SimTest

Niklaus E. Zimmermann
Swiss Federal Research Institute WSL
CH-8903 Birmensdorf

niklaus.zimmermann@wsl.ch

Introduction

SimTest is a simple program to analyze the accuracy of probabilities of simulated habitat distributions against observed species/community occurrences. ***SimTest*** calculates a number of different accuracy measures to support various goals in testing simulations of species distributions in the fields of ecology and conservation biology. The program reads a simple input file and generates a series of output files and a summary screen. ***SimTest*** is written in Fortran 90, and runs under **DOS**, **Win98/ME/NT/2000**, **Unix** and **Mac** operating systems. Most of the measures used in this program are described in detail in Fielding & Bell (1997), and in Fielding (2000).

User instruction

Copy ***SimTest*** into the directory where you'd like to perform your analyses. Alternatively, you can place it in a different directory and add this directory to your system path (DOS, Win98, NT). Under Mac and Win98/NT, you may also create a shortcut (=alias) to your data analysis directory instead.

SimTest requires a file called <simtest.dat>. This file contains the observed presence and absence variable, and the simulated occurrence probabilities as shown in the text box. The file <simtest.dat> must contain (1) an **integer ID**, (2) the **observed presence-absence values**, and (3) the **simulated probabilities** in the order described here. Per line, the values can be separated either by comma, by space(s) or by tab(s).

```
 1,  1,  0.617
 2,  1,  0.746
 3,  1,  0.875
 4,  1,  0.391
 5,  1,  1.000
 6,  0,  0.786
 7,  0,  0.000
 8,  0,  0.369
 9,  0,  0.545
10,  0,  0.125
```

Box: Simtest.dat

Once this file is generated or exported to the analysis directory, simply start ***SimTest*** either in a **DOS** -window (by typing <simtest> (and hitting the return key), or by clicking on the ***SimTest*** icon. The program automatically reads the file <simtest.dat> and generates the analysis output.

The output of ***SimTest*** consists of four files. Additionally, the program generates a summary screen, wrapping up the most important accuracy measures and some data structure informations. The four output files, generated by ***SimTest*** are called `<abcd.dat>`, `<accuracy.dat>`, `<rocplot.dat>`, and `<shortacc.dat>`. The first file has calculated confusion matrix values (*a*, *b*, *c*, *d*) for cut-off levels ranging from 0.00 to 1.00 (in steps of 0.01). The second file contains a series of accuracy measures for cut-off levels ranging from 0.00 to 1.00. The third file includes all data necessary to draw a ROC plot for illustrating the AUC statistics. The fourth file gives a short overview of all optimized accuracy measures (including cut-off levels). It has virtually the same layout and information, as is printed to the screen.

Theory

When testing simulated species distributions against field observations, two issues are of primary interest: (1) the optimized cut-off level to predict "presence" given the simulated probabilities of occurrence, and (2) the accuracy of the statistical model for this optimized (and other) cut-off level(s). The accuracy can be assessed through a number of different measures. The optimization of the cut-off level is then dependent on the statistics chosen. Therefore, ***SimTest*** calculates different measures, including their respective optimized cut-off levels.

Some measures are dependent of a specific cut-off level (e.g. *correct classification rate, Kappa*) some others are not. They rather describe general structural details of the data set (e.g. *prevalence*, *overall diagnostic power*), or they represent a threshold independent approach to assess the accuracy of a model (e.g. *AUC statistics*)

		Observed	data
		Presence	**A**bsence
Simulated	**P**	**a** [1] **correct** ***True positive***	**B** [2] **Incorrect** ***False positive***
data	**A**	**c** [3] **incorrect** ***False negative***	**d** [4] **correct** ***True negative***

Box: A generic confusion matrix of simulated vs. observed presence (**P**) and absence (**A**) data.

Most of the statistics calculated in ***SimTest*** is based on a *2 x 2* confusion matrix (see box), summarizing the comparison of simulated against observed presence and absence data (P/A). The four possible cases when comparing simulations against field data are labeled "**a**" through "**d**", as indicated in the box. "**N**" will describes the sum of all observations (=a+b+c+d). This notation will be used in all subsequent description of statistical accuracy measures.

Available accuracy measures

When running ***SimTest***, the following (and additional) measures of accuracy are calculated and exported to a series of files. Table 1 summarizes the measures, which are based on a *2 x 2* confusion matrix:

Table 1: *2 x 2* confusion matrix derived measures. N is the number of cases (a+b+c+d)

Measure	**Calculation**	**Abbreviation**
Prevalence	(a+c)/N	*Prv*
Overall diagnostic power	(b+d)/N	*ODP*
Correct classification rate	(a+d)/N	*CCR*
Misclassification rate	(b+c)/N	*MCR*
Sensitivity	a/(a+c)	*Ss*
Specificity	d/(b+d)	*Sp*
Positive predictive power	a/(a+b)	*PPP*
Negative predictive power	d/(c+d)	*NPP*
False positive rate	b/(b+d)	*fpos*
False negative rate	c/(a+c)	*fneg*
Odds ratio	(a*d)/(c*b)	*OR*
Kappa statistics	$\frac{(a+d) - (((a+c)*(a+b)+(b+d)*(c+d))/N)}{N - (((a+c)*(a+b)+(b+d)*(c+d))/N)}$	*k*

Threshold independent measures

Two values are calculated that give a threshold independent summary of the structure of the data simulated and tested. Prevalence (*Prv*) indicates the proportion of observed presences [P = (a+c)] of the overall sum of observations [N]. Overall diagnostic power (*ODP*) is basically the opposite of prevalence (*ODP*=1-*Prv*), since it stands for the proportion of observed absences, [A = (b+d)] given the overall number of [N] observations. Thus, the two measures give an indication of the overall structure of the evaluation data set.

It is important to consider the prevalence, when discussing and evaluating accuracy measures. If one uses a data set that has a prevalence of 0.02, then a model that predicts absence for all evaluation points would yields a *CCR* of 98% (!), even though the positive predictive power (*PPP*) is null.

Threshold-dependent measures of accuracy

The simplest measure of prediction accuracy from the confusion matrix is the proportion of cases that are classified correctly, the *CCR*: (a+d)/(a+b+c+d). This is a measure used in many ecological studies (e.g., BRENNAN *ET AL.,* 1986; CAPEN *ET AL.*, 1986; DONÁZAR *ET AL.*, 1993; VERBYLA & LITVAITIS, 1989). A drawback of this measure is that it doesn't necessarily provide very useful information if the prevalence (*Prv*) of a data set is very low (meaning that *d* is very high). This is, however, often the case in ecological studies. If we are then interested in how good the model predicts on locations of true observations, we risk to loose this information because of the dominance of *d*. We might want to (additionally) consider other accuracy measures.

For example sensitivity (*Ss*) is a measure of the proportion of positive cases that are correctly classified, it takes no account of the false positives. Conversely specificity (*Sp*) is primarily concerned with the false positive errors. Some of the measures in Table 1 are sensitive to the prevalence (*Prv*) of positive cases. For example, even the simple correct classification rate (*CCR*) is affected by the prevalence. This can be demonstrated by an alternative computational route (RUTTIMAN, 1994).

$$CCR = Prv.\text{sensitivity} - (1 - Prv).\text{specificity}$$

Consequently, it is important to avoid a number of potential pitfalls when these performance measures are interpreted in an ecological context (FIELDING & BELL, 1997). In addition, it may be appropriate to consider measures that incorporate misclassification costs or measure improvement over chance. For example, if one group has a high prevalence it is possible to achieve a high *CCR* by the simple expedient of assigning all cases to the most common group. For example, if the prevalence of positive cases was 0.01 a *CRR* of 0.99 is possible if all cases are labelled as negatives. It is also important to use measures that quantify agreement and not association. For example a classifier that got all cases wrong would show perfect association but no agreement. Calculating a c^2 statistic for the following two confusion matrices yields the same value ($c^2 = 162$).

5	95
95	5

95	5
5	95

Positive predictive power (*PPP*) assesses the probability that a case is observed as "presence" if the model classifies it as "presence". Negative predictive power (*NPP*) assesses the probability that a case is observed as "absence" if the model classifies it as "absence". Thus, the two measures assess the probabilities that reality is really what the model simulates (presence/absence). This is not the same as sensitivity and specificity. Speaking in the same terms, sensitivity is the conditional probability that a true "presence" is classified correctly.

Finally, two measures specifically address simulation errors. The false positive rate (*fpos*) denominates the proportion of cases that are simulated as "presence", even though they are "absence" in reality. This type of error is sometimes also called commission or "type I error". Conversely, the false negative rate (*fneg*) denominates the proportion of cases that are simulated as "absence", even though they are "presence" in reality. This type of error is sometimes also called omission or "type II error".

Kappa statistics

Kappa (k), which is the proportion of specific agreement, is often used to assess improvement over chance. LANDIS AND KOCH (1977) suggested that $k < 0.4$ indicates poor agreement, whilst a value above 0.4 is indicative of good agreement. However, k is sensitive to the sample size and it is unreliable if one class dominates. The Tau (t) coefficient (MA & REDMOND*,* 1995) is a related measure but it depends on *a priori* knowledge of the prevalence rather the *a posteriori* estimate used by k. Although the more recent *NMI* measure does not suffer from these problems it shows non-monotonic behaviour under conditions of excessive errors (FORBES, 1995).

Originally, Kappa statistics was designed to evaluate n x n confusion matrices, not the simple 2 x 2 confusion matrix as cited above (see e.g. COHEN, 1960). This is e.g. necessary when evaluating the accuracy of a simulated vegetation map against a field derived map. While comparing vegetation maps MONSERUD & LEEMANS (1992) proposed the following scale for assessing the agreement with kappa statistics:

Kappa	Agreement
0.00 - 0.05:	none
0.05 - 0.20:	very poor
0.20 - 0.40:	poor
0.40 - 0.55:	moderate
0.55 - 0.70:	good
0.70 - 0.85:	very good
0.85 - 0.99:	excellent
0.99 - 1.00:	perfect

ROC plot and AUC statistics

An alternative to determine optimized thresholds and then calculate accuracy measures thereof, is to use the whole information contained within the original raw score (individual probabilities per observation point) and to calculate measures that are independent of any threshold. The Receiver Operating Characteristic (*ROC*) plot is a threshold independent measure that was developed as a signal processing technique. The term refers to the performance (the operating characteristic) of a human or mechanical observer (the 'receiver') involved in assigning cases into dichotomous classes (DELEO & CAMPBELL, 1990; DELEO, 1993). The technique has been applied widely to clinical problems (ZWEIG & CAMPBELL, 1993) and there has been some recent interest from the machine learning community (PROVOST & FAWCETT, 1997; BRADLEY, 1997). MARSDEN & FIELDING (in press) used ROC plots in an ecological context.

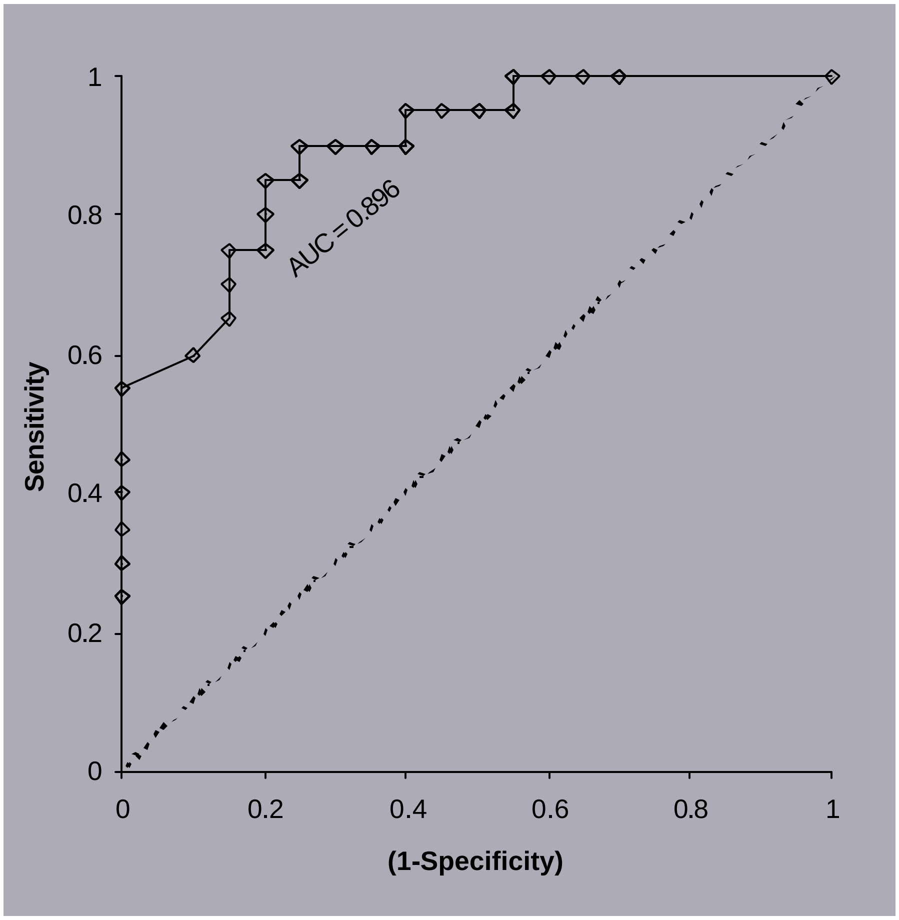

Fig. 1: ROC-Plot of an example data set. The solid line shows an AUC of 89.6%, while the dashed line shows a chance performance (of 50% AUC).

A ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the *y* axis against their equivalent (1-specificity) values (false positive fraction) for all available thresholds on the *x* axis (Figure 1). The area under the ROC function (*AUC*) is usually taken as the index of performance because it provides a single measure of overall accuracy that is independent of any particular threshold (DELEO, 1993). The value of the AUC is between 0.5 and 1.0. If the value is 0.5 the scores for two groups do not differ, while a score of 1.0 indicates no overlap in the distributions of the group scores (Figure 1). A value of 0.8 for the AUC indicates that, for 80% of the time, a random selection from the positive group will have a score greater than a random selection from the negative class (DELEO, 1993). A value of 0.5 for the AUC is equivalent to selecting classes using a random event such as the result of a coin toss.

References

BRADLEY, A.P., 1997. The use of the area under the ROC curve in the estimation of machine learning algorithms. *Pattern Recognition* **30**: 1145-1159.

BRENNAN, L. A., BLOCK, W. M. & GUTIÉRREZ, R. J. 1986. The use of multivariate statistics for developing habitat suitability index models. In: *Wildlife 2000: Modelling habitat relationships of terrestrial vertebrates,* ed. J. A. Verner, M. L. Morrison and C. J. Ralph, pp. 177-182. University of Wisconsin Press, Madison.

CAPEN, D. E., FENWICK, J. W., INKLEY, D. B. & BOYNTON, A. C. 1986. Multivariate models of songbird habitat in New England forests. In: *Wildlife 2000: Modelling habitat relationships of terrestrial vertebrates,* eds. J. A. Verner, M. L. Morrison and C. J. Ralph, pp. 171-175. University of Wisconsin Press, Madison.

DELEO, J.M., 1993. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: *Proceedings of the Second International Symposium on Uncertainty Modelling and analysis*, pp. 318-325. IEEE, Computer Society Press, College Park, MD.

DELEO, J.M. & CAMPBELL, G., 1990. The fuzzy receiver operating characteristic function and medical decisions with uncertainty. In: *Proceedings of the First International Symposium on Uncertainty Modelling and analysis*. IEEE, Computer Society Press, College Park, MD.

DONÁZAR, J. A., HIRALDO, F. & BUSTAMANTE, J. 1993. Factors influencing nest site selection, breeding density and breeding success in bearded vulture (*Gypaetus barbatus*). *Journal of Applied Ecology* **30**: 504-514.

FIELDING, A.H. 2000. How should accuracy be measured? Manuscript in press.

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**(1): 38-49.

FORBES, A. D. 1995. Classification algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring* **11**: 189-206.

LANDIS, J. R. & KOCH, G. C. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**: 159-174.

MA, Z. & REDMOND, R. L. 1995. Tau coefficients for accuracy assessment of classifications of remote sensing data*. Photogrammetric Engineering and Remote Sensing* **61**: 435-439.

MARSDEN, S. & FIELDING, A.H., In press. Habitat associations of parrots on the islands of Buru, Sera and Sumba. *Journal of Biogeography* **00**(0): 00-00.

MONSERUD, R.A. & LEEMANS, R. 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modeling* 62: 275-293.

PROVOST, F, & FAWCETT, T., 1997. Analysis and visualization of classifier performance: comparison under imprecise class cost distributions. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 43-48. AAAI Press.

VERBYLA, D. L. & LITVAITIS, J. A. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management* **13**: 783-787.

ZWEIG, M.H. & CAMPBELL, G., 1993. Receiver-Operating Characteristic (ROC) Plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**: 561-577.