




metrica: an R package to evaluate prediction performance of regression and classification point-forecast models

Adrian A. Correndo ^{1¶}, Luiz H. Moro Rosso ², Carlos H. Hernandez ¹, Leonardo M. Bastos ³, Luciana Nieto ¹, Dean Holworth⁴, and Ignacio A. Ciampitti ¹

¹ Department of Agronomy, Kansas State University, Manhattan, KS, USA. ² Private Consultant, Brasil. ³ Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA. ⁴ CSIRO Agriculture and Food, Australia. ¶ Corresponding author

DOI: [10.21105/joss.04655](https://doi.org/10.21105/joss.04655)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: Øystein Sørensen 

Reviewers:

- [@wiljnich](#)
- [@simonpcouch](#)

Submitted: 29 July 2022

Published: 25 November 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary



The *metrica* R package (Correndo et al., 2022) is an open-source software designed to facilitate the quantitative and visual assessment of prediction performance of point-forecast simulation models for continuous (regression) and categorical variables (classification). The package ensembles a series of 80+ functions that account for multiple aspects of the agreement between predicted and observed values. Without the need of advanced skills on programming, *metrica* enables users to automate the estimation of multiple prediction performance metrics including goodness of fit, error metrics, error decomposition, model efficiency, indices of agreement, and to produce stylish data visualization outputs. This article introduces *metrica*, an R package developed with the main objective of contributing to transparent and reproducible evaluation of point-forecast models performance.

Statement of need

Evaluating the prediction quality is a crucial step for any simulation model, for which a myriad of metrics and visualization techniques have been developed (Tedeschi, 2006; Wallach et al., 2019; Yang et al., 2014). Nonetheless, to conduct a comprehensive assessment of the predicted-observed agreement in R (R Core Team, 2021), users normally have to rely on multiple packages, and even on self-defined functions, which increases the risk of involuntary mistakes due to the need of fluctuating syntax and data wrangling.

As the reproducibility of data analysis continues to be a challenge for science (Seibold, 2022), developing open source software like *metrica* offers a step toward a transparent and reproducible process to assist researchers in evaluating models performance. We decided to create *metrica* in R (R Core Team, 2021) due to its substantial role in data science (Thieme,

2018). Under its open-source philosophy, R empowers the democratization of statistical computing (Hackenberger, 2020) by hosting and globally distributing cutting-edge algorithms through the Comprehensive R Archive Network (CRAN).

Finally, it is noteworthy that in the area of agricultural sciences, although point-forecast simulation models such as the Agricultural Production Systems sIMulator (APSIM) (Holzworth et al., 2014, 2018) count with tools to facilitate the integration into R through packages such as *apsimx* (Miguez, 2022), the assessment of its prediction quality is not yet integrated for R users. Therefore, we aim for *metrica* to offer users of simulation models for agriculture, plant, and soil sciences community a toolbox for assessing the performance of regression and classification point-forecast models.

Package features

For regression models, *metrica* includes four plotting functions (scatter, tiles, density, & Bland-Altman plots) using *ggplot2* (Wickham, 2016), and 48 prediction performance metrics. For classification models (two-class or multi-class), it includes one function to visualize a confusion matrix, and 26 functions of prediction scores. The full list of metrics with description, formula, and literature sources is presented in the package documentation at:

- [Regression metrics vignette](#).
- [Classification metrics vignette](#).

To the best of our knowledge, *metrica* covers several functions not supported, or partially supported by similar R packages (or components) designed for model evaluation such as *yardstick* (Kuhn & Vaughan, 2022) from *tidymodels* (Kuhn & Wickham, 2020), the measuring performance components from *caret* (Kuhn, 2022) or *mlr3* (Lang et al., 2019), *Metrics* (Hamner & Frasco, 2018), *hydroGOF* (Zambrano-Bigiarini, 2020), *cvms* (Olsen & Zachariae, 2021), *scoringutils* (Bosse et al., 2020), or *performance* (Lüdecke et al., 2021). Unique features include:

- one of the most extensive collections of prediction performance metrics for regression and classification models in R.
- working under both vectorized (calling variables with `$`) or tabulated forms (Wickham et al., 2019).
- controlling the output format as a list (`tidy = FALSE`) or as a table (`tidy = TRUE`).
- for classification, functions automatically recognizing two-class or multi-class data; and specifically for multi-class cases, several metrics can be estimated for each class (`atom = TRUE`) (Ferri et al., 2009), (Ben-David, 2007), including balanced and imbalanced scenarios (Kubat et al., 1997).
- for regression, implementing a symmetric linear regression (standardized major axis-SMA-, (Warton et al., 2006)) to describe: i) pattern of the bivariate relationship with linear parameters (`B0_sma`, `B1_sma`), and ii) degree of predicted-observed agreement by using SMA-line to decompose the mean-squared-error (MSE) into lack of accuracy (MLA, PLA, RMLA) and lack of precision (MLP, PLP, RMLP) components (Correndo et al., 2021).
- offering MSE decomposition approaches described by (Kobayashi & Salam, 2000) (SB, SDSD, LCS), and (Smith & Rose, 1995) (Ub, Uc, Ue).
- including multiple indices of agreement and model efficiency such as: i) index of agreement *d* (Willmott, 1981), and its modified *d1* (Willmott et al., 1985) and refined *d1r* (Willmott et al., 2012) variants, ii) Nash–Sutcliffe model efficiency (NSE) (Nash & Sutcliffe, 1970) and its improved variants *E1* (Legates & McCabe Jr., 1999), *Erel* (Krause et al., 2005), and Kling-Gupta model efficiency (KGE) (Kling et al., 2012), iii) Robinson’s index of

agreement (RAC) (Robinson, 1957, 1959), iv) Ji & Gallo agreement coefficient (AC) (Ji & Gallo, 2006), v) Duvellier's lambda (Duvellier et al., 2016), vi) distance correlation (dcorr) (Székely et al., 2007), or vii) maximal information coefficient (MIC) (Reshef et al., 2011)), among others.

- importing files from APSIM Classic with `import_apsim_out()`, and from APSIM Next Generation with the `import_apsim_db()` function.

Using the functions

There are two core arguments to all *metrica* functions: (i) `obs` (O_i ; observed, a.k.a. actual, measured, truth, target, label), and (ii) `pred` (P_i ; predicted, a.k.a. simulated, fitted, modeled, estimate) values.

For regression, specific functions such as `scatter_plot()` require defining the axis orientation (e.g. predicted vs. observed -PO- or observed vs. predicted -OP-). For two-class models, the `pos_level` argument serves to indicate the alphanumeric order of the "positive level". For multi-class classification, some functions present the `atom` argument (TRUE / FALSE), which controls the output to be an overall average estimate across all classes (default), or class-wise.

Example 1: Regression (continuous variables)

Figure 1 is an output example of regression performance analysis using the `scatter_plot()` function for the native dataset called `wheat`.

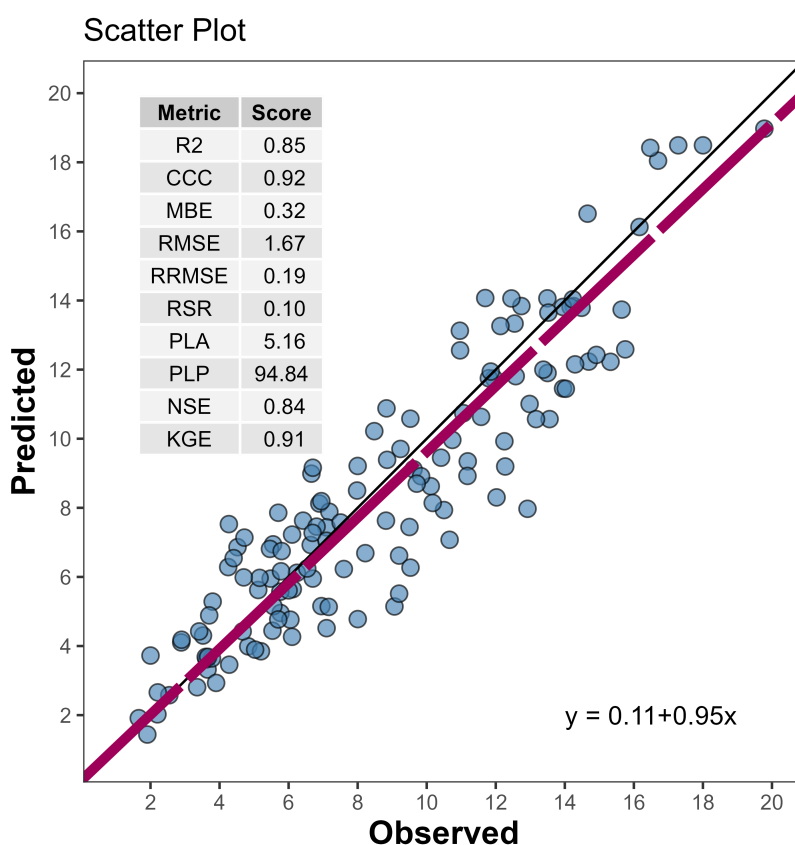


Figure 1: Predicted vs. Observed scatter plot using `metrica::scatter_plot()`.

Example 2: Classification (categorical variables)

Figure 2 is an output example of classification performance analysis using the `confusion_matrix()` function for the native dataset called `maize_phenology`.

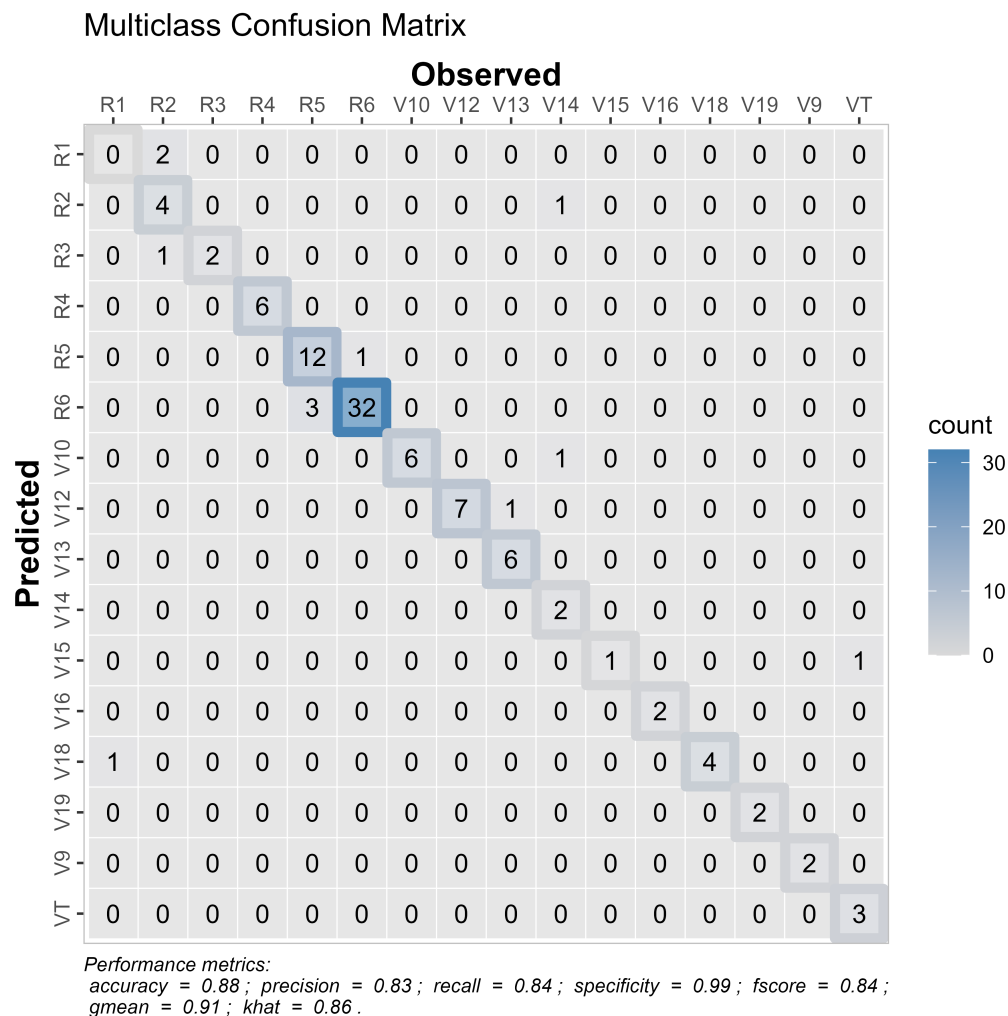


Figure 2: Confusion matrix plot using `metrca::confusion_matrix()`.

Documentation & License

The complete documentation and vignettes of *metrca* are available online at <https://adriancorrendo.github.io/metrca/>. The package is under the MIT License (<https://opensource.org/licenses/MIT>). Source code is available at GitHub (<https://github.com/adriancorrendo/metrca>) along with its corresponding section to report issues and suggestions (<https://github.com/adriancorrendo/metrca/issues>).

Acknowledgements

Authors gratefully acknowledge the financial support from the Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL) at Kansas State University through funding United States Agency for International Development (USAID) under the Cooperative Agreement (Grant number AID-OAA-L-14-00006).

References

- Ben-David, A. (2007). A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence*, 20, 875–885. <https://doi.org/10.1016/j.engappai.2007.01.001>
- Bosse, N. I., Gruson, H., Funk, S., EpiForecasts, & Abbott, S. (2020). *scoringutils: Utilities for scoring and assessing predictions*. <https://doi.org/10.5281/zenodo.4618017>
- Correndo, A. A., Hefley, T. J., Holzworth, D. P., & Ciampitti, I. A. (2021). Revisiting linear regression to test agreement in continuous predicted-observed datasets. *Agricultural Systems*, 192, 103194. <https://doi.org/10.1016/j.agsy.2021.103194>
- Correndo, A. A., Moro Rosso, L. H., Schwalbert, R., Hernandez, C., Bastos, L. M., Nieto, L., Holzworth, D., & Ciampitti, I. A. (2022). *metrica: Prediction performance metrics*. <https://CRAN.R-project.org/package=metrica>
- Duveiller, G., Fasbender, D., & Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports*, 6, 19401. <https://doi.org/10.1038/srep19401>
- Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30, 27–38. <https://doi.org/10.1016/j.patrec.2008.08.010>
- Hackenberger, B. K. (2020). R software: Unfriendly but probably the best. *Croat Med. J.*, 29;61(1), 66–68. <https://doi.org/10.3325/cmj.2020.61.66>
- Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation metrics for machine learning*. <https://CRAN.R-project.org/package=Metrics>
- Holzworth, D., Huth, N., deVoil, P., Zurcher, E., Herrmann, N., McLean, G., Chenu, K., van Oosterom, E., Snow, V., Murphy, C., Moore, A., Brown, H., Whish, J., Verrall, S., Fainges, J., Bell, L., Peake, A., Poulton, P., Hochman, Z., ... Keating, B. (2014). APSIM – evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62, 327–350. <https://doi.org/10.1016/j.envsoft.2014.07.009>
- Holzworth, D., Huth, N., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N., Zheng, B., & Snow, V. (2018). APSIM next generation: Overcoming challenges in modernising a farming systems model. *Environmental Modelling & Software*, 103, 43–51. <https://doi.org/10.1016/j.envsoft.2018.02.002>
- Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering & Remote Sensing*, 72(7), 823–833. <https://doi.org/10.14358/PERS.72.7.823>
- Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277. <https://doi.org/10.1016/j.jhydrol.2012.01.011>
- Kobayashi, K., & Salam, M. U. (2000). Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, 92(2), 345–352. <https://doi.org/10.2134/agronj2000.922345x>
- Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5, 89–97. <https://doi.org/10.5194/adgeo-5-89-2005>
- Kubat, M., Matwin, S., & others. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *ICML*, 97, 179.
- Kuhn, M. (2022). *caret: Classification and regression training*. <https://CRAN.R-project.org/package=caret>

- Kuhn, M., & Vaughan, D. (2022). *yardstick: Tidy characterizations of model performance*. <https://CRAN.R-project.org/package=yardstick>
- Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. <https://www.tidymodels.org>
- Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*. <https://doi.org/10.21105/joss.01903>
- Legates, D. R., & McCabe Jr., G. J. (1999). Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233–241. <https://doi.org/10.1029/1998WR900018>
- Lüdtke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Miguez, F. (2022). *Apsimx: Inspect, read, edit and run 'APSIM' "next generation" and 'APSIM' classic*. <https://CRAN.R-project.org/package=apsimx>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Olsen, L. R., & Zachariae, H. B. (2021). *cvms: Cross-validation for model selection*. <https://CRAN.R-project.org/package=cvms>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062), 1518–1524. <https://doi.org/10.1126/science.1205438>
- Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review*, 22(1), 17–25. <https://doi.org/10.2307/2088760>
- Robinson, W. S. (1959). The geometric interpretation of agreement. *American Sociological Review*, 24(3), 338–345. <https://doi.org/10.2307/2089382>
- Seibold, S. A. D., Heidi AND Czerny. (2022). Correction: A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. *PLOS ONE*, 17(5), 1–1. <https://doi.org/10.1371/journal.pone.0269047>
- Smith, E. P., & Rose, K. A. (1995). Model goodness-of-fit analysis using regression and related techniques. *Ecological Modelling*, 77(1), 49–64. [https://doi.org/10.1016/0304-3800\(93\)E0074-D](https://doi.org/10.1016/0304-3800(93)E0074-D)
- Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>
- Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural Systems*, 89(2), 225–247. <https://doi.org/10.1016/j.agsy.2005.11.004>
- Thieme, N. (2018). R generation. *Significance*, 15(4), 14–19. <https://doi.org/10.1111/j.1740-9713.2018.01169.x>
- Wallach, D., Makowski, D., Jones, J. W., & Brun, F. (2019). Chapter 9 - model evaluation. In D. Wallach, D. Makowski, J. W. Jones, & F. Brun (Eds.), *Working with dynamic crop models (third edition)* (Third Edition, pp. 311–373). Academic Press. <https://doi.org/10.1016/B978-0-12-811756-9.00009-5>

- Warton, D. I., Wright, I. J., Falster, D. S., & Westoby, M. (2006). Bivariate line-fitting methods for allometry. *Biological Reviews*, 81(2), 259–291. <https://doi.org/10.1017/S1464793106007007>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golem, G., Haye, A., Henry, L., Hester, J., Kuhn, M., Lapeere, M., Miller, E., Munn, S. M., Müller, K., Roldán, J., Schloer, B., Sievert, C., ... Hiroaki Yutani. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2(2), 184–194. <https://doi.org/10.1080/02723646.1981.10642213>
- Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddes, J. J., Klink, K. M., Legates, D. R., O'Donnell, J., & Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans*, 90(C5), 8995–9005. <https://doi.org/10.1029/JC090iC05p08995>
- Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, 32(13), 2088–2094. <https://doi.org/10.1002/joc.2419>
- Yang, J. M., Yang, J. Y., Liu, S., & Hoogenboom, G. (2014). An evaluation of the statistical methods for testing the performance of crop models with observed data. *Agricultural Systems*, 127, 81–89. <https://doi.org/10.1016/j.agsy.2014.01.008>
- Zambrano-Bigiarini, M. (2020). *hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series*. <https://doi.org/10.5281/zenodo.839854>