How confident are We in Our Models? Unidentified Dataset Dependencies may Inflate Machine Learning Performance Metrics

Taryn Scharf[1], Matthew Daggitt[2], Luc Doucet[1], Christopher Kirkland[1]
[1]Curtin University, Perth, Australia, [2]University of Western Australia, Perth, Australia

Machine learning algorithms are increasingly used to develop predictive models across Earth Science including mineral systems. The accuracy of reported model performance is important if models are to reliability inform decision making. However, unidentified data dependencies can contribute to 'data-leakage' between training and testing datasets. In effect such leakage promotes model overfitting. This may cause a model to have artificially higher performance on the test set and reduced generalisability.

Geological datasets are often hierarchical, with multiple analyses (e.g. assay data) drawn from a single sample (e.g. drill core), and multiple samples themselves relating to a single phenomenon (e.g. mineralised region). Here we demonstrate the effects of hierarchical data structures on algorithm performance by simulating a dataset of samples comprising multiple analyses that are related through latent variables (e.g. individual analyses linked to a common compositional sample). With this simulated dataset, we illustrate conditions that exacerbate the impact of data-leakage in hierarchical datasets. We then demonstrate this phenomenon in real-world datasets.

This work highlights that dependencies are often present in geological datasets. If these dependencies are unidentified and untreated, they may result in over-optimistic estimates of model performance.