Geophysical Research Abstracts
Vol. 19, EGU2017-7899, 2017
EGU General Assembly 2017

Life beyond MSE and R^2 — improving validation of predictive models with observations

Andreas Papritz and Madlene Nussbaum

ETH Zurich, Institute of Biogeochemistry and Pollutant Dynamics, Department of Environmental Systems Science D-USYS, Zurich, Switzerland (papritz@env.ethz.ch)

Machine learning and statistical predictive methods are evaluated by the closeness of predictions to observations of a test dataset. Common criteria for rating predictive methods are bias and mean square error (MSE), characterizing systematic and random prediction errors. Many studies also report R^2-values, but their meaning is not always clear (correlation between observations and predictions or MSE skill score; Wilks, 2011). The same criteria are also used for choosing tuning parameters of predictive procedures by cross-validation and bagging (e.g. Hastie et al., 2009). For evident reasons, atmospheric sciences have developed a rich box of tools for *forecast verification*. Specific criteria have been proposed for evaluating deterministic and probabilistic predictions of binary, multinomial, ordinal and continuous responses (see reviews by Wilks, 2011, Jollie and Stephenson, 2012 and Gneiting et al., 2007). It appears that these techniques are not very well-known in the geosciences community interested in machine learning. In our presentation we review techniques that offer more insight into proximity of data and predictions than bias, MSE and R^2 alone. We mention here only examples: (i) Graphing observations vs. predictions is usually more appropriate than the reverse (Piñeiro et al., 2008). (ii) The decomposition of the Brier score score (= MSE for probabilistic predictions of binary yes/no data) into reliability and resolution reveals (conditional) bias and capability of discriminating yes/no observations by the predictions. We illustrate the approaches by applications from digital soil mapping studies.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, **69**, 243–268.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York, second edition.

Jolliffe, I. T. and Stephenson, D. B., editors (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley-Blackwell, second edition.

Piñeiro, G., Perelman, S., Guerschman, J., and Paruelo, J. (2008). How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling*, **216**, 316–322.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press, third edition.