VOL. 11, NO. 3, FEBRUARY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SEMIVARIOGRAM MODELING USING MIXTURE SEMIVARIOGRAM MODEL

K. N. Sari[1], O. Neswan[1], U. S. Pasaribu[1] and A. K. Permadi[2]
[1]Department of Mathematics and Natural Sciences, Bandung Institute of Technology, Indonesia
[2]Department of Petroleum Engineering, Bandung Institute of Technology, Indonesia
E-Mail: kurnia@math.itb.ac.id

ABSTRACT

In semivariogram modeling, data characteristics greatly influence the steps involved in the semivariogram modeling. Homogeneous data will be better modeled with simple semivariogram models such as: exponential, Gaussian, and spherical models. Meanwhile, heterogeneous data are expected to be better modeled with a mixture semivariogram models. A mixture model is a combination of several simple semivariogram models of a certain proportion. The proportion for each model can be determined from the mean squared error (MSE). If the MSE value is smaller, then the proportion of the corresponding simple models will be greater. Even though the mixture is more complicated, the model can be an alternative in semivariogram modeling which allows to give MSE values that is not much different than MSE values yielded by using the simple models.

Keywords: mixture model, least square, semivariogram, proportion.

1. INTRODUCTION

In geostatistics, we observe the sequence of random variables $\{Z(s)|s \in D\}$ where s is set of locations and D is random set in R^d, d = 1,2,3. Z(s) is a stochastic process with index parameters such as location. The developments of stochastic process in spatial modeling quite rapidly such as Markov process. Integrating Markov decision processes (MDP) with geographic information system (GIS) was used to examine the financial optimality of floods disaster risk reduction in Queensland, Australia in 2010/2011 [1]. Markov approximations method also used for spatial prediction (kriging) and gave more accuracy compared with the covariance tapering and the process convolution method [2].

One tool to measure the variance from the difference between two spatial locations that are separated by a certain distance. Semivariograms can be classified based on the presence and the absence of the influence of the angle between a pair of locations; respectively called anisotropic and isotropic semivariogram. The model has 3 parameters, i.e. nugget effect, sill, and range. Nugget effect is the initial semivariance when the autocorrelation is at its highest or just the uncertainty where distance (d) is close to 0, sill is the horizontal assymptote of the semivariance, and range is lag distance where the sill is reached. There are 7 simple semivariogram models, i.e. nugget effect, linear, spherical, exponential, power functions, Gaussian, and hole effect [3]. The models and the parameters are illustrated in Figure-1.

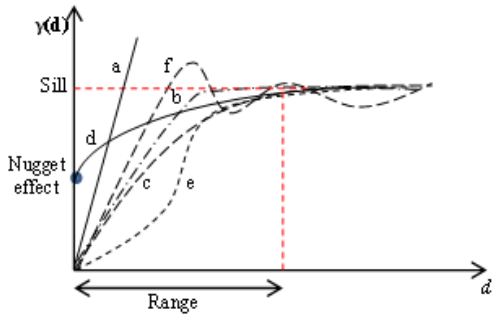


Figure-1. Plot of semivariogram models ($\gamma(d)$) and their parameters (nugget effect, sill, and range). There models are a. Linear, b. Spherical, c. Exponential, d. Power function, e. Gaussian, f. Hole effect [3].

There are several other models that were developed from the models above as cubic model, prismato-magnetik, and prismato-gravimetrik. That model has similar properties the parabolic nature at the point of origin with Gaussian model. The cardinal sine have periodic properties, similar with hole effect model.

The main problems in semivariogram modeling are to estimate the value of the model parameters and to decide which model would be the most suitable model for the data. Iguzquiza and Dowd [4] compared the inference methods for estimating semivariogram model parameters and their uncertainty for the case of small data sets. To avoid subjectivity in fitting models to experimental semivariograms, ordinary least squares, weighted least squares, and generalized least squares are often used. Uncertainty evaluation in this indirect method is done using computationally intensive resampling procedures such as the bootstrap method.

Moreover, Sari *et al* [3] used bootstrap least square method for estimating the parameter vector of the semivariogram models. This method will be applied to estimate the parameters after resampling the errors of the model. The selection of the resulting semivariogram models from bootstrap method will be affected by the number of distance lags, the precision level of the range partitions, the number of bootstrap iterations, and the given reference model. The estimation with bootstrap method with the same model as the reference converges faster with the maximum iteration of 50. The exponential and Gaussian models are sufficiently good in the estimation for the models with the same references. Meanwhile, the estimation yielded from spherical model is quite far from the reference exponential and Gaussian models.

However, in the previous modeling, the real data hasn't been used at all, so that the above three models can be less appropriate to the existing data which is usually indicated by the increase in the value of the mean squared error (MSE). Waterman [5] used different method to analyze the data of gold and copper ore deposits at Graspberg, Papua, which has outliers and not symmetrical. He proposed weighted jackknife-ordinary kriging in the estimation of the deposits.

In this paper, we propose the mixture semivariogram model which is a liniear combination of some simple models such as exponential, Gaussian, and spherical. This model uses proportion for each simple model, which is determined by the value of the MSE from each estimated model. Even though this mixture model is more complicated, it is expected to give better result, along with the MSE value which is smaller than when the simple models are used.

2. MIXTURE SEMIVARIOGRAM MODELING

2.1 Notations

The following notations will be used in this paper:

d - Distance between pair of locations

$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$ - Parameters vector (nugget effect, sill, and range)

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)'$ - Estimated model-parameters vector (nugget effect, sill, and range estimator)

$\hat{\gamma}(d) = \begin{bmatrix} \hat{\gamma}(d_1) \\ \vdots \\ \hat{\gamma}(d_n) \end{bmatrix}$ - Experimental semivariogram

$\gamma(d, \boldsymbol{\theta}) = \begin{bmatrix} \gamma(d_1, \boldsymbol{\theta}) \\ \vdots \\ \gamma(d_n, \boldsymbol{\theta}) \end{bmatrix}$ - Semivariogram model

p_j - The proportion of the model j

2.2 Parameters estimation

In this paper, the experimental variogram from the real data will be fitted to some semivariogram models, that are exponential (exp), Gaussian (gauss), and spherical (sph). These models are respectively formulated as follow:

$$\hat{\gamma}_{\exp}(d) = \hat{\theta}_0 + \hat{\theta}_1 \left(1 - \exp\left(-\frac{d}{\hat{\theta}_2}\right)\right) \quad (1)$$

$$\hat{\gamma}_{\text{gauss}}(d) = \hat{\theta}_0 + \hat{\theta}_1 \left(1 - \exp\left(-\left(\frac{d}{\hat{\theta}_2}\right)^2\right)\right) \quad (2)$$

$$\hat{\gamma}_{\text{sph}}(d) = \hat{\theta}_0 + \hat{\theta}_1 \left(1.5\left(\frac{d}{\hat{\theta}_2}\right) - 0.5\left(\frac{d}{\hat{\theta}_2}\right)^3\right) \quad (3)$$

In general, $\hat{\theta}_0$ dan $\hat{\theta}_1$ are formulated respectively as $\hat{\theta}_0 = \frac{a}{n} - \hat{\theta}_1 \frac{b_j}{n}$ and $\hat{\theta}_1 = \frac{c_j - \frac{ab_j}{n}}{d_j - \frac{(b_j)^2}{n}}$, where $a = \sum_{i=1}^{n} \gamma(d_i)$. Meanwhile, b_j, c_j, and d_j are in accordance with the model of the estimates (j = 1 (exp), 2 (Gauss), and 3 (sph)). For exponential model, $b_1 = \sum_{i=1}^{n} \left(1 - e^{-\hat{\alpha} d_i}\right)$, $c_1 = \sum_{i=1}^{n} \gamma(d_i)\left(1 - e^{-\hat{\alpha} d_i}\right)$, and $d_1 = \sum_{i=1}^{n} \left(1 - e^{-\hat{\alpha} d_i}\right)^2$ with $\hat{\alpha} = \frac{1}{\hat{\theta}_2}$. While for gaussian model, $b_2 = \sum_{i=1}^{n} \left(1 - e^{-(\hat{\alpha} d_i)^2}\right)$, $c_2 = \sum_{i=1}^{n} \gamma(d_i)\left(1 - e^{-(\hat{\alpha} d_i)^2}\right)$ and $d_2 = \sum_{i=1}^{n} \left(1 - e^{-(\hat{\alpha} d_i)^2}\right)^2$. And for spherical model, $b_3 = \sum_{i=1}^{n} \left(1.5(\hat{\alpha} d_i) - 0.5(\hat{\alpha} d_i)^3\right)$, $c_3 = \sum_{i=1}^{n} \gamma(d_i)\left(1.5(\hat{\alpha} d_i) - 0.5(\hat{\alpha} d_i)^3\right)$ and $d_3 = \sum_{i=1}^{n} \left(1.5(\hat{\alpha} d_i) - 0.5(\hat{\alpha} d_i)^3\right)^2$.

From each model estimates, we can calculate the mean squared errors, MSE, i.e. the average of the squares of the difference between the experimental semivariogram and the semivariogram model. In addition, there will be semivariogram modeling using mixture model, which can be written as:

$$\hat{\gamma}^*(d) = p_1 \hat{\gamma}_{\exp}(d) + p_2 \hat{\gamma}_{\text{gauss}}(d) + p_3 \hat{\gamma}_{\text{sph}}(d) \quad (4)$$

where p_j, j = 1 (exp), 2 (Gauss), 3 (sph) is the proportion for each model.

Those proportions are determined from two steps, those are simulating using reference model (exp, Gauss, and sph) and approaching real data using experimental semivariogram. Through simulation, the determined reference model will be fitted to 3 simple semivariogram models then some proportion combinations from each model are determined. The purpose is to know which proportion combination will give the smallest MSE value. Meanwhile, for the real data approach, the proportion is determined based on the MSE values of each simple model. The model with smaller MSE value will contribute a greater proportion for the mixture model.

3. RESULTS AND DISCUSSIONS

3.1 Simulation

From simulation, the proportions in mixture model are determined by simulating the reference models (exp, gauss, and sph). For that, the following steps are executed:

a. Consider one of the reference semivariogram models: (1) exp, (2) gauss, or (3) sph with the parameters vectors respectively $\boldsymbol{\theta} = \left(\theta_{0\text{ref}}, \theta_{1\text{ref}}, \theta_{2\text{ref}}\right)$.
b. Select p for forming semivariogram function.
c. Compute the expected semivariogram $\hat{\gamma}(d_i), i = 1, 2, ..., m$.
d. Take ε_i from the normal distribution with mean 0 and variance 1.
e. Add $\varepsilon_i, i = 1, 2, ..., m$ to $\hat{\gamma}(d_i)$.
f. Estimate the parameter vector, i.e. $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right)$ by using least square method with the semivariogram models, with respectively: (1) exp, (2) gauss, or (3) sph as the estimation models.
g. Construct 3 semivariogram models $\hat{\gamma}^*(d_i)$ (exp, gauss, and sph) with $\hat{\boldsymbol{\theta}}$ as the input for f.
h. Compute the mean square error (MSE), formulated as $\frac{1}{m}\sum_{i=1}^{n}\left(\hat{\gamma}^*(d_i) - \hat{\gamma}(d_i)\right)^2$.
i. Determine the proportions p_1, p_2, dan p_3 for exp, gauss, and sph models.
j. Formulate the mixture model as a linear combination of the simple semivariogram models as written in (2).
k. Compute MSE for the mixture model, which is formulated as $\frac{1}{m}\sum_{i=1}^{n}\left(\hat{\gamma}^*(d_i) - \hat{\gamma}(d_i)\right)^2$.
l. The best model is the one with the smallest MSE.

With the simulation steps above, the following is the simulation result in determining the proportions for three simple semivariogram models in constructing mixture model, along with the corresponding MSE values.

Table-1. The simulation results of the proportions of the simple semivariogram models in constructing mixture model.

Ref. Model	MSE			Mixture Model	
	Exp	Gauss	Sph	Ratio	MSE
Exp	22	325	1248	1:1:1	84
	113	173	835	2:2:1	**22**
	129	192	861	1:1:0	114
	100	242	996	2:1:0	88
	38	299	1233	1:1:2	225
Gauss	1195	245	7500	1:1:1	1338
	681	560	4215	2:2:1	158
	693	623	3999	1:1:0	529
	1204	266	7312	1:2:0	307
	898	221	6397	1:1:2	1628
Sph	3418	5035	1268	1:1:1	2220
	2724	6370	2762	1:1:2	1542
	2971	4263	1010	1:1:3	1243
	3428	5052	1292	2:1:2	1828
	3179	6153	2376	4:1:1	2444
	3531	5348	1481	3:1:2	2041
	3562	5269	1347	2:1:3	1597

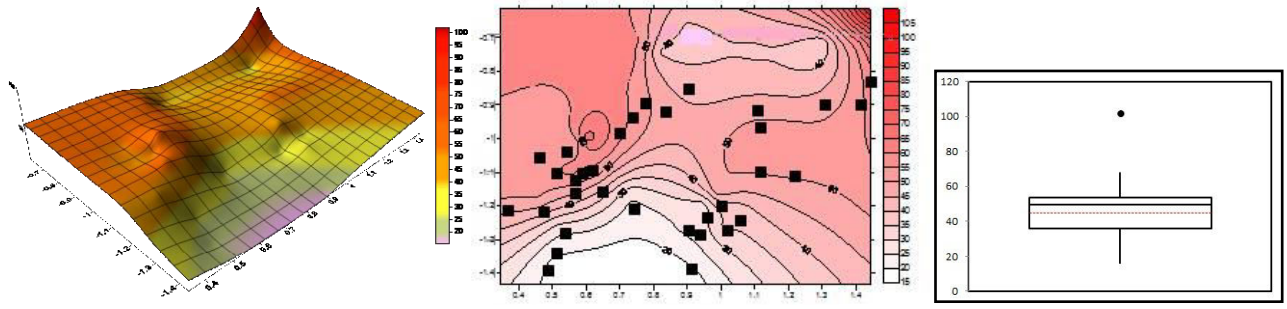
From Table-1, it can be concluded that:

a. From the three semivariogram models above, the exponential one is the simplest model mathematically, followed by the Gaussian which is a square form of exponential model? Spherical is a polynomial model with the orders of 1 and 3, different with the two previous models; in other words, it can be considered to be more complicated.
b. By assigning greater proportion to the original model, the MSE value will be smaller. For the case when the reference model is spherical, the greatest proportion for spherical model will give the smallest MSE value. In line with this, if the reference models are exponential and Gaussian, the greater proportion for the same references will give the smaller MSE.
c. A simple model can be fitted with the more complicated model, but it can increase the MSE value. As a consequence, a simple model will be better fitted by another simple model, or mixture model with the smaller proportion for the more complicated model.
d. Meanwhile, a complicated model will be better fitted by a mixture model where the proportion of the same model is greater than the proportions of the simpler models. In estimating more complicated model, a high proportion for simple models will increase the MSE.
e. The mixture model is surely more complex than the three other models, but it is a new idea in modeling the semivariogram given many software that facilitates the determination of the model parameters. Moreover in general, it can provide a smaller MSE value unless it is fitted with a spherical model.

3.2 Data

The data used will be the reservoirs at Jatibarang field. Jatibarang reservoir has special characteristics. There are volcanic stones with fractures and low sulfur content. The volcanic layer is the largest oil producer among the Jatibarang reservoirs. This reservoir is located in the north of West Java and the oil field area has an elongated position $\pm$10 kilometers north-south and $\pm$16 kilometers west-east. Since 1969, there have been $\pm$200 opened and in 1998 the production reached a cumulative production of nearly 13 million m^3 [6].

About 132 wells were observed, and the information about the depths and *k-fracture* that show the oil permeability were provided. The data will be used is *k-fracture* information. The wells with sufficiently homogeneous *k-fracture* conditional to the wells' depth were selected. From 132 wells, 33 wells were taken at a depth of (271,326] with the contour of *k-fracture values* and its projections to the ground surface illustrated respectively in Figure-2a and Figure-2b. The descriptive statistics of the data are summarized in Table-2 and outliers are shown in the boxplot in Figure-2.



(a) (b) (c)

Figure-2. (a) 3D-contour and (b) 2D-contour of the 33 selected wells' *k-fracture* at a depth of (210, 270).

Table-2. Numerical summary of the 33 selected wells' *k-fracture* at a depth of (210, 270).

Descriptive statistics			
Count	33	Variance	287.46
Sum	1483.46	Standard Error	2.95
Average	44.95	Skewness	0.83
Minimum	16.09	Kurtosis	2.68
Maximum	102.01	25th Percentile	35.87
Range	85.92	50th Percentile	49.73
Standard Deviation	16.95	75th Percentile	53.78

From the numerical summary, the wells has the *k-fracture* average 44.95 with adequately high dispersion, 16.95. there are 23 wells (68% of 33 wells) having *k-fracture* included on the interval (28.0, 61.9). From Figure-2c, there is a well having the highest *k-fracture,* reaching 102.01, behaving as an outlier. From above contour, the *k-fracture* increases along with the direction of northeastern, and decreases along with the direction of southern. Then the *k-fracture* data from the 33 wells will be processed to determine the best semivariogram model.

3.3 Semivariogram modeling

After getting the descriptive statistics, semivariogram modeling is conducted by doing these steps:

a. Compute the experimental semivariogram with the number of the distance pairs from 33 wells determined from Sturgess' rule, so that the pairs of distances and experimental semivariogram (d_i, γ_i), i = 1, 2,...,n where n represents the number of pairs.

b. Estimate the parameter vectors $\hat{\boldsymbol{\theta}}_j = \left(\hat{\theta}_{0j}, \hat{\theta}_{1j}, \hat{\theta}_{2j}\right)'$ using least square method for j-th model where j = 1, 2,...,r and r represents the number of semivariogram models to be fitted. Here are 3 models to be used, those are 1(exp), 2 (gauss), dan 3 (sph).

c. Compute MSE from the difference between experimental semivariogram and each model using the following formula:

$$\text{MSE}_j = \frac{\sum_{i=1}^{n}\left(\gamma_i - \hat{\gamma}_{ji}\right)^2}{n}.$$

d. Compute the proportion of each model (p_j) for the mixture semivariogram model by assigning smaller proportion for the bigger MSE value, so that the proportion can be formulated as the following:

$$k_j = \frac{\sum_{j=1}^{m}\text{MSE}_j}{\text{MSE}_j} \text{ and } p_j = \frac{k_j}{\sum_{j=1}^{m} k_j}, j = 1,2,3.$$

e. Compute MSE from the difference between the experimental semivariogram and the mixture model,

$$\text{MSE}_{\text{mix}} = \frac{\sum_{i=1}^{n}\left(\hat{\gamma}_i - \hat{\gamma}_{\text{mix}_i}\right)^2}{n}.$$

f. Compare the MSE values between the semivariogram modeling with simple models and mixture model.

3.4 Results

The estimation result of parameter vector and the comparation graphs between the experimental semivariogram and its corresponding models values can be seen in Table-3 and Figure-3. Meanwhile, the estimation of the models' parameters and their MSE values can be seen in Table-4.

Table-3. The experimental semivariogram and the corresponding semivariogram model.

d	γ(d)	γ exp	γ gauss	γ sph	γ mix
0.0639	45.7	-232.1	-88.7	-358.7	-204.1
0.1916	148.9	23.4	2.6	1.8	10.4
0.3193	238.2	232.3	164.2	316.3	224.4
0.4470	256.4	403.2	362.4	554.4	421.1
0.5747	236.6	542.8	561.7	685.3	582.2
0.7024	311.1	657	734.9	701.1	697.6
0.8301	517.9	750.3	868.4	701.1	785.7
0.9578	778.2	826.7	960.7	701.1	851.2
1.0855	843.1	888.8	1018	701.1	897.5
1.2132	1778.5	939.9	1050.8	701.1	929.9

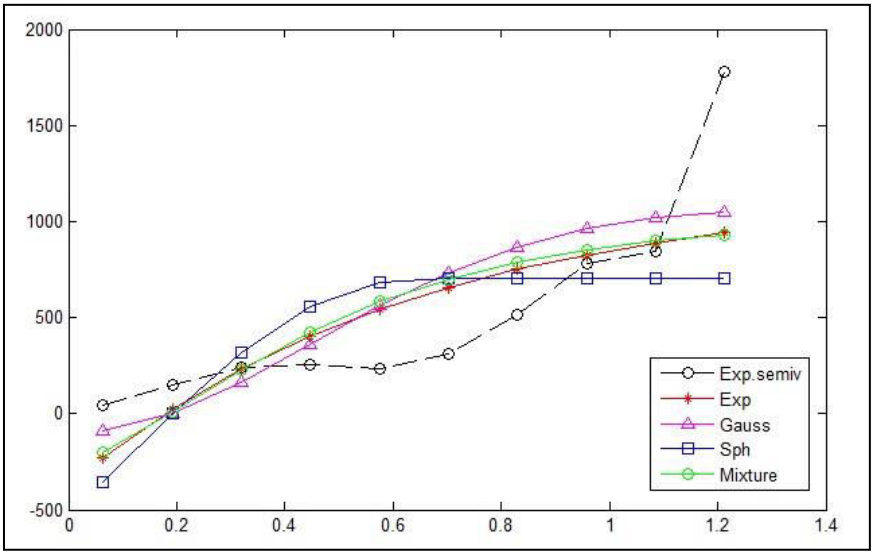


Figure-3. The Experimental Semivariogram and Semivariogram Model (exponential, Gaussian, spherical, and mixture).

Table-4. The parameter vector and MSE for three semivariogram models (exp, gauss, sph, and mixture).

The parameter vector and MSE value			
Exp	**Gauss**	**Sph**	**Mixture**
(-380.7, 1549.2, 0.60) 1.0896 x 10^5	(-100.6, 1181.7, 0.60) 1.0578 x 10^5	(-546.8, 1247.9, 0.60) 1.8540 x 10^5	- 1.1779 x 10^5

These 4 models in Table-4 can be written as follows:

$$\hat{\gamma}_{\exp}(d) = -380.7 + 1549.2\left(1 - \exp\left(-\frac{d}{0.6}\right)\right) \quad (3.1)$$

$$\hat{\gamma}_{\text{gauss}}(d) = -100.6 + 1181.7\left(1 - \exp\left(-\left(\frac{d}{0.6}\right)^2\right)\right) \quad (3.2)$$

$$\hat{\gamma}_{\text{sph}}(d) = \begin{cases} -546.8 + 1247.9\left(1.5\left(\frac{d}{0.6}\right) - 0.5\left(\frac{d}{0.6}\right)^3\right), d \le 0.6 \\ -546.8 + 1247.9 \qquad\qquad\qquad\qquad , d > 0.6 \end{cases} \quad (3.3)$$

$$\hat{\gamma}_{\text{mix}}(d) = 0.3820\,\hat{\gamma}_{\exp}(d) + 0.3935\,\hat{\gamma}_{\text{gauss}}(d) + 0.2245\,\hat{\gamma}_{\text{sph}}(d) \quad (3.4)$$

From Figure-3, there are 10 groups of distance lags obtained. Based on the model-fitting result, the exponential model has the least MSE value among all the other models used (gauss, sph, and mixture), reaching 1.0896x10^5. Meanwhile, the mixture model has MSE, i.e. 1.1779x10^5, bigger than the MSE from gaussian model for about 1.0578x10^5. But, the difference of MSE between mixture model and exponential model are not much different. Meanwhile, spherical model have the greatest MSE value. So, this model is less well in modeling the experimental semivariogram. The mixture model is formed from three simple models with the largest proportion is a Gaussian model, i.e. 0.3935, then followed by exponential and spherical model.

4. CONCLUSIONS

Mixture model can be an alternative in semivariogram modeling, which gives MSE values that is not much different from exponential and Gaussian model. For any data that can be modeled with simple models, mixture model can give smaller MSE value with a slight difference. This makes simple model more preferable than the more complicated mixture model.

REFERENCES

[1] R. Espada, A.Apan, and K.McDougall. 2014. Spatial Modelling of Natural Disaster Risk Reductiion Policies with Markov Decision Processes. Applied Geography. 53: 284-298.

[2] D. Bolin and F. Lindgren. 2013. A Comparison between Markov Approximations and Other Methods

for Large Spatial Data Set. Computational Statistics and Data Analysis. 61: 7-21.

[3] K.N. Sari, et al. 2015. Estimation of the Parameters of Isotropic Semivariogram Model through Bootstrap, Journal of Applied Mathematical Sciences. 9(103): 5123-5137.

[4] E.P. Iguzquiza and P. A. Dowd. 2013. Comparison of Inference Methods for Estimating Semivariogram Model Parameters and Their Uncertainty: The Case of Small Data Sets. Computers and Geosciences. 50: 154-164.

[5] S. Waterman. 2003. Weighted Jackknife-OK dalam Penaksiran Sumber Daya Mineral, Dissertasion, Institut Teknologi Bandung.

[6] Damayanti. 2003. Spatial Point Process untuk Prediksi Distribusi Peluang Permeabilitas Reservoir Jatibarang, Final project, Bandung Institute of Technology.