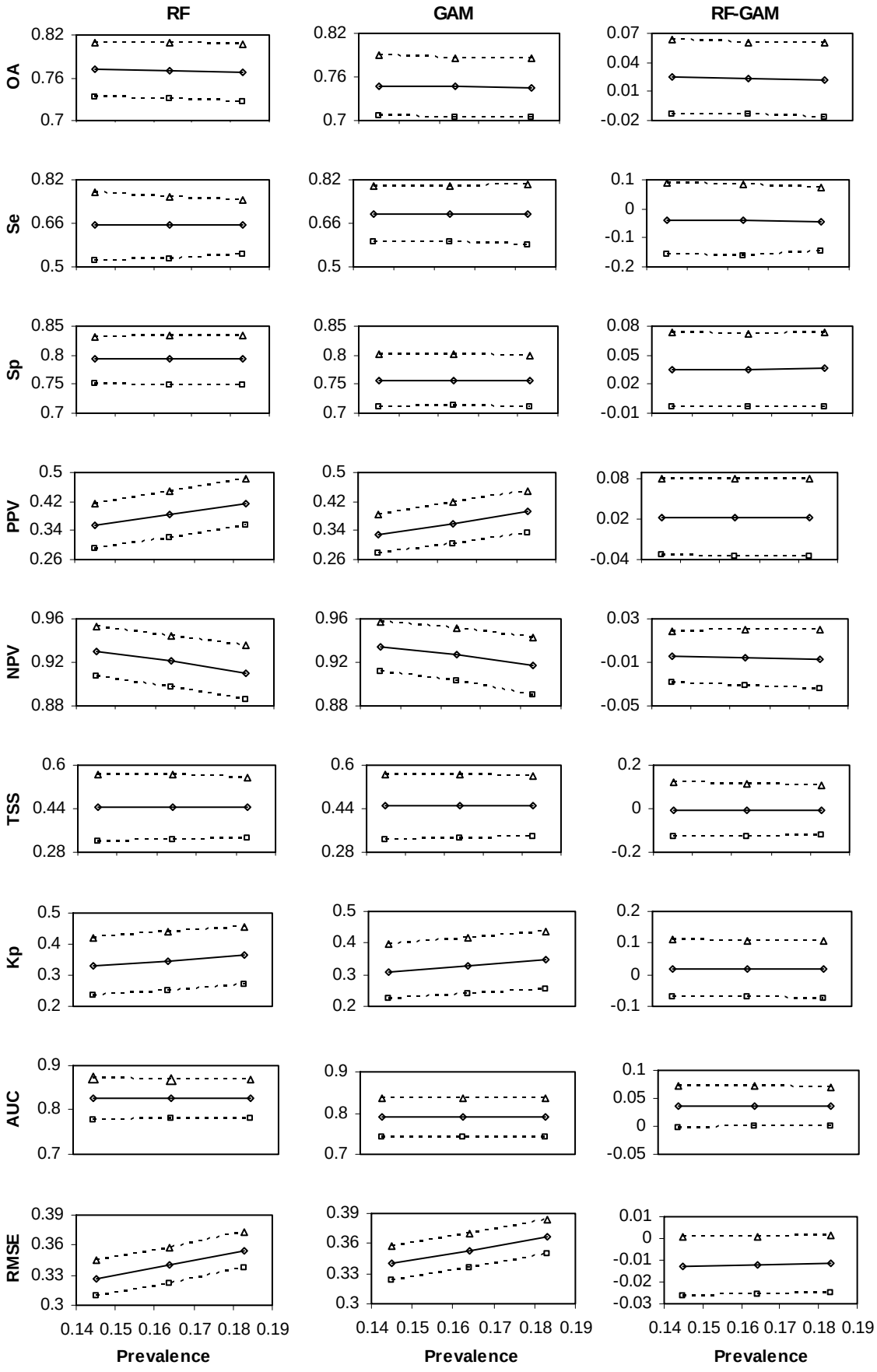


**Figure 2.** The 95% confidence interval for each index and the difference between models RF and GAM.

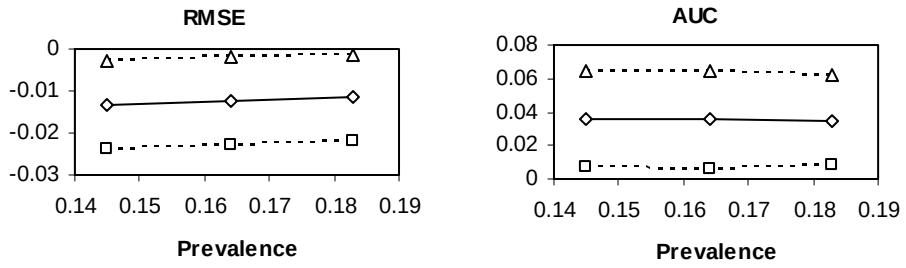


**Figure 3.** The 90% confidence interval for the difference between models RF and GAM.

## REFERENCES

Ash, A., and M. Shwartz, (1999), $R^2$: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine*, 18, 375-384.

Bossuyt, P.M.M. (2008), Interpreting diagnostic test accuracy studies. *Seminars in Hematology*, 45, 189-195.

Breiman, L. (2001), Random forest. *Machine Learning*, 45, 5-32.

Caruana, R., and A. Niculescu-Mizil, (2004), Data mining in metric space: an empirical analysis of supervised learning performance criteria. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, August 22–25, 2004, Seattle, Washington, USA. pp. 69-78.

Couto, P. (2003), Assessing the accuracy of spatial simulation models. *Ecological Modelling*, 167, 181-198.

DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, (1988), Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845.

Fielding, A.H., and J.F. Bell, (1997), A review of methods for the measurement of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 38-49.

Guisan, A., and N.E. Zimmermann, (2000), Predictive habitat distribution models in ecology. *Ecological Modelling*, 135, 147-186.

Jolliffe, I.T. (2007), Uncertainty and inference for verification measures. *Weather and Forecasting*, 23, 637-650.

Koopman, P.A.R. (1984), Confidence intervals for the ratio of two binomial proportions. *Biometrics*, 40, 513-517.

Liu, C., P. Frazier, and K. Kumar, (2007), Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606-616.

Lloyd, C.J., and M.V. Moldovan, (2007), Exact one-sided confidence limits for the difference between two correlated proportions. *Statistics in Medicine*, 26, 3369-3384.

Medina, L.S., and D. Zurakowski, (2003), Measurement variability and confidence intervals in medicine: why should radiologists care? *Radiology*, 226, 297-301.

Miao, W., and J.L. Gastwirth, (2004), The effect of dependence on confidence intervals for a population proportion. *The American statistician*, 58, 124-130.

Molodianovitch, K., D. Faraggi, and B. Reiser, (2006), Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal*, 48, 745-57.

Newcombe, R.G. (1998), Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17, 873-890.

Parviainen, M., M. Luoto, T. Ryttäri, and R.K. Heikkinen, (2008), Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. *Journal of Biogeography*, 35, 1888-1905.

Pearce, J., and S. Ferrier, (2000), Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133, 225-245.

Prasad, A.M., L.R. Iverson, and A. Liaw, (2006), Newer classification and regression tree techniques: Bagging and Random Forests for ecological predictions. *Ecosystems*, 9, 181-199.

Qin, G., and L. Hotilovac, (2008), Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test**.** *Statistical Methods in Medical Research*, 17**,** 207-221**.**

Riddle, D.L., and P. W. Stratford, (1999), Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. *Physical Therapy*, 79, 939-948.

Shapiro, D. (1999), The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, 8, 113-134.