DIGITAL SOIL MAPPING WITH REGRESSION TREE CLASSIFICATION APPROACHES BY RS AND GEOMORPHOMETRY COVARIATE IN THE QAZVIN PLAIN, IRAN

Mousavi [1], S.R., Sarmadian*[1], F., Rahmani, A[1]., Khamoshi, S E. [1]

[1] Dep. of Soil Science and engineering, University of Tehran, Iran - (a. rahmani, fsarmad, r_mousavi, khamoshierfan)@ut.ac.ir

KEY WORDS: Random Forest, Boosting decision tree, Soil Mapping, Data mining

ABSTRACT:

Digital soil mapping applies soil attributes, Remote sensing and Geomorphometrics indices to estimate soil types and properties at unobserved locations. This study carried out in order to comparison two data mining algorithms such as Random Forest (RF) and Boosting Regression tree (BRT) and two features selection principal component analysis (PCA) and variance inflation factor (VIF) for predicting soil taxonomy class at great group and subgroup levels. A total of 61 soil profile observation based on stratified random determined and digged in area with approximately 16660 hectares.19 RS indices and geomorphometrics covariates derivated from Landsate-8 imagery and DEM with 30 meters' resolution in ERDAS IMAGINE 2014 and SAGA GIS version 7.0 software's. Also to run four Data mining algorithms scenarios (PCA-RF, VIF-RF, PCA-BRT, VIF-BRT) from "Randomforest" and "C.5" packages were used in R studio software. 80% and 20% from soil profiles were applied for calibrating and validating. The results showed that in PCA and VIF approaches, eight covariates such as (Relative slope position, diffuse insolation, modified catchment, normalized height, RVI, Standard height, TWI, Valley depth) and six covariates (NDVI, DVI, Catchment area, DEM, Salinity index, Standard height) were selected. The validation results based on overall accuracy and kappa index for scenarios at great group level indicated that 88,93,62, 54 and 75,83,51,45 percentages and for subgroup level had 70, 77, 54, 47 and 60, 71, 43, 37 percentages, respectively. Generally, VIF-RF had accuracy rather than from other scenarios at two categorical level in this study area.

* Corresponding author

1. INTRODUCTION

Soil mapping is required as a prerequisite for agricultural land management, but according to statistics, about 75% of Iran's soils have a shortage detailed information at 1:25000 scale (Roozitalab, 2018). Digital soil mapping (DSM) has been widely used as a cost-effective method for generating soil maps (Padarian et al. 2019). Also DSM has now been widely used globally for mapping soil classes and properties (Arrouays et al., 2014). In particular, DSM has been used to map soil type in Iran (Taghizadeh et. al 2015). The incorporation of remote sensing (RS) data as well as digital elevation model (DEM) data and derivatives thereof have been used in DSM studies (Boettinger,2010). The use of Landsat spectral data has been specially in arid and semi-arid area to estimate some soil properties (Taghizadeh et. al 2015 ؛Padarian et al.,2019). But selection of the best covariates for modeming of map soils is one of the challenges before using of mathematical and statistical methods for soil predicting. As soon as different data mining methods should have been used such as variance inflation factor (Dormann et al. 2013) and principal component analysis (Brungard et al. 2015). Tree-based methods are atypical statistical models – they do not utilise distributions, likelihoods or design matrices; metrics typically associated with modelling. Regression Trees are tree-based models that have been widely used in DSM (Taghizadeh-Mehrjardi et al., 2016). Random Forest may also be used for both regression and classification purposes (Dharumarajan et al., 2017). Random Forest operates via a resampling approach or boosting, where for regression, the prediction is the average of the individual tree outputs, whereas in classification, the trees vote by majority on the correct classification mode (Grimm et al., 2008). Bosting is a combination of false algebra in the field of machine learning that is used to reduce imbalance and variance. This method is used in supervised learning and is a family of machine learning algorithms. This method is to transform weak learning systems into strong based on the combination of different class results. In fact, using a boosting method, a sequence of decision trees is developed. Each tree tries to reduce the error rate of the wrong classification. The C5.0 algorithm uses the cynical pruning method to remove the wrong classification.

Each tree tries to reduce the error rate of the wrong classification (DeFriesand Chan, 2000). The advanced feature of the C5.0 algorithm is the use of the Boosting method (Kuhn and Johnson, 2013).

2. PROPOSED METHOD

2.1. Study area and soil sampling

An area in the Qazvin plain of Iran, across 36 1° and 36 9° N, and 50 14° and 50 21° E was chosen (Fig. 1). It covers nearly 16660 ha. Piedmont (45%), Plain (44.58%), Peneplain (9.29%) and Hilland (1.13%) are the dominant landscape units in this area. The mean elevation of the area is 1287 m a.s.l., and the slope variation is zero to 25%. Mean annual precipitation is 257 mm, and temperature is 14.37 °C. Based on Iranian soil moisture and temperature regime map and synoptic meteorological station of Qazvin (2015) the soil moisture and temperature regimes are dry xeric, weak aridic, aquic and thermic, respectively. 61 pedons with 750-m intervals and using the stratified random sampling method were excavated in various Geoform map (Zinck et al., 2016) units of studied area (1: 50,000 scale) based on a semi-detailed soil survey (Rossiter 2000). Then, the pedons were described according to the ‘‘field book for describing and sampling soils’’ (Schoeneberger et al. 2012). After that soil samples were taken from all identified genetic horizons, air-dried, and transform to soil geneses and classification laboratory of Tehran University for determining physiochemical properties. Finally, the pedons were classified based on American soil taxonomy (Soil Survey Staff 2014) up to subgroup level.

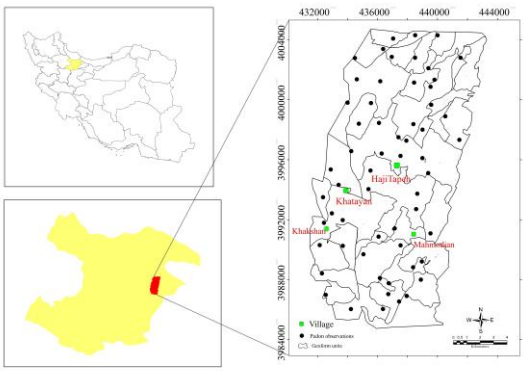

Figure. 1 Location of the study area with Pedon observation

2.2. Environmental covariates

Digital elevation model (DEM) with 12.5 m spatial resolution was used for derivation primary and secondary Terrain attributes including dem, slope, aspect, Relative Slope Position, Diffuse Insolation, Modified Catchment, Normalized Height, Standard Height, Total wetness index, Valley Depth, Mrvbf, Catchment area, Mid slope position, Vertical distance and Flow accumulation were obtained from SAGA GIS software.

The normalized difference vegetation index (NDVI), Difference vegetation index (DVI) Salinity index (SI), Ratio-based Vegetation Indices (RVI) in Erdas Imagine 2014 that were gathered from one scene of the Landsat 8 operational land imager (OLI) that was acquired on June 2018 with the grid resolution of 30 × 30 m with lowest cloud cover.

2.3. Data mining and spatial prediction

For feature selection, in this study was used from VIF and PCA data mining methods in Minitab.16 version and R Studio software. After selection of the best covariate for modelling then relation between soil class in great group and subgroup levels and covariates was applied "random forest" and "C.5" package in R Studio 1.0.136 version. Four scenarios from two feature selection and regression modelling including PCA-RF, VIF-RF, PCA-BRT, VIF-BRT were considered. Training the models was done with 80% of the data (i.e., 49 pedons) and their validation was tested by the remaining 20% of the dataset (i.e., 12 pedons) that were split randomly. The accuracy of the predicted soil classes was determined using error matrices. Then, map accuracy indicators including overall accuracy, kappa index (K) and adjusted kappa are calculated according to the following equations:

$$OA = \sum_{i=1}^{n} Xij / N \quad (1)$$

$$Kappa = \quad (2)$$

$$N\sum_{i=1}^{n} Xij - \sum_{i-1}^{n}(Xio - Xoi) / N^2 - \sum_{i=1}^{n}(Xio - Xoi)$$

where n is the number of rows (and therefore columns) in the matrix, Xij is the count in a diagonal cell where row and column i meet (i.e., correct classifications), Xio is the row total, Xoi is the column total, and N is the total number of observations.

3. RESULTS

The soil classification results based on Pedon description was found 13 class at subgroup and eight class at great group level with Fluventic Haploxerepts and Haploxerepts as a dominate class at two level respectively. According to data mining methods eight covariate including Relative slope position, diffuse insolation, modified catchment, normalized height, RVI, Standard height, TWI, Valley depth and six covariates such as NDVI, DVI, Catchment area, DEM, Salinity index, Standard height by using PCA and VIF were selected respectively. At great group level was obtained 70, 77, 54, 47 percentages of OA and 60, 71, 43, 37 percentages kappa index at four scenario PCA-RF, VIF-RF, PCA-BRT, VIF-BRT respectively and spatial distribution of great group and subgroup created by VIF-RF shown in Fig.2 and 3, also at great group level the validation results based on OA and kappa index showed that 88,93,62, 54 and 75,83,51,45 that was similar to subgroup level for mentioned scenario. So based on two validation statistics (OA and Kappa) VIF-RF scenario had the higher value from other scenarios at two soil taxonomy level in this study. Based on table 1 and 2 at the great group taxonomic level Haploxerepts and Haplosalids with 37.28% and 0.47% had the maximum and minimum area percentage of observed soils, also Fluventic Haploxerepts and Lithic Xerorthents with 31.25% and 0.20% had the highest and lowest area percentage.

Random forests identify important covariates by generating multiple classification trees using bootstrap sampling, randomly scrambling the covariates in each bootstrap sample, and reclassifying the bootstrap sample. The misclassification error between the bootstrap sample using the scrambled covariate is then compared to the misclassification error of the original covariate (Peters et al., 2007) so in regard to RF model was the best data miner algorithm in this study thus it can have used as an important relatively method according to mean decrease accuracy (MDA) and mean decrease Gini (MDG). In RF model based on MDA and MDG at the best scenario (VIF- RF) shown in (fig .4) base on relative important.

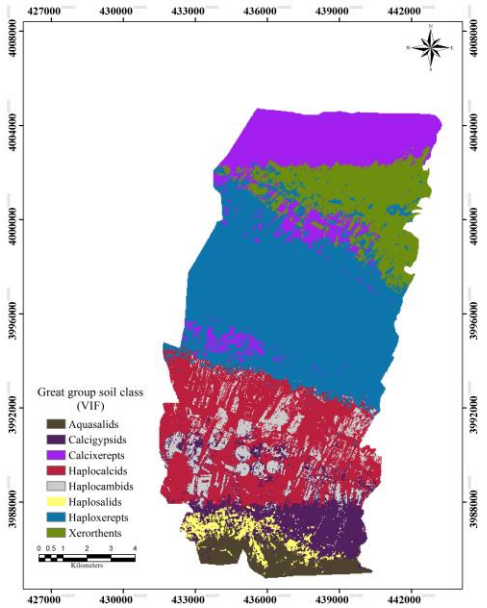

Figure 2. Spatial distribution of the soil great groups derived from RF model

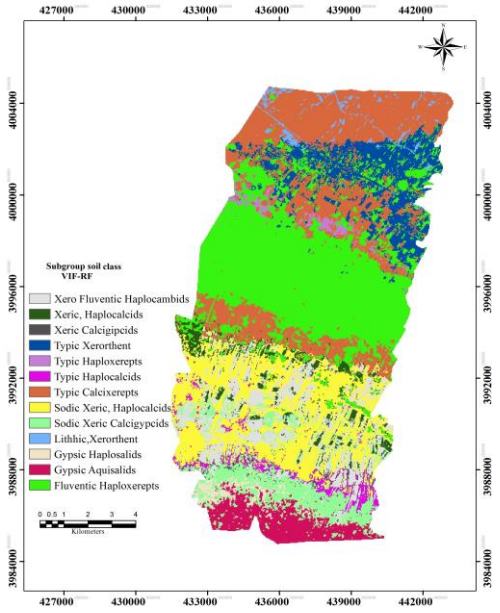

Figure 3. Spatial distribution of the soil subgroups derived from RF model

Table1: Soil great group class area

Number	Soil class	Area (ha)	Area (%)
1	Aquisalids	837.1	5.03
2	Calcigypsids	1200	7.22
3	Calcixerepts	2569	15.45
4	Haplocalcids	2527.04	15.20
5	Haplocambids	1497.7	9.01
6	Haplosalids	77.77	0.47
7	Haploxerepts	6199	37.28
8	Xerorthent	1720.39	10.35
Total	---	16628	100

Table 2: Soil subgroup class area

Number	Classification	Area (ha)	Area (%)
1	Fluventic Haploxerepts	5241.88	31.52
2	Gypsic Aquisalids	163.18	0.98
3	Gypsic Haplosalids	160.5	0.97
4	Lithic Xerorthents	32.97	0.20
5	Sodic Xeric Calcigypsids	1207.58	7.26
6	Sodic Xeric Haplocalcids	1760.74	10.59
7	Typic Calcixerepts	3227.57	19.41
8	Typic Haplocalcids	734.14	4.42
9	Typic Haploxerepts	88	0.53
10	Typic Xerorthents	1487.4	8.95
11	Xeric Calcigypsids	545.88	3.28
12	Xeric Haplocalcids	1079.64	6.49
13	Xero Fluventic Haplocambids	898.52	5.40
Total	---	16628	100

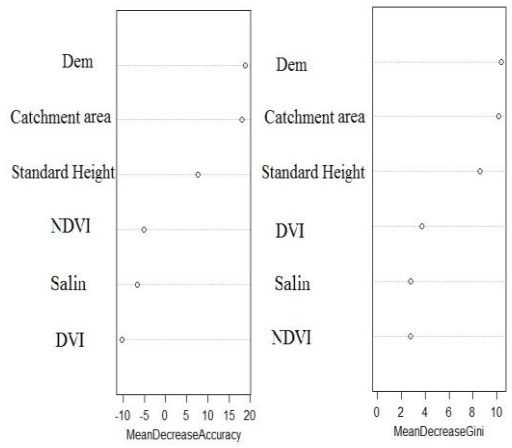

Figure 4. Relative important covariates for subgroup level based on MDA and MDG factors-VIF-RF scenario

4. CONCLUSIONS

Generally, in this study the Geomorphometry covariate had more important relative to remote sensing indices based on the best scenario (VIF-RF) and two relatively important (MDA and MDG), since the spatial distribution of great group and subgroup soil map had visual and statistical validation also VIF-RF can be as a good data mining algorithm in future study in arid and semi-arid regions at family and series soil taxonomy.

REFERENCES

Boettinger, J.L., 2010. Environmental covariates for digital soil mapping in the western USA. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, Dordrecht, pp. 17–27.

Brungard, C.W, Boettinger, J.L, Duniway, M.C, Wills, S.A, Edwards, T.C., 2015. Machine learning for predicting soil classes in three semiarid landscapes. Geoderma 239–240:68–83.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., and Munkemuller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

DeFries, R. S., Chan, J. C. W., 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment, 74(3), 503-515.

Dharumarajan, S., Hegde, R., Singh, S. K., 2017. Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. Geoderma Regional, 10, 154-162.

Grimm, R., Behrens, T., Marker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. Geoderma, 146(1-2), 102-113.

Kuhn, M., Johnson, K., 2013. Applied predictive modelling (Vol. 26). New York: Springer.

Padarian, J., Minasny, B., McBratney, A. B., 2019. Using deep learning for digital soil mapping. Soil, 5(1), 79-89.

Peters, J., De Baets, B., Verhoest, N.E., Samson, R., Degroeve, S., Becker, P.D., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol Model 207:304–318.

Rossiter, D.G, 2000. Methodology for soil resource inventories. Soil Science Division, International institute for Aerospace Survey and Earth Science (ITC). 2nd revised version.

Roozitalab, M. H., Siadat, H., Farshad, A., 2018. The Soils of Iran. Springer International Publishing.

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Soil Survey Staff., 2012. Field book for describing and sampling soils, 3nd version. Natural Resources Conservation Service. National Soil Survey Center, Lincoln.

Soil Survey Staff., 2014. Keys to soil taxonomy. 12th edn. USDA Natural Resources Conservation Service, Washington,

DC.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., & Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma, 253, 67-77.

Zinck, J. A., Metternicht, G., Bocco, G., Del Valle, H. F., .2016. An Integration of Geomorphology and Pedology for Soil and Landscape Studies. Springer.