PREDICTIVE SOIL MAPPING
USING MACHINE LEARNING ALGORITHMS

George Garbacea

Predictive Soil Mapping (PSM) is based on applying statistical and/or machine learning techniques to fit models for the purpose of producing spatial and/or spatiotemporal predictions of soil variables, i.e. maps of soil properties and classes at different resolutions. It is a multidisciplinary field combining statistics, data science, soil science, physical geography, remote sensing, geoinformation science and a number of other sciences (Scull et al, 2003; McBratney et al, 2003; Henderson et al, 2004; Boettinger et al, 2010; Zhu et al, 2015).

The main differences between predictive vs traditional expert-based soil mapping are that the production of maps is based on using statistical methods to ensure objectivity

of maps (including objective uncertainty assessment vs expert judgment), and PSM is driven by automation of the processes so that overall soil data production costs can be reduced and updates of maps implemented without requirements for large investments.

Various methods could be used for fitting quantitative relationships between soil properties and dependent variables. These include generalized linear models, classification and regression trees, neural networks, fuzzy systems and geostatistics methods (McBratney, 2003).

In this study we test witch machine learning algorithms (MLA) can give the best prediction using soil samples that had been collected by national forest inventory (NFI) department in Romania. The environmental variables used as dependent variables were selected based on correlation with the predicted parameter running Multiple Linear Regression Analysis for points and predictor grids. From soil database we selected soil pH as the predicted parameter to compare the results with geostatistical modeling that had been previously calculated. We tested 4 machine learning algorithms to test their performance.

Table 1 – RMSE of Machine learning algorithms used in predicting soil pH:

Algorithms	**RMSE**
GSIF_RF	0.25
RF	0.47
RT	0.94
GLM	0.92

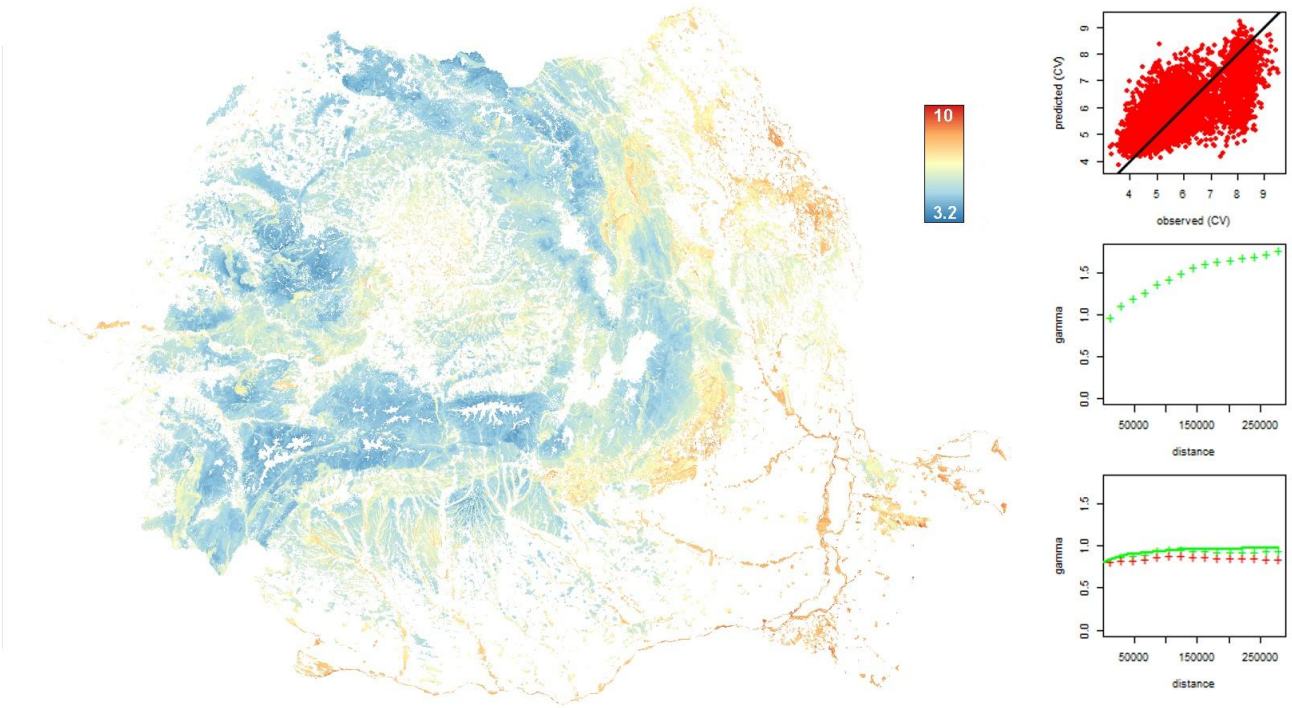

Fig. 1 Predicted map of soil pH

The results show (Table 1) that Random Forest (GSIF_RF) is by far the best algorithm in predicting soil pH (Fig. 1), recording a low RMSE in comparison with other algorithms. RF algorithm also records a lower RMSE in comparison with geostatistical methods as regression kriging (0.98). MLA algorithms can be a much better alternative for interpolation than traditional interpolation methods.

References

McBratney B., Minasny B., Mendonça Santos M. L. (2003). On digital soil mapping. Geoderma, ISSN: 0016-7061, Vol: 117, Issue: 1, Page: 3-52

Boettinger J.L., Howell D.W., Moore A.C., Hartemink A.E., Kienast-Brown S. (2010). Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Progress in Soil Science, vol 2. Springer

Henderson B.L., Bui E.N., Moran C.J., Simon D. (2004). Australia-wide predictions of soil properties using decision trees. Geoderma 124(3-4):383–398

Scull P., Franklin J., Chadwick O.A., McArthur D. (2003). Predictive soil mapping: a review. Progress in Physical Geography 27(2):171–197

Zhu A., Liu J., Du F., Zhang S., Qin C., Burt J., Behrens T., Scholten T. (2015). Predictive soil mapping with limited sample data. European Journal of Soil Science 66(3):535–547, DOI 10.1111/ejss.12244