This document is an extract from the PhD thesis entitled “Can satellites help organic crops certification?”, pages 90-99, Antoine DENIS, 2018, ULIEGE, https://orbi.uliege.be/handle/2268/226209.

This extract presents the statistical analysis method used for the assessment of the discriminating power of some indices to discriminate between 2 categories (organic vs conventional), and in particular, how the concepts of the **Area Under the Receiver Operating Characteristic Curve, i.e. ROC-AUC** and of the **P-value of Mann-Whitney-Wilcoxon statistical test** were used to this end.

5.1.2.3 Statistical analysis: assessment of the discriminating power of indices

5.1.2.3.1 ROC-AUC

The Area Under the Receiver Operating Characteristic Curve, i.e. ROC-AUC, is used as an indicator of the **separation** between on the one hand, conventional fields, and, on the other hand, organic and organic in conversion fields that were considered as part of the same population given their similarity in terms of index value.

The ROC-AUC is a common method to assess classifier performance (Fawcett, 2006; Flach et al., 2011; Haibo He and Garcia, 2009) to predict a binary outcome (Seshan et al., 2013).

The ROC-AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance (i.e. organic field) higher than a randomly chosen negative instance (i.e. conventional field) (Fawcett, 2006; Flach et al., 2011).

Computation of the ROC-AUC

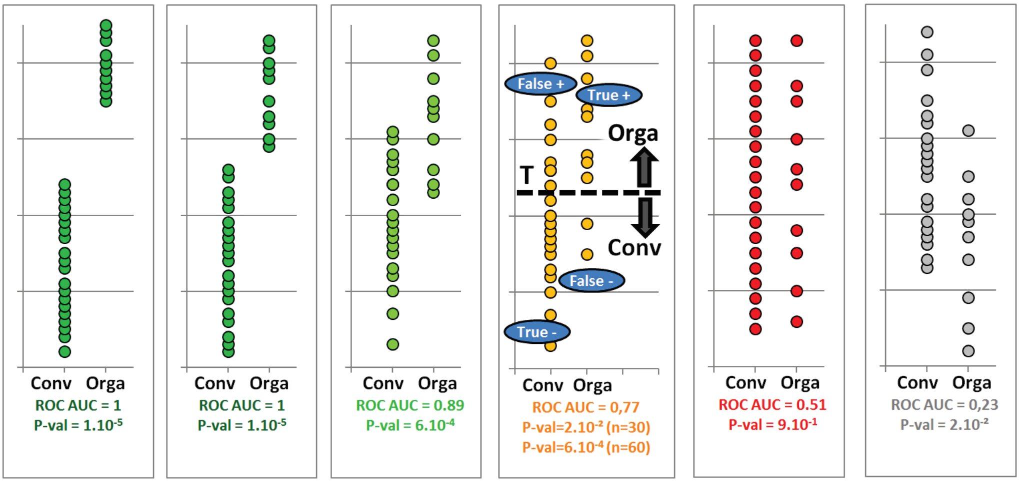
Let us consider a series of (satellite) indices for which the performance of classification in 2 classes (organic and conventional) has to be assessed (Figure 25, page 91). A classification threshold (T) (Figure 25, 4th separability plot) can be defined for one index. This classification threshold classifies the index values in 4 categories which are traditionally reported into a "confusion matrix" or "contingency table" (Figure 25, middle). These 4 categories correspond to, in the case of this study, and by associating by convention organic fields to "positive", and conventional fields to "negative":

- "**True Positive (TP)**": organic fields correctly classified
- "**True Negative (TN)**": conventional fields correctly classified
- "**False Negative (FN)":** organic fields incorrectly classified as conventional
- "**False Positive (FP)**": conventional fields incorrectly classified as organic

These 4 categories can then be combined in a series of manners among which:

- The "**True Positive Rate**" (TPR) or "**Sensitivity**": the ratio of the organic fields correctly classified (TP) by the total number of organic fields.
- The "**False Positive Rate**" (FPR) or "**1-Specificity**": the ratio of conventional fields incorrectly classified (FN) by the total number of conventional fields.

The **ROC curve** is the curve resulting from the graphical representation of the "Sensitivity" as a function of "1-Specificity" when varying the classification threshold "T" over and beyond the range of index values. It illustrates the classification performance of a classifier (in this case, a satellite index) when the classification threshold varies (Fawcett, 2006).



CONFUSION MATRIX For the 4th separability plot at classification threshold "T"		VALIDATION		USER	
		ORGA +	CONV -	User Precision	Commission error
CLASSIFICATION	ORGA +	8 True +	7 False +	8/15 = 0.53%	7/15 = 0.47%
	CONV -	2 False -	13 True -	13/15 = 87%	2/15 = 13%
PRODUCER	Producer precision	8/10 = 80% True Positive Rate (TPR) Sensitivity	13/20 = 65% Specificity	Global accuracy 21/30 = 70%	
	Omission error	2/10 = 20%	7/20 = 35% False Positive Rate (FPR) 1-Specificity		Global error 7/30 = 30%

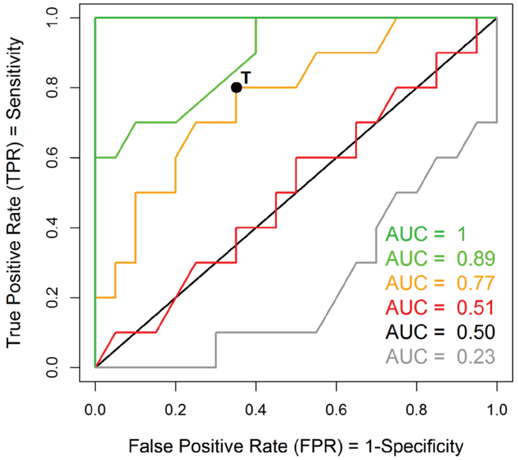


Figure 25 : Examples of separability plots for 6 (dummy) indices presenting different levels of separability, for 1 dataset of 20 conventional and 10 organic fields (n=30) with corresponding ROC-AUC and MWW p-value (above), corresponding ROC curves (right), and confusion matrix corresponding to the classification threshold "T" used in the 4th separability plot (middle).

P-val for n=30 and n=60 (duplication) are given in the 4th plot. **AUC** = Area Under Receiver Operating Characteristic Curve; **P-val** = MWW p-value; **n** = number of fields; **+** = positive = orga = organic; **-** = negative = conv = conventional.

The **ROC-AUC** is the Area Under the ROC Curve in that graphical representation. It is thus independent of any classification threshold.

The **ROC-AUC values** vary between 0 and 1:

- A ROC-AUC of 1 (Figure 25, 1st and 2nd separability plot, dark green ROC curve) corresponds to a perfect classification of the instances of the 2 groups, i.e. a perfect separation between the 2 groups, with no overlapping points.
- A ROC-AUC of 0.5 (Figure 25, 5th separability plot, red or black ROC curves) corresponds to a random classification or to a classification that could be achieved by an uninformative classifier, i.e. a perfect mix between the 2 groups.
- A ROC-AUC smaller than 0.5 (Figure 25, 6th separability plot, gray ROC curves) corresponds to a counterproductive classifier classifying worse than a random classification. This corresponds nevertheless to a certain level of separation between the instances of the 2 groups.*
- A ROC-AUC of 0 corresponds to a “perfect misclassification” and, as for a ROC-AUC of 1, to a perfect separation between the 2 groups.*

*The ROC-AUC analysis considers and classifies by convention values above the classification threshold T in the “positive class”, i.e. in the organic class in the case of this study, independently of the relative values of the 2 groups. The ROC-AUC values give thus information on the relative values of the 2 groups, with:

- ROC-AUC smaller than 0.5 when conventional fields present generally higher values than organic ones,
- ROC-AUC higher than 0.5 when organic fields present generally higher values than conventional ones.

The ROC-AUC, given it is based on the rates of good or bad classification, do not reflect the magnitude of the difference between the 2 groups and gives, for example, an AUC of 1 for a complete separation, independently of the magnitude of that separation (Figure 25, 1st and 2nd separability plot).

One should note that it is possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier (Fawcett, 2006), in which case their ROC curves cross.

The ROC-AUC presents the 2 main advantages that:

(i) It is **insensitive to the number of observations** (fields) compared (Figure 25, 4th separability plot: ROC-AUC of 0.77 for n = 30 or n = 60). This property enables to use it to compare classifier performance over situations presenting significantly different number of observations, as it is the case in this study for the robustness analysis with datasets varying from a total of 23 to 66 fields (Table 10).

(ii) It is **insensitive to changes in class distribution** (Fawcett, 2006; Saito and Rehmsmeier, 2015) i.e. changes in the relative number of observations in each of the 2 classes, conversely to, for example, the overall classification accuracy (Haibo He and Garcia, 2009). This property enables to use it to assess the separability among differently imbalanced datasets (i.e. presenting different number of observations in the 2 groups), as those encountered in this study presenting imbalances varying between 48%/52% (rather balanced) up to 76%/24% (imbalanced) (Table 10).

Table 10 : Number and percentage of fields of each management mode in the 15 situations studied in the case studies 1 (2010) and 2 (2011).

Crop	Country	Year	Date	Satellite	# Conv.	# Orga.	# Total	% Orga.	% Conv.
Maize	Germany	2010	8th July	Landsat-5	20	12	32	0.38	0.63
			10th August	WorldView-2	14	9	23	0.39	0.61
			21st Sept.	KOMPSAT-2	17	11	28	0.39	0.61
			24th Sept.	SPOT-4	13	12	25	0.48	0.52
		2011	4th June	RapidEye	47	19	66	0.29	0.71
			27th June	RapidEye	45	17	62	0.27	0.73
			24th August	RapidEye	47	19	66	0.29	0.71
Wheat	Germany	2010	5th June	PROBA-1	10	19	29	0.66	0.34
			22th June	PROBA-1	10	19	29	0.66	0.34
			29th June	KOMPSAT-2	23	20	43	0.47	0.53
			8th July	Landsat-5	23	21	44	0.48	0.52
	Germany	2011	4th June	RapidEye	36	15	51	0.29	0.71
			27th June	RapidEye	34	12	46	0.26	0.74
	France		26th April	RapidEye	8	26	34	0.76	0.24
			16th May	RapidEye	8	23	31	0.74	0.26
				MIN	8	9	**23**	**0.26**	**0.24**
				MAX	47	26	**66**	**0.76**	**0.74**

= number; **Conv.** = Conventional fields; **Orga.** = Organic fields; **%** = percentage of total.
Darker blue corresponds to higher Total. **Green/yellow** corresponds to more balanced/imbalanced datasets.

The ROC-AUC was computed in R software with the ROCR package (Sing et al., 2005, 2015) with the "prediction" and "performance" functions.

For the interpretation of ROC-AUC, the analysis grid presented in Table 11 will be used.

Table 11 : Level of discrimination associated to ROC-AUC values used for the analysis.

ROC-AUC		Level of discrimination
1	0	Perfect
[0.95 - 1.00[	]0.00 - 0.05[	Very high
[0.90 - 0.95[	[0.05 - 0.10[	
[0.85 - 0.90[	[0.10 - 0.15[	High
[0.80 - 0.85[	[0.15 - 0.20[	
[0.75 - 0.80[	[0.20 - 0.25[	Moderate
[0.70 - 0.75[	[0.25 - 0.30[	
[0.60 - 0.70[	[0.30 - 0.40[	Slight
[0.50 - 0.60[	[0.40 - 0.50[	Not

5.1.2.3.2 P-value of Mann-Whitney-Wilcoxon statistical test

In this study, the **p-value** of the Mann-Whitney-Wilcoxon (MWW) non-parametric statistical hypothesis test (Mann and Whitney, 1947; Wilcoxon, 1945) (also called Mann-Whitney *U* test or Wilcoxon rank-sum test) is used to (i) measure the **statistical significance of the difference** between organic (organic and organic in conversion) and conventional fields from a range of indices, and (ii) to **select the indices** enabling the **best discrimination between the 2 crop management modes**[15].

The test was carried out in R software through the "wilcox.test" function, as a 2 sided test, with a confidence level of 0.95, i.e. with a significance level α of 0.05.

[15] The selection of the most discriminant indices among a given dataset could have been done through the ROC-AUC instead of the MWW p-value as both parameters deliver the same ranking of the indices. However this was not the case for the sole reason that ROC-AUC was not yet implemented when the selection of most discriminant indices was carried out.

The statistical hypotheses

The R function “wilcox.test” tests the following null (H_0) and alternative (H_a) hypotheses (The R Core Team, 2017):

- H_0: the distributions of the 2 populations differ by a location shift. This location shift is set to 0 in this study.
- H_a: the distributions of the 2 populations differ by some other location shift.

The location shift (Figure 26) is estimated through the Hodges-Lehmann estimator (Hodges and Lehmann, 1963; Signorell, 2017; The R Core Team, 2017) which is defined as the median of all pairwise differences of the values of the 2 sampled populations (adapted from formula given in Everitt and Skrondal, 2010).

Some authors state that that test reduces to comparing medians (H_0: medians are the same, H_a: medians differ.) in each of the following 2 conditions:

- If the 2 population have distribution of exactly the same shape (Moore et al., 2013) (Figure 26, right).
- If it can be assumed that each of the 2 population distributions is symmetrical (Kirk, 2017, 2008; Rosenkranz, 2009). In that case, as the distributions are symmetrical, it is also equal to comparing means.

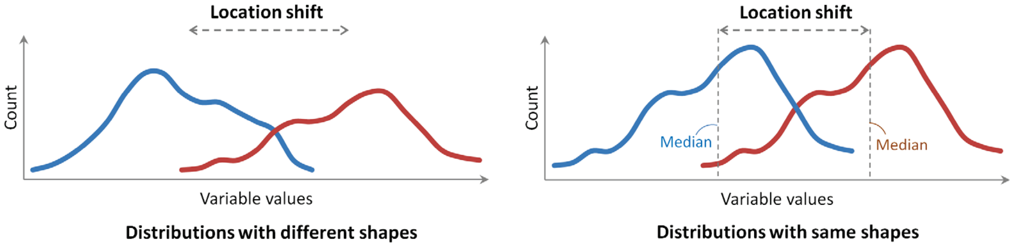


Figure 26 : 2 examples of location shift of 2 non normal distributions with different shapes (left) and same shapes (right).

Conditions for the application of the Mann-Whitney-Wilcoxon statistical test

This test assumes that the 2 populations present **continuous distributions** (Johnson and Bhattacharyya, 2010; Kirk, 2008; Moore et al., 2013). Some other authors state however that such test can be applied on semi-quantitative ordered data (Frontier et al., 2007).

This test is appropriate for testing **independent samples** from 2 populations (Johnson and Bhattacharyya, 2010; Kirk, 2008; Moore et al., 2013). Independent samples are samples for which knowing the outcome of one give no information about the other (Everitt and Skrondal, 2010), i.e. samples that have no effect on one another, or that are not correlated (Easton and McColl, 2004).

The MWW test statistic is based on the **ranks** of observations rather than on their numerical values (Kirk, 2008; Moore et al., 2013). The ranks of observations correspond to their positions in the list of all observations ordered by increasing order (Moore et al., 2013). Therefore this test makes no assumption regarding the **shape** of the population distributions (Johnson and Bhattacharyya, 2010; Kirk, 2008; Moore et al., 2013). In this sense it is a **distribution-free test** (Kirk, 2008), and, in particular, it does not require the assumption of normal distributions or equality of variances (Kirk, 2008; Moore et al., 2013). This test does not assume that the 2 population distributions have the same shape (Moore et al., 2013).

As a **non-parametric test**, this test does not test a hypothesis about one of the parameters of the distribution of the sampled populations (Kirk, 2008).

Justification of the choice of this test

This test was selected because this study deals with comparisons of independent samples from 2 populations of which the normality of distribution is not always met.

Computational method

A test statistic is computed and compared to its probability distribution under the null hypothesis in order to get the p-value.

The MWW test computational procedure may vary depending on the samples size, the presence of ties (observations presenting the same value), the choice of the test statistic and the software used. A very brief insight of the computational procedure used in this study with R software is presented below.

The MWW test statistic “W” used by the “wilcox.test” R function is defined as:

$$W = W_A = \sum R_A - \frac{n_A(n_A + 1)}{2}$$

Where,

- $\sum R_A$ is the sum of the ranks of the observations belonging to the first group (group A)
- n_A is the number of observations in the group A
- $\frac{n_A(n_A+1)}{2}$ corresponds to the sum of the ranks of group A if all observations of the group A would occupy the first positions of the ordered list including group A and B, i.e., positions 1, 2, 3,…, n_A.

The probability distribution of the test statistic W under the null hypothesis becomes approximately normal as the 2 sample sizes increase (Johnson and Bhattacharyya, 2010; Moore et al., 2013). In that case, a Z statistic based on W and that has a probability distribution approximately normal can be used (Johnson and Bhattacharyya, 2010; Kirk, 2008). In the presence of ties, the Z statistic is modified by the so called “continuity correction”.

In this study some indices presented “tied observations”. In such case the “wilcox.test” R function automatically uses the test statistic Z that takes into consideration the “normal approximation with continuity correction” and is defined as (R software, n.d.):

$$Z = \frac{\sum R_A - \frac{n_A(n_A+1)}{2} - \frac{n_A n_B}{2} - \pm 0.5}{\sqrt{\frac{n_A n_B}{12}\left((n_A+n_B+1) - (\frac{\sum(t_i^3 - t_i)}{(n_A n_B)(n_A+n_B-1)}\right)}}$$

Where,

- Z, is the MWW test statistic in the case of normal approximation with continuity correction
- n_A, n_B, are the numbers of observations in the group A, B respectively
- R_A, are the ranks of the group A
- ±0.5, “±” is used because 0.5 takes the same sign as the sign of $(W_A - \frac{n_A n_B}{2})$
- t_i, is the number of tied observations of a particular value i

Despite not all indices encountered in this study presented tied observations, the “normal approximation with continuity correction” was used to evaluate all indices in order to apply exactly the same test on all indices and ensure the comparability of their results.

Interpretation and utilization of the p-value of the statistical test

The p-value of a statistical test is the probability of obtaining a value of the test statistic equal to or more extreme than that actually observed, assuming that the null hypothesis is true (Johnson and Bhattacharyya, 2010; Kirk, 2008; Moore et al., 2013).

It measures the strength of the evidence against H_0, with smaller p-value providing stronger evidence against H_0 (Johnson and Bhattacharyya, 2010; Moore et al., 2013), and meaning the result is more highly statistically significant (Johnson and Bhattacharyya, 2010). Small p-values are evidence against H_0 because they say that the observed result would be unlikely to occur if H_0 were true (Moore et al., 2013).

The p-value of a statistical test is highly affected by the sample size (Figueiredo Filho et al., 2013; Sullivan and Feinn, 2012), with bigger sample size, all other things constant, providing smaller p-value (Figueiredo Filho et al., 2013), conversely to the ROC-AUC. An example of this effect is illustrated in the 4th plot of the Figure 25 where a dataset of 30 values gives a p-value of 0.02, and a dataset of 60 values which is a simple duplication of the former dataset gives a p-value of 0.000 6. With a sufficiently large sample, even very small difference that may be meaningless, or marginal effects, tend to be statistically significant (Figueiredo Filho et al., 2013; Sullivan and Feinn, 2012). Given the sensitivity of the p-value to the sample size, the selection of more discriminant indices based on the p-value is carried out exclusively among a given dataset corresponding to a fixed number of fields.

Notes

It has to be noted that such statistical test based on the ranks ignores the **magnitude** information contained in observations (scores, observed numerical values) (Kirk, 2008) and consequently do not reflect the magnitude of difference between populations, similarly to ROC-AUC (Figure 25, 2 first plots).

As a rank test, it is less affected by **extreme values** than test based on observed numerical values.

Good examples of the use, computational method and interpretation of the Mann-Whitney-Wilcoxon statistical test are available in Bellera *et al.* (2010), Frontier *et al.* (2007), Johnson and Bhattacharyya (2010), Kirk (2008) and Moore et al (2013).

For a given dataset with a given number of observations, the MWW test statistic W is equivalent to the ROC-AUC (Bamber, 1975; Hanley and McNeil, 1982; Mason and Graham, 2002), but this relation breaks when considering datasets of different sizes as the MWW W statistic depends on the sample size while ROC-AUC does not.

5.1.2.4 Graphical representation

3 types of graphical representations were used to present the discrimination power of the measurements and indices computed.

Plot of the discrimination power of indices

The discrimination power of all spectral indices was plotted by associating their ROC-AUC values to a color scale and using as plot coordinates the spectral range of the spectral bands used to compute these indices. This results in a 1 dimension plot for “single spectral band” indices and in 2 dimensions plot for “2 spectral bands ratio” indices. Similar representation was used for a part of the spatial heterogeneity indices computed.

These plots enable to easily compare the discriminating power of all spectral regions and identify the most discriminant ones.

Hyperspectral signatures

The hyperspectral signature of wheat and maize were plotted from *in situ* (ASD) and satellite (CHRIS-PROBA) hyperspectral reflectance measurements.

Separability plots

The separability achieved between organic and conventional fields by the different *in situ* and satellite indices is represented by "Box-and-whisker" plot, latter called "separability plot" (Figure 27). Such plot shows, for each management mode of a given crop, under the form of:

Empty circles	Values of the indices
Black filled circles	Mean value
Arrows	Standard deviation
Box	InterQuartile Range (IQR) delimited by the 1st (Q1) and 3rd (Q3) quartiles, with the median indicated by a horizontal line
"Whiskers"	The most extreme data points which are at a distance of less than 1.5 times the IQR from the edges of the IQR box (i.e., Q1 - 1.5 * IQR and Q3 + 1.5 * IQR).
Top-left corner	ROC-AUC value with a colored background referring to ROC-AUC color scale (Table 11, page 94).
Top-right corner	P-value of the MWW statistical test, with stars representing its level of statistical significance: No star = not significant * = p-value Є [0.05-0.01[** = p-value Є [0.01-0.001[*** = p-value ≤ 0.001
Conv / Orga / Orga C	Conventional / Organic / Organic in Conversion
Numbers below x labels	Number of fields belonging to each management mode

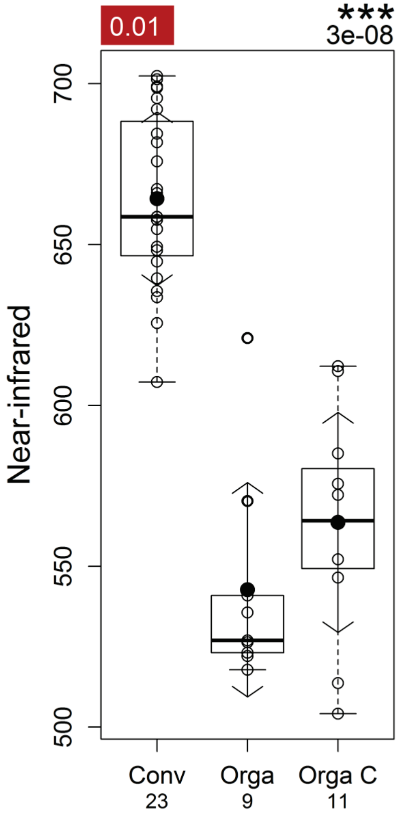


Figure 27 : Example of separability plot with complete legend.

References

Bamber Donald, 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12, 387–415. https://doi.org/10.1016/0022-2496(75)90001-2

Bellera Carine A., Julien Marilyse, Hanley James A., 2010. Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights. Journal of Statistics Education 18, 17.

Easton Valerie J., McColl John H., 2004. Statistics glossary. Web document. URL http://www.stats.gla.ac.uk/steps/glossary/sampling.html#randsamp

Everitt B. S., Skrondal A., 2010. The Cambridge Dictionary of Statistics. 4th Edition. Cambridge University Press.

Fawcett Tom, 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Figueiredo Filho Dalson Britto, Paranhos Ranulfo, da Rocha Enivaldo C., Batista Mariana, da Silva Jr. José Alexandre, D. Santos Manoel L. Wanderley, Marino Jacira Guiro, 2013. When is statistical significance not significant? Brazilian Political Science Review.

Flach Peter, Hernández-Orallo José, Ferri Cèsar, 2011. A coherent interpretation of auc as a measure of aggregated classification performance. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA.

Frontier Serge, Davoult Dominique, Gentilhomme Valérie, Lagadeuc Yvan, 2007. Statistique pour les sciences de la vie et de l’environnement : cours et exercices corrigés.

Haibo He, Garcia E.A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hanley J A, McNeil B J, 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hodges J. L., Lehmann E. L., 1963. Estimates of Location Based on Rank Tests. The Annals of Mathematical Statistics 34, 598–611. https://doi.org/10.1214/aoms/1177704172

Johnson Richard A., Bhattacharyya Gouri K., 2010. Statistics. Principles and Methods. Sixth Edition., John Wiley. ed. Laurie Rosatone.

Kirk Roger E., 2017. Personal communication of Roger E. Kirk, Distinguished Professor of Psychology & Statistics Master Teacher Department of Psychology & Neuroscience Baylor University Waco, Texas 76798.

Kirk Roger E., 2008. Statistics. An Introduction. Fifth Edition., Thomson Wa. ed. Michele Sordi.

Mann H B, Whitney D R, 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other 50–60. https://doi.org/10.1214/aoms/1177730491

Mason S. J., Graham N. E., 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128, 2145–2166. https://doi.org/10.1256/003590002320603584

Moore David S., Notz William I., Fligner Michael A., 2013. The Basic Practice of Statistics. Sixth Edition., W. H. Free. ed. Ruth Baruth, New York.

R software, n.d. Source code of the “wilcox.test” function of R software.

Rosenkranz Gerd K., 2009. A note on the Hodges-Lehmann estimator. Pharmaceutical Statistics n/a-n/a. https://doi.org/10.1002/pst.387

Saito Takaya, Rehmsmeier Marc, 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10, e0118432. https://doi.org/10.1371/journal.pone.0118432

Seshan Venkatraman E., Gönen Mithat, Begg Colin B., 2013. Comparing ROC curves derived from regression models. Statistics in Medicine 32, 1483–1493. https://doi.org/10.1002/sim.5648

Signorell Andri, 2017. Package ‘DescTools’. Tools for Descriptive Statistics.

Sing T., Sander O., Beerenwinkel N., Lengauer T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Sing Tobias, Sander Oliver, Beerenwinkel Niko, Lengauer Thomas, 2015. Package ‘ROCR’. Visualizing the Performance of Scoring Classifiers.

Sullivan Gail M., Feinn Richard, 2012. Using Effect Size—or Why the P Value Is Not Enough. Journal of Graduate Medical Education 4, 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

The R Core Team, 2017. R: A Language and Environment for Statistical Computing. Reference Index. Wilcoxon Rank Sum and Signed Rank Tests. wilcox.test {stats}.

Wilcoxon Frank, 1945. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1, 80–83. https://doi.org/10.2307/3001968