COMPARISON OF THREE MODELS FOR PREDICTING THE SPATIAL DISTRIBUTION OF SOIL ORGANIC CARBON IN BOALEMO REGENCY, SULAWESI

Eloise Mason[1] and Yiyi Sulaeman[2]*

[1] National Engineering School of Agronomy and Food Sciences, Nancy, France
[2] Indonesian Agency for Agricultural Land Resource Research and Development, Agency for Agricultural Research and Development, Jl. Tentara Pelajar 12, Cimanggu, Bogor, West Java, Indonesia

ABSTRACT

Information on the spatial distribution of soil organic carbon content is required for sustainable land management. But, creating this map is time consuming and costly. Digital soil mapping methodology make use legacy soil data to create provisional soil organic carbon map. This map helps soil surveyors in allocating next soil observation. This study aimed: (i) to develop predictive statistical soil organic carbon models for Sulawesi, and (ii) to evaluate the best model between the three obtained models. Boalemo Regeny in Gorontalo Province (Sulawesi) was selected as studying area due to abundant legacy soil data. The study covered dataset preparation, model development, and model comparison. Dataset of soil organic carbon at 6 different depths as target was established from 176 soil profiles and 7 terrain parameters were selected as predictors. Soil-landscape models for each soil depth were created using regression tree, conditional inference tree, and multiple linear regression technique. Result showed that model performance differed among 3 modelling techniques and soil depths. The tree models were better than the multiple linear regression model as they have the lowest RMSE index. The best model in the mountanious area seems to be the regression tree model, whereas in the plains it may be the conditional inference tree. In creating provisional map, several model should be developed and the median of predicted value is used as provisional map.

Keywords: Digital soil mapping, multiple linear regression, regression tree, soil-landscape model, soil organic carbon map

INTRODUCTION

Sulawesi is a fertile land and is a cacao and rice national production centre. For sustainable land management, it is very important to assess the spatial distribution of soil organic carbon (SOC) content. However, data in Sulawesi are limited and difficult to be obtained, especially in the mountainous area. It is therefore useful to develop other technique to assess useful soil organic carbon data. Using digital soil mapping is a solution to improve soil mapping in poorly accessible areas (Moore *et al*., 1991; Florinsky, 1998). Lagarcherie and McBratney (2007) defines digital soil mapping as "the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables". The statistical analysis is used to create predictive models of soil properties, thus requiring less human intervention than traditional soil mapping techniques. Therefore, digital soil mapping is cost-efficient and time-saving.

Nowadays the main source of data for soil prediction is digital elevation model (DEM) created from radar data. DEM-based prediction of SOC content relies on finding relationships between SOC and environmental variables to build statistical models (Odeh *et al*., 1994; Gessler *et al*., 1995; Thompson *et al*., 1997; Arrouays *et al*., 1998). To describe this relationship, seven different environmental factors can be used (McBratney *et al*., 2003). They can all be decomposed into separate layers and mapped separately. Based on Jenny's soil-forming factors (1941), these factors are: s (soil, other or previously measured attributes of the soil at a point), c (climate), o (organisms, vegetation or human activity or fauna), r (topography, including terrain attributes and classes), p (parent material), a (age or elapsed time), n (space, spatial or geographic position). In this study, only the relief factor will be used, since the spatial distribution of SOC is mainly affected by topography (Florinsky, 2002), especially in upper layer.

Many researchers develop techniques for mapping SOC or soil organic matter (SOM). Kempen *et al.* (2011) present technique to integrate pedological knowledge and geostatistical mapping using soil map and soil profile as input. Kumar *et al.* (2012) used geostatistical hybrid approach (geographically weighted regression kriging and regression kriging) for mapping SOC stock using elevation, slope gradient, precipitation and temperature as input. They showed that geographically weighted regression kriging was better technique for estimating SOC stock across regional scale.

In Indonesia, Sulaeman *et al.* (2012) proposed general framework for applying digital soil mapping using legacy data. This includes 3 main steps i.e. (i) dataset preparation by collecting previous soil data and auxiliary information, (ii) soil-landscape model development, and (iii) model application to derive digital soil properties map.

The objectives of this study are: (i) to develop predictive statistical SOC models for Sulawesi, and (ii) to evaluate the best model between the three obtained models.

This research is developed within the context of limited data while provide baseline soil-landscape models that could be improve in the future. Soil-landscape models are important input for deriving provision soil map that can assist in allocating new observation and zoning recomendation domain, especially in Indonesia.

*) *Penulis Korespondensi: Telp. +62251-8323012; Email. y.sulaeman@gmail.com*

MATERIAL AND METHODS

Study Area

Our study area is Boalemo Regency in Gorontalo Province (Figure 1). The regency is located in Northern Sulawesi Province, at longitude of 122°15′ E and latitudes of 0°42′ N with a total land area of about 2,567 km^2. Boalemo is a good experimental study site as its landscape is much diversified. It has coastal areas as well as mountainous areas. Elevation goes from 0 to 1,870 meters above sea level. Soil profiles have been studied in order to characterize the SOC.

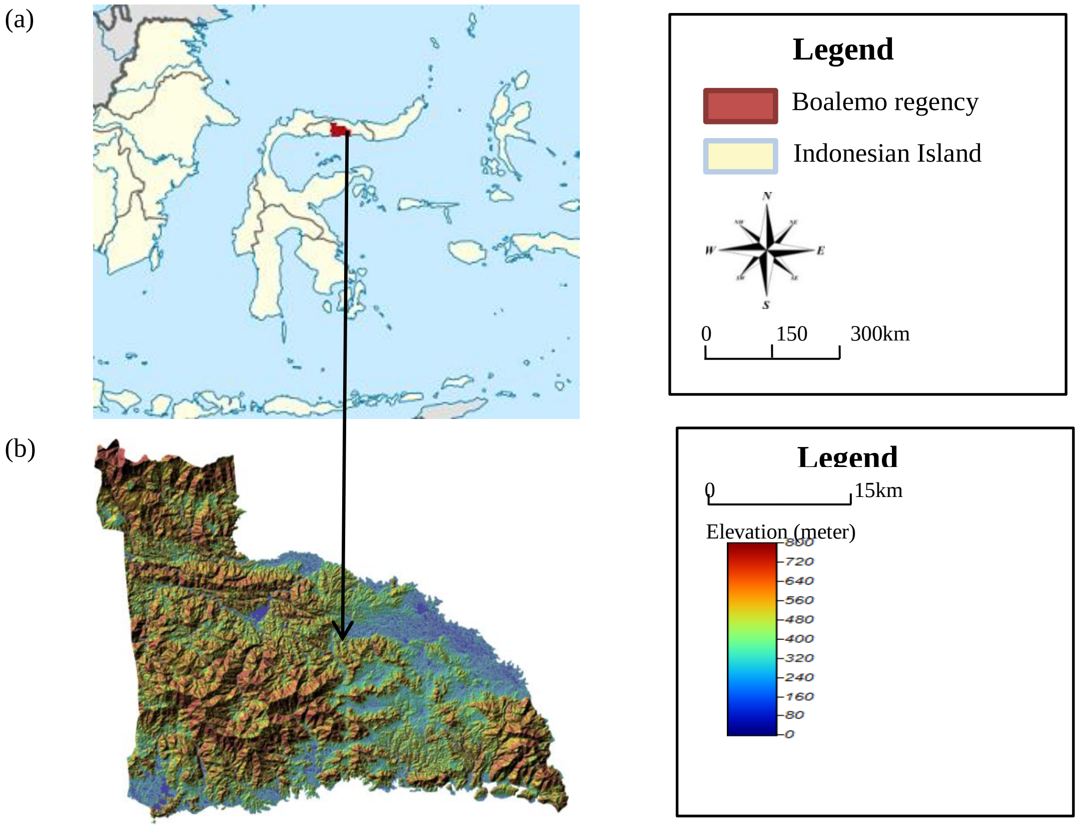


Figure 1. Study area on (a) Sulawesi and (b) on an elevation map extracted from DEM with SAGA

Soil-Landscape Dataset Preparation

Dataset preparation includes: (i) collection of soil data and environmental data; (ii) data harmonization; (iii) deriving terrain parameter from DEM; and (iv) data integration.

(1) collection of soil data and environmental data: the soil data can be in the form of existing soil maps with legends and/or soil observations, which are complete soil profile description or soil laboratory test data with clear geographical position.

(2) data harmonization: some soil observations come from different projects, and were created at different time by different surveyors. Geo-reference system and soil depth observations will so, also be different. All the data was harmonized using spline-tool software.

(3) deriving terrain parameter from DEM: the digital elevation model was used to derive seven relief variables (from SAGA GIS version 2.0.8), which were slope gradient, plan curvature, profile curvature, landforms, hill-shading, wetness index and elevation. The topographic variable’s description is presented in Table 1 (Tesfa *et al*., 2009). Each of these hydrologic or landform parameters are useful as they can be predictive variable because the relief has a great influence on soil formation.

(4) data integration : the covariates of the relief factors derived from the DEM are joined to the soil organic carbon data set for each soil data point.

Table 1. Description of topographic variables derived from DEM

Relief variables	Description
Elevation	Elevation above sea level. It classifies the local relief.
Hill-shading	Angle between the surface and the incoming light beams.
Landforms	Elevation of one cell compared to the elevation of the 8 adjacent cells.
Plan curvature (kh)	Curvature of the surface perpendicular to the direction of the maximum slope. A positive value indicates divergence of the substance flows; a negative value indicates convergence of the substance flows.
Profile curvature (kv)	Curvature of the surface in the direction of maximum slope. A negative value indicates a deceleration of the flow;a positive value indicates flow acceleration.
Slope gradient	Angle of inclination to the horizontal at a given point on the land surface. It affects the velocity of the surface and subsurface flow
Saga Wetness Index	Catchment area calculation, which does not consider the flow as very thin film.

Equal-Area Spline Algorithm

This study employed the equal-area spline algorithm used by Malone *et al*. (2009) and Odgers *et al*. (2012), which is a generalization of the algorithm developed by Bishopet *et al*. (1999). To generate a spline, two pieces of information are required:

i. Soil property values for a number of layers in the soil profile. In practice, the layers are usually soil horizons; the soil property values are assumed to represent the bulk mean value of the soil property across the layer.
ii. The upper and lower boundaries of the input layers. The layers do not need to be contiguous with depth (i.e., 0–7 cm, 10–20 cm) but they should not overlap.

The equal-area spline consists of a series of quadratic polynomials fitted piecewise through the input layers and is linear between layers. We used Spline Tool version 2 (http://www.asris.csiro.au\downloads\GSM\SplineTool_v2.zip) to calculate soil organic matter for 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm, 100-200 cm as Global Soil Map specifications (Global Soil Map, 2011).

Soil-Landscape Model Development

The dataset containing SOC as dependent variable and covariate as predictors was obtained from 176 soil profile sites. R software is used to perform statistical analysis on soil data to obtain three predictive models of SOC: regression tree, conditional inference tree, and multiple linear regression. The three different models will be compared and the best model will be selected to predict SOC content for each of the following depth: 0-5 cm (OC1); 5-15 cm (OC2) ; 15-30 cm (OC3) ; 30-60 cm (OC4) ; 60-100 cm (OC5) ; 100-200 cm (OC6) ; 0-30 cm (OC7).

Preliminary Data Treatment

The data needs to be normally distributed to fit statistical models. A plot of the data is obtained to determine the normality (Figure 2). Currently, it is not normal (Figure 2.a), so we have to force it to be normal using log (Figure 2.b) before starting to fit the models. The output values will also be in natural logarithm.

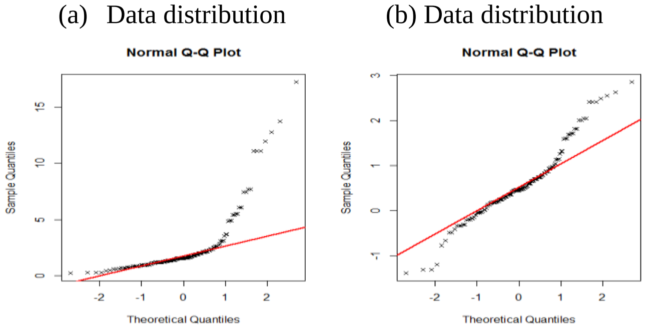

(a) Data distribution (b) Data distribution

Figure 2. Data distribution when (a) not using log, (b) using log

Regression Tree (RT) Model

Regression trees can help the prediction of SOC content at different depth. The success of this model relies on its ability to deal with non-linearity; which is useful because the interactions between the SOC content (a response variable) and topographic variables are often conditional on other variables. The stronger the relationship between soil properties and available environmental variables is, the stronger the model will be. The data is splitted into nodes in a binary way that creates the tree structure. The split ends when the node becomes too homogenous or when the number of observations is too little. The regression tree model was fitted using the **rpart** function in the RPART package for R.

To avoid obtaining tree with over fitted data, a suitable size tree with a minimized cross-validated error has to be carefully selected. The complexity parameter (Cp) associated with the smallest cross-validated error is used to prune back the tree to a more suitable size. The most appropriate Cp considering the **xerror** can be obtained using a test in R software. The result of a test is shown in Table 2 and describes the relationship between different possible complexity parameter and their corresponding **xerror**.

Conditional Inference Tree (CIT) Model

Conditional inference trees are trees created by binary splits and early stops. Tree growth is based on statistical stopping rules; therefore, we do not focus on pruning as we did with the regression tree model. The early stop overcomes the over-fitting problem of trees. To build the tree, first is set the global null hypothesis as the independence between the SOC content response and the environmental input variables. The hypothesis is tested and the algorithm stops if the hypothesis is not rejected. If rejected, the environmental variable with the strongest association to SOC content is selected.

Table 2. Complexity parameter and their corresponding xerror for the prediction of log OC7

Run	CP	nsplit	rel error	Xerror	xstd
1	0.1083	0	1.000	1.010	0.154
2	0.0507	1	0.891	1.163	0.179
3	0.0377	2	0.841	1.135	0.165
4	0.0267	4	0.765	1.107	0.158
5	0.0150	5	0.738	1.149	0.164
6	0.0114	6	0.724	1.163	0.158
7	0.0106	8	0.701	1.193	0.159
8	0.0103	9	0.690	1.196	0.157
9	0.0088	10	0.680	1.198	0.156
10	0.0020	11	0.671	1.244	0.160

Note: Here, the best Cp is 0.027, as the xerror = 1.107 is the smallest (omitting xerror when nsplit=0 because a tree without any splits is created)

The data is then partitioned into a binary split with the best split point for the selected variable. These steps are repeated for the new partitions until the hypothesis can not be rejected. The conditional inference tree model was fit using the ctree function in the PARTY package for R.

Multiple Linear Regression (MLR) Model

The output result of the MLR is an equation with SOC content as the response variable and the relief parameters as factors. The 7 different relief parameters (Table 1) are considered as independent variables.

Model Comparison

These 3 different models were built and statistically compared. The choice of the best model is based on the estimation of the standard deviation of the residual error (root mean square error, RMSE). The model with the smallest RMSE is considered as the best for the prediction of SOC. The RMSE value represents the sample standard deviation of the differences between predicted values and observed values. RMSE is a good way to measure accuracy, but only to compare errors between different models. The RMSE value is defined as the square root of the mean square error. The output results from each model is then exported and compiled in the SAGA software to create predictive maps.

RESULTS AND DISCUSSION

Results from Models

For each soil depth, a model is set up separately using the three modelling techniques. All of the results values are reported in natural logarithm. The statistics showed different equation to predict SOC content. Each predictive model equation has its own algorithm with specific parameter coefficients. Not all parameters are taken into account for the prediction of SOC content. Multiple linear regression analysis establishes a functional relationship between SOC data and all derived parameters. The following equation describes SOC content at depth 0-5cm (log OC1) :

$$\log OC1 = 0.328629 - 0.00119 * elevation + 0.033028 * hillshade - 0.101734 * landforms + 230.25 * kh + 131.21 * kv + 2.19 * slope + 0.029 * wetness\ index$$

All derived parameters are used in this equation. The SOC distribution patterns should be therefore more detailed. As for the regression tree and conditional inference treemodels, the number of input factors used in the tree depends on the depth (Table 3). The output results for the RT and CIT model are both trees (Figure 3.) from which algorithm can be extracted.

Accuracy Assessment of Model

Once all three models are obtained, it is useful to determine if the model are accurate. Towards that purpose, the result of the performance of RMSE and the number of parameters (variables) used in each predictive tested model are compared. Table 3 shows the values for these accuracy indicators.

Table 3 indicates that the RMSE is always higher in the MLR model than the two tree models, whatever depth. This means that the MLR is less accurate and does not seem to be the best model. The RT model and CIT model both have a similar lower RMSE. It is not possible to distinguish these two models when looking at the accuracy. The differentiation may be done visually (Figure 4).

Comparison of Models Maps

It seems interesting to focuse on one particular soil depth (0-5cm) and compare visualy the maps representing the different models between them for that same area and soil depth. Some models are much more detailed and precise than others (Figure 4 and 5). The spatial distribution patterns of SOC content obtained with the CIT model (Figure 4 and 5) is quite similar to the other two models, but doesn't seem as precise when zooming in. Some local variations of SOC data are not detected by this method. Only four parameters are used to describe the conditional inference tree model. Whereas all seven relief parameters are taken into account for the multiple linear regression. Indeed, visually, the MLR map is very detailed.

When comparing the RT and CIT map, it seems that the RT model map is surprisingly more detailed than the CIT one. The two model have both four parameters in their respective algorithm to predict OC1. But one of the parameters differ from a model to another. The elevation factor is taken into account in the regression tree algorithm, whereas and in the CIT algorithm it is the wetness index. This explains why in plains, the CIT map is more detailed in than in the mountainous area (Figure 4.a).

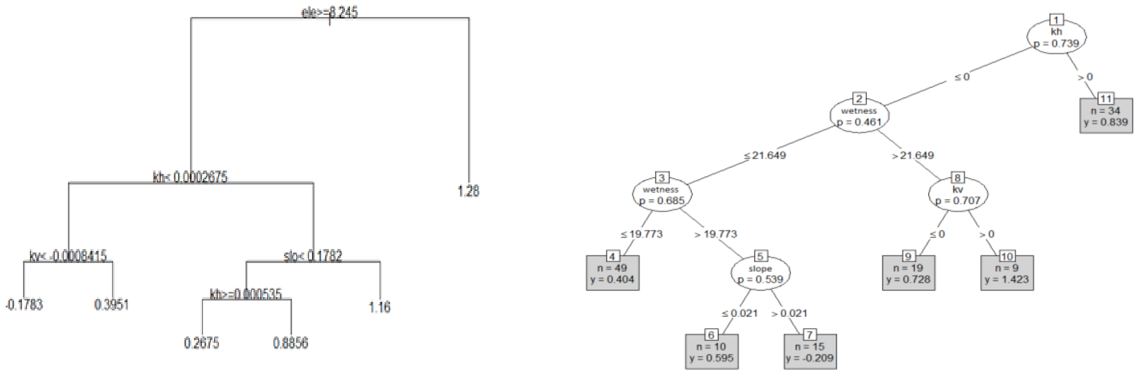


Figure 3. SOC conten at top soil (log OC1) (a) RT model and (b) CIT model obtained in R

Table 3. Comparison of the performance of each predictive model for the different depth

	model[ϕ]	R^2	RMSE	Input number	Input factors[ψ]
OC1	M1	0.240	0.694	4	a/d/e/f
	M2	0.047	0.760	7	a/b/c/d/e/f/g
	M3	0.230	0.699	4	d/e/f/g
OC2	M1	0.260	0.650	3	a/d/e
	M2	0.053	0.740	7	a/b/c/d/e/f/g
	M3	0.260	0.650	4	b/d/e/g
OC3	M1	0.220	0.650	2	a/e
	M2	0.046	0.730	7	a/b/c/d/e/f/g
	M3	0.240	0.640	4	b/e/f/g
OC4	M1	0.210	0.750	4	a/d/e/g
	M2	0.049	0.890	7	a/b/c/d/e/f/g
	M3	0.210	0.750	3	e/f/g
OC5	M1	0.220	0.930	4	a/b/d/g
	M2	0.063	1.017	7	a/b/c/d/e/f/g
	M3	0.250	0.910	3	e/f/g
OC6	M1	0.230	1.080	5	a/d/e/f/g
	M2	0.062	1.200	7	a/b/c/d/e/f/g
	M3	0.220	1.100	3	e/f/g

[ϕ] M1 = RT model, M2 = MLR model, M3 = CIT model

[ψ] a = elevation ; b = hillshade ; c = landforms ; d = plan curvature ; e = profile curvature ; f = slope ;g = wetness index

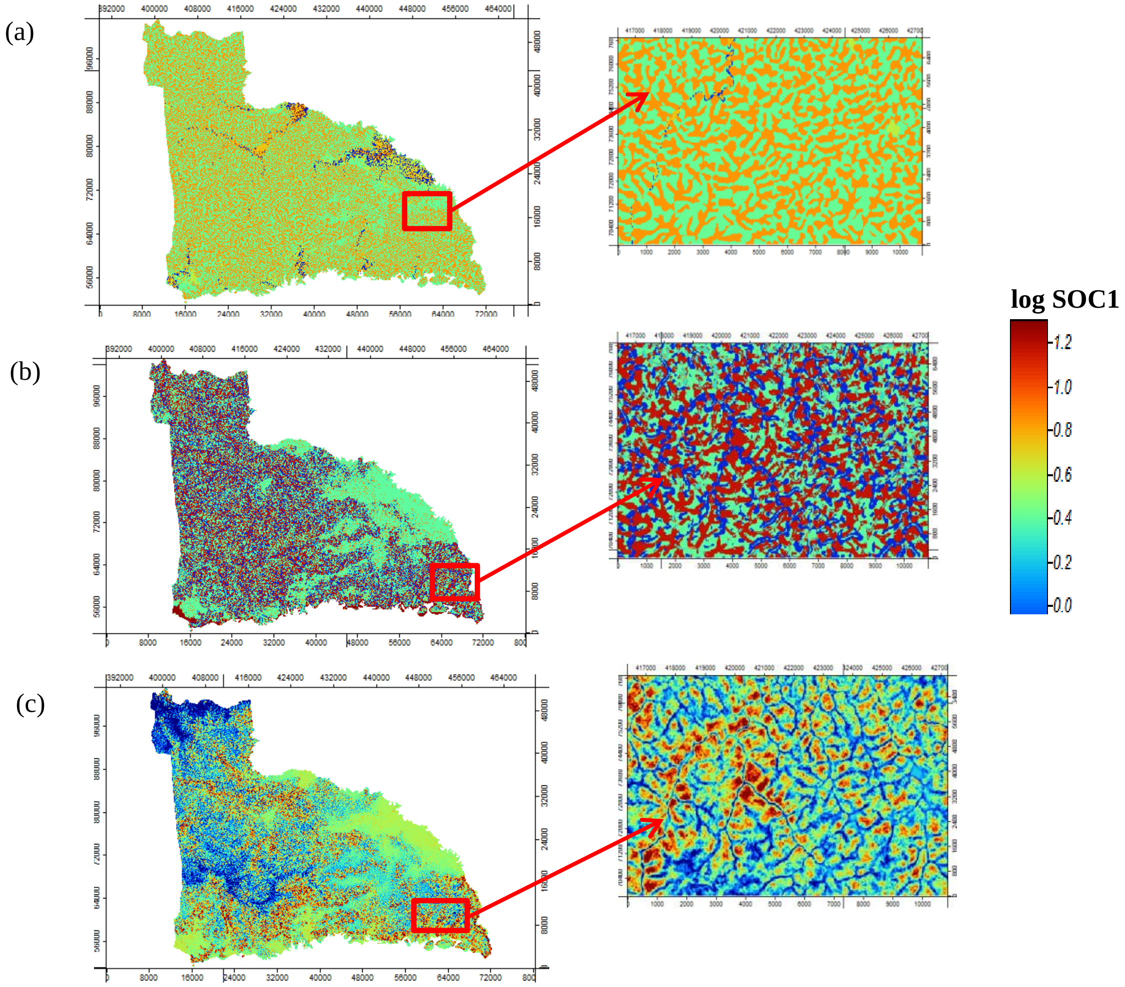


Figure 4. Spatial distribution of SOC content in Boalemo Regency with detailed zoom at depth 0-5cm with (a) CIT model, (b) RT model, and (c) MLR model

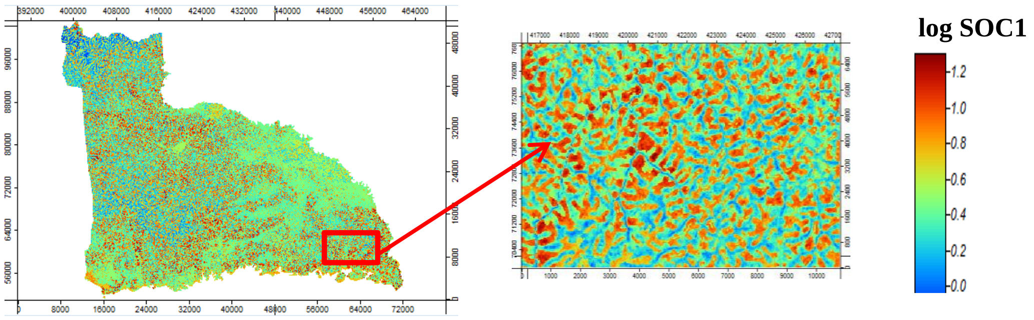


Figure 5. Spatial distribution of SOC content in Boalemo Regency with detailed zoom at depth 0-5cm with mean of the three models

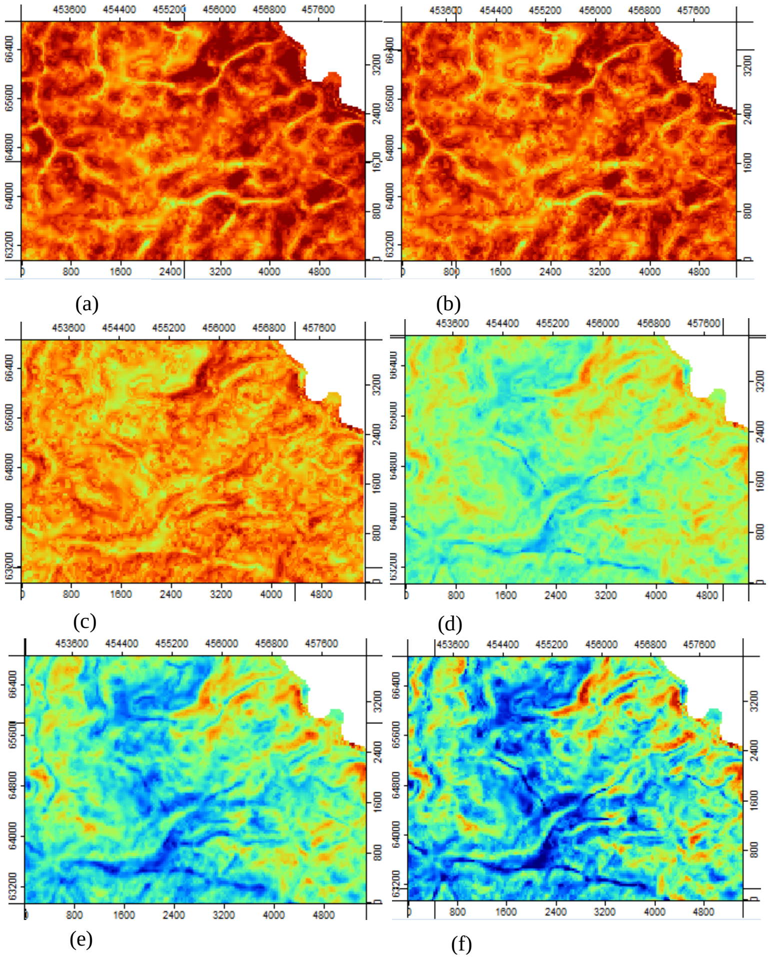
It is the opposite in the RT map. The areas with relief are more detailed than in the flat area (Figure 4.b). Therefore, when studying plains, the conditional inference model seems more suitable than the regression tree model, more suited for area with higher elevation.

Role of the Depth on SOC Value

As said before, the multiple regression linear model is the most detailed model. All derived relief parameters are exploited in the model equation. Therefore it seems the best model to obtain a detailed analyses of the SOC content at different depth. The interpretation is done visualy but also by comparing the RMSE for each depth.

Figure 6 shows the spatial distribution of SOC content at different depth. Visualy (Figure 6), it looks clear that the deeper the soil, the less SOC content is available. For example, at depth 15-30 cm (log OC3), organic carbon values range from 0.0 to 0.95. As at depth 100-200 cm (log OC6), values goes from 0.0 to 0.68. Every area presents a diminution of its SOC content with depth. The areas with the lowest SOC value at deep soil profile also have the lowest value for the top soil.

The SOC content distribution patterns do not vary with depth. It is always the same areas with the highest OC1 value whatever depth. As we go along deeper in the soil, the accuracy measured by the RMSE (Table 3) does not vary from depth 0-30 cm, but raises from depth 30-60 cm. It seems logical, the deeper, the less the topography has an influence on the SOC content.



(a) (b) (c) (d) (e) (f)

Figure 6. Spatial distribution of SOC at one random area from the study site obtained with the MLR model at different depth (a) log OC1, (b) log OC2, (c) log OC3, (d) log OC4, (e) log OC5, (f) log OC6

CONCLUSION

DEM and its topographic derivates were used for the construction of different models of RT, CIT and MLR. Three simple predictive models were generated and implemented with SAGA software, to predict organic carbon content in the soil at depth 0-200 cm, at unobserved locations, in Boalemo regency. The RT and CIT models are better than MLR model as they have the lowest RMSE index. The best model in the mountainous area seems to be the RT model, whereas in the plains it may be the CIT tree.

Further studies are to be done, to test if our selected best models using the lowest RMSE index and visual comparison are really the best predictions. To test the accuracy of these models, the soil organic carbon matter content from the field at unobserved locations will be compared to the values from our predictive models. The models will have to be improved if our predictive data doesn't correspond to the data from the field.

ACKNOWLEDGEMENT

This study was funded by Indonesian Agency for Agricultural Research and Development (IAARD), Ministry of Agriculture. We thank Ms. Fitri Widiastuti and Mr. Jaelani of ICALLRD who assisted author in data collection and data preparation.

REFERENCES

Arrouays, D., J. Daroussin, J.L. Kicin, and P. Hassika. 1998. Improving topsoil carbon storage prediction using a digital elevation model in temperate forest soils of France. *Soil Science*, 163: 103–108.

Bishop, T.F.A., A.B. McBratney, and G.M. Laslett. 1999. Modelling soil attribute depthfunctions with equal-area quadratic smoothing splines. *Geoderma*, 91: 27–45. http://dx.doi.org/10.1016/S0016-7061(99)00003-8.

Florinsky, I.V. 1998. Combined analysis of digital terrain models and remotely sensed data in landscape investigations. *Progress in Physical Geography*, 22: 33–60.

Florinsky, I.V., R.G. Eilers, G. Manning, and L.G. Fuller. 2002. Prediction of soil properties with digital terrain modelling. *Environmental Modelling and Software*, 17 (in press).

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems*, 9: 421–432.

Global Soil Map. 2011. Specifications, Version 1 Global Soil Map.net products. Release 2.1.

Jenny, H. 1941. *Factors of Soil Formation*. McGraw-Hill, New York, NY.

Kempen, B., D.J. Brus, J.J. Stoorvogel. 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma*, 162: 107-123.

Kumar, S., R. Lal, and D. Liu. 2012. A geographically weighted regression krigging approach for mapping soil carbon stock. *Geoderma*, 189-190: 627-634.

Lagacherie, P., A.B. McBratney, and M. Voltz. 2007. *Digital Soil Mapping – An Introductory Perspective*. Developments in Soil Science, 31. Elsevier, Amsterdam.

Malone, B.P., A.B. McBratney, B. Minasny, and G.M. Laslett. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154: 138–152. http://dx.doi.org/ 10.1016/ j.geoderma.2009.10.007.

McBratney, A.B., M. Santos, and Minasny. 2003. On digital soil mapping. *Geoderma*, 117: 3–52.

Moore, I.D., P.E. Gessler, G.A. Nielsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57: 443-452.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, 63: 197–214.

Odgers, N.P., Z. Libohova, and J.A. Thompson. 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. *Geoderma*, 189–190: 153–163.

Sulaeman, Y., Hikmatullah, D.A. Suriadikarta, M. Sarwani, A. Sutandi, dan B. Barus. 2012. Aplikasi pemetaan tanah digital untuk pemetaan sifat tanah menunjang rekomendasi pemupukan. *In* Wigena *et al*. (eds.) Prosiding Seminar Nasional Teknologi Pemupukan dan Pemulihan Lahan Terdegradasi Bogor, 29-30 Juni 2012. Balitbangtan, Bogor, pp 73-82.

Sulaeman, Y., B. Minasny, A.B. McBratney, and M. Sarwani. 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. *Geoderma*, 192: 77-85.

Tesfa, T.K., D.G. Tarboton, D.G. Chandler, and J.P. McNamara. 2009. Modeling soil depth from topographic and land cover attributes. *Water Resour. Res.*, 45: W10438. doi:10.1029/ 2008WR007474.

Thompson, J.A., J.C. Bell, C.A. Butler. 1997. Quantitative soil-landscape modeling for estimating the areal extent of hydromorphic soils. *Soil Science Society of America Journal*, 61: 971–980.