SPATIAL PREDICTION OF SOIL CLAY CONTENT USING RANDOM FOREST KRIGING IN HIEP HOA DISTRICT, BAC GIANG PROVINCE, VIETNAM

Doan Thanh Thuy[1]

[1] Department of Land Information System, Faculty of Land Management, Vietnam National University of Agriculture, Trau Quy, Gia Lam, Ha Noi

Email: doanthanhthuy209@gmail.com

ABSTRACT

In Vietnam, it is essential to manage land efficiently in order to meet the increasing requirements of food and fuels while national farm land is limited. Thus, it is essential to have accurate, quantitative and up-to-date soil information for supporting sustainable land management. Due to the lacks of funds, conventional soil mapping methods in Vietnam are limited in terms of both the level of spatial detail, spatial coverage and the accuracy of soil attributes. Therefore, there is an urgent need to apply Digital Soil Mapping (DSM) method which embraces a set of quantitative methods due to its lower cost requirement for sampling. Information on the spatial variability of soil texture including soil clay content in a landscape is very important for agricultural and environmental use. This research aims to assess hybrid spatial model called Random Forest Kriging (RFK) with the use of auxiliary variables based on machine learning algorithms for predicting soil clay content in Hiep Hoa district, Bac Giang province, Vietnam. Four derivatives extracted from the digital elevation model, together with the map of NDVI, PVI which were derived from satellite image and information derived from the map of land use, geology were used as predictors in RFK modeling. Predicting ability of the model was assessed with 5 folds cross validation approach to calculate Root Mean Square of Error (RMSE), Mean Error (ME). According to the findings of the research, the RFK model predicted Clay content with a relatively high accuracy (ME=0.13% and RMSE=4.924% for Clay content in the range of 8.3-49.7%). The model also indicated that most important environmental variables which affect the variation of Clay content in Hiep Hoa are elevation and parent materials. We found RFK to be an effective prediction method and recommend this method for any future soil mapping activities in Vietnam.

Key words: Digital soil mapping, Random Forest Kriging, soil clay content.

I. INTRODUCTION

Soil spatial variability is a well-known phenomenon of the soil system and has been recognized for many years. Variability of soil properties in a landscape is a result of differences in soil-forming factors, landforms, or geomorphic elements and soil use and management. (1941) proposed a famous equation which he suggested soil as a function of climate, organisms, relief, parent material and time: ***S= f (c, o, r, p, t,...)***. Clay content variability is an important properties of soil because it involved in almost every reaction in soils which affects plant growth. Both chemical and physical properties of soils are controlled to a very large degree by properties of clay, and an understanding of clay properties is essential if we are to arrive at a full understanding of soil plant relationships (Buehrer, 1952). A better understanding of clay variability is vital for proper crop and land management and is also useful for environmental

assessment and modeling of nutrient and pesticide leaching in the soil (K. Adhikari et al., 2013).

One of the most helpful and functional tools to understand soil clay content variability is mapping. However, maps of those properties are very limited in Vietnam due to low budget for soil sampling and laboratory analyzing. Thus it is essential to find a DSM method to provide maps of Clay content with sufficiently resolution. Modern users of soil geo-information require maps at detailed scales. The technological and theoretical advances in the last 20 years have led to a number of new methodological improvements in the field of soil mapping. Most of these belong to the domain of a new emerging discipline – **pedometrics** – for the quantitative, (geo)statistical production of soil geoinformation. Pedometrics is strongly focused on **predictive** or **digital soil mapping (DSM)**. DSM embraces a set of quantitative mapping methods that have developed from more traditional soil mapping techniques. There were various case studies that demonstrated the application of DSM methods in mapping soil properties and classes, updating soil attribute maps or mapping soil features (Carré et al., 2002; Jafari et al., 2012; Kempen et al., 2009; Yang et al., 2011). Mc Kenzie et al (1999) used generalized linear models to predict soil clay content using environmental variables such as geomorphic unit, local relief, as predictors. Shen et al (2013) produced map of clay content in Glacial till soil using near infrared reflectance spectroscopy as predictors with partial least squares regression. Hengl et al (2015) used a hybrid method called Random forest kriging (RFK) to map soil properties of Africa including clay content at 250m resolution. Their study confirmed that the random forest kriging consistently outferformed the regression kriging for all soil properties. Therefore, the aims of this study is to using Random forest kriging method to map soil clay content of Hiep Hoa district at a large resolution which is 10m. The resultant map could be an important document to support sustainable agricultural production of the study area. Moreover, this techniques can allowing the application of DSM in other similar landscapes of Vietnam.

II. MATERIALS AND METHODS

2.1. Study area

The study area is Hiep Hoa district of Bac Giang province in Vietnam, covers an area of 203 km^2. It is located in the Red River Delta region – one of the highest rice production of Vietnam (Mussgnug et al., 2006), about 50 kilometers to the east of Hanoi (Figure 1). Hiep Hoa is characterized by flat to slightly terraced topography, elevation ranges from 5 to 100m. Hiep Hoa district's economy is primarily dedicated to agriculture which accounts for 63% of the economic and 67% of the total area. The main production of the district are rice, vegetable, livestock, poultry and aquaculture. It is essential to have an appropriate understanding about the properties of the soil in Hiep Hoa in order to maintain a sustainable agriculture in Hiep Hoa.

There are 100 soil samples which were taken at a depth of 0 to 15cm in 2015. The sampling scheme were designed based on existing soil map, geology map of Hiep Hoa district. Figure 1 shows the spatial distribution of soil samples. These samples were then analyzed in the laboratory following the Vietnamese standard for Soil Quality which were published by Department of Science and Technology.

2.2. Environmental covariates

Terrain, vegetation and parent materials were chosen to be the environmental conditions which characterize the Clay content of the topsoil in Hiep Hoa district. There are 5 environmental variables were generated including elevation, slope gradient, parent materials,

land use types and normalized difference vegetation index (NDVI). Information of environmental data in detail were listed in (Table 1)

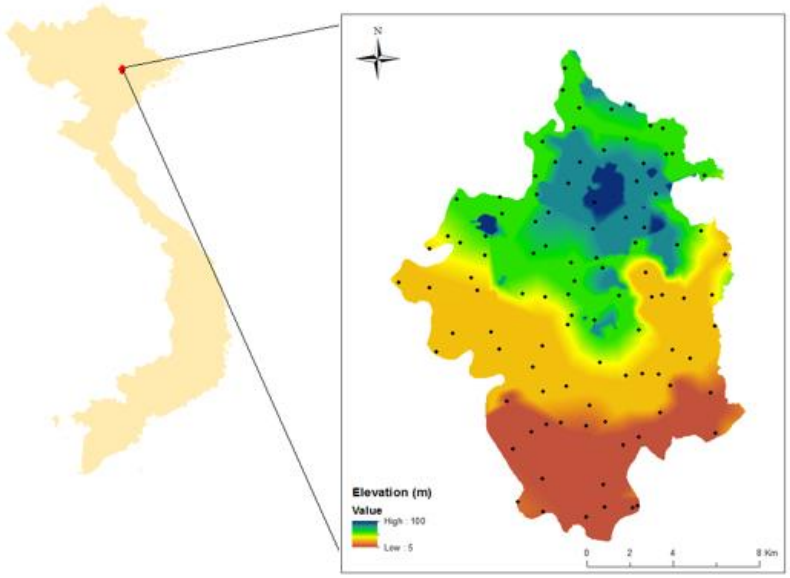

Figure 1: Locations of soil samples in Hiep Hoa district

Table 1: Environmental covariates

Category of data	Variables	Original resolution/scale
Terrain	Elevation	Vector, 1:10,000
	Slope gradient	Raster, 10m
Parent materials	Geology	Vector, 1: 200,000
Vegetation	Normalized Difference Vegetation Index (NDVI)	Raster, 10m
	Land use	Vector, 1:10,000

2.3. Random forest kriging

In the last decade, a number of 'hybrid' interpolation techniques, which combine kriging and use of auxiliary information, has been developed and tested. Here, two main paths can be recognized: co-kriging and kriging with machine learning algorithms (Hengl et al., 2015; McBratney et al., 2000). Common machine learning algorithms are: artificial neural networks, support vector machines, classification and regression trees, and random forest. In this paper we specifically evaluate the applicability of the random forests algorithm which were explained in very details by Breiman (2001) for soil mapping. This is for two main reasons: (1) it has been proven in numerous studies (Hengl et al., 2015; Kuhn et al., 2013; Strobl et al., 2009) that the random forests algorithm can outperform linear regression and (2) unlike linear regression, random forests has no requirements considering the probability distribution of the target variable and can fit complex non-linear relationships in p+ 1-dimensional space (where p is the number of covariates).

A limitation of using random forests however, is that the model is usually only effective within the range in covariate values exhibited by the training data (Statnikov et al., 2008).

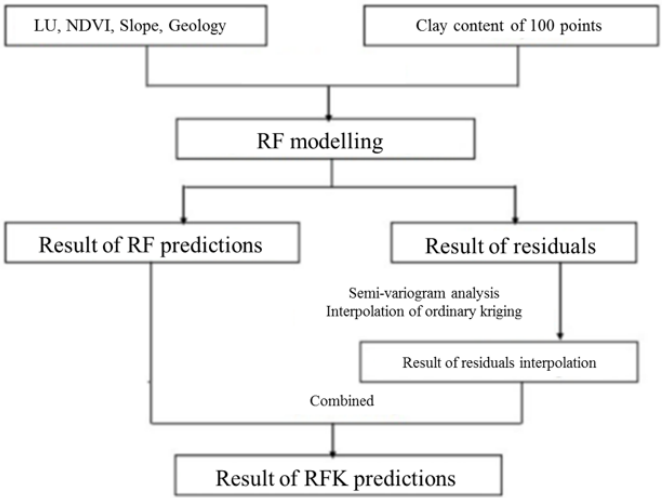

Figure 2: Flowchart of Random forest kriging method

In general terms, RFK approach follow the flowchart shown in figure 2. All modeling and validation processes were done in R project using fit.gstatModel function from the GSIF package.

2.4. Validation

The performance of RFK model were evaluated using 5-fold cross-validation. In this process, the total dataset was randomly divided into 5 subsets, 4 of which were used for calibration and 1 subset was used for validation. This procedure were repeated 5 times, each time 1/5 group is used for calculating Mean Error (ME), Root Mean Squared Error (RMSE), Residual Prediction Deviation (RPD) and coefficient of determination (R^2) in turn. Finally, the average ME and RMSE were calculated to assess the prediction accuracy of the RFK model. The formula of ME and RMSE are as below (Williams, 1987):

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_{pred} - Y_{obs}) \qquad RMSE = \frac{1}{n}\sum_{i=1}^{n}(Y_{pred} - Y_{obs})^2$$

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{pred} - \overline{Y})^2}{\sum_{i=1}^{n}(Y_{obs} - \overline{Y})^2} \qquad \text{RPD} = \frac{S.D.}{RMSE} * \sqrt{\frac{n}{n-1}}$$

Where n is the number of validation points, Y_{pred} is the clay content predicted by RFK model and Y_{obs} is the clay content measured at that point.

III. RESULTS AND DISCUSSIONS

3.1. Random Forest Kriging modelling

As mentioned earlier, a random forest was constructing with the input of elevation, geological unit, land use types, slope gradient, NDVI and measured clay content from calibration dataset. As can be seen from Figure 3, geological unit and the elevation are the most important environmental variables that affected the spatial variability of soil clay content. However, slope, NDVI and especially land use are very poor predictors of clay content. These findings correspond with observations by Balstrøm et al. (2013) that geology are good predictors of soil clay content on a national scale in Denmark. Figure 4 depicts the experimental variogram of clay content. Each point of the variogram was the result of the average

semivariance of the point pairs falling within the specific lag defined during variogram calculation.

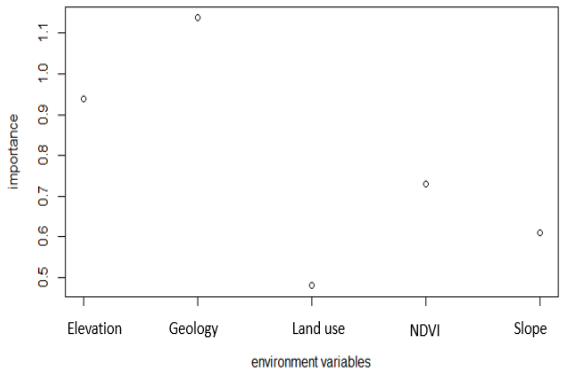

Figure 3: Importance plots for prediction of soil clay content by RFK

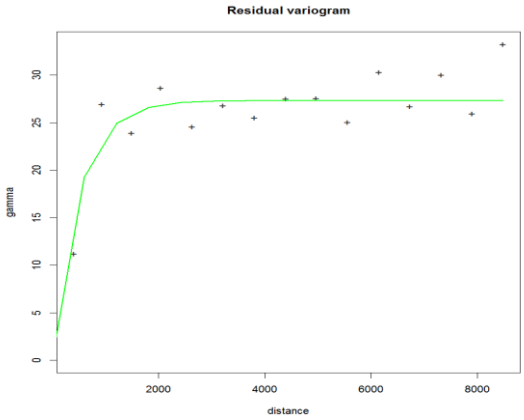

Figure 4: Variogram of residuals of soil clay content by RFK

3.2. Prediction accuracy

The map of soil clay content as predicted by RFK was shown in figure 5. The map shows a general trend for clay content to be lower toward the east and higher toward the west of Hiep Hoa district. Most of the areas in the east have a clay content less than 10%, whereas clay content increases to as much as 49% in southwestern area.

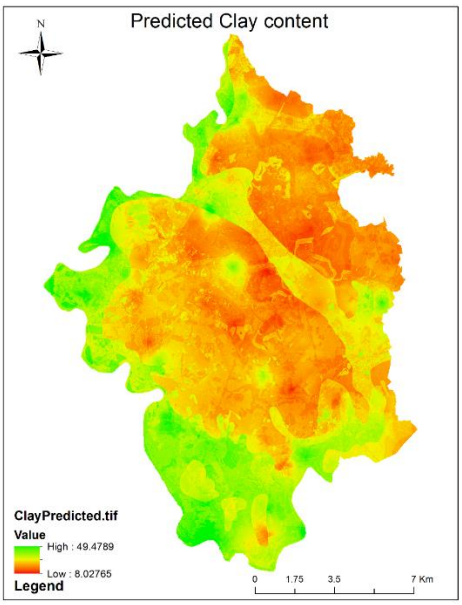

Figure 5: Distribution of Clay content predicted by RFK

Table 2: Prediction accuracy of RFK for clay content

ME	0.13
RMSE	4.92
R^2	0.68
RPD	1.81

Table 2 indicates the performance of the RFK model to predict soil clay content of Hiep Hoa district. The RFK was able to capture 68% variability (R^2=0.68) with an RMSE of 4.92. The RPD is 1.81 which indicates that the RFK had moderate prediction performance as suggested by (Williams (1987)). On overall, Random forest proved to be successful in predicting soil clay content in Hiep Hoa district with relatively high coefficient of determination, low ME (0.13) and moderate RPD.

Thus RFK can be a promising approach to map the spatial variability of soil properties in Vietnam. More area need to be test with variety of landscape and relief.

III. REFERENCES

Balstrøm, T., Breuning-Madsen, H., Krüger, J., Jensen, N. H., & Greve, M. H. (2013). A statistically based mapping of the influence of geology and land use on soil pH A case study from Denmark. *Geoderma, 192*, 453–462.

Breiman, L. (2001). RANDOM FORESTS.

Buehrer, T. F. (1952). Role of Chemical properties of clays in soil science *Clays Clay Technol Bull* (pp. 169:167-176).

Carré, F., & Girard, M. C. (2002). Quantitative mapping of soil types based on regression kringing of taxonomics distances with landform and land cover attributes. *Geoderma, 111*, 241-263.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., . . . Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PLoS One, 10*(6), e0125814. doi: 10.1371/journal.pone.0125814

Jafari, A., Finke, P., Van de Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA-great group in arid Zarand region, Iran, comparing logistic regression approaches to predict diagnostic horizons and soil types. *European journal of Soil science*.

Jenny, H. (1941). *Factors of soil formation - a system of quantitative pedology*: New York: McGraw-Hill.

K. Adhikari, R.B. Kheir, M.B. Greve, & M.H. Greve. (2013). Comparing Kriging and Regression Approaches for MappingSoil Clay Content in a Diverse Danish Landscape. *Soil Science, 178*(9).

Kempen, B., Brus, D. J., Heuvelink, G. B. M., & Stoorvogel, J. J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma, 151*, 311-326.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling*: Springer.

McBratney, A. B., Odeh, I. O. A., Bishop, T. F. A., Dunbar, M. S., & Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma, 97*, 293-327.

McKenzie, N. J., & Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma, 89(1-2)*(67-94).

Mussgnug, F., Becker, M., Son, T. T., Buresh, R. J., & Vlek, P. L. G. (2006). Yield gaps and nutrient balances in intensive, rice-based cropping systems on degraded soils in the Red River Delta of Vietnam. *Field crops research, 98*, 127-140.

Statnikov, A., Wang, L., & Aliferis, C. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *Bmc Bioinformatics, 9*(1).

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods, 14* (4), 323.

Williams, P. C. (1987). Variables affecting near-infrared reflectance spectroscopic analysis. In P. C. Williams & K. Norris (Eds.), *Near-Infrared Technology in the Agricultural and Food Industries* (pp. 146-166). St. Paul, MN: American Association of Cereal Chemists.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.-X., Hann, S., Burt, J. E., & Qi, F. (2011). Updating Conventional Soil Maps through Digital Soil Mapping. *Soil Science Society of America, 75*(3), 1044-1053.

Zhang-Quan, S., Ying-Jie, S., Li, P., & Yu-Gen, J. (2013). Mapping of Total Carbon and Clay Contents in Glacial Till Soil Using On-the-Go Near-Infrared Reflectance Spectroscopy and Partial Least Squares Regression. *Pedosphere, 23*(3), 305-311.