Geophysical Research Abstracts
Vol. 19, EGU2017-2149, 2017
EGU General Assembly 2017

Uncertainty in Random Forests: What does it mean in a spatial context?

Jens Klump and Francky Fouedjio
CSIRO, Mineral Resources, Kensington, Australia (jens.klump@csiro.au)

Geochemical surveys are an important part of exploration for mineral resources and in environmental studies. The samples and chemical analyses are often laborious and difficult to obtain and therefore come at a high cost. As a consequence, these surveys are characterised by datasets with large numbers of variables but relatively few data points when compared to conventional big data problems. With more remote sensing platforms and sensor networks being deployed, large volumes of auxiliary data of the surveyed areas are becoming available. The use of these auxiliary data has the potential to improve the prediction of chemical element concentrations over the whole study area.

Kriging is a well established geostatistical method for the prediction of spatial data but requires significant pre-processing and makes some basic assumptions about the underlying distribution of the data. Some machine learning algorithms, on the other hand, may require less data pre-processing and are non-parametric. In this study we used a dataset provided by Kirkwood et al. [1] to explore the potential use of Random Forest in geochemical mapping. We chose Random Forest because it is a well understood machine learning method and has the advantage that it provides us with a measure of uncertainty.

By comparing Random Forest to Kriging we found that both methods produced comparable maps of estimated values for our variables of interest. Kriging outperformed Random Forest for variables of interest with relatively strong spatial correlation. The measure of uncertainty provided by Random Forest seems to be quite different to the measure of uncertainty provided by Kriging. In particular, the lack of spatial context can give misleading results in areas without ground truth data.

In conclusion, our preliminary results show that the model driven approach in geostatistics gives us more reliable estimates for our target variables than Random Forest for variables with relatively strong spatial correlation. However, in cases of weak spatial correlation Random Forest, as a nonparametric method, may give the better results once we have a better understanding of the meaning of its uncertainty measures in a spatial context.

References

[1] Kirkwood, C., M. Cave, D. Beamish, S. Grebby, and A. Ferreira (2016), A machine learning approach to geochemical mapping, Journal of Geochemical Exploration, 163, 28–40, doi:10.1016/j.gexplo.2016.05.003.