

Summary measures for binary classification systems in animal ecology

Szilárd NEMES and Tibor HARTEL

Mihai Eminescu Trust, Str. Cojocarilor 2, 545400 Sighișoara, Romania.
Correspondence: Sz. Nemes, E-mail: nemessz@gmail.com; T. Hartel, E-mail: asobeka@gmail.com

Abstract. Ecological studies often result in dichotomous, binary outcomes of the response variables (e.g. the presence or absence of the studied species). In such cases, logistic regression is used to model organismal response to various environmental factors. Receiver Operating Characteristic (ROC) curve, a relatively old technique, is a standard procedure for assessing classifier's performance in various fields of science and is increasingly used in ecology. In this note, we present the idea and sketch the mathematics behind the ROC curve, discussing its utility and interpretation.

Key-words: AUC-ROC, model performance, animal occurrence.

Introduction

Empirical ecological research results in a high variability of data types, sometimes with dichotomous, binary outcomes (e.g. presence or absence of species or studied phenomenon). Regression analysis is one of the most powerful statistical tools. Since the development of Generalized Linear Models (Nelder & Wedderburn 1972), regression analysis can also deal with other than normally distributed and continuous outcomes. Binary responses are modelled mostly with logistic regression. The coefficients of the model are estimated by numerical methods and tested using their asymptotic properties (Pawitan 2001).

In addition to assessing the significance of an individual variable or all parameters estimated, it is useful to have a single number to summarize the overall appropriateness of the model. Such measure eases the comparison of competing models as well. Model comparison can be done by information theoretic criteria (AIC or BIC) or by performing a global test (i.e. the Likelihood Ratio test) (Hilborn & Mangel 1997). Both are extensively used in model selection but their usage as a

single numerical measure of model suitability is impractical. Scientists often desire a single scalar measure that allows the comparison of not only different models in one study, but also of models from different studies as well. Such a measure must have a straightforward universal interpretation and a proper scale with easily interpretable values. Both information theoretic criteria and global significance tests fail to meet this criterion. Global significance tests provide a relatively simple and powerful tool for comparing alternative models, but only nested models can be compared. Information theoretic criteria are suitable for comparing both nested and non-nested models, but their interpretation becomes problematic when different models have different support sets (are built on different data sets). The outcome of both methods can be theoretically any real number that lacks a straightforward interpretation as a single numerical measure. Instead, their interpretation depends on the context and two identical values might lead to different conclusions in different situations. The most common measure of model goodness of fit that meets the above stated criteria is the coefficient of determination - R^2 . The coeffi-

cient of determination originates from the Least Squares Regression methods and is defined as the percentage of the total variation in the response variable that can be attributed to the relationship with the predictors. Numerous counterparts of this index were constructed for binary response (reviewed by Long 1997), despite the fact that R^2 is not suitable to judge the effectiveness of regression models with binary responses (Cox & Wermuth 1992). Binary response variables make the assessment of the model's discriminatory performance - the extent to which a model successfully separates the positive and negative observations and classifies them correctly - more feasible. Single measures of association, such as regression coefficients or odds ratios, do not meaningfully describe the predictor variables' ability to classify observations (Pepe et al. 2004). Reporting only odds (or hazard ratio) might intuitively overestimate the predictors' prognostic capacity; therefore there is a need for external validating measures (Dunkler et al. 2007). Such measures ease the comparison of competing models as well. Receiver Operating Characteristic (ROC) curve, a relatively old technique, is a standard procedure for assessing classifier's performance in various fields of science (e.g. biomedical research, pattern recognition, machine learning, psychology) and is increasingly used in ecology (e.g. Huntley et al. 2004, Arntzen 2006, Broennimann et al. 2007, Hartel et al. 2007, 2010a,b, Hartley et al. 2007, Arntzen & Espregueira-Themudo 2008, Rödder et al. 2008). ROC curves are conceptually simple with an easy to follow and understand mathematical algorithm. Studies that model the animal occurrence, in relation to habitat and landscape features are actually still scarce in Romania and Eastern Europe (see e.g. Hartel et al. 2008, 2010a,b). Given this circumstance, we believe that it is timely to present the idea and sketch the mathematics behind the ROC curve and discuss its utility and interpretation.

The receiver operating characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) is a graphical technique used for visualizing and selecting classifiers based on their performance. Formally, each event is mapped to set the positive and negative outcomes (Fawcett 2006). Tests used for binary classification do not produce binary values but a continuous one, T_i , and a threshold c is applied in order to predict the class of the outcome. Generally, high values of the test are assumed to indicate the event of interest ($T_i \geq c$: positive outcome, p), while low values the absence of the event ($T_i < c$: negative outcome, n). Given a binary classifier and the event of interest there are four possible situations:

- i) The event is positive and it is classified as positive, True Positive (TP),
- ii) The event is positive and it is classified as negative, False Negative (FN),
- iii) The event is negative and it is classified as negative, True Negative (TN),
- iv) The event is negative and it is classified as positive, False Positive (FP).

The performance of a binary classifier at a given threshold can be represented as a two-by-two contingency table, the confusion matrix (Fig. 1). From the confusion matrix the two following metrics are calculated: Sensitivity (SN), the percentage of True Positive results, and Specificity (SP) the percentage of True Negative results.

Sensitivity and Specificity are estimated as follows

$$SN = \frac{TP}{TP + FN} \quad \text{and} \quad SP = \frac{TN}{TN + FP}.$$

It should be noted that Sensitivity depends only on positive observations while Specificity on negative observations alone. As a consequence, Sensitivity and Specificity ROC curves can be derived without worrying about the proportional representation of the two groups in the sample.

		actual value	
		P	N
prediction outcome	P	True Positive	False Positive
	N	False Negative	True Negative

Figure 1. Two-by-two confusion matrix of a binary classifier. The observations along the major diagonal represent the correct decisions, the minor diagonal the errors - the confusion between different classes.

A ROC curve captures in a single graph the trade-off between the sensitivity and specificity of the test over its entire range. Thus the ROC-curve plots SN vs. 1-SP as the threshold varies over its entire range

$$\{(SN\{c\}, 1-SP\{c\}) : -\infty < c < \infty\},$$

each data point on the plot representing a particular setting of the threshold (c) and each threshold setting defines a particular set of True Positive and False Negative results, consequently a particular pair of SN and 1-SP values (Lasko 2005). Sensitivity and specificity are synonymous to the True Positive and False Positive rates (TPR and FPR) defined as

$$TPR\{c\} = P(T_i \geq c | Y_i = 1)$$

$$\text{and } FPR\{c\} = P(T_i \geq c | Y_i = 0),$$

the ROC-curve being a two dimensional plot of $\{(FPR\{c\}, TPR\{c\}) : -\infty < c < \infty\}$ (Ma et al. 2006).

Fawcett (2006) points out several important points in the ROC space (Fig. 2):

i) (0,0), the lower left point that represents the strategy of never issuing positive classification (i.e. $c > \max(T_i)$),

ii) (1,1) the upper right point representing the strategy of unconditionally issuing only positive classification (i.e. $c = \min(T_i)$),

iii) (0,1) represents the perfect classification with no error.

A ROC curve generated from a finite data set is a step function, a two dimensional representation of the classification performance. In scientific publications it is not always feasible to reproduce ROC curves, thus numerical indices are commonly used to summarize the curves. These numerical indices have to convey important information about the ROC curve and about the test that the ROC curve was constructed for. Although there are several summary indices, one in particular, the area under the ROC curve (AUC), dominates the practical applications (Hanley and McNeil 1982). The area under a ROC curve is defined as

$$AUC = \int_0^1 ROC(t) dt.$$

Generally the test with higher AUC score is considered the better. If test A is uniformly better than test B in the sense that $ROC_A \geq ROC_B$ then their AUC statistics is ordered as well $AUC_A \geq AUC_B$ (Pepe 2003) (Fig. 3). However, the inverse is not true, a higher AUC value does not necessarily imply higher ROC value at certain sensitivity and specificity, nor equal AUC values assure equal classification accuracy as the threshold varies over its entire range (Fig. 4). Formal comparisons can be made by the means of statistical significance testing (Hanley & McNeil 1982, DeLong et al. 1988, Pepe 2003, Rosset 2004). Two less encountered summary measures of ROC curves are the partial AUC ($pAUC$), that restricts the attention at certain specificity intervals, and the Kolmogorov-Smirnov statistic (KS). The latter is the maximum vertical distance between the ROC curve and the 45° line, the line of the uninformative test. KS ranges between 0 - uninformative test and 1 - the ideal test.

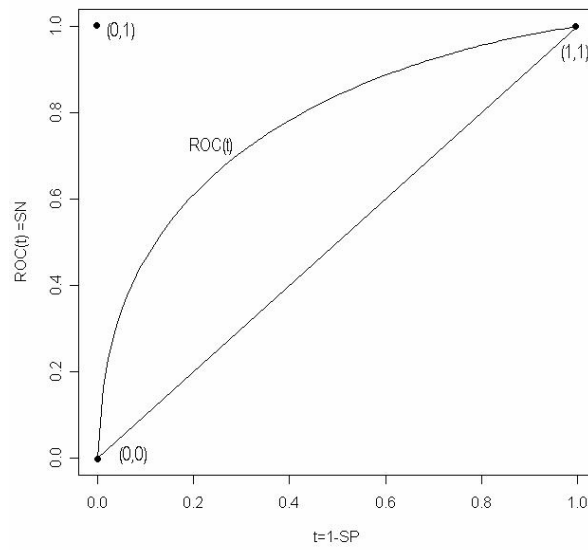


Figure 2. An example of a ROC curve and the three specific points of the ROC curve with no positive classification (0,0), perfect classification (0,1) and just positive classification (1,1).

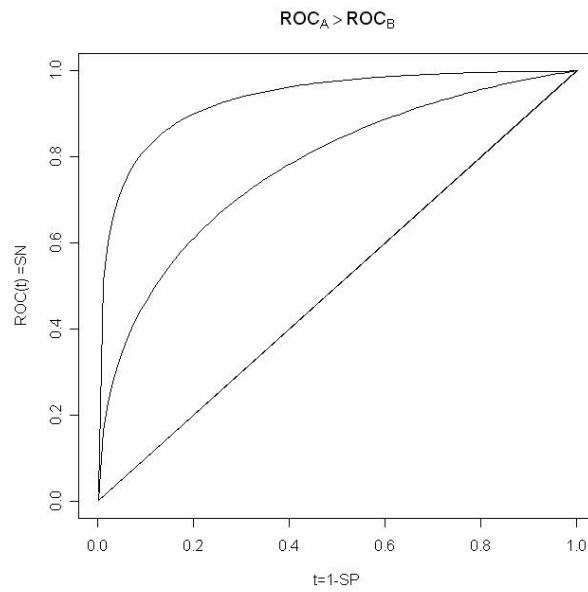


Figure 3. ROC curves of two tests A and B, with test A being uniformly better than B, leading to a strictly higher AUC value. ($AUC_A = 0.936$, $AUC_B = 0.777$).

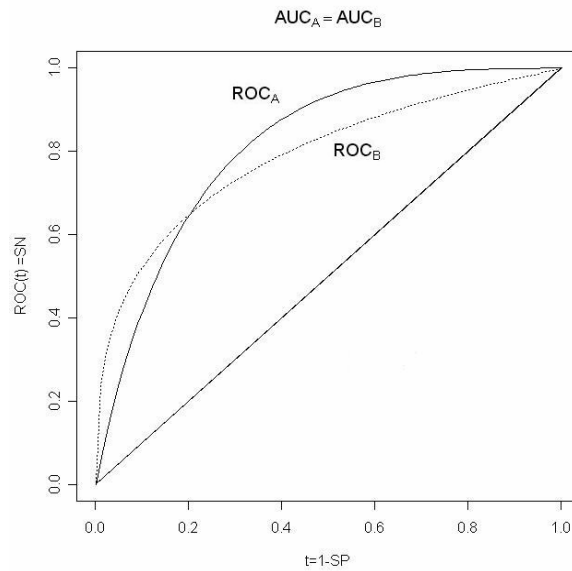


Figure 4. ROC curves of two tests with equal AUC values (0.789) but with unequal accuracy at specific sensitivity and specificity points.

Interpreting of ROC - AUC

If ROC is a straight line between the (0,0) and (1,1) points of the ROC space ($AUC = 0.5$), then the constructed binary classification model has no information about the response variable's class and thus prediction is completely random. It is expected that the ROC curve will lie between the three key points identified by Fawcett (2006), which is $AUC > 0.5$. In this case the model has information about the response variable, and possesses a certain predictive power. Theoretically it is possible that the ROC curve lies in the lower triangle of the ROC space ($AUC < 0.5$). In this case the classifier has information and predictive power but there is a structural error in the model and the information is wrongly utilized.

Conceptually AUC has several interpretations:

i) The probability that the test will produce a value for a randomly chosen event subject is

greater than the value for a randomly chosen not-event subject

$$AUC = P(T_i > T_j | x_i = 1, x_j = 0),$$

ii) The average sensitivity for all values of specificity and vice versa,

iii) A perfect test has an AUC of 1.0 whereas random chance gives an AUC of 0.5.

Even though the first interpretation has appealing mathematical properties the second interpretation has more ecological relevance.

The main purpose of ROC-AUC is to offer an objective technique for selecting classifiers based on their discriminatory performance. Since the publication of Fielding & Bell (1997), ROC and specially AUC tend to be increasingly utilized in the ecological literature, although this technique is not widely accepted and its usage is criticized (Lobo et al. 2007). Criticism of ROC-AUC mostly addresses the properties of AUC as scalar measure of model appropriateness. Lobo et al. (2007) emphasize the

problem of model goodness of fit. They state that a poorly fitted model might possess a good discriminatory power and, on the other end, a well-fitted model might have rather poor discriminatory power if the probabilities of presence are only moderately higher than the probabilities of absence. Both scenarios might be true. A poorly fitted model can have good discriminatory power, but improving the fit might improve the AUC value as well. Plus, AUC values do not assess nor estimate the model goodness of fit, proper model fitting is the researcher's responsibility. A full regression model that incorporates both significant and non-significant variables will have higher AUC value but also parameter estimation will have lower accuracy and should be avoided for reasons of efficiency and interpretability. As in ecology the magnitude and sign of regression parameters have high importance, it is paramount to construct proper models (see Faraway 2005, 2006 for a practical approach on this topic), build the ROC curve and estimate the associated AUC value after. Poorly fitted models or faulty research design is not a drawback of ROC-AUC but a serious flaw of the study in subject.

AUC values are often used to describe the models and not to compare them. Generally it is considered that classifiers with AUC values between 0.5 and 0.7 have low accuracy, between 0.70 and 0.90 have moderate and over 0.9 have high accuracy (Swets 1988, Streiner & Cairney 2007). This classification is rather subjective and can be somewhat misleading. If researchers are interested only in restricted areas of the ROC space it is possible that a classifier with high AUC value performs worse than one with a low AUC value. In this case it would be proper to compare the specific areas of the two ROC curves. A low AUC value does not necessarily suggest a bad or poor model; rather it simply suggests that, beside the accounted predictors, other factors also exercise influence on the response variable.

AUC values may be used as model informative descriptors. In ecological studies, the value of AUC is influenced by the variables used to describe the organism habitats and the spatial scale considered. If the studied organism is generalist, it may tolerate a wide range of ecological conditions and thus, may be less sensitive to the variation of explanatory variables used to predict its occurrence. This will lead to a low value of the AUC for the set of predictor variables used in a given study although there may be significant relationships with some variables. Similarly, low AUC values can be obtained when the spatial scale considered is wrong. If the studied organism is specialist (i.e. prefers a narrow range of habitats) and the researchers capture these habitats in the set of explanatory variables used to predict the organism occurrence, the AUC values will be high. High AUC values, however, will not always suggest that models are convincing. For example, Hartel et al. (2010a) found high model fit in predicting the occurrence of *Pelobates fuscus* (AUC = 0.80). However, since the logistic regression gave only negative associations with the habitat parameters, the model was not considered straightforward. It was assumed that key variables for this species (e.g. the soil type) were omitted from the analysis. Therefore the researcher expert knowledge is important in interpreting the AUC values.

ROC-AUC may serve the same purpose as Information Criteria or Global test, namely to help the researcher to find better models. ROC curves and AUC values facilitate comparisons not only between nested models (comparable with all three methods) or non-nested models built on the same data set (comparable with Information Criteria and ROC-AUC) but also comparisons of models that address the same problems and are built on different data sets (comparable only with ROC-AUC). It is desirable that whenever researchers use logistic regression (or other similar techniques)

they also provide the model's AUC value. This value, like its counterpart from least squares regression (R^2), can suggest whether there is need for considering other factors to increase the discrimination power (AUC) or the explained variance (R^2). Providing the AUC values also facilitates comparisons between different studies. Ecological data sets tend to contain many zero values; zero inflated models (e.g. Zero Inflated Poisson regression or Zero Inflated Negative Binomial regression) do inference of this kind of data (Heilbron 1994, Martin et al. 2005). A first step in zero inflated data analysis is to discriminate the data in two groups, one with true and the other with false zeros. An AUC value helps the researcher to assess the classification performance, thus consequently offering valuable information about the constructed zero-inflated model's reliability.

In this paper we argue that ROC graphs and their associated AUC values combined with ecological knowledge (e.g. Austin 2007) are useful tools for the evaluation of model performance and may serve as a guide for ecologists in evaluating their efficiency. They provide an objective ground to decide whether one or other classification algorithm (or testing procedure) performs better.

References

- Arntzen, J.W. (2006): From descriptive to predictive distribution models: a working example with Iberian amphibians and reptiles. *Frontiers in Zoology* 8: 3.
- Arntzen, J.W., Espregueira-Themudo, G. (2008): Environmental parameters that determine species geographical range limits as a matter of time and space. *Journal of Biogeography* 35: 1177-1186.
- Austin, M. (2007): Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling* 200: 1-19.
- Broennimann, O.U., Treier, A., Müller-Schärer, H., Thuiller, W., Peterson, A.T., Guisan, A. (2007): Evidence of climatic niche shift during biological invasion. *Ecology Letters* 10: 701-709.
- Cox, D.R., Wermuth, N. (1992): A Comment on the Coefficient of Determination for Binary Responses. *American Statistician* 46: 1-4.
- DeLong, E.R., DeLong, D. M., Clarke-Pearson, D.L. (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837-845.
- Dunkler, D., Michiels, S., Schemper, M. (2007): Gene expression profiling: Does it add predictive accuracy to clinical characteristics in cancer prognosis. *European Journal of Cancer* 43: 745-751.
- Faraway, J. (2005): *Linear Models with R*. Chapman and Hall/CRC.
- Faraway, J. (2006): *Extending the Linear Models with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC.
- Fawcett, T. (2006): An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861-874.
- Hanley, J.A., McNeil, B.J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 142: 29-36.
- 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.
- Hartel, T., Moga, C.I., Öllerer, K., Sas, I., Demeter, L., Rusti, D., Balog, A. (2008): A proposal towards the incorporation of spatial heterogeneity into animal distribution studies in Romanian landscapes. *North-Western Journal of Zoology* 4: 173-188.
- Hartel, T., Nemes, Sz., Cogalniceanu, D., Öllerer, K., Schweiger, O., Demeter, L., Moga, C.I. (2007): The effect of fish and aquatic habitat complexity on amphibians. *Hydrobiologia* 583: 173-182.
- Hartel, T., Schweiger, O., Öllerer, K., Cogalniceanu, D., Arntzen, J.W. (2010a): Amphibian distribution in a traditionally managed rural landscape of Eastern Europe: probing the effect of landscape composition. *Biological Conservation* doi:10.1016/j.biocon.2010.02.006.
- Hartel, T., Nemes, Sz., Öllerer, K., Cogalniceanu, D., Moga, C.I., Arntzen, J.W. (2010b): Using connectivity metrics and niche modeling to explore the occurrence of the Northern crested newt (Amphibia, Caudata) in a traditionally managed landscape. *Environmental Conservation* doi:10.1017/S037689291000055X.
- Hartley, S., Harris, R., Lester, P.J. (2006): Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology Letters* 9: 1068-1079.
- Heilbron, D.C. (1994): Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* 36: 531-547.
- Hilborn, R., Mangel M. (1997): *The Ecological Detective. Confronting Models with Data*. Monographs in Population Biology 28. Princeton University Press. New Jersey.
- Huntley, B., Green, R.E, Collingham, Y.C., Hill, J.K., Willis, S.G., Bartlein, P.J., Cramer, W., Hagemeyer, W.J.M.,

- Thomas, C.J. (2004): **The performance of models relating species geographical distribution to climate is independent of trophic level.** *Ecology Letters* 7: 417-426.
- Lasko, T.A., Bhagwat, J.G., Zou, K.H., Ohno-Machado, J. (2005): The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics* 38: 404-415.
- Lobo, J.M., Jiménez-Valverde, A., Real, R. (2007): AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145-151.
- Long, J.S. (1997): *Regression Models for Categorical and Limited Dependent Variables.* Advanced Qualitative Techniques in the Social Sciences Series 7. SAGE Publications.
- Ma, S., Song, X., Hunag, J. (2006): Regularized binormal ROC method in disease classification using microarray data. *BMC Bioinformatics* 7: doi:10.1186/1471-2105-7-253.
- Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P. (2005): Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8: 1235-1246.
- Nelder, J., Wedderburn, R. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society, Series A* 132: 370-384.
- Pawitan, Y. (2001): *In all Likelihood: Statistical Modelling and inference Using Likelihood.* Oxford Science Publications.
- Pepe, M.S. (2003): *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford Science Publications.
- Pepe, M.S., Janes, H., Longton, H., Leisenring, W., Newcomb, P. (2004): Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic or screening marker. *American Journal of Epidemiology* 159: 882-890.
- Rosset, S. (2004): Model selection via the AUC. In: *Proceedings of the 21st International Conference in Machine Learning.* Banff, Canada.
- Rödger, D., Solé, M., Böhme, W. (2008): Predicting the potential distribution of two alien invasive Housegeckos (Gekkonidae: *Hemidactylus frenatus*, *Hemidactylus mabouia*). *North-Western Journal of Zoology* 4: 236-246.
- Streiner, D.L., Cairney, J. (2007): What's under the ROC? An introduction to Receiver Operating Characteristic Curves. *Canadian Journal Psychiatry* 52: 121-128.
- Swets, J. (1988): Measuring the accuracy of diagnostic systems. *Science* 240: 1285-1293.

Submitted: 26 February 2010
/ Accepted: 04 June 2010

Published Online: 24 July 2010