Modelling of silt content using geographically weighted regression

H Pramoedyo1*, S Riza2, D Ardianti1 and A Oktaviarina3,4

1 Department of Statistics, Brawijaya University, Malang, Indonesia
2 Department of Soil Science, Brawijaya University, Malang, Indonesia
3 Department of Mathematics, Brawijaya University, Malang, Indonesia
4 Department of Mathematics, Surabaya State University, Surabaya, Indonesia

* hennyp@ub.ac.id

Abstract. Multiple linear regression is a method used to model or predict an object that sees the relationship between a dependent variable and a group of independent variables. Geographically Weighted Regression (GWR) is the development of multiple linear regression involving geographical factors. In this study, both methods were used in the study to analyze one of the soil elements, namely the silt soil texture. Through the Digital Elevation Model (DEM) data, the topographic variables used in the study are Eastness Aspects (Ae), Northness Aspects (An), Slope (S), UnspHERicity Curvature (M), Vertical Curvature (Kv), Horizontal Curvature (Kh), Accumulation Curvature (Ka) and Elevation (Elv). The results showed that the GWR model with fixed Gaussian weighting better than the multiple linear regression model. $R^2$ value of GWR was 57%, greater than the multiple regression, which was 55%. And the SSE of GWR and multiple regression value were 2014.69 and 2177.19 respectively.

1. Introduction

Extensive soil mapping is needed as the main information in land resources management [1]. Several studies successfully attempted to interpolate soil mapping soil properties [2], indicating the importance of developing soil properties modelling. Fine-textured soils, including the silt content, are essentials soil properties that know its importance in the physical protection of organic matter and thus one of the keys in the land management system [3]. Silt and the others soil particles, is influenced by topographic variability which modifies water flow and material distribution thus produce a soil pattern in a landscape [4]. Therefore, the relationship among silt content and those topographic properties is a base of this study.

Spatial statistics are growing rapidly in line with the needs of the time and widely used in various fields. One of method in spatial statistics is geographically weighted regression (GWR). GWR is based on non-parametric technique of locally weighted regression developed in statistics for curve fitting and smoothing application [5]. In GWR, a regression model can be fitted at each observation location in the data set [6]. The objective of this study is to predict the silt content through the GWR modelling.

2. Methods

The topsoil of 0-10 cm depth based on 50 samples was taken randomly on Kalikonto watershed, Malang, in June-July 2020. Silt content was then derived from the laboratory analysis was used as the primary
Data in this study. Variables that used in this study are eastness aspects (Ae) as X1, northness aspects (An) as X2, slope (S) as X3, unsphericity curvature (M) as X4, vertical curvature (Kv) as X5, horizontal curvature (Kh) as X6, accumulation curvature (Ka) as X7, and elevation (Elv) as X8 and Silt soil texture as Y. Before obtaining the geographically weighted regression model, some procedure is done by checking spatial aspects, create multiple linear regression, determining optimal spatial weights, build GWR model, estimating GWR parameters and performing the partial test.

3. Results and discussion
It will be investigated whether the data has spatial aspects. The first spatial test to be performed is Morans I test. The test is used to determine the spatial dependence on data. The origin of this test is standardizes the variables by subtracting the sample mean, and then deflating by an appropriate factor [7].

| Morans I Test | p-value |
|---------------|---------|
| Morans I      | 0.0000  |

The p-value is less than 0.5, as shown in the table 1 above. This means that there is a spatial dependence on the data.

Breuch Pagan is the second spatial test used to investigate whether the data has spatial heterogeneity [8]. P-value result from the test is less than 0.5, so that there is heterogeneity in the data. It can be seen in table 2 below:

| Breuch Pagan Test |
|-------------------|
| Breuch Pagan      | P-Value       |
| Shapiro Wilk      | 27,709        |
|                   | 0.00005328    |

It can be concluded that data fulfill the spatial aspects since data has spatial dependency and spatial heterogeneity.

In this research, fixed gaussian spatial weight was compared with adaptive bisquare spatial weight. The smaller CV was used to build GWR model.

| Weight Spatial |
|----------------|
| Bandwith       | CV minimum    |
| Fixed gaussian | 22043,220     |
| Adaptive bisquare | 0,677        |
|                 | 3767,201      |
|                 | 3798,31       |

The smaller CV minimum was Fixed Gaussian as seen in Table 3. So, Fixed Gaussian weight spatial was used to make GWR model.

| Parameter GWR Estimation |
|--------------------------|
| Parameter               | Minimum | Maximum | Global |
|                          |        |         |        |
| β intercept              | 27,5163| 27,7998 | 27,500 |
| β1                       | 1,14747| 1,3448  | 1,1832 |
| β2                       | -0,4801| -0,2678 | -0,3682|
| β3                       | 2,9984 | 3,1486  | 3,0438 |
| β4                       | -4,4882| -3,9209 | -4,2890|
| β5                       | -5,5011| -4,9580 | -5,2969|
| β6                       | 1,5558 | 2,1242  | 1,8954 |
| β7                       | -4,1397| -4,0602 | -4,1780|
| β8                       | 7,0855 | 7,5849  | 7,2919 |
Table 4 shows estimation value of GWR parameter. $X_6$ produces the greatest impact on the model, with interval $7.0855 \leq \beta_6 \leq 7.5849$. The opposite condition, $X_5$ gives the smallest influence to the model with interval parameter $-5.5011 \leq \beta_5 \leq -4.9580$. The value of $R^2$ in GWR model is 0.57, it means that performance model in explaining variables is 57%.

In GWR, each location has a different parameter estimation of the GWR silt model. For example, the GWR silt model will be show at location 5 as follows:

$$\hat{Y}_{silt} = 27.562 + 1.257X_1 - 0.402X_2 + 3.095X_3 - 4.309X_4 - 5.316X_5 + 1.947X_6 - 4.082X_7 + 7.287X_8$$

Based on the model formed, there are 4 coefficients which are positive, which means that the silt element will increase if the variable also increases. If the variable $X_1$ increases by one percent, the number of silt elements will increase by 1.257 provided that the other variables are constant and the effect of the location around the observed point, namely point 5, is considered constant. This variable ($X_1$) show where the particles are distributed, generally affected by the flow direction or wind, which can deposit a particle in some areas. From this variable we can understand that the silt is deposited in the east area of the watershed. The number of silt content increases by 3.095 if the $X_3$ variable increases by one percent provided that other variables are constant and the effect of the location around the observed point, namely point 5, is considered constant. The slope variable ($X_3$) is represent that silt distribution is affected by the surface erosion-flow.

The horizontal curvature ($X_6$) have positive value in this area. The horizontal curvature can be recognize as ridge in the real topography [9]. One percent increase in the $X_6$ variable by one percent and other variables are constant and the effect of the location around the observed point, namely point 5, is considered to be constant and will also increase the number of silt elements by 1.947. Moreover, the number of silt content increases by 3.095 if the $X_3$ variable increases by one percent and other variables are constant and the effect of the location around the observed point, namely point 5 is considered constant. This two-variable explained that the silt content is distributed in high areas. The main causes why this occur is the Kawi Mountain eruptions in 2014 [10].

The remaining 3 variables have a negative coefficient. It can be interpreted that the number of silt elements will decrease by 0.402 with an increase in $X_2$ by one percent and other variables are constant and the effect of the location around the observed point, namely point 5 is considered constant. From those explanation, in can recognize that the soil in the south-facing area is sily that in the north. The reason why the silt distribution is also can be interpreted by the hill aspect is because silt is lighter than other particles that makes may be transported by strong winds or by storm winds [11]. The increase in $X_4$ by one percent also reduces the number of silt elements by 4.309 on the condition that the variable another constant and the effect of the location around the observed point, namely point 5 is considered constant. Then the decrease of 4.082 silt elements will occur with an increase of one percent of the $X_7$ element and other variables are constant and the effect of the location around the observed point, namely point 5 is considered constant. Concerning the soil formation and the silt content distribution, our findings can explain the relationship between soil properties and soil formation.

The $Y$ predicted value generated in the GWR silt model at point 5 is 23.80351, or the difference between 6.19649 and the factual value is 30. The difference in these values can occur because of the inclusion of regional weighting elements in the GWR model.

The significance test of model parameters at each location was carried out by partially testing the parameters to determine which parameter had a significant effect on the response variable.
Table 5. Variables Significant.

| Location     | Variables          |
|--------------|--------------------|
| 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 | X4, X5 & X8 |

Table 5 shows that only three variables significant in all location, as $|t_{count}| > 2.02$. The other variables are not significant because of $|t_{count}| < 2.02$.

4. Conclusion

GWR is a better alternative model for data that contains spatial aspects. This also applies to the modelling of the Silt soil texture. The value of $R^2$ in GWR model is 0.57, it means that performance model in explaining variables is 57%. Significant variables in all location are X4, X5 & X8.

References

[1] Saraiva Koenow Pinheiro H, de Carvalho Junior W, da Silva Chagas C, Helena Cunha dos Anjos L and Ray Owens P 2018 Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions Artic. Rev Bras Cienc Solo 42 170167

[2] Behrens T, Schmidt K, Zhu A X and Scholten T 2010 The ConMap approach for terrain-based digital soil mapping Eur. J. Soil Sci. 61 133–43

[3] Hassink J 1997 The capacity of soils to preserve organic C and N by their association with clay and silt particles Plant Soil 191 77–87

[4] Gessler P E, Moore I D, McKenzie N J and Ryan P J 1995 Soil-landscape modelling and spatial prediction of soil attributes Int. J. Geogr. Inf. Syst. 9 421–32

[5] Brunsdon C, Fotheringham A S and Charlton M E 1996 Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity Geogr. Anal. 28 281–98

[6] Fischer M and Getis A 2010 Handbook of Applied Spatial Analysis (New York: Springer)

[7] Kelejian H H and Prucha I R 2001 On the asymptotic distribution of the Moran I test statistic with applications Econometrica 104 219–57

[8] Breusch T S and Pagan A R 1979 A Simple Test for Heteroscedasticity and Random Coefficient Variation Econometrica 47 1287–94

[9] Florinsky I V. 2012 Digital Terrain Analysis in Soil Science and Geology

[10] Dibyosaputro S, Dipayana G A, Nugraha H, Pratiwi K and Valeda H P 2015 Lahar at Kali Konto after the 2014 Eruption of Kelud Volcano, East Java : Impacts and Risk 29 59–72

[11] Yang F, Zhang G L, Yang F and Yang R M 2016 Pedogenetic interpretations of particle-size distribution curves for an alpine environment Geoderma