Fitting the variogram model of nickel laterite using root means square error in Morowali, Central Sulawesi

Benny Anggara, Irfan Marwanza*, Masagus Ahmad Azizi and Wiwik Dahani, Subandrio

Department of Mining Engineering, Trisakti University, Jl. Kyai Tapa No. 1, RT. 6/RW. 16, Kel. Grogol, Kec. Grogol Petamburan, Kota Jakarta Barat, DKI Jakarta, 11440, Indonesia

*Corresponding author’s e-mail: irfanm@trisakti.ac.id

Abstract. The nickel commodity is getting popular due to its role as one of the raw materials for battery manufacture. It is estimated that this trend will continue for the next 2 - 3 years and reaching its peak when the factories that process the raw material for electric vehicle batteries are established. For this reason, the nickel mining companies are competing to explore new nickel deposits. The research location is a nickel mine in Sulawesi. The purpose of this study was to determine the most suitable Nickel variogram model based on root means square error (RMSE). To obtain an accurate number of resources, it is necessary to apply an accurate and validated estimation method to gain data that are in line with the actual conditions. Therefore, this study uses a geostatistical method that takes into account the spatial relationship of each data using a variogram which is validated by the cross-validation method and RMSE. From the results of the RSME analysis, the most suitable variogram model for nickel content in the limonite and saprolite layers is the exponential variogram model. In addition, the values of root mean square error for nickel content in the limonite and saprolite layers were 0.022 and 0.098 respectively.

1. Introduction

The potential for nickel resources and reserves in Indonesia is still quite large. From data as of July 2020, the reserves are still reaching 11,887 million tons [1]. These resources and reserves are scattered in various regions in Indonesia. Geologically, Sulawesi Island is the most associated with nickel. Nickel production in Indonesia can still compete in the global market. Moreover, nickel ore in Indonesia can be managed well by smelters that have operated in Indonesia. Although the resources and reserves of nickel are considered to be abundant, the government, in this case, the Geological Agency, and national & international private nickel mining companies are still trying to find new potential locations. Currently, the government continues to encourage the energy transition in the transportation sector from fossil fuels to new and renewable energy. This means that the government will promote the use of electric transportation in the future. In other words, the energy source for those electric transportations is from the battery. In addition, one of the battery components is nickel, resulting in nickel being the favourite of mining investors nowadays.

The purpose of this study was to determine the most suitable Nickel grade variogram model based on a root means square error (RMSE). The research location of this study is one of the national companies engaged in nickel mining. In its exploration activities, resource modelling and estimation play a very important role in validating geological data obtained from exploration activities. These geological data are modelled and estimated to gain the grade and tonnage of the resources.

The resource estimation can be carried out using the geostatistical method. It has the advantage of taking into account the spatial relationship of each data so that the continuity of the spread can be found
with the help of a variogram. A valid variogram is highly important to produce a correct estimation. Therefore, the variogram must be validated by the cross-validation method. In the variogram, it is necessary to determine the variogram model that best fits the real data conditions. For this reason, a RSME analysis is carried out to determine the most suitable variogram model.

The characterization of the RSME has been discussed for more than a century. However, it is continuing to grow. Several recent studies on this issue include, for example, the variogram analysis. Researchers have discussed it in various topics, including the moment method, extrapolation simulation, probability approach, and others. This analysis is also often applied in the field of medicine or genetics but is rarely used in the mining sector, especially in the estimation of mineral resources [2].

After obtaining the variogram model with the smallest error level, the values can be used in the kriging analysis. In the resource estimation with the geo-statistics, the method applied is ordinary kriging. In the ordinary kriging method, each data is related to one another and can predict the value of the location that does not have real data. Furthermore, this method can express the geometric characteristics of sediment and its distribution so that it can be used to estimate the resources. In this study, a RSME analysis is carried out to determine the most suitable variogram model, while the kriging estimation is to determine the tonnage of nickel resources.

2. Literature Review

2.1. Descriptive statistics

Descriptive statistics are related to describing and summarizing data. In describing data, it generally uses two methods: the measure of central tendency and the measures of variation [3]. The measure of central tendency is applied to describe the midpoint position of the frequency distribution from a data group, including the values of mean, median, and mode. The values of mean, median, and mode can provide information related to the central tendency of the data. Furthermore, this measure is also to find out the values of range, variance, and standard deviation. Those data can be analysed and visualized with tables and graphs, making it easy to analyse the distribution of the data. For example, with a histogram, researchers can know the value of the skewness of distribution, in which a positive value means that the peak is on the left and decreases to the right, while a negative value means the contrary. Furthermore, the normal skewness is in the range between -2 and 2. Meanwhile, for the level of frequency and its distribution (Kurtosis), the value of 0 indicates a normal state, the value above 0 indicates a high-frequency curve (centralized data), and the value below 0 indicates the gentle and scattered curve. Furthermore, the normal kurtosis is in the range value between -2 and 2 [4].

Another method in descriptive statistics is the measures of variation. It is to find out to what extent the data differ or deviate from their central value or measures of dispersion. Several types of dispersions are range, mean deviation, standard deviation, and coefficient of variation. The coefficient of variation (COV) can describe the uncertainty of a data set. The higher the COV is, the higher the uncertainty which affects the error of the estimate will be.

2.2. Spatial statistics

Geo-statistics accept the concept that each point in a deposit represents an example of a certain distribution in space. However, the distribution of other points may differ from one another in a deposit, both in terms of mean and variant. Mathematical models are the basis for assumptions in the deterministic relationship between sample locations and their variables. Despite that, no model can explain variations from natural phenomena. Furthermore, the use of interpolation and spatial techniques will result in errors. Geo-statistics plays a role in estimating the error based on a model with the un-bias principle and known variants or errors [5].

Ordinary kriging is based on the assumption of random variations and depends on the spatial aspect. The underlying random process is intrinsic stationery with constant means and variants that only depend on the separation of distances and directions between places, and are not on absolute positions [6]. The weighting method of kriging is based on distance and spatial correlation between sample points. Before estimating the resources using the kriging method, a variogram must be fitted first [7]. Fitting a variogram is the first step in geostatistical calculations to determine the spatial variation between data. In the variogram model, the obtained values are the sill variance, nugget effect, and range [8,9]. The sill variance is the average variant of the sample in general. Meanwhile, the nugget effect is the amount of
variation in a short distance. Furthermore, the range is the value describing the distance between interconnected data. A key element of a good assessment is not only to produce block grades but should also provide an indication of the extent to which these grades may differ from actual grades. The calculation commonly used in variograms is Matheron’s method of moments, using the following formula as shown in Equation 1.

\[ \hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{m(h)} (z(X_i) - z(X_i + h))^2 \]  

(1)

Where:
\( \hat{\gamma}(h) \): the estimated variogram for vector h
\( N(h) \): number of data pairs
\( h \): 1d, 2d, 3d (d = average spacing of data points)
\( z(x_i) \): data value in point i
\( z(x_i+h) \): data value in points joining h and xi

Before conducting estimation using the ordinary kriging method, it is necessary to determine the variogram model that is most suitable to be used as an estimation parameter [10]. The models of variogram commonly used are as shown in Equations 2-6 [11].

1. Nugget
\[ \hat{\gamma}(h) = \begin{cases} 0 & \text{if } h = 0 \\ \end{cases} \]  

(2)

2. Spherical model
\[ \hat{\gamma}(h) = \begin{cases} c - 0.5 \left( \frac{h}{a} \right)^3 & \text{if } h \leq a \\ \end{cases} \]  

(3)

3. Exponential Model
\[ \hat{\gamma}(h) = c \left( 1 - \exp \left( -\frac{3h}{a} \right) \right) \]  

(4)

4. The Gaussian Model
\[ \hat{\gamma}(h) = c \left( 1 - \exp \left( -\frac{3h^2}{a} \right) \right) \]  

(5)

5. Power
\[ \hat{\gamma}(h) = c \cdot h^\omega \quad 0 < \omega < 2 \]  

(6)

Where:
\( \hat{\gamma}(h) \): the estimated variogram for vector h
\( h \): the range of lag
\( a \): the value of the range
\( c \): the value of sill
\( \omega \): the form parameter

2.3. Cross-Validation

Cross-validation is used to test the validity of the variogram model to produce accurate kriging results. Cross-validation allows its user to compare the estimated value with the real value using the information available in the sample data set. At first, the sample value is temporarily omitted from the sample data set. After that, these values are estimated using the remaining sample. The estimation is then compared with the real value so that the average error value, the correlation coefficient value, and the linear regression value can be found out. Because of that, the user can be found out how close the estimation results are to the sample value. In geostatistical analysis, cross-validation was first performed by Clark [12], in which the results show that if cross-validation is used to compare estimation methods, it can justify the use of kriging as an estimation method. Clark pointed out that using cross-validation to select semivariogram models may be acceptable and very useful, but perhaps not sensitive enough [13].
2.4. Kriging
Kriging is a geostatistical method for interpolating the values of different regional variables, which consists of ordinary kriging (OK), universal kriging, indicator kriging, co-kriging, and others [14]. Ordinary kriging provides an unsampled point estimate based on the weighted average of the neighbouring points observed in a certain area. Ordinary kriging is part of kriging, determined based on the variogram spatial structure parameters that measure the relationship between the squared differences of paired samples and their range. The ordinary kriging method takes data samples and models the relationship between the variance in value and the inter-point range. The spatial autocorrelation between measured sample points is examined using a semivariogram/covariance. Semivariograms are used to model the spatial relationships in the dataset and to find the most suitable model [15].

The equation for ordinary kriging is presented in Equation 7, where: \( \hat{Z}(s_0) \) is the predicted location, \( \lambda_i \) is the unknown weight of the measured point pair at location \( i \), \( Z(s_i) \) is the measured value of the pair of points at the \( i \)-th location, and \( N \) is the sum of the measured values of the pair of points multiplied by the range \( h \).

\[
\hat{Z}(s_0) = \sum_{i=0}^{N} \lambda_i Z(s_i)
\] (7)

2.5. Root mean square error (RMSE)
Root mean square error (RMSE) has been widely used as a standard statistical metric in measuring the performance of a model in certain studies, which in this case is the error model for variograms in geostatistical studies. In geoscience, it is common to find the use of RMSE as a standard metric to make model errors [16–19].

This RMSE is also known as root mean square deviation (RMSD). The value of RMSE can be used to determine the type of variogram model that is most suitable by paying attention to the resulting error value. The variogram model which has the smallest RMSE value will be used as a variogram model for kriging estimation. Calculation of the RMSE can be conducted using Equation 8, presented as follows [20].

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \hat{Z}(x_i) - Z(x_i) \right)^2}
\] (8)

Where:
- \( \hat{Z}(x_i) \) : estimated value
- \( Z(x_i) \) : sample value
- \( n \) : number of samples

RMSE describes the short-term performance of the model by looking at the actual difference between the estimated value and the measured value [21]. The smaller the value is, the better the performance of the model will be. The weakness of this calculation is that several large errors in the obtained value can result in a significant increase in RMSE. In other words, the test does not differentiate between underestimation and overestimation [11].

3. Methods
The methods used in this study were descriptive and quantitative. The descriptive method describes a condition of an event that is influenced by a factor so that the event may occur. In this study, the descriptive method was carried out by examining the data of assay, lithology, collar, and core survey of the exploration results in the research location. Meanwhile, in the quantitative method, secondary data were processed into models, which were carried out mathematically by software. The research stages applied in this study were as follows.

3.1. Collecting and evaluating initial data
In this study, the results of nickel exploration drilling were used. The data were taken from a nickel company whose mining business license area is administratively located in Morowali, Central Sulawesi.
3.2. Defining testing scenarios

3.2.1. Descriptive statistics test. This test was to get a description of the data from the values of mean, maximum-minimum, range, standard deviation, skewness, kurtosis, variance, and coefficient of variation. After that, the normality test and the goodness-of-fit test were carried out.

3.2.2. Testing and analysing the variograms. Variograms were used to describe spatial variability. The type of variogram chosen in this study was the experimental variogram. Furthermore, modelling and fitting were carried out based on the shape of the variogram. To test the validity of the data, the cross-validation technique was used to examine the results of the variogram calculation.

The estimated value was obtained from the kriging method towards the candidate variogram models. In this study, the researcher used ordinary kriging. In addition, this study also examined different search parameters. A plausible set of scenarios, consisting of a variogram model, type of kriging, and search strategy, were used in the cross-validation test.

3.3. Cross-validation

Cross-validation was carried out for the search scenario and then its results were compared. Each sample point was left out of the data set in turn and was estimated with the remaining data and the appropriate scenario parameters. In the estimation, errors were calculated as the estimated value minus the actual value. Meanwhile, the expected result was that the probabilistic model is accurate and precise while minimizing uncertainty.

3.4. Analysing root mean square errors (RMSE)

RMSE was calculated by squaring the errors (the difference between the estimated value and the actual value), then finding its mean by summing up all squared errors and then dividing it by the number of data (n).

3.5. Estimating and classifying nickel resources

The estimation of nickel resources was conducted using ordinary kriging, while its classification was by the kriging efficiency (KE).

4. Results and discussion

4.1. Data

The data used in this study was the result of exploration drilling for nickel laterite in the western block of the research location. Those drilling data were from 102 drill holes with an average drill hole spacing of ± 50 meters and an average depth of 24.91 meters. Those holes were spread over the coordinates 9662272.288 N - 9662598.293 N and 414951.271 E - 415997.8 E (Figure 1).

Figure 1. Research location map and distribution of drill holes.
4.2. Basic statistical analysis

The values of nickel contents from the drilling data were analysed using basic statistical analysis to determine the mean, median, standard deviation, skewness, kurtosis, and coefficient of variance (CoV). From data processing, the mean of nickel elements was 0.987. Its median in the limonite zone was 0.93, meaning that 50% nickel content was more than 0.93 and the rest 50% was less than 0.93. Furthermore, its standard deviation in this zone was 0.31, meaning that the nickel content in the limonite zone was not spread far from its mean, in which the standard deviation is a measure of the value of the data distribution to the mean value. Apart from that, the value of skewness was 1.59, indicating the asymmetry of the data distribution. The skewness value found in this study was positive, illustrating that the nickel data distribution curve was more right-leaning. The kurtosis value is a measure that describes the ductility of the data distribution. In this study, the nickel content in the limonite zone had a kurtosis value of 8.37, which means that the data has a high peak distribution (Figure 2a). The coefficient of variance in nickel contents in the limonite zone was 0.31, meaning that the nickel content in this zone was normal and homogeneous because the CoV value was < 0.5.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure2.png}
\caption{(a) Histogram graph for nickel contents in limonite zone, and (b) Histogram graph for nickel contents in saprolite zone.}
\end{figure}

4.3. Spatial statistical analysis

In the experimental variogram for the nickel limonite layer, the applied parameter was a lag of 65 with a spread of 22.5° and the direction with a maximum range value of 92° as shown in Figure 3a. Meanwhile, in the experimental variogram for the nickel saprolite layer, the applied parameter was a lag of 65 with a spread of 22.5° and the direction with a maximum range value of 90° as shown in Figure 3b.
In the limonite layer, the range of nickel content that still had an effect was 134.094 meters. The average value of variation (sill) in nickel limonite contents was 0.079, while a variation value at close range (nugget effect) was 0.02. The ratio of the range of the major axis to the semi-major axis was 1.244 so that the range on the semi-major axis was 107.9 meters. Meanwhile, the ratio of the range of the major axis to the minor axis was 7.554 so that the range on the minor axis was 17.8 meters (Figure 4a).

In the saprolite layer, the experimental variogram of nickel content was also modelled. After conducting the modelling, it was found that the continuity or distance of nickel data that was still influential (range) was 167.837 meters. The average value of variation (sill) in nickel saprolite contents was 0.38, while a variation value at close range (nugget effect) was 0.01. The data of sill provided information that the nickel content in the saprolite layer varied more than the nickel content in the limonite zone. The ratio of the range of the major axis to the semi-major axis was 2.247 so that the range on the semi-major axis was 74.5 meters. Meanwhile, the ratio of the range of the major axis to the minor axis was 15.652 so that the range on the minor axis was 14.4 meters (Figure 4b).

4.4. Cross-validation
The results of cross-validation for nickel contents in the limonite layer resulted in a mean error value of 0. The correlation coefficient was 0.79, which is high. Furthermore, the gradient value from linear regression was 0.6 (Figure 5a). Meanwhile, the results of cross-validation for the nickel contents in the saprolite layer resulted in a mean error value of -0.01. The correlation coefficient was 0.85, which is high. Furthermore, the gradient value from linear regression was 0.72 (Figure 5b). The value of the mean error was small, the value of the correlation coefficient was high, and the gradient value of the linear regression was close to 1. Those indicated that the results of the variogram were valid and can be used as a parameter in the estimation.
4.5. Root mean square error (RMSE)

The value of root mean square error was used to determine the most suitable variogram model to be used as a parameter in the estimation \cite{22}. A root mean square error (RMSE) is the magnitude of the error rate of the prediction results, where the smaller (close to 0) the RMSE value is, the prediction results will be more accurate \cite{23}. From the root mean square error analysis results, the most suitable variogram model for nickel limonite and nickel saprolite was the exponential model, in which the RMSE values for nickel limonite and nickel saprolite were 0.022 and 0.098, respectively. Table 1 shows the results of the RMSE analysis for nickel content in the limonite and saprolite layers.

Table 1. Root mean square with variogram model for nickel contents in limonite and saprolite layers.

| Zone   | Variable | Variogram Model | Nugget | Sill  | Spatial Ratio (Nugget/Sill) (%) | RMSE  |
|--------|----------|-----------------|--------|-------|---------------------------------|-------|
| Limonite | Ni       | Spherical       | 0.02   | 0.08  | 25                              | 0.023 |
|         |          | Exponential     | 0.02   | 0.08  | 25                              | 0.022 |
|         |          | Gaussian        | 0.03   | 0.07  | 43                              | 0.024 |
| Saprolite | Ni      | Spherical       | 0.03   | 0.36  | 8.3                             | 0.101 |
|         |          | Exponential     | 0.01   | 0.38  | 2.6                             | 0.098 |
|         |          | Gaussian        | 0.08   | 0.31  | 26                              | 0.138 |

5. Conclusions

Fitting the right variogram model is a must, which will later be used as input data in resource estimation (kriging) with a minimum error rate. This study shows that, after conducting trial and error on the variogram, the best summary statistics are finally fitted. From the modelling results, the obtained values for effect nugget, sill, and spatial ratio were 0.02, 0.08, and 25% respectively. The result of the correlation coefficient was high, namely 0.85. Furthermore, the gradient value from linear regression was 0.72. In other words, the value of the mean error was small, the value of the correlation coefficient was high, and the gradient value of the linear regression was close to 1. Therefore, the result of the variogram was considered to be valid. The variogram model that was the most suitable for nickel content in the limonite and saprolite layers was the exponential model with RMSE values of 0.022 and 0.098 respectively.

References

\[1\] Ministry of Energy and Mineral Resources 2020 Downstream processing creates added value and strengthens economy [www.esdm.go.id]
\[2\] Gustafsson C 1996 Identification of new RNA modifying enzymes by iterative genome search using known modifying enzymes as probes *Nucleic Acids Res.* 24 3756–62
\[3\] Yi G Y 2017 Statistical analysis with measurement error or misclassification: Strategy, method
and application (New York, NY: Springer New York)

[4] Schabenberger O and Gotway C A 2017 Statistical Methods for Spatial Data Analysis (Boca Raton: Chapman and Hall/CRC)

[5] Oliver M A and Webster R 2015 Basic steps in geostatistics: The variogram and kriging (Cham: Springer International Publishing)

[6] Marwanza I, Azizi M A, Nas C, Permatasari C I, Dahani W and Subandrio 2019 The use of geostatistics in setting the spread of claystone in coal mines Int. J. Sci. Technol. Research 8 3607–12

[7] Marwanza I, Dahani W, Subandrio, Azizi M A, Kurniawati R and Farhan I 2020 Variogram modeling of lime saturation factor on limestone quarry Int. J. Sci. Technol. Research 9 1679–84

[8] Desassis N and Renard D 2013 Automatic variogram modeling by iterative least squares: Univariate and multivariate cases Math. Geosci. 45 453–70

[9] Li Z, Zhang X, Clarke K C, Liu G and Zhu R 2018 An automatic variogram modeling method with high reliability fitness and estimates Comput. Geosci. 120 48–59

[10] Bohling G 1005 Introduction to geostatistics and variogram analysis (Kansas)

[11] Clark R D 2003 Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics J. Comput. Aided. Mol. Des. 17 265–275

[12] Clark I 1986 The art of cross-validation in geostatistical application 19th Application of Computers and Operations Research in the Mineral Industry ed R V. Ramani (Pennsylvania: Society of Mining Engineers, Inc) pp 211–20

[13] David M 2012 Geostatistical ore reserve estimation (New York, NY: Elsvier)

[14] Narciso J, Araújo C P, Azevedo L, Nunes R, Costa J F and Soares A 2019 A Geostatistical Simulation of a Mineral Deposit using Uncertain Experimental Data Minerals 9 247

[15] McKeen S, Wilczak J, Grell G, Djalalova I, Peckham S, Hsie E-Y, Gong W, Bouchet V, Menard S, Moffet R, McHenry J, McQueen J, Tang Y, Carmichael G R, Pagowski M, Chan A, Dye T, Frost G, Lee P and Mathur R 2005 Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004 J. Geophys. Res. 110 D21307

[16] Savage N H, Agnew P, Davis L S, Ordóñez C, Thorpe R, Johnson C E, O’Connor F M and Dalvi M 2013 Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation Geosci. Model Dev. 6 353–72

[17] Chai T, Kim H-C, Lee P, Tong D, Pan L, Tang Y, Huang J, McQueen J, Tsidulko M and Stajner I 2013 Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO2 measurements Geosci. Model Dev. 6 1831–50

[18] Buonaccorsi J P 2010 Measurement error: Models, methods, and applications (Boca Raton: CRC Press)

[19] Ouabo R E, Sangodoyin A Y and Ouondiran M B 2020 Assessment of ordinary kriging and inverse distance weighting methods for modeling chromium and cadmium soil pollution in E-Waste Sites in Douala, Cameroon J. Heal. Pollut. 10 200605

[20] Kalogirou S A 2012 Solar thermal systems: Components and applications – introduction Comprehensive Renewable Energy (Elsevier) pp 1–25

[21] Chai T and Draxler R R 2014 Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature Geosci. Model Dev. 7 1247–50

[22] Wang W and Lu Y 2018 Analysis of the mean absolute error (MAE) and the root Mean square error (RMSE) in assessing rounding model IOP Conf. Ser. Mater. Sci. Eng. 324 012049

[23] Willmott C J and Matsuura K 2005 Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance Clim. Res. 30 79–82

Acknowledgments
This research was supported by the Department of Mining Engineering, Faculty of Earth and Energy Technology, Trisakti University, Jakarta, Indonesia. In addition, the researchers would like to thank the editors & reviewers who provided useful and constructive input. Furthermore, the researchers are also
grateful to the organizer of the 1st International Seminar of Mineral and Coal Technology (ISMCT 2021).