ITM Web of Conferences **58**, 04007 (2024) https://doi.org/10.1051/itmconf/20245804007
The 6th IICMA 2023

Comparison Between Iterative Least Square and Nonparametric Epanechnikov Kernel in Semivariogram Modeling, Case study: Urban Land Cover in East Java Province

Kurnia Novita Sari[1*], *Yonathan Yeremy* Budiman[1], *Udjianna Sekteria* Pasaribu[1], and Abdullah Sonhaji[1]

[1]Departement of Mathematics, Institut Teknologi Bandung, Bandung, West Java, Indonesia

Abstract. Landcover is an example of spatial data that contains location coordinate information along with the variables measured at each location, namely height, slope, and curvature. The spatial relationship between locations can be measured using a semivariogram. Semivariogram is a statistic used to measure the spatial correlation of pairs of locations separated by a certain distance and angle. In estimating semivariogram parameters, namely, nugget, sill, and range, there are two methods, namely parametric via iterative least squares and nonparametric via kernel functions. These two methods will be compared for semivariogram modeling of built-up land in East Java province. The best method is selected based on the smallest SSE value. For each physical factor, the best model with the smallest SSE is the Epanechnikov kernel function, with 3.6 for the elevation SSE, $8.84 \cdot 10^{-7}$ for the slope SSE, and $2.15 \cdot 10^{-21}$ for the curvature SSE. So, it is concluded in this case that the nonparametric kernel Epanechnikov method is much better than the parametric method using Gaussian and exponential models.

1 Introduction

Indonesia had a birth rate of 2.18 per woman in 2022 [1], so there is an increasing population in Indonesia, which increases the need for residential land, which intersects with the socio-economic dynamics of society. Two relevant terms are land cover and land use. Land cover is a biophysical attribute of the earth's surface in the form of plants, buildings, water bodies, and so on [2]. Meanwhile, land use is an area on the earth's surface with particular land cover as an area for activities. Land cover changes are highly dependent on physical and socio-economic factors. Biological factors include height, slope, and curvature. Meanwhile, socioeconomic factors include population density, distance to the nearest road, distance to the National Capital, distance to residential areas, and the center of business activities [3]. Land cover in a province in Indonesia is divided into eight classes, namely: (1) Forest, (2) Protected/conservation areas, (3) Urban land, (4) Grasslands, (5) Plantations, (6) Wetland agriculture, (7) Dryland farming, and (8) Ponds. Land cover data is obtained from satellite

* Corresponding author: kurnia@math.itb.ac.id

images with a resolution of 30 × 30 meters per pixel with a total of 53,283,768 pixels, so it is classified as location-based big data, called spatial data.

The relationship between variable values represented by the factors above at each observation location can be analyzed using semivariogram modeling. This modeling uses variogram statistics, which is a variance diagram of the difference in observation values between locations at a distance of h, with the general form of the semivariogram model being as follows:

$$C_0 + C \cdot f\left(\frac{|h|}{a}\right) \quad (1)$$

Semivariogram modeling has various applications such as estimating the spatial behavior of agricultural soil fertility by kriging interpolation and parameter estimation using the methods of moments and restricted maximum likelihood [4], comparison of the Matheron, Cressie-Hawkins, and Dowd models for the prediction of theft in the city of Bandung [5], modeling the spread of the insect Bradysia ocellaris on oyster mushroom cultivation using isotropic spherical models [6], and estimating travel demand variables for transportation planning using genetic algorithms and geostatistical methods to predict travel demand variables to optimize the calculation and matching of semivariograms [7].

The parameters of the semivariogram model can be estimated using various methods, such as the least squares method [8], the maximum likelihood method [9], and non-parametric methods such as the bootstrap method [10]. Furthermore, many other researchers have developed land cover modeling, such as [3] using a divisive hierarchical algorithm in cluster analysis to group land change areas to study land types in East Java and group them based on physical and social factors multivariate analysis, and [11] used a binary logistic regression method applied to variables that influenced land cover change, including physical and social factors such as slope, curvature, population density, and so on.

In this study, a comparison of the Gaussian and exponential semivariogram models will be carried out to model the spatial correlation of the urban land cover data using the iterative least squares method and the kernel function method using the Epanechnikov kernel function to estimate the theoretical semivariogram model. Next, it will be seen which model has the smallest SSE from the semivariogram model and experimental semivariogram, so a model will be obtained to predict the maximum distance of changes in urban land cover with the smallest SSE among the models being compared.

2 Methodology

Variogram is a statistic used in geostatistics tools that is useful for describing spatial correlation between variables measured between locations. The variogram is a variance diagram of the difference in variable values between pairs of sites and is formulated as follows:

$$2\gamma(h) = E[Z(u+h) - Z(u)]^2 \quad (2)$$

With $\gamma(h)$ is a semivariogram. The experimental variogram is an approximate value obtained from field sampling based on the spatial correlation between two variables separated by a certain h distance. Matheron (1965) defined the experimental semivariogram as follows:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left(Z(u_i + h) - Z(u_i)\right)^2 \quad (3)$$

With

u_i : location of sample points containing location coordinates

$Z(u_i)$: observation value at the location u_i

h : distance between two location points measured using Euclidean distance

u_i and $u_i + h$: pairs of sample points with distance h

$N(h)$: the number of pairs of locations with distance h

To find the value of the semivariogram, many data pairs are divided into several classes using the Sturges equation $K = 1 + 3.3 \log n$ with n is the sample size and K is the number of class intervals. The use of this equation is generally under the normal distribution assumption.
The semivariogram modeling steps are described as follows:

1. Determining the Experimental Semivariogram
   1) Compute $u_i = (x_i, y_i)$ which states the coordinates of the i-th location with i =1, 2, …, n in meters.
   2) Build a distance matrix $D = [d_{ij}]$, with $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is a Euclidean distance
   3) Determining d_{max} and lag distance h_i, with $N(h_i)$ is the amount of data in the-ith lag
      a) Determining the number of lag with Sturges formula $K = 1 + 3.3 \log N$ with $N = \binom{n}{2}$ where n is the number of pairs of distinct locations.
      b) Computing the length of the lag, i.e. $\frac{d_{max} - d_{min}}{K}$
      c) Computing $N(h_i)$, namely the number of pairs of locations at lag i
   4) Calculate Matheron's experimental semivariogram for the lag of the-ith distance:

$$\gamma(h_i) = \frac{1}{2N(h_i)} \sum \left(Z(u_i) - Z(u_j)\right)^2 \tag{4}$$

2. Semivariogram Parameter Estimation Using the least squares method

In this section, the least squares estimation method is used in the theoretical semivariogram model to estimate the parameter estimator of the semivariogram model. This approach uses the estimation method in the following linear regression equation

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i \tag{5}$$

with

$$\widehat{\beta_0} = \frac{(\sum_{i=1}^{n} x^2)(\sum_{i=1}^{n} y) - (\sum_{i=1}^{n} xy)(\sum_{i=1}^{n} x)}{(n \sum_{i=1}^{n} x^2) - (\sum_{i=1}^{n} x)^2} \text{ and } \widehat{\beta_1} = \frac{n(\sum_{i=1}^{n} xy) - (\sum_{i=1}^{n} x)(\sum_{i=1}^{n} y)}{(n \sum_{i=1}^{n} x^2) - (\sum_{i=1}^{n} x)^2} \tag{6}$$

Based on the form of the model and parameter estimates obtained, the theoretical semivariogram model can be written as follows:

$$\gamma(h_i) = \widehat{C_0} + \hat{C} \cdot f\left(\frac{|h_i|}{a}\right) \tag{7}$$

In the least squares method, the sum of the squared errors of the experimental semivariogram and model semivariogram will be minimized, i.e.,

$$s(h_i) = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} \left(\gamma(h_i) - \hat{\gamma}(h_i)\right)^2 \tag{8}$$

With and $\hat{\gamma}(h_i)$ respectively are experimental semivariogram and semivariogram model. First, we will look for parameters C_0, C, and a for the semivariogram model that depends on the model to be selected. Then, $s(h_i)$ will be minimized by differentiating $s(h_i)$ for each parameter as follows:

$$\frac{\partial s(h_i)}{\partial C} = \sum_{i=1}^{n} 2 \cdot \left[\gamma(h_i) - C_0 - C \cdot f\left(\frac{|h_i|}{a}\right)\right]\left[-f\left(\frac{|h_i|}{a}\right)\right] = 0 \tag{9}$$

$$C \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right) = \sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right) - C_0 \sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right) \tag{10}$$

$$\hat{C} = \frac{\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)}{\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} - C_0 \frac{\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)}{\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} \tag{11}$$

Then perform a second derivative test to ensure that $\hat{C}$ minimize the SSE, i.e.:

$$\frac{\partial^2 s(h_i)}{\partial C^2} = 2 \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right) > 0 \tag{12}$$

Because the second derivative at $s(h_i)$ with respect to C is always positive, then it can be ascertained that the parameter estimate $\hat{C}$ in equation (11) is an estimate that minimizes $s(h_i)$. Next, the parameter values for C_0 which minimizes $s(h_i)$ will be estimated. Differentiating $s(h_i)$ with respect to C_0, then we get:

$$\frac{\partial s(h_i)}{\partial C_0} = \sum_{i=1}^{n} 2 \cdot \left[\gamma(h_i) - C_0 - C \cdot f\left(\frac{|h_i|}{a}\right)\right](-1) = 0 \tag{13}$$

$$C_0 \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} \gamma(h_i) - C \sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right) \tag{14}$$

$$\hat{C}_0 = \frac{\sum_{i=1}^{n} \gamma(h_i)}{n} - C \frac{\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)}{n} \tag{15}$$

Then perform a second derivative test

$$\frac{\partial^2 s(h_i)}{\partial {C_0}^2} = 2 \sum_{i=1}^{n} 1 = 2n > 0 \tag{16}$$

Because the second derivative on $s(h_i)$ with respect to C_0 is always positive, then the estimate of $\hat{C}_0$ in Equation (15) is an estimate that minimizes $s(h_i)$. Next, $\hat{C}$ which minimizes $s(h_i)$ will be estimated. Substitute Equation (15) into (11), then we get

$$\hat{C} = \frac{\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)}{\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} - \left(\frac{\sum_{i=1}^{n} \gamma(h_i)}{n} - \hat{C} \frac{\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)}{n}\right)\left(\frac{\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)}{\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)}\right) \tag{17}$$

$$\hat{C} = \frac{\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)}{\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} - \frac{\left[\sum_{i=1}^{n} \gamma(h_i)\right]\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} + \hat{C} \frac{\left(\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right)^2}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} \tag{18}$$

$$\hat{C}\left(1 - \frac{\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]^2}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)}\right) = \frac{n\left[\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)\right] - \left(\sum_{i=1}^{n} \gamma(h_i)\right)\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} \tag{19}$$

$$\hat{C} \frac{n\left[\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]^2}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} = \frac{n\left[\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} \gamma(h_i)\right]\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n \sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)} \tag{20}$$

$$\hat{C} = \frac{n\left[\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} \gamma(h_i)\right]\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n\left[\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]^2} \tag{21}$$

Then we get $\hat{C}$ which depends on $\hat{a}$. Then, $\hat{C}_0$ is obtained by substituting equation (21) to (15), as follows:

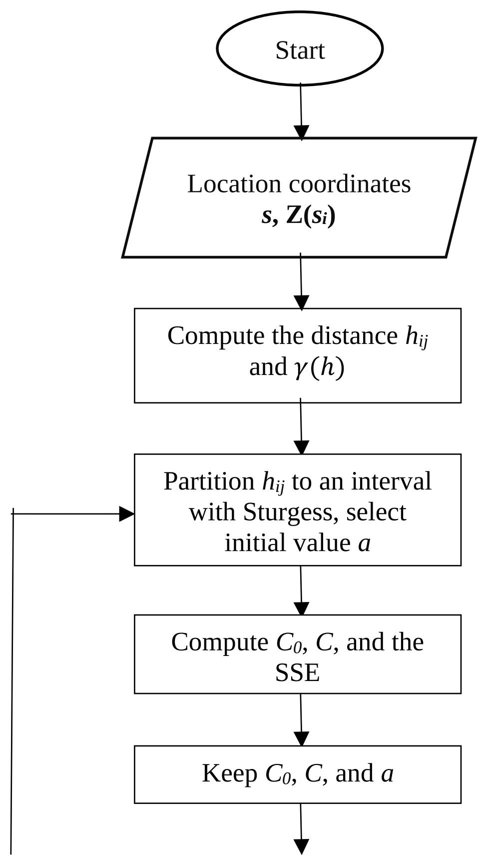
$$\hat{C}_0 = \frac{\sum_{i=1}^{n} \gamma(h_i)}{n} - \left(\frac{n\left[\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} \gamma(h_i)\right]\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n\left[\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]^2}\right)\left(\frac{\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)}{n}\right) \tag{22}$$

$$\hat{C}_0 = \frac{\left[\sum_{i=1}^{n} \gamma(h_i)\right]\left[\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} \gamma(h_i) f\left(\frac{|h_i|}{a}\right)\right]\left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]}{n\left[\sum_{i=1}^{n} f^2\left(\frac{|h_i|}{a}\right)\right] - \left[\sum_{i=1}^{n} f\left(\frac{|h_i|}{a}\right)\right]^2} \tag{23}$$

Then we get $\hat{C}_0$ which also depends on $\hat{a}$. Next, for calculating $\hat{a}$, The least squares method algorithm will be carried out as follows:

1. Distance partition
   a. Divide the distance between locations from 0 to h_{maks} with the distance between partitions being 10^{-2}
   b. Because the value $a \neq 0$, then choose a value that is quite close to zero as the initial value of a, for example 10^{-10}
2. Compute the value of C_0 and C
   a. Insert the initial value of â in Equations (21) and (23) to get $\hat{C}_0$ and $\hat{C}$
   b. Insert the values of $\hat{C}_0$ and $\hat{C}$ into the theoretical semivariogram model.
   c. Compute the SSE using experimental semivariogram and theoretical semivariogram values.
3. Do a looping to the algorithm as follows:
   a. Change the value of a one-step after initialization, namely $10^{-10} + 10^{-2}$ to h_{maks}
   b. Compute C_0, C, and SSE with the number of repetitions being $h_{max} \cdot 10^2$
   c. Compare the SSE values
   d. If the current SSE value is smaller than the previous SSE, then keep the smaller SSE along with the values $\hat{C}_0$, $\hat{C}$ and $\hat{a}$ of the smaller SSE

The following is a flow diagram of the semivariogram modeling



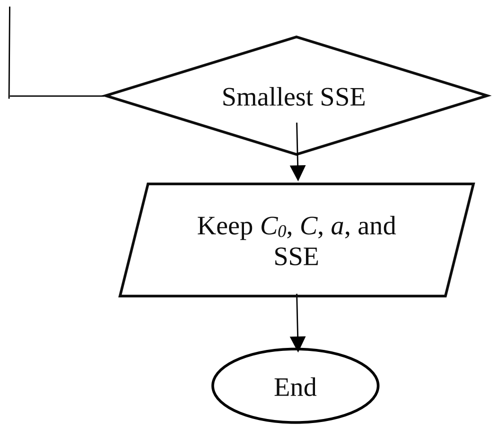


Fig. 1. Flow diagram of the semivariogram modeling

3. Semivariogram Estimator Using the kernel function method

In this section, the nonparametric kernel function method is used to estimate the theoretical semivariogram from the experimental semivariogram, and we will show the derivation of the uniform kernel function formula and give the general form of the kernel function that will be used. Consider the sequence of random variables $X_1, X_2, \cdots, X_n$ that are mutually independent and identically distributed as a random variable X, with the distribution function being $F(x) = P(X \leq x)$ which is absolutely continuous, and its probability density function is $f(x) = \frac{dF(x)}{dx}$. We can estimate the distribution function $F(x)$ with with $F_n(x)$, which is the average of the number of observations $\leq x$ among the sequence of random variables $X_1, X_2, \cdots, X_n$ [12], which can be written as the sum of the following indicator function:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} \tag{24}$$

Next, we know that $f_{n(x)} = \frac{dF_n(x)}{dx}$, and it is known that the symmetric difference quotient formula for approximating the first derivative is

$$F'(x) \approx \frac{F(x+h) - F(x-h)}{2h} \tag{25}$$

For sufficiently small values of h [13], we have

$$f_n(x) = F_n'(x) \tag{26}$$

$$= \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{1}_{x-h \leq X_i \leq x+h} \tag{27}$$

$$= \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{1}_{\left|\frac{x-X_i}{h}\right| \leq 1} \tag{28}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K(p) \tag{29}$$

With $K(p)$, which is defined as follows:

$$K(p) = \begin{cases} \frac{1}{2}, & |p| \leq 1 \\ 0, & |p| > 1 \end{cases} \tag{30}$$

with $p = \frac{x - X_i}{h}$, then, it is clear that $K(p)$ is a uniform density function on the interval $[-1,1]$. Note that the function $f_{n(x)}$ is an estimate of the density function $f(x)$ that calculates the weight of the observations at a point x. If there are many observations X_i that are close to point x, then $f_n(x)$ has a reasonably large value, but conversely, if only a few observations X_i are close to point x, then $f_n(x)$ has a small value. The parameter h in the function $f_n(x)$ controls the smoothness level of a curve. The function $f_n(x)$ is also known as kernel estimator, which is generally written as:

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \tag{31}$$

with $K(.)$ is a kernel function, h referred to as bandwidth, X_i is the independent variable in the data, x is the value to be estimated, and n is the amount of data. In this paper, we will use the Epanechnikov kernel function with $K(p) = \begin{cases}\frac{3}{4}(1-p^2), & |p| \leq 1 \\ 0, & |p| > 1\end{cases}$, which will be processed using NumXL using the available data to obtain the kernel semivariogram numerically.

3 Data and results of semivariogram modelling

Land cover data for urban areas was obtained using the QGIS tool, with 106,655 data obtained. Furthermore, the data samples were taken from 106,655 data using systematic random sampling. The sample size is determined using the Slovin equation below

$$n = \frac{N}{1+Ne^2} \tag{32}$$

With n is the number of samples taken, N is the number of population, and e is the margin of error with a 95% confidence interval. By calculating the Slovin Equation on the land cover data of urban areas, it is obtained that the number of samples to be used is

$$n_s = \frac{106655}{1+106655\times 0.05^2} \approx 400 \tag{33}$$

Furthermore, descriptive statistics from the 400 samples were produced in Table 1 as follows:

Table 1. Descriptive statistics from elevation, slope, and curvature data

Des. Stat	**Elevation**	**Slope**	**Curvature**
Mean	123.5261	3.522868	$2.748 \cdot 10^{-6}$
Standard Error	8.078256	0.157731	$3.525 \cdot 10^{-5}$
Median	67.00664	2.638235	$3.815 \cdot 10^{-5}$
Standard Deviation	161.5651	3.15462	$7.049 \cdot 10^{-4}$
Variance	26103.29	9.951624	$4.969 \cdot 10^{-7}$
Kurtosis	6.49887	16.55634	3.2662026
Skewness	2.374301	3.256353	-0.75174

Range	1081.136	29.46804	0.0057572
Minimum	0	0.122341	-0.003631
Maximum	1081.136	29.59039	0.0021263
Sum	49410.43	1409.147	0.001099
Count	400	400	400

Note that for elevation, the mean and median have quite a big difference, where the mean is much larger than the median. This indicates that the elevation distribution is skewed to the right. For slope, the mean is also larger than the median, indicating a right-skewed distribution of slopes. For curvature, note that the mean and median have very small values close to zero. Still, it can be seen that the median is greater than the mean, which indicates that the curvature distribution is slightly skewed to the left because the mean and median are almost zero.

Meanwhile, the skewness for elevation and slope is quite large in the positive direction, which indicates the data will be skewed to the right, whereas for curvature, there is a skew to the left, which is not too large. Next, note that the kurtosis of elevation and slope is much greater than 3, so the tails of the dataset for elevation and slope are much thicker than the normal distribution, but for curvature, the tails are slightly thicker than the normal distribution.

The following is a boxplot of elevation, slope, and curvature data

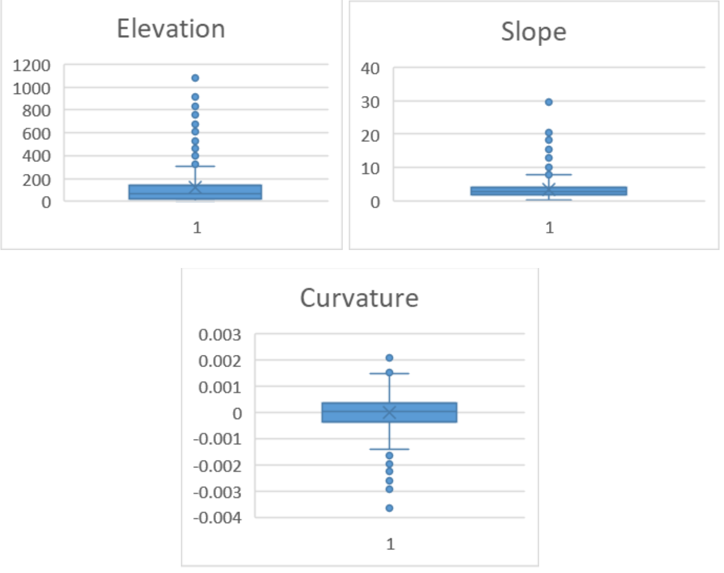


Fig. 2. Boxplot of Slope, Elevation, and Curvature

It can be seen from figure 2 that the boxplot of elevation and slope data is skewed to the right with many upper outliers, which mean that the data is skewed to the right. The boxplot of curvature is symmetric but has many lower outliers, which mean that the data is skewed to the left.

The following are the experimental semivariogram contour results for elevation, slope, and curvature

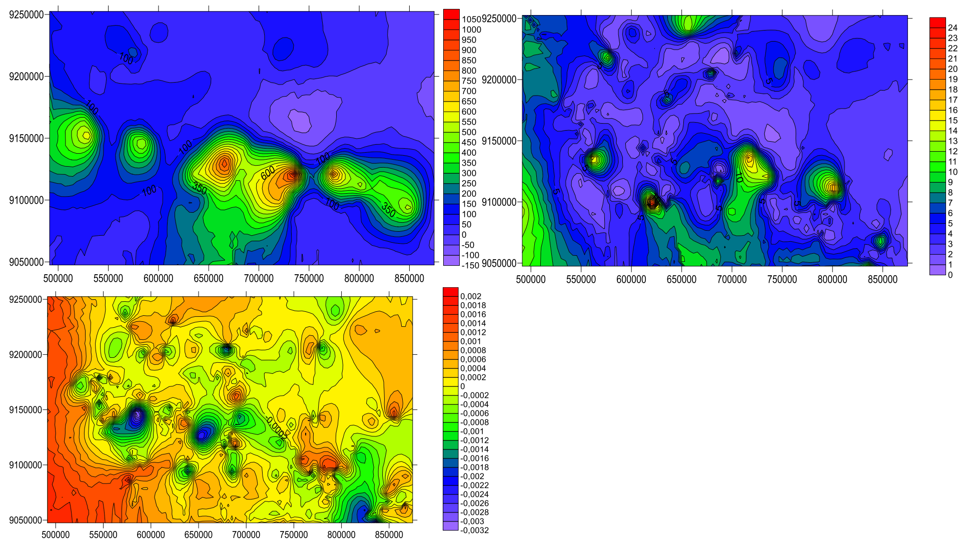


Fig. 3. Contour for elevation, slope, and curvature

From Figure 3, the elevation and slope contours are quite similar to most of the low-value locations (in blue) and a small portion of medium-value (in green) in the central part of East Java. Meanwhile, on the contrary, the curvature contour has medium and high curvature values (yellow and red) in most areas. The central part of East Java shows moderate to low curvature values (green and blue). From the contour pattern, there is a similar pattern between elevation and slope, while curvature is the opposite of the two previous variables.

By using the Sturgess equation, 17 intervals are obtained. The following is a table of distance lag results, the amount included in the distance lag (***N*(*h*)**), and experimental semivariograms for slope, elevation, and curvature.

Table 2. Table of results for distance lag, number of distance lags, and experimental semivariogram of slope, elevation, and curvature

Interval	**N(h)**	**Experimental Semivariogram**		
		Slope	**Elevation**	**Curvature**
		0	0	0
[0, 22544.03387)	3663	7.085960615	6739.732285	$4.42776 \cdot 10^{-7}$
[22544.03387, 45088.06773)	5800	10.34332584	21383.54075	$5.41788 \cdot 10^{-7}$
[45088.06773, 67632.1016)	9029	9.995825897	27860.24457	$4.84642 \cdot 10^{-7}$
[67632.1016, 90176.13547)	9927	8.92705003	29608.37401	$4.46144 \cdot 10^{-7}$
[90176.13547, 112720.1693)	9866	10.87987132	36968.73353	$5.05113 \cdot 10^{-7}$

[112720.1693, 135264.2032)	8499	11.68876685	37777.45526	$4.93878 \cdot 10^{-7}$
[135264.2032, 157808.2371)	7712	10.28036758	29854.10482	$4.81693 \cdot 10^{-7}$
[157808.2371, 180352.2709)	5897	11.31731521	24547.67548	$4.94563 \cdot 10^{-7}$
[180352.2709, 202896.3048)	4534	9.396507554	28889.42977	$4.94 \cdot 10^{-7}$
[202896.3048, 225440.3387)	4289	9.087832281	16211.83075	$5.80904 \cdot 10^{-7}$
[225440.3387, 247984.3725)	3674	9.742873146	13162.69423	$4.94003 \cdot 10^{-7}$
[247984.3725, 270528.4064)	2960	8.672341016	11545.87616	$4.79502 \cdot 10^{-7}$
[270528.4064, 293072.4403)	1560	8.048589106	13439.81664	$6.2421 \cdot 10^{-7}$
[293072.4403, 315616.4741)	1223	10.13716953	9827.895208	$6.21402 \cdot 10^{-7}$
[315616.4741, 338160.508)	897	6.476326543	8794.203236	$4.6448 \cdot 10^{-7}$
[338160.508, 360704.5419)	235	7.209771691	15829.9678	$5.6793 \cdot 10^{-7}$
[360704.5419, 383248.5757]	35	31.80662491	11831.6654	$6.73331 \cdot 10^{-7}$

The following is a comparison graph between the experimental semivariogram and the exponential and Gaussian model semivariogram

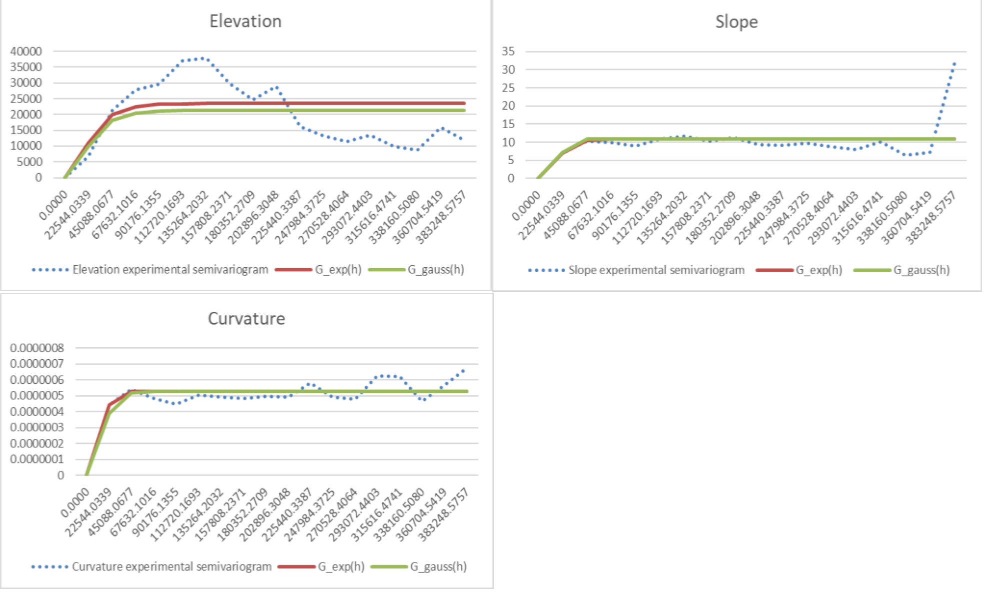


Fig. 4. Comparison chart between the experimental semivariogram and the semivariogram model

The graphical comparison in figure 4 between the exponential semivariogram model, the Gaussian semivariogram model, and the experimental semivariogram shows that the two models have fairly close values except for the semivariogram model for elevation.

The following figure 5 is a comparison graph between the experimental semivariogram and the Epanechnikov kernel semivariogram with the chosen bandwidth of 9.

Fig. 5. Comparison of the experimental semivariogram and the Epanechnikov kernel semivariogram

From the results of semivariogram modeling, by selecting three models, namely exponential, Gaussian, and Epanechnikov kernel, we get the parameter estimates and SSE that can be seen in Tables 3 and 4.

Table 3. Estimating the experimental semivariogram of elevation, slope, and curvature using the exponential and Gaussian semivariogram models.

	Elevation	**Slope**	**Curvature**
Exp	$(C_0 = 0; C = 23495.63; a = 17954.16)$	$(C_0 = 0.035; C = 10.876; a = 10864.7)$	$(C_0 = 0; C = 5.28 \cdot 10^{-7}; a = 6167.3)$
	SSE = 1.47214×10^9	SSE = 496.5923	SSE = 6.6×10^{-14}
Gauss	$(C_0 = 0; C = 21328.24; a = 17914.07)$	$(C_0 = 5.7 \cdot 10^{-4}; C = 10.875; a = 10978.7)$	$(C_0 = 1.95 \cdot 10^{-12}; C = 5.28 \cdot 10^{-7}; a = 8345.72)$
	SSE = 142709	SSE = 497.0222	SSE = 6.59×10^{-14}

Table 4. Estimation of the semivariogram model using Kernel Epanechnikov function.

Elevation	**Slope**	**Curvature**
SSE = 3.6	SSE = $8.84 \cdot 10^{-7}$	SSE = $2.15 \cdot 10^{-21}$

Note that the SSE of the kernel semivariogram model can reach values of $8.84 \cdot 10^{-7}$ and $2.15 \cdot 10^{-21}$, different from the SSE of semivariogram parametric models such as exponential and Gaussian. This shows that the kernel function, a nonparametric estimator, is much better than parametric estimator methods such as the iterative least squares method, because the nonparametric semivariogram model using Epanechnikov kernel function have a much smaller SSE than the exponential and Gaussian parametric semivariogram.

4 Conclusions and suggestionss

From the modeling done by comparing exponential, Gauss, and kernel functions and taking the smallest SSE, it is recommended to choose the semivariogram model using the Epanechnikov kernel function because the SSE is much smaller than the SSE of the exponential and gauss semivariogram models.

The authors thank FMIPA ITB, who helped the author finance this journal's publication. The author also thanks Julius Susanto for valuable discussions while writing this scientific paper.

References

1. Badan Pusat Statistik, Statistik Indonesia 2023. BPS-Statistics Indonesia (2023)
2. E. F. Lambin, B. L. Turner, H. J. Geist, S. B. Agbola, A. Angelsen, J. W. Bruce, O. T. Coomes, R. Dirzo, G. Fischer, C. Folke, P. S. George, K. Homewood, J. Imbernon, R. Leemans, X. Li, E. F. Moran, M. Mortimore, P. S. Ramakrishnan, J. F. Richards, H. Skanes, W. Steffen, G. D. Stone, U. Svedin, T. A. Veldkamp, C. Vogel, and J. Xu, Global Environmental Change, **11**, 4, 261–269, (2001)
3. L. K. Amanah. Analisis Klaster Pada Data Perubahan Lahan Melalui Algoritma Hirarki Divisif Studi Kasus: Provinsi Jawa Timur. Master's thesis, Institut Teknologi Bandung, (2022).
4. A. L. G. Oliveira, J. P. Lima, T. L. Brasco, and L. R. Amaral, Computers and Electronics in Agriculture, **196**, (2022)
5. K. N. Sari, U. S. Pasaribu, and R. Trismayangsari, Indonesia Journal, **12**, 1, 9–15, (2020).
6. K. N. Sari and R. Amelia. Statistics analysis of distribution of Bradysia Ocellaris insect on oyster mushroom cultivation. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. AIP Publishing LLC, (22 December 2015).
7. S. S. Rocha, C. S. Pitombo, L. H. M. Costa, and S. de França Marques, Transportation Research Interdisciplinary Perspectives, **10,** 100369 (2021)
8. S. N. Lahiri, Y. Lee, and N. Cressie, Journal of Statistical Planning and Inference, **102**, 1–2, 65–85 (2002)
9. A. Stein, J. Flacke, and J. Zevenbergen, Journal of Geographical Information Science, **34**, 7, 1451–1474 (2020)
10. K.N. Sari, U.S. Pasaribu, O. Neswan, A.K. Permadi, Applied Mathematical Sciences, **9**, 103 (2015)
11. U.S. Pasaribu, R. Vitriana, A. Deliar, I. Sumarto, IOP Conference Series: Earth and Environmental Science, **500** (2020)
12. A. W. van der Vaart. Asymptotic Statistics. (Cambridge University Press, Cambridge, United Kingdom, 1998).
13. P. D. Lax and M. S. Terrell. Calculus With Applications. (Springer, New York State, United States of America, 2014).