ISSN (Online): 2350-0530
ISSN (Print): 2394-3629

International Journal of Research -GRANTHAALAYAH
February 2021, Vol 9(2), 64 – 71
DOI: https://doi.org/10.29121/granthaalayah.v9.i2.2021.3112

GEOGRAPHICALLY WEIGHTED REGRESSION AND MULTIPLE LINEAR REGRESSION FOR TOPSOIL TEXTURE PREDICTION

Henny Pramoedyo [*1], Sativandi Riza [2], Afiati Oktaviarina [1, 4], Deby Ardianti [3]

[1, 2, 3] Department of Statistics, Brawijaya University, Malang, Indonesia
[4] Department of Mathematics, Surabaya State University, Surabaya, Indonesia

DOI: https://doi.org/10.29121/granthaalayah.v9.i2.2021.3112

Article Type: Research Article

Article Citation: Henny Pramoedyo, Sativandi Riza, Afiati Oktaviarina, and Deby Ardianti. (2021). GEOGRAPHICALLY WEIGHTED REGRESSION AND MULTIPLE LINEAR REGRESSION FOR TOPSOIL TEXTURE PREDICTION. International Journal of Research - GRANTHAALAYAH, 9(2), 64-71. https://doi.org/10.29121/granthaalayah.v9.i2.2021.3112

Received Date: 15 January 2021

Accepted Date: 23 February 2021

Keywords:
Regression
Geographically Weighted Regression
Soil Texture Modelling
Terrain Analysis
Digital Elevation Model

ABSTRACT

Land resource management requires extensive land mapping. Conventional soil mapping takes a long time and is expensive; therefore, geographic information system data as a predictor in soil texture modeling can be used as an alternative solution to shorten time and reduce costs. Through digital elevation model data, topographic variability can be obtained as an independent variable in predicting soil texture. Geographically weighted regression is used to observe the effects of spatial heterogeneity. This study uses a data set of 50 observation points, each of which had soil particle-size fraction attributes and eight local morphological variables. The covariates used in this study are eastness aspects, northness aspects, slope, unsphericity curvature, vertical curvature, horizontal curvature, accumulation curvature, and elevation. Prediction using geographically weighted regression shows more results compared to multiple linear regression models. The spatial location can affect product Y, with the R2 value of 0.81 in the sand fraction, 0.57 in the silt fraction, and 0.33 in the clay fraction.

1. INTRODUCTION

Soil texture is influenced by topographic variability, which modifies water flow and material distribution to produce a soil pattern in a landscape [1]. Mapping of soil texture is needed as the main source of information in land resource management [2]. Soil mapping is conducted using conventional methods, which require large amounts of time and high costs. This results in minimal information regarding the broad spatial distribution of soil textures. In studies on soil texture mapping, many methods are utilized, including modeling [2], [3], [4], which produces soil texture mapping efficiently and accurately.

The combination of statistical modeling and GIS is an alternative solution to shorten the time and reduce costs. Hence, GIS data can be used as predictor variables in modeling [5], including GIS data for topographic variability to predict soil texture, which is the digital elevation model (DEM) [6]. Through DEM data, topographic variability can be obtained as an independent predictor of soil texture.

The simplest modeling, when there are two or more predictor variables, is multiple regression analysis. Multiple linear regression can model or predict an object by looking at the relationship between the dependent variable and a group of independent variables [7]. However, in regression analysis, several assumptions must be met. This regression is applied to modeling data that are influenced by spatial aspects or geographic conditions, and there will be assumptions that are difficult to fulfill that lead to spatial heterogeneity [8]. Spatial heterogeneity is a condition defined by different conditions from one location to another [9]. Additionally, this study uses geographically weighted regression (GWR) to observe the effects of spatial heterogeneity. GWR is based on a non-parametric technique of a locally weighted regression developed in statistics for curve fitting and smoothing [10]. Then, we compare the results of simple multi-linear regression with modeling using GWR. This study expects to produce a soil texture prediction model with high accuracy.

2. MATERIALS AND METHOD

The topsoil at a depth of 0-10 cm based on 50 randomly selected samples was taken from the Kalikonto watershed, in Malang, during June-July 2020. Soil texture content was then derived from the laboratory analysis and used as the primary data in this study. This was because soil texture is a combination of three particle-size fractions (PSFs): sand, silt, and clay. Modeling is conducted on the three PSFs, which are the Y variables. The X variables 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.

2.1. DATA SETS

This study's data sets consisted of 50 observation points, each of which had soil PSF attributes, and eight local morphological variables (LMV), which showed curvature diversity of a topography [11]. The LMV was obtained from the formula shown in Table 1. However, to obtain this variable, an analysis of the DEM data was performed to obtain the value derived from the elevation, which is the DEM digital number value. To obtain the derived value of the elevation, the following formula is used [12]:

$$p = \frac{z_3 + z_6 + z_9 - z_1 - z_4 + z_7)}{6w}$$

$$q = \frac{z_1 + z_2 + z_3 - z_7 - z_8 - z_9)}{6w}$$

$$r = \frac{z_1 + \ z_3 + z_4 + z_6 + z_7 + z_9 - 2(z_2 + z_5 + z_8)}{3w^2}$$

$$s = \frac{z_3 + z_7 - z_1 - z_9}{4w^2}$$

$$t = \frac{z_1 + \ z_2 + z_3 + z_7 + z_8 + z_8 - 2(z_4 + z_5 + z_6)}{3w^2}$$

Where z is the elevation, and w is the cell size in pixels. We apply a 3x3 window calculation to perform this analysis.

Table 1: Formula to obtain the LMV [11]..

Covariates	Formula
eastness aspects (Ae)	$A_e = sin\left[-90[1 - sin(q)](1 - \lvert sin(p)\rvert) + 180[1 + sin(p)] - \frac{180}{\pi}\ sin(p)arcos\left(\frac{-q}{\sqrt{p^2 + q^2}}\right)\right]$

northness aspects (An)	$A_n = cos\left[-90[1 - sin(q)](1 - \lvert sin(p)\rvert) + 180[1 + sin(p)] - \frac{180}{\pi} sin(p) arcos\left(\frac{-q}{\sqrt{p^2 + q^2}}\right)\right]$
slope (S)	$S = arctan\sqrt{p^2 - q^2}$
unsphericity curvature (M)	$M = \sqrt{H^2 - K}$
vertical curvature (Kv)	$K_v = \frac{p^2 r + 2pqs + q^2 t}{(p^2 + q^2)\sqrt{(1 + p^2 + q^2)^3}}$
horizontal curvature (Kh)	$K_h = \frac{q^2 r - 2pqs + p^2 t}{(p^2 + q^2)\sqrt{1 + p^2 + q^2}}$
accumulation curvature (Ka)	$K_a = \frac{(q^2 r - 2pqs + p^2 t)(p^2 r + 2pqs + q^2 t)}{[(p^2 + q^2)(1 + p^2 + q^2)]^2}$
elevation (Elv)	Direct DEM's pixel value

2.2. MULTILINEAR REGRESSION ANALYSIS

Multilinear regression analysis is the development of a simple regression analysis that explains and describes the relationship between the response variable and more than one predictor variable [13]. The regression equation model that can be formed with n observations and p predictor variables can be written as follows [7]:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + e_i$$

Where:

i	:	Observation – ith with i = 1, 2, …, n
X_{ik}	:	Observation – ith in – kth predictor with k =1, 2, …, p
β_0	:	The intercept value for all observations
β_k	:	kth predictor value
e_i	:	Observation – ith error

Before starting the analysis, we performed several assumption tests as a standard procedure in regression analysis. We conducted the normality test, heterogeneity test, and non-multicollinearity test.

2.3. GEOGRAPHICALLY WEIGHTED REGRESSION

In the spatial aspect, we tested the spatial autocorrelation by using the test statistic Moran's I, based on the following hypotheses [14]:

Hypotheses:

$H_0 : I = I_0$ (no spatial correlation).
$H_1 : I \neq I_0$ (there is a spatial correlation),
if H_0 true test statistic,

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad i \neq j$$

and

$$I_0 = E(I) = -\frac{1}{n - 1}$$

Where is the mean of , is the element of weighted matrix, is Moran's index, is the expected value of Moran's index, and is the number of samples.

The Breusch–Pagan test was used to test the spatial heterogeneity, based on the following hypotheses [15]:

Hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots .. = \sigma_j^2 = \sigma^2$$

H_1 : there are at least one j where $\sigma_j^2 \neq \sigma^2$

If H_0 true test statistic,

$$BP = \left(\frac{1}{2}\right) f^T Z (Z^T Z)^{-1} Z^T f + \left(\frac{1}{T}\right) \left[\frac{e^T W e}{\sigma^2}\right]^2 \sim \chi^2_{(p+1)}$$

Where f is $(f_1, f_2, \ldots, f_n)^T$ is e, is the galat vector, is the weighting matrix, is the matrix containing the standard predictor variable, and T is $Tr[W^T W + W^2]$.

The GWR model considers geographic factors and produces local estimators of the parameter model for each point or location [16]. The GWR model is as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i \; ; \; i = 1,2, \ldots. , n$$

Where y_i is the observed value of the ith predictor variable, x_{ik} is the kth predictor variable's observed value, $\beta_0(u_i, v_i)$ is the regression model intercept value, $\beta_k(u_i, v_i)$ is the kth predictor variable regression coefficient, and $\varepsilon_{i,}$ is the i-error.

The weighted least square method is used to estimate the parameter of the GWR model that produces different weighting in each location. The following is the parameter estimation for the GWR model [16]:

$$\hat{\beta}(u_i, v_i) = [X' W(u_i, v_i) X]^{-1} X' W(u_i, v_i) Y$$

From equation (5), the parameter coefficient of the GWR model for each location has different values.

The weighting forms by kernel function are divided into fixed kernel and adaptive kernel (Fotheringham). The fixed kernel function has the same bandwidth in all locations.[17]

$$w_{ij} = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]$$

Where is the bandwidth, is the adaptive bandwidth, and is the Euclidean distance
with,

$$d_{ij} = \sqrt{\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2}$$

Where is the coordinate point in location, and is the coordinate point in location.
Additionally, is optimum bandwidth with the cross validation (CV) method

$$CV(b) = \sum_{i=1}^{n} (y_i - \hat{y}_i(b))^2$$

Where n is the number of samples, and is the estimated value of

Partial testing in the GWR parameter model is used to determine which predictor variable influences the response variable for each location. Based on the following hypotheses:

Hypotheses:

$H_0 : \beta_j(u_i, v_i\ ) = 0$
$H_1 : \beta_j(u_i, v_i\ ) \neq 0$

The statistics test can be written as [16]:

$$\frac{\hat{\beta}_j(u_i, v_i\ )}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{(n-p-1)}$$

Where,

$$\sigma^2 = \frac{e'e}{n-p-1}$$

and c_{jj} is a diagonal matrix element CC^T

$$C = (X^T W X)^{-1}$$

Reject H_0 if the test statistic $|t| > t_{(\frac{\alpha}{2}, n-p-1)}$

3. RESULTS AND DISCUSSION

3.1. MULTILINEAR REGRESSION RESULT

For the sand model, the equation for the multiple linear regression model obtained is as follows:

$$\hat{Y} = 46{,}1293 - 2{,}4214Ae + 1{,}0453An - 3{,}1944S + 6{,}1095M + 2{,}3279Kv + 0{,}5892Kh - 3{,}5746Ka - 9{,}4836Elv$$

Based on the model obtained, An, M, Kv, and Kh have a positive relationship to the sand soil fraction. Meanwhile, Ae, S, Ka, and Elv have a negative relationship with the sand soil fraction. For example, the lower the Ka value, the lower the sand soil fraction. This multiple linear regression model produces an R2 value of 0.6285, which means that the study's independent variables simultaneously affect the sand soil fraction of 62.85%, and other variables outside the research variables influence the remaining 37.15%.

The equation for the silt model obtained is as follows:

$$\hat{Y} = 27{,}5042 + 1{,}1831Ae - 0{,}3682An + 3{,}0436S - 4{,}2889M - 5{,}2967Kv + 1{,}8952Kh - 4{,}1777Ka + 7{,}2921Elv$$

Based on this model, An, M, Kv, and Ka have a negative relationship with the silt soil fraction. Meanwhile, Ae, S, Kh, and Elv have a positive relationship with the sand soil fraction. For example, the lower the Ka value, the silt soil fraction will increase. This multiple linear regression model produces an R2 value of 0.5503, which means that the study's independent variables simultaneously affect the sand soil fraction by 55.03%, and the remaining 44.97% is influenced by other variables outside the research variables.

For the clay model, the equation for the multiple linear regression model obtained is as follows:

$$\hat{Y} = 26{,}3665 + 1{,}2383Ae - 0{,}6771An + 0{,}1508S - 1{,}8206M + 2{,}9689Kv - 2{,}4844Kh + 7{,}7523Ka + 2{,}1915Elv$$

Based on this model, An, M, and Kh have a negative relationship with the clay fraction. Meanwhile, Ae, S, Kv, Ka, and Elv positively correlate with the clay soil fraction. For example, the lower the Ka value, the higher the clay soil fraction. The multiple linear regression model produces an R2 value of 0.3034, which means that the independent variables simultaneously affect the sand soil fraction of 30.34% and the remaining 69.66% for other variables outside the research variables. The above models met the standard test for multiple regression analysis.

3.2. GWR ANALYSIS RESULT

Based on the results of the spatial dependence test in this study, the p-value of the three types of soil is smaller than $\alpha = 0.05$; therefore, a spatial dependence on observations exists. Likewise, with the results of the heterogeneity test in the three PSFs, spatial heterogeneity exists. Therefore, based on testing the spatial aspect, spatial dependence on observations and spatial heterogeneity exist, so the multiple linear regression method is not appropriate for describing the phenomenon of soil types. Therefore, it is better to use a model that accommodates the location factor of the observation.

The first step in GWR modeling is to determine the optimal bandwidth and minimum CV by using fixed Gaussian spatial weighting. The minimum CV and bandwidth results are shown in Table 2.

Table 2: Minimum CV and bandwidth

	p-value		
	Sand	Silt	Clay
CV minimum	3459.44	3767.20	3594.77
Bandwidth	4645.38	22043.22	22043.22

Then the GWR result is shown in Table 3.

Table 3: GWR Model

PSF	Variable	Coefficient		
		Min	Max	Global
Sand	*X intercept*	41.5669	48.4123	46.1294
	X1	-2.9916	-1.5447	-2.4215
	X2	-0.5120	1.6084	1.0453
	X3	-4.3435	-1.2016	-3.1946
	X4	-0.6242	5.5791	6.1098
	X5	-2.6798	3.4149	2.3281
	X6	-1.6738	3.4592	0.5892
	X7	-9.0681	1.5708	-3.5741
	X8	-12.2402	-4.1612	-9.4836
Silt	*X intercept*	27.5163	27.7998	27.500
	X1	1.14747	1.3448	1.1832
	X2	-0.4801	-0.2678	-0.3682
	X3	2.9984	3.1486	3.0438
	X4	-4.4882	-3.9209	-4.2890
	X5	-5.5011	-4.9580	-5.2969
	X6	1.5558	2.1242	1.8954
	X7	-4.1397	-4.0602	-4.1780
	X8	7.0855	7.5849	7.2919
Clay	*X intercept*	26.1564	26.5512	26.3665
	X1	1.1501	1.2355	1.2383
	X2	-0.7888	-0.5473	-0.6771

	X3	-0.0714	0.2663	0.1508
	X4	-2.038	-1.5131	-1.18208
	X5	2.8305	2.9345	2.9688
	X6	-2.8077	-2.1079	-2.4846
	X7	7.4888	8.1884	7.7521
	X8	2.0445	2.2463	2.1917

Table 4: MLR and GWR models comparison

Model	Determination Coefficient		
	Sand	Silt	Clay
MLR	0.63	0.55	0.30
GWR	0.81	0.57	0.33

Based on Table 4, the value of the R^2 GWR model for the three types of soil is greater than the value of the multiple regression R^2, meaning that the GWR model is better for modeling the existing data.

4. CONCLUSION

Prediction using GWR shows more results compared to multiple linear regression models. The spatial location can affect product Y, with the R2 value of 0.81 in the sand fraction, 0.57 in the silt fraction, and 0.33 in the clay fraction.

SOURCES OF FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The author have declared that no competing interests exist.

ACKNOWLEDGMENT

This research is sponsored by grants professors and doctors, faculty of mathematics and natural sciences, University of Brawijaya. The authors are grateful to the anonymous referees for a careful checking of the details and for helpful comments that improved the overall presentation of this paper.

REFERENCES

[1] P. E. Gessler, I. D. Moore, N. J. McKenzie, and P. J. Ryan, “Soil-landscape modelling and spatial prediction of soil attributes,” Int. J. Geogr. Inf. Syst., vol. 9, no. 4, pp. 421–432, 1995.

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

[3] V. L. Mulder, S. de Bruin, M. E. Schaepman, and T. R. Mayr, “The use of remote sensing in soil and terrain mapping — A review,” Geoderma, vol. 162, no. 1–2, pp. 1–19, Apr. 2011.

[4] M. Ließ, B. Glaser, and B. Huwe, “Uncertainty in the spatial prediction of soil texture,” Geoderma, vol. 170, pp. 70–79, Jan. 2012.

[5] C. da S. Chagas, W. de Carvalho Junior, S. B. Bhering, and B. Calderano Filho, “Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions,” CATENA, vol. 139, pp. 232–240, Apr. 2016.

[6] J. B. Lindsay, J. M. H. Cockburn, and H. A. J. Russell, “An integral image approach to performing multi-scale topographic position analysis,” Geomorphology, vol. 245, pp. 51–61, 2015.

[7] J. P. Holcomb, N. R. Draper, H. Smith, J. O. Rawlings, S. G. Pantula, and D. A. Dickey, “Applied Regression Analysis Applied Regression Analysis: A Research Tool,” Am. Stat., 1999.

[8] A. S. Fotheringham and T. M. Oshan, “Geographically weighted regression and multicollinearity: dispelling the myth,” J. Geogr. Syst., 2016.

[9] J. Tamas, P. Reisinger, P. Burai, and I. David, “Geostatistical analysis of spatial heterogenity of Ambrosia artemisiifolia on Hungarian acid sandy soil,” in Journal of Plant Diseases and Proctectio, Supplement, 2006.

[10] [10] C. Brunsdon, S. Fotheringham, and M. Charlton, “Geographically Weighted Regression,” J. R. Stat. Soc. Ser. D (The Stat., vol. 47, no. 3, pp. 431–443, 1998.

[11] P. A. Shary, “Land surface in gravity points classification by a complete system of curvatures,” Math. Geol., vol. 27, no. 3, pp. 373–390, 1995.

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

[13] J. I. Daoud, “Multicollinearity and Regression Analysis,” J. Phys. Conf. Ser., vol. 949, no. 1, 2018.

[14] M. Fischer and A. Getis, Handbook of Applied Spatial Analysis. New York: Springer, 2010.

[15] L. Anselin, Spatial econometrics: methods and models, vol. 4. Springer Science & Business Media, 2013.

[16] A. S. Fotheringham, C. Brunsdon, and M. Charlton, Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons, 2003.

[17] C. Chasco, “Modeling spatial variations in household disposable income with Geographically Weighted Regression,” no. January, 2007.