EGU2020-15729
https://doi.org/10.5194/egusphere-egu2020-15729
EGU General Assembly 2020
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Evaluating the accuracy of equivalent-source predictions using cross-validation

Leonardo Uieda[1] and Santiago Soler[2,3]
[1]University of Liverpool, School of Environmental Sciences, Department of Earth, Ocean and Ecological Sciences, Liverpool, United Kingdom of Great Britain and Northern Ireland (leouieda@gmail.com)
[2]CONICET, Argentina
[3]Instituto Geofísico Sismológico Volponi, Universidad Nacional de San Juan, Argentina

We investigate the use of cross-validation (CV) techniques to estimate the accuracy of equivalent-source (also known as equivalent-layer) models for interpolation and processing of potential-field data. Our preliminary results indicate that some common CV algorithms (e.g., random permutations and k-folds) tend to overestimate the accuracy. We have found that blocked CV methods, where the data are split along spatial blocks instead of randomly, provide more conservative and realistic accuracy estimates. Beyond evaluating an equivalent-source model's performance, cross-validation can be used to automatically determine configuration parameters, like source depth and amount of regularization, that maximize prediction accuracy and avoid over-fitting.

Widely used in gravity and magnetic data processing, the equivalent-source technique consists of a linear model (usually point sources) used to predict the observed field at arbitrary locations. Upward-continuation, interpolation, gradient calculations, leveling, and reduction-to-the-pole can be performed simultaneously by using the model to make predictions (i.e., forward modelling). Likewise, the use of linear models to make predictions is the backbone of many machine learning (ML) applications. The predictive performance of ML models is usually evaluated through cross-validation, in which the data are split (usually randomly) into a training set and a validation set. Models are fit on the training set and their predictions are evaluated using the validation set using a goodness-of-fit metric, like the mean square error or the R^2 coefficient of determination. Many cross-validation methods exist in the literature, varying in how the data are split and how this process is repeated. Prior research from the statistical modelling of ecological data suggests that prediction accuracy is usually overestimated by traditional CV methods when the data are spatially auto-correlated. This issue can be mitigated by splitting the data along spatial blocks rather than randomly. We conducted experiments on synthetic gravity data to investigate the use of traditional and blocked CV methods in equivalent-source interpolation. We found that the overestimation problem also occurs and that more conservative accuracy estimates are obtained when applying blocked versions of random permutations and k-fold. Further studies need to be conducted to generalize these findings to upward-continuation, reduction-to-the-pole, and derivative calculation.

Open-source software implementations of the equivalent-source and blocked cross-validation (in progress) methods are available in the Python libraries Harmonica and Verde, which are part of the Fatiando a Terra project (www.fatiando.org).