Evaluating the potential of digital soil mapping to map soil types in Bac Ninh province, Vietnam

Doan Thanh Thuy
Department of land information system, Faculty of Land management, VNUA

ABSTRACT

There has been considerable expansion in the application of digital soil mapping (DSM) techniques because it could help to save much time and costs for collecting and analyzing soil data points compared to conventional methods. This research aims to assess the potential of mapping soil types in a Northern region of Vietnam based on the comparison between two DSM methods: Multinomial Logistic Regression (MLR) and Artificial Neural Networks (ANNs). Eight predictive variables were derived from the ancillary data including land use, altitude, slope, NDVI, PVI, RVI, Topographic Wetness Index and SAGA Wetness Index. MLR and ANNs models were constructed to predict soil classes at 2 levels: WRB-Reference Soil Group and intermediate level of Soil Group between Reference Soil Group and the full WRB soil name. The map quality was indicated by the soil map purity estimated with an independent validation dataset. The diversity indices were calculated to assess the information content of the resultant maps. Selection of the best model is based on the soil map purity, the Shannon's entropy and a combined index. At both taxonomic levels, MLR yields higher map purity than ANNs. When the taxonomic level changed from Reference Soil Group level to intermediate level the map purity decreases while the value of the diversity indices increases. Therefore, soil mapping using MLR in predicting Reference Soil Group will be more efficient. However, at intermediate level, the model predicts higher diversity of soil map and thus the informative value estimated by the combined index is higher.

1. Introduction

Soil remains one of the most important, yet most abused natural resources on the planet; indeed a responsible management of soil resources plays a critical role in the survival and prosperity of many nations around the world (White 2005). The understanding of soil properties and behavior strongly support sustainable land management. One of the most helpful and functional tools to study soil science is soil mapping. Many countries have been involved in making maps of their soils to determine the range of soil types in their territory, where they occur and how they can be used efficiently.

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 – **geometrics** – for the quantitative, (geo) statistical production of soil geo-information. There were various case studies that demonstrated the application of digital soil mapping (DSM) methods in mapping soil properties and classes, updating soil attribute maps or mapping soil features (Carré and Girard 2002; Kempen, Brus et al. 2009; Jafari, Finke et al. 2011; Yang, Jiao et al. 2011). Because traditional methodologies are costly and time-consuming, the use of DSM methods has increased and has resulted in improvements in soil survey and classification steps, also allowing the application of the results in other similar landscapes (Resende 2000).

Vietnam is a developing country; agriculture has played a key role in the economy coupled with the dramatic development of industry and service. These create a huge pressure in using land resources. Although soil mapping has made certain progress, conventional soil maps produced in the past decades are the major data sources for information on the spatial variation of soil. They are limited in terms of both the level of spatial detail and the accuracy of soil attributes as well as high requirements of costs and time. Thus, this project aims to propose a digital soil mapping method which is capable of mapping soil types of Vietnam in more detail and requires lower costs for soil survey.

2. Objective

The objectives of this study is to (1) predict the map of soil types in Hiep Hoa district using MLR and ANNs model and (2) compare the accuracy of these model in predicting the soil types at 2 levels: Soil Group and Intermediate level of soil group.

3. Materials and methods

3.1. Study area

The DSM methods are applied in Bac Ninh province in the Northern part of Vietnam. Bac Ninh is located at 21^o 05’ N latitude and 106^o10’ E longitude and have natural land of about 82,300 ha. It’s a part of Red river delta which has a rather level and flat terrain; mainly sloping from North to South and West to East. The terrain is not much dissected, field areas are 3-7m high and hill and mountain areas are 300-400m high above sea level.

3.2. Data collection

3.2.1. Soil point data

The point dataset was collected during a soil survey project in 2010 and contains 537 observations. At the selected locations, soil profiles were made to describe and classify according to the WRB classification system. The soil was classified in 2 levels: the Reference Soil Groups (RSG) and the qualifiers which describe in detail the properties of the RSG by adding a set of uniquely defined qualifiers (WRB 2006). There were five WRB Reference Soil Groups found in the surveyed area: Fluvisols (402 samples), Acrisols (58 samples), Arenosols (7 samples), Gleysols (15 samples) and Plinthosols (55 samples). At lower level, the classification results in 30 different soil categories, which was considered a too high number for digital soil mapping because of the low presence of samples in many of the categories. Therefore, soil data points were classified into an intermediate level which was based on some properties relevant for soil management: base saturation status as indicator of soil fertility, texture and appearance of hard layer in the soil profile. As a result, 15 intermediate level of soil units were classified as summarized in table 1.

Table 1. Presence of soil profiles at the intermediate categorical level

No	Intermediate level classification	Number of soilprofiles	Properties
1	Acrisol00000	4	Acrisols having no special property
2	Acrisol00001	11	Acrisols having a hard-subsurface horizon (plinthic horizon) which make it more difficult to work on this soil
3	Acrisol10000	21	Acrisols having a low base saturation (dystric qualifier),thus with higher fertilizers need
4	Acrisol10001	22	Acrisols having both hard subsurface horizon (plinthic horizon) and a low base saturation
5	Arenosol10000	7	Arenosols having a low base saturation
6	Fluvisol0001000	9	very wet Fluvisols having reducing condition within 50cm of the soil surface
7	Fluvisol0010000	9	a wet Fluvisols that have reducing condition between 50cm and 100cm from the soil surface
8	Fluvisol0100010	42	Fluvisols have high base saturation and texture of silt, silt loam, silty clay loam or silty clay
9	Fluvisol1000000	171	Fluvisols have low base saturation
10	Fluvisol1000010	38	Fluvisols have low base saturation and texture of silt, silt loam, silty clay loam or silty clay
11	Fluvisol1000100	126	Fluvisols have low base saturation and a hard-subsurface horizon
12	Fluvisol0100000	2	Fluvisols have high base saturation
13	Fluvisol0100001	5	Fluvisols have high base saturation and texture of loamy fine sand or coarser
14	Gleysol10000	15	Gleysols have low base saturation
15	Plinthosol10000	55	Plinthosols have low base saturation

3.2.2. Digital elevation model (DEM)

Topography is one of the most important factors which affects the soil formation, thus it may determine the soil types in an area. Therefore, a DEM of the area at the grid resolution of 25m was created by digitizing the topographic map of the region which has a contour interval of 20m. The DEM was used to derive four terrain attributes using the Saga GIS: Altitude, Slope, Topographic wetness index and SAGA wetness index (Olaya 2004).

3.2.3. Remote Sensing indices

The SPOT image has a resolution of 20m, and was used to compute remote sensing indices such as Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI) and Perpendicular Vegetation Index (PVI) by using ArcGIS. As a result, three raster maps at a resolution of 20m were derived: NDVI map, RVI map and PVI map. Subsequently, these maps were rescaled into a resolution of 25m in order to obtain the same map extent and grid size as the DEM – derived attributes maps. This was done in ArcGIS.

3.2.4. Land use map

A land use map of Bac Ninh province at a scale of 1:25,000 in 2010 was produced to be a source of ancillary data. Because the observations were obtained only in the agricultural area, the following three main land use types were encountered in the study area: two rice per year (LUC), one rice per year (LUK) and annual crops (BHK).

3.3. Multinomial logistic regression

Multinomial logistic regression was used to model the relationships between the Reference Soil Group or the intermediate level soil groups (categorical dependent variables) and the terrain attributes, remote sensing indices and land use types in the research area (quantitative predictors) using the “nnet’ package of R. This model belongs to the family of generalized linear models and is used when with categorical response variable. Suppose that we want to model the probability π_{ij} that observation *i* in each *j*th class of the m soil groups $j = 1 \ldots m.$ In the model for predicting soil groups, the Fluvisols (*j*=1) is taken as the reference class due to its dominance in the soil point data (402 of 537 samples). In the MLR model for more detail level, the Fluvisol1000000 is the reference class for the same reason (171/537 samples). Consequently, the base probability π_{i1} is computed as the residual probability after the other classes $\pi_{i2} \ldots \pi_{im}$ have been modeled. In R-project, we use the *predict* function to provide the probability of all the classes (which of course sum to 1).

3.4. Artificial neural network

ANNs are a standard technique in the range of artificial intelligence and data mining in general. They are thus designed to learn rules from examples. In R-project, ANNs was run using the “neuralnet” package (Fritsch and Guenther 2012)..

The application of an ANNs consists of two stages. During the first stage, the network is trained, meaning that it learns the conditions on which a certain soil group occurs using the calibration data set. Each input unit (cell or neuron) of the ANNs represents a prediction variable: terrain attributes, remote sensing indices and land use units. The output unit represents the Soil Groups or the intermediate level of Soil Groups. The connection between neurons are described by the weight w_i $(w_{i1} \ldots w_{in})$. The adjustment of these weights depends on the learning process. As each attribute combination (in terms of pixels of a grid map) is put into the network in succession, the weights are adjusted iteratively if the predicted output does not match the output of a training data set. The other network parameters including the optimum iteration learning rates, the number of hidden layer and transfer function were adjusted after the stage of learning to train the network. During the second stages, the learned knowledge in terms of the calibrated weights can be applied to the whole study area, for which the same input parameters (terrain attributes, remote sensing indices and land use maps) are available but no soil map has been surveyed. The network then predicts the soil units based on the learned weights. (Behrens., Förster et al. 2005)

3.5. Validation

The quality of a soil map can be determined by comparing the prediction at the calibration sites with the observed values. An independent validation data set of 53 observations was selected randomly from the data set. The predictions based on the dataset excluding the validation dataset are then compared with independent validation data which were not used in the modeling.

For assessing the quality of the predicted soil maps, the map purity was used based on the confusion matrix (Brus, Kempen et al. 2011). Table 2 shows an error matrix: the row margins (the area covered by the map units) of the matrix are known, whereas the column margins (the areas covered by the true classes) are unknown, and must be estimated from the samples.

The overall purity is defined as the proportion of the mapped samples in which the predicted soil class, which is the soil class as depicted on the map, equals the true soil class as determined on validation points. In other words, it is the proportion correctly classified:

$$p=\sum_{u=1}^{U}\frac{A_{UU}}{A}$$

Where U denotes the number of classes, A_{UU} denotes the number of correctly classified observations of map unit *u* and A denotes the total number of observations in the study area. A good map has a value for map purity close to 1 (Finke 2011).

Table 2: Confusion matrix

		Field				
		1	2	...	U	$\sum$
	1	A_{11}	A_{12}	...	A_{1U}	A_{1+}
	2	A_{21}	A_{22}	...	A_{2U}	A_{2+}
Map	.	.	.	...	.	.
	.	.	.	...	.	.
	.	.	.	...	.	.
	U	A_{U1}	A_{U2}	...	A_{UU}	A_{U+}
	$\sum$	A_{+1}	A_{+2}	...	A_{+U}	A

A_{ij} = number of observations mapped as class C_i with observed soil class C_j

3.6. Soil diversity indices

The diversity of a map indicates the amount of information depicted on the map: a high diversity corresponds to high information content. In this research, three pedodiversity indices including Shannon's entropy H', richness S and evenness were calculated for each predicted map.

- Richness (S): is the number of soil classes that exists in an area.
- Shannon's entropy: is the most commonly used measurement of pedodiversity (Ibáñez, De-Alba et al. 1998; Guo, Gong et al. 2003). Where $p_i$is the proportion of area found in *i*-th unit over the total area of the map. $H=-\sum_{i=1}^{S}p_i\times\ln p_i$
- Evenness (E) refers to the relative abundance of each soil class in the area.

$$E=\frac{H'}{H_{\max}}=\frac{H'}{\ln S}$$

If each soil class is equally abundant, the evenness has high value and inversely, an area in which the abundance of soil classes differ greatly has low evenness (McBratney and Minasny 2007).

3.7. Combined Index practical management

In terms of management practices, the goal of soil mapping is to construct a map with high purity that adequately represents soil diversity. Therefore, the combination of map purity and Shannon's entropy is an important index to assess the soil mapping's performance. The combined index for accuracy and depicted diversity was defined by multiplying H' and map purity.

4. RESULTS

4.1. The soil maps modeled by multinomial logistic regression

Multinomial logistic regression predicts the soil classes directly from the predictors. Figure 1 shows the occurrence of Reference Soil Group predicted by MLR. As can be seen from the map, Fluvisols is the dominant class over the area. This can be explained by the fact that Bac Ninh is located in Duong River. Acrisols, Arenosols and Plinthosols are predicted with a very limited proportion by MLR method. However, the model did not predict any Gleysols even though we have samples belong to this group, too. Acrisols occurs in high landscape position in the study area as compared to the topographic map (Figure 3), which is a good prediction of the model because this soil group is often associated with hilly or undulating topography in wet tropical climates (FAO 2001).

Figure 2 illustrates the distribution of the intermediate level of soil group predicted by MLR. The model predicted more detailed soil classes: 11 soil classes appear in the resultant map. Nevertheless, there is still no occurrence of Gleysols which lead to the missing information of the model similar to the soil group prediction.

Fluvisol1000000 – Fluvisols have low base saturation - occurs in most area of Bac Ninh. Generally this is the fertile alluvial soil, distributed over different types of terrain, but due to the long exploitation for cultivation without appropriate land treatment reduces the soil fertility. Fluvisols have high base saturation and fine texture (Fluvisol0100010) appears in both sides following the Duong river. This soil class has high fertility because the river annually deposits a certain amount of sediment to the area around it. The model also results in the distribution of Acrisol00000 over the hilly region but in a more extensive area as compared to the Reference Soil Group level. The prediction of the MLR model for other soil classes concerns very small area.

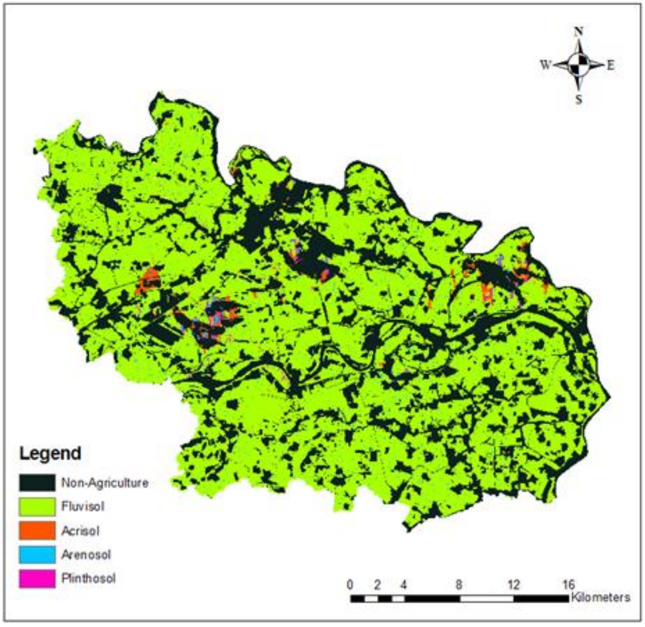

Figure 1: Map of Reference Soil Group predicted by MLR

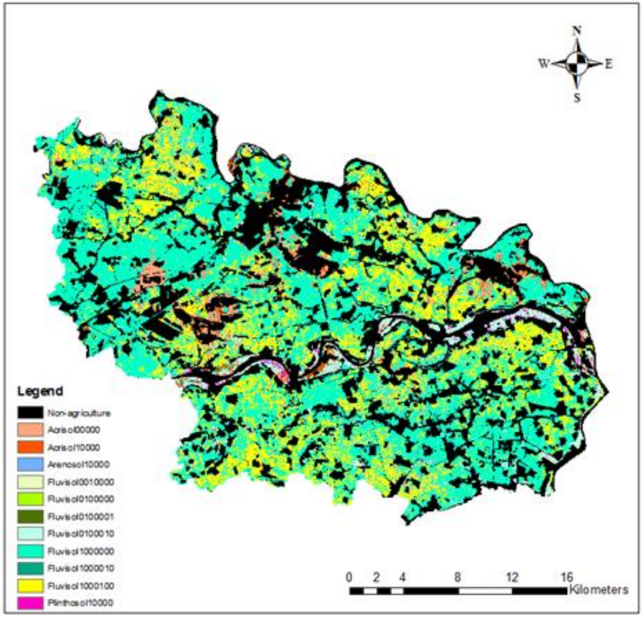

Figure 2: Map of intermediate level of Soil Group predicted by MLR

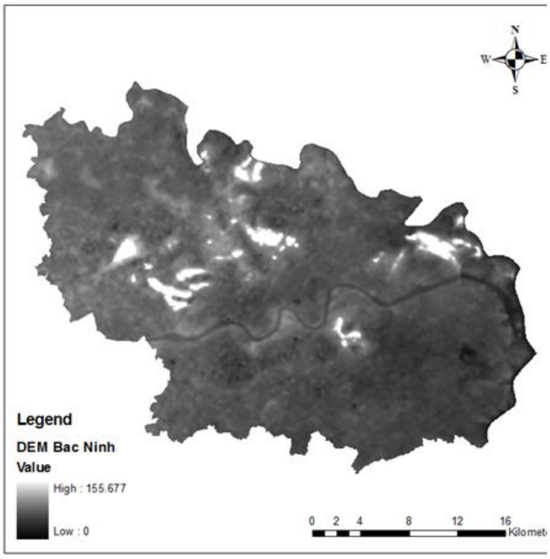

Figure 3: Digital Elevation Model of Bac Ninh

4.2. The soil maps modeled by artificial neural network

Figure 4 shows the map of Reference Soil Group constructed by ANNs model. Three out of the five Reference Soil Groups were predicted by the model: Fluvisols, Acrisols and Plinthosols. ANNs predicted Fluvisols in about 98% of the total area, which was also attributed to the unequal presence of the soil types in the observation data: more than 400 samples were Fluvisols in a 537 points dataset.

In terms of ANNs for predicting intermediate level of Soil groups (Figure 5), the model predicted six soil classes belong to the same Soil groups with higher level: Acrisols, Fluvisols and Plinthosols. Similarly, the soil classes belonging to both Gleysols and Arenosols were not classified by the model. This map shows similar pattern with the map produced by MLR: Fluvisols have low base saturation cover most of the area (78.8%), Fluvisols have high base saturation and fine texture located in both sides following the Red river, and Acrisols distribute in hilly regions.

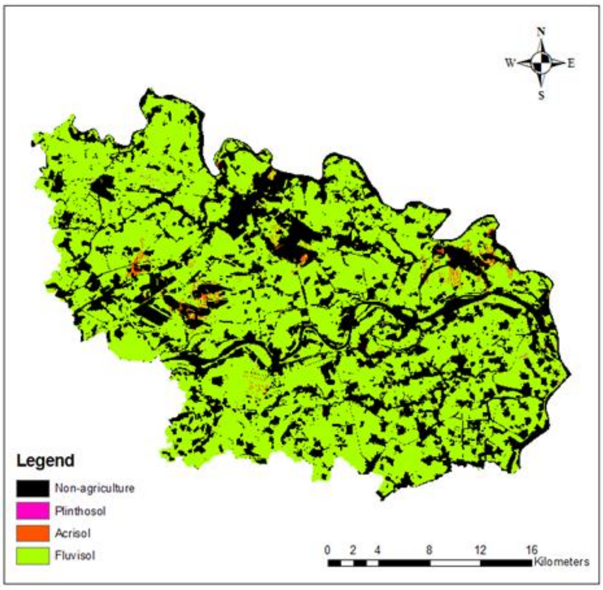

Figure 4: Map of Reference Soil Group predicted by ANNs

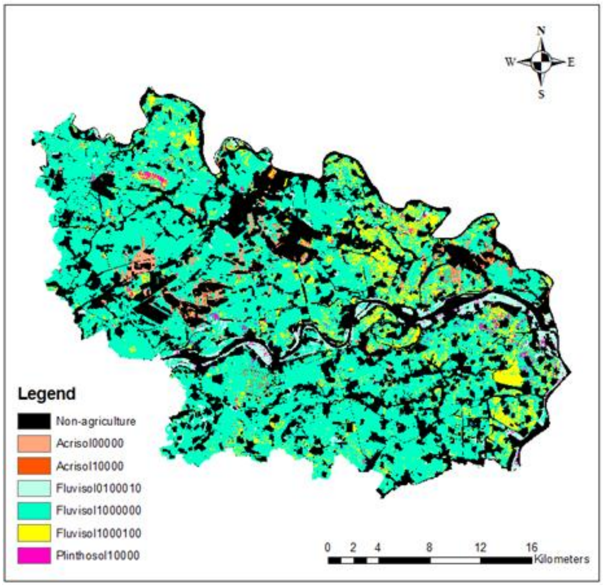

Figure 5: Map of intermediate level of Soil Group predicted by ANNs

4.3. Comparison of predictive methods

4.3.1. Soil map purity

Table 3 presents the estimated purity of the soil maps predicted by Multinomial Logistic Regression and Artificial Neural Networks at both levels. Both of the two methods get the same map purity value (0.73) at the high level of soil class. This indicates a good performance of both methods in predicting Reference Soil group.

As expected, in terms of lower level of soil class, the map purity drops dramatically to 0.39 and 0.37 for MLR and ANNs, respectively. MLR have slightly higher purity in predictive map than that of ANNs. Descending in the classified level introduces more properties that might be related to local conditions and natural selection, thus can lead to the complexity of the system (Toomanian, Jalalian et al. 2006). Therefore, some properties might not be included in the applied covariates and disconnection occurs between soil classes and covariates at lower level. Digital soil mapping relies on the relationships between soil samples and environmental factors of the target area. Weak relationships will result in weak prediction as seen in the performance of both methods at intermediate level of Soil groups. Jafari et al (2013) also found that soil map purity decreased toward the lower taxonomy category. Another reason is the low contrasting soil units at lower level. Olaniyan and Ogunkunle (2007) reported that soil mapping units with high purity included very contrasting soil types.

Table 3: Map purity, diversity indices and combined index of maps predicted by MLR and ANNs

	Level	Map purity	Richness	Shannon H'	Evenness E	Purity * Shannon
MLR	Reference Soil Group	0.73	5	0.20	0.12	0.15
	Intermediate level of Soil Group	0.39	15	1.21	0.44	0.41
ANNs	Reference Soil Group	0.73	5	0.07	0.04	0.05
	Intermediate level of Soil Group	0.37	15	0.77	0.28	0.29

4.3.2. Soil diversity

Table 3 shows the Richness, Shannon index and the Evenness of the resultant maps from two methods at both taxonomic levels of soil units. It is clear to see that with increasing number of soil units from the Reference Soil Group to the intermediate level, the diversity and the evenness rise sharply. The greater number of soil units correspond to the higher the diversity at the lower taxonomic level.

At the same taxonomic level, MLR always yields a higher value of the Shannon's index than ANNs. With the same Richness, the higher values of H' from MLR compared to that of ANNs indicate that higher soil diversity was MLR. The lower level of classification acquires a higher value of Shannon's index: 1.21 for MLR and 0.77 for ANNs. This could be attributed to the increasing number of soil map units at this level, thus induce the diversity of the predicted map. Similar with Reference Soil Group, the diversity is higher in maps made with MLR than with ANNs.

The diversity indices including richness, Shannon's index and evenness represent the deterministic soil complexity(Jafari, Ayoubi et al. 2013). For that reason, the increase of entropy in the study area from Reference Soil Group to lower level indicates higher complexity of the soil system. Besides, an increase in entropy associated with the larger number of different soil classes influences the prediction ability of the model. When the system complexity increases, there are more different soil classes in the area, thus the model should be trained for larger number of soil classes. It means that there are fewer observations per class for training of the model. This raises the uncertainty of the prediction for each soil classes and soil map purity decreases for the intermediate level of Soil groups. The soil diversity is a reflection of the intricacy of soil maps and may therefore influence the soil map purity(Minasny, McBratney et al. 2010).

The combined index defined by multiplying Shannon's entropy and map purity increases from the Reference Soil Group level to intermediate level in both MLR and ANNs approaches. However, MLR show higher value at both levels in comparison with ANNs as illustrated in table 3.

In terms of management practices, we need a soil map with high purity that adequately represents soil diversity. The pedodiversity measurements are related to the density of soil map or presence of various soil units (Jafari, Ayoubi et al. 2013). Soil mapping methods should acquire high map purity and also, it should represent the real soil diversity. In this research, although there are small differences in map purity between those two predictive methods, MLR shows higher pedodiversity at both mapping levels than ANNs does. Therefore, it seems that soil mapping will be more efficient by using Multinomial Logistic Regression than Artificial Neural Network. In MLR methods, the map purity at Reference Soil Group level is much higher than that value at intermediate level of Soil groups. Therefore, the model performs much better in predicting Soil groups. However, at lower level, the model predicts better diversity of the soil map and thus the informative value estimated by the combined index of the intermediate level maps is higher.

5. Conclusion

Some main conclusions can be drawn from the results of this study:

1. The Multinomial Logistic Regression could successfully be used to directly predict soil types map.

2. The soil map purity shows an opposite trend to that of the mapped soil diversity: as the purity decreases from Soil Groups to intermediate level of Soil groups, the soil diversity increases.

3. Based on the map purity and the combined index, Multinomial Logistic Regression performed better for predicting soil types than Artificial Neural Networks. Soil mapping at the level of Reference Soil Group acquires a high map purity and a low diversity.

4. To improve the model performance, more observations are needed for Acrisols, Plinthosols, Arenosol and especially Gleysols to avoid the abundance of Fluvisol over the dataset

References

1. Behrens., T., H. Förster, et al. (2005). "Digital soil mapping using artificial neural networks." Journal of Plant Nutrition and Soil Science **168**(1): 21-33.

2. Brus, D. J., B. Kempen, et al. (2011). "Sampling for validation of digital soil maps." European Journal of Soil Science **62**: 394–407.

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

4. FAO (2001). Lecture notes on the Major Soils of the World.

5. Finke, P. (2011). Syllabus for the course *Soil prospection and classification* in the *Physical Land Resources* program.

6. Fritsch, S. and F. Guenther (2012). Package "neuralnet". Training of neural network.

7. Guo, Y., P. Gong, et al. (2003). "Pedodiversity in the United States of America." Geoderma **117**: 99-115.

8. Ibáñez, J. J., S. De-Alba, et al. (1998). "Pedodiversity and global soil patterns at coarse scales (with discussion)." Geoderma **83**: 171-192.

9. Jafari, A., S. Ayoubi, et al. (2013). "Selection of taxonomic level for soil mapping using diversity and map purity indices: A case study from an Iranian arid region." Geomorphology.

10. Jafari, A., P. Finke, et al. (2011). "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.

11. Kempen, B., D. J. Brus, et al. (2009). "Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach." Geoderma **151**: 311-326.

12. McBratney, A. B. and B. Minasny (2007). "On measuring pedodiversity." Geoderma **141**: 149-154.

13. Minasny, B., A. B. McBratney, et al. (2010). "Global pedodiversity, taxonomic distance, and the World Reference Base." Geoderma **155**: 132-139.

14. Olaya, V. F., Ed. (2004). A gentle introduction to Saga GIS Gottingen, Germany.

15. Resende, R. J. T. P. (2000). Characterizations of the Physical Environment of Coffee Areas of the South of Minas Through SPRING, University of Lavras, UFLA, MG, Brazil

16. Toomanian, N., A. Jalalian, et al. (2006). "Pedodiversity and pedogenesis in Zayandeh-rud Valley, Central Iran." Geomorphology **81**: 376-393.

17. White, R. E., Ed. (2005). Principles and Practice of Soil Science: The Soil as a Natural Resource, Wiley-Blackwell.

18. WRB, I. W. G. (2006). World Refernce Base for soil resources 2006. Rome, FAO.

Yang, L., Y. Jiao, et al. (2011). "Updating Conventional Soil Maps through Digital Soil Mapping." Soil Science Society of America **75**(3): 1044-1053.