Kriging on comparison of Original and Outlier-free data

Nur Ali AmriAbdul Aziz JemainWan Fuad Wan Hassan

Citation: **1614**, 929 (2014); doi: 10.1063/1.4895326
View online: http://dx.doi.org/10.1063/1.4895326
View Table of Contents: http://aip.scitation.org/toc/apc/1614/1
Published by the American Institute of Physics

Kriging on Comparison of Original and Outlier-free Data

Nur Ali Amri[a,b], Abdul Aziz Jemain[b] and Wan Fuad Wan Hassan[c]

[a] *Mining Engineering Department, Faculty of Mineral and Technology, UPN "Veteran" Yogyakarta, Indonesia*
[b] *School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, Selangor, Malaysia.*
[c] *Geological Department, University of Malaya 50603 Kuala Lumpur, Malaysia.*

Abstract. Kriging prediction on gold grade with the contamination of outliers, which are grouped into two zones i.e. rich and poor zones, performed with classical and robust semivariogram comparison. Both are based on ordinary least squared fitting of spherical model. The ordinary kriging method based on robust semivariogram is used in both Original and Outlier-free data. The first scenario conducted on the Original, while the second performed on the Outlier free. Both are fitted based on the classical method and the assumption of omnidirectional.

Keywords: Semivariogram, Outlier, OLS Fitting, Ordinary Kriging.

INTRODUCTION

One of the important factors in the mining operations is the calculation of resources or reserves. A variety of methods are offered, but the most widely used is geostatistics, which is a method based on the model of semivariogram (some authors called variogram) and kriging prediction. Semivariogram is the main tool of regionalized variable theory which quantifies the size and intensity of spatial variation. Semivariogram provide a basis for the optimum interpolation through of kriging method [1].

There are two semivariogram models, namely the empirical (or experimental) and theoretical. The empirical model is a model of discrete function which is used to execute the observation data. While the theoretical semivariogram (e.g., spherical, exponential, linear, etc.) is a continuous function obtained by fitting (empirical against theoretical), in which the output (among others) is the sill and the range. Both (sill and range) has a very important role in the kriging method. Sill is a situation which is between samples within an area (or region) is no longer spatially correlated. While the maximum distance of influence is called the range.

Empirical semivariogram can be solved with the classical model [2], or the robust model [3]. Classical models used for non-skewed data, whereas in general, for skewed data is used robust model [4].

Kriging is a method, in geostatistics which is used to predict the values of a regionalized of random variable at which unsampled data, and mainly based on sill and range parameters.

MATERIALS AND METHODS

In general, the minerals or metals component (including gold) inside the vein is not evenly distributed. Similarly with the grade, which is often very erratical [5 and 6]. The existence of minerals, in this study is the gold deposits that are in the vein, is also not evenly distributed. In one location (within a region) there is a deposit which has a high grade (rich zone), whereas at other locations, in the same region there is a moderate or even low grade (poor zones). Therefore, not necessarily that the calculation of reserves (mainly) conducted simultaneously, assuming that the influence area can apply to all locations.

Sample data such gold grade (in g/t Au) which had been sedimented in the veins quartz, derived from the results of assaying of 138 drilling samples, which are located in the Ciurug Pongkor, the area owned by PT. Aneka Tambang UBPE Pongkor, West Java, Indonesia. Based on the results of statistical processing, the distribution of grade clustered in two zones, namely, in the west is a high-grade zone, while the east is a zone with a low grade.

The first scenario is to calculate (based on robust model) the semivariogram data groups of 138 data. Outliers should be excluded from the analysis [7]. Based on these considerations, since the data distribution is skewed, the second scenario performed which is to eliminate the outliers from the data calculation and analysis. Therefore, there are 10 data outliers then in the second scenario which is outlier-free data sets, eventually living to 128. Both the results of these calculations serve as the comparison basis to the next step.

The 2014 UKM FST Postgraduate Colloquium
AIP Conf. Proc. 1614, 929-935 (2014); doi: 10.1063/1.4895326

The entire data set, both 138 and 128, located in an area which is extent of 1496×364 m^2, or about 560000 m^2 (if taken of outermost points). The main purpose of this paper is to calculate the average grade based on kriging method. Reserve calculation is not done, because it is associated with the company's permission.

Geostatistics

Geostatistics is a part of statistics which is considered the regionalized variables as reliability of the random sample. Mapping Z: Ω→R called a random variable if the area of function Z is a countable set of $\{z_1, z_2, \ldots\}$ and $\{\omega: Z(\omega)=z_j\} \in \psi$ for all j≥1, where ψ consists of all subsets of Ω [8].

Regionalized variables can be simply described as a variable which is distributed in space. Journel & Huijbregts [9] stated that mineralization phenomena (including gold grade distribution in two dimensions) can be characterized by the spatial distribution of regionalized variables values.

Empirical Semivariogram

Semivariogram is a squared-difference function of expectation of the sample pair on the relative orientation [10, 11], i.e.

$$\gamma(\mathbf{h}) = \frac{1}{2} E[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})]^2. \tag{1}$$

Z (**s**) is the value of a variable which is situated in a location **s**, while Z(**s**+**h**) is the value of a variable which is situated in a location within **h** distance of **s**. Empirical semivariogram then presented as [2]

$$\gamma(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_i+\mathbf{h})]^2. \tag{2}$$

If $\mathbf{h}=\mathbf{s}_j - \mathbf{s}_i$, then Eq. 2 can be simplified (and this form will serve as a reference in writing of the next formula) to

$$\gamma(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2. \tag{3}$$

To minimize the occurrence of atypical observations has transformed Box and Cox [10] by assuming the principle of normality and chi-squared distribution of four square root of $(Z(\mathbf{s}+\mathbf{h})-Z(\mathbf{s}))^2$, in order to obtain the robust semivariogram, namely

$$\bar{\gamma}(\mathbf{h}) = \left(\frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^{1/2} \right)^4 \Big/ \left(0.457 + \frac{0.494}{|N(\mathbf{h})|} \right), \tag{4}$$

$N(\mathbf{h}) \equiv \{(i,j); \mathbf{s}_j - \mathbf{s}_i = \mathbf{h}\}$ is sample pair observation which is separated at a distance **h**. If *n* is the number of data observations, than the number of observation points is $n(n-1)/2$.

Semivariogram Fitting

Geostatistically, semivariogram fitting is a procedure to find the best continued-curve of points (by minimizing the sum of squares, of course) interpreted as the process of matching the experimental semivariogram versus theoretical semivariogram. The ultimate goal is to obtain a precise information of semivariogram parameters among others, nugget, sill and range. In many mining literature, theoretical semivariogram fitting also called as a *structural analysis* [12].

Semivariogram fitting is an important step of spatial estimations, because this will determine the weight of kriging [13]. Fitting method, here using of ordinary least squares (OLS), as [10] and [14],

$$\min \sum_{i=1}^{k} [\hat{\gamma}_z(\mathbf{h}_i) - \gamma_z(\mathbf{h}_i;\theta)]^2. \tag{5}$$

In case where the estimator is robust, then $\hat{\gamma}_z(.)$ is replaced with $\bar{\gamma}(.)$.

Sill, which is geologically defined as igneous intrusions concordant to bedding rock [15], geostatistically defined as a condition in which the semivariogram has reached a stationary condition. Range is the maximum distance where a semivariogram has reached the sill, and mathematically written as $\lim_{|\mathbf{h}|\to\infty}\gamma(|\mathbf{h}|)=\gamma_\infty<\infty$. While nugget is a condition where the extrapolation curve leadings to **h**=0 (the origin) does not produce $\gamma(0)=0$, but $\gamma(0)=c_0$.]. Or, $\gamma(\mathbf{h})\to c_0>0$ for $\mathbf{h}\to 0$.

Theoritical Semivariogram

Sill and range parameters as in empirical semivariogram of Eq. 3 and Eq. 4, to be fitted with theoretical semivariogram, which in this paper is using spherical model [16]

$$\gamma(\mathbf{h})=\begin{cases} c_0+c\left\{\dfrac{3\mathbf{h}}{2a}-\dfrac{1}{2}\left(\dfrac{\mathbf{h}}{a}\right)^3\right\}, & \mathbf{h}\le a \\ c_0+c, & \mathbf{h}>a. \end{cases} \tag{6}$$

Here, a is the range (of influence), c_0 as nugget and c_0+c is the sill. The data execution used the geoR's library of R package [17].

Ordinary Kriging

Ordinary kriging (OK) is a linear method with weighted averages of each value in a location which is, certainly related to the semivariogram. This method is useful to predict the value at a certain unsampled point, by comparing to the similar data on the other points that have been known. Suppose that the $Z(\mathbf{s}_1)$, $Z(\mathbf{s}_2)$, …, $Z(\mathbf{s}_n)$ is a collection of grade points which are at spatial points $\mathbf{s}_1$, $\mathbf{s}_2$, …, $\mathbf{s}_n$. If $\mathbf{s}_0$ is a point to be predicted in which, each predictor random variable $Z(\mathbf{s}_i)$ has the same probability distribution at all locations, than the predicted value is [10]

$$\hat{Z}(\mathbf{s}_0)=\sum_{i=1}^{n}w_i Z(\mathbf{s}_i) \tag{7}$$

which is a weighted average of the sample values, provided unbias $\sum_{i=1}^{n}w_i=1$.

Kriging will minimize the mean squared error, with

$$\min\sigma_e^2=E[Z(\mathbf{s}_0)-\hat{Z}(\mathbf{s}_0)]^2=\min\left[Z(\mathbf{s}_0)-\sum_{i=1}^{n}w_i Z(\mathbf{s}_i)\right]^2 \tag{8}$$

In the case where Z is an intrinsically stationary process, then

$$\sigma_e^2=E\left[Z(\mathbf{s}_0)-\sum_{i=1}^{n}w_i Z(\mathbf{s}_i)\right]^2=2\sum_{i=1}^{n}w_i\gamma(\mathbf{s}_0-\mathbf{s}_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}w_i w_j\gamma(\mathbf{s}_i-\mathbf{s}_j) \tag{9}$$

Minimization can be done, as long as (w_1, w_2, …, w_n) eligible to limit $\sum_{i=1}^{n}w_i=1$. Thus, minimization can be written as,

$$\min 2\sum_{i=1}^{n}w_i\gamma(\mathbf{s}_0-\mathbf{s}_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}w_i w_j\gamma(\mathbf{s}_i-\mathbf{s}_j)-2\lambda\left(\sum_{i=1}^{n}w_i-1\right), \tag{10}$$

where λ is a Lagrange multiplier. After dividing the last equation by w_1, w_2, …, w_n and λ, and also equalize the partial derivatives, respectively of the parameters equal to zero $\dfrac{\partial\sigma_e^2}{\partial w_i}=0=2\gamma(\mathbf{s}_0-\mathbf{s}_i)-\sum_{j=1}^{n}w_j\gamma(\mathbf{s}_i-\mathbf{s}_j)-2\lambda,$ then obtained

$$\sum_{j=1}^{n}w_j\gamma(\mathbf{s}_i-\mathbf{s}_j)+2\lambda=2\gamma(\mathbf{s}_0-\mathbf{s}_i),\quad \forall i=1,\ldots,n. \tag{11}$$

The partial derivative

$$\frac{\partial \sigma_e^2}{\partial \lambda} = 0 = 2\left(\sum_{i=1}^{n} w_i - 1\right) \rightarrow \sum_{i=1}^{n} w_i = 1,$$

is a boundary condition as the described above.

If the matrix $\mathbf{W}=(w_1, w_2, \ldots, w_n, \lambda)$ and $\boldsymbol{\gamma}=(\gamma(\mathbf{s}_0\text{-}\mathbf{s}_1), \gamma(\mathbf{s}_0\text{-}\mathbf{s}_2), \ldots, \gamma(\mathbf{s}_0\text{-}\mathbf{s}_n))'$,

$$\boldsymbol{\Gamma} = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j), & i = 1,2,\ldots,n,\ j = 1,2,..n, \\ 1, & i = n+1,\ j = 1,\ldots,n, \\ 1, & j = n+1,\ i = 1,\ldots,n, \\ 0, & i = n+1,\ j = n+1. \end{cases} \tag{12}$$

Eq (12) then be written succinctly, as $\boldsymbol{\Gamma}\mathbf{W} = \boldsymbol{\gamma}$. While the weights $w_1, w_2, \ldots, w_n$ and λ can be calculated,

$$\mathbf{W} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}. \tag{13}$$

Variance of a predictor written as,

$$\sigma_e^2 = \sum_{i=1}^{n} w_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + \lambda \tag{14}$$

RESULTS AND DISCUSSIONS

The initial step in spatial analysis is to check the raw data for the presence of trends and outliers [18, 10 and 7]. Based on assumption that gold grade values more than 15 g/t Au, there is 10 data outlier which then eliminated of the calculation. These considerations is also linked to the value of CV=0.89 (already approaching to one), the kurtosis value which is too high (12.41), and the difference between the mean and maximum value which is also large (i.e. 42.250-70.55 g/t Au).

Empirical semivariogram calculations performed with the assumption of omnidirectional, with a lag of 50 m, the maximum distance of 1500 m, and the angle tolerance of 22.5. Semivariogram of Original data is calculated using robust model, while for Outlier-free data is using classical models.

TABLE (1). Semivariogram parameters for original and outlier-free data.

No	**Distance** (m)	**Semivariogram** (g/t Au)		**Pairs**	
		Original	Outlier-free	Original	Outlier-free
1	50	14.231	6.672	487	404
2	100	16.684	9.069	675	579
3	150	22.661	11.804	825	680
4	200	21.768	11.254	801	672
5	250	21.621	13.190	715	610
6	300	20.185	12.812	620	535
7	350	18.101	11.047	582	510
8	400	21.661	14.323	560	491
9	450	22.397	13.474	505	442
10	500	31.874	18.744	481	415
11	550	30.743	18.271	459	399
12	600	31.174	16.355	424	361
13	650	28.864	15.134	386	330
14	700	17.379	11.058	302	279
15	750	31.454	19.704	278	247
16	800	32.926	22.280	254	228
17	850	55.257	32.864	217	185
18	900	73.678	41.744	197	164
19	950	59.721	32.261	169	144
20	1000	55.405	32.360	130	112

21	1050	38.193	24.735	120	107
22	1100	37.068	27.112	81	74
23	1150	46.395	29.364	60	53
24	1200	69.893	33.360	41	33
25	1250	66.640	42.970	30	25
26	1300	17.091	14.432	23	22
27	1350	75.624	19.816	10	8
28	1400	15.267	11.776	13	12
29	1450	29.646	16.600	6	5
30	1500	1.513	1.513	2	2

Semivariogram and Fitting of Original Data

The empirical semivariogram execution of Original data as TABLE (1) produces as much as 9543 point pairs, where most couples are 825 pairs occurred on the third lag (150 m). The smallest lag distance is 50 m, and the largest is 1500, so overall there are 30 lag. Based on the calculation is obtained that the smallest pair is 2 points which is located on the thirtieth lag. The greatest semivariogram value for Original data is 69.893, and the smallest is 1.513 (both in $(g/t)^2$ Au).

Using OLS fitting spherical models (as Eq. 6), then in the location that has 457.500 m^2 wide, obtained the sill value is 43.489 $(g/t)^2$ Au, and the range is 1020.992 m (FIGURE 1 (a)). The greatest semivariogram value for the Original data is 69.893 $(g/t)^2$ Au. Thus, the equation of theoretical semivariogram based on spherical model is

$$\gamma(\mathbf{h}) = \begin{cases} 43.489\left\{\dfrac{3\mathbf{h}}{2\times 1020.992} - \dfrac{1}{2}\left(\dfrac{\mathbf{h}}{1020.992}\right)^3\right\}, & \mathbf{h} \le 1020.992 \\ 43.489, & \mathbf{h} > 1020.992. \end{cases}$$

Semivariogram and Fitting of Outlier-free

The empirical semivariogram computation on the Outlier-free data is using the classical semivariogram models, and here produce of 8128 pairs of points. As the Original data, Outlier-free data is calculated at angle tolerance of 22.5, which the lag of 50 and a maximum distance is 1500. The calculation of empirical semivariogram as TABLE (1) is using the robust model, with lag of 50 meters, a maximum distance is 1500, and produces 8128 pairs of points $[n(n–1)/2]$. As for the Original, the Outlier-free data, here gained of 30 lag.

The greatest semivariogram value is 42.97 $(g/t)^2$ Au, obtained on 25 pair of points with the distance of 1250 (or at lag of 25). The smallest pair of points, i.e. 2 pairs, occurs on the last lag (distance 1500) where, further to the east, the number drill samples become fewer. Greatest point pair is 680, which occurs at a distance of 150 m, or in the third lag.

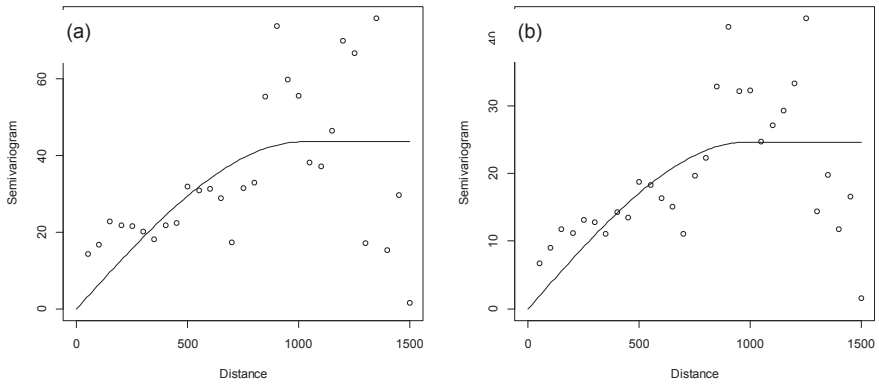

FIGURE 1. Spherical Fitting of (a) Original with Sill=43.489 $(g/t)^2$ Au, Range=1020.992, n=138 and (b) Outlier-free with sill=24.684 $(g/t)^2$ Au, range=988.593 m and n=128.

Based on OLS fitting spherical models, obtained the calculation that for Outliers-free, produces of sill=24.684 $(g/t)^2$ Au, which began in range of 988.593 m (FIGURE 1 (b)). The equation, then become

$$\gamma(\mathbf{h}) = \begin{cases} 24.684\left\{\dfrac{3\mathbf{h}}{2\times 988.593} - \dfrac{1}{2}\left(\dfrac{\mathbf{h}}{988.593}\right)^3\right\}, & \mathbf{h} \le 988.593 \\ 24.684, & \mathbf{h} > 988.593. \end{cases}$$

Ordinary Kriging

The calculation of OK for both groups of data (Original and Outlier-free), with total area of 547.500 m^2 based on prediction of block size (5×5) m^2. On Original, ordinary kriging produces a lower prediction (about 30%) where the mean error is –2.119 (g/t) Au. While in the Outlier-free, OK also predicts a lower value, although the mean error is smaller (about 24%), i.e. –1.431 g/t Au (TABLE (2)). If seen from the mean error or percent decline predictions, the Outlier-free produces better predictions than the Original.

TABLE (2). Original and Outlier-free Statistic and OK Predicted Parameters (in g/t Au).

Location	*n*	Sources	Kurtosis	Min	1st Qu	Median	Mean	3rd Qu	Max
Original	138	True	12.410	0.450	2.713	5.750	7.055	9.220	42.250
		Predicted	9.803	0.013	1.956	3.323	4.936	6.831	38.610
		Error					-2.119		
Outlier-free	128	True	2.446	0.450	2.650	5.275	5.823	8.362	14.800
		Predicted	2.774	0.379	2.059	3.426	4.392	6.401	14.570
		Error					-1.431		

The distribution for Original and Outlier-free which is generated by the OK prediction can be seen in FIGURE 2. Seen that the distribution of predicted for Outlie-free (FIGURE 2 (b)) is more evenly distributed than the OK predicted results for the Original. In fact, for the "great grade" which is about 15 g/t Au (which looks yellow to cream) move almost reaching to the border of the eastern region (FIGURE 2 (a)).

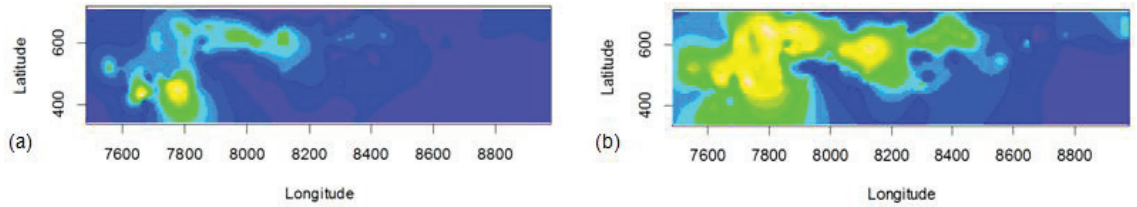

FIGURE 2. OK Predicted Distribution based on Spherical Model of (a) Original Data (Sill=43.489, Range=1020.992) and (b) Outlier-free Data (Sill=24.684, Range=988.593).

CONCLUSION

Ordinary kriging which is based on OLS semivariogram fitting of spherical models produces a lower value than the true data. However, elimination of outlier data can increase the percentage or mean prediction erros when compared with the Original data, which is significantly skewed.

ACKNOWLEDGEMENTS

Thanks to PT. Aneka Tambang UBPE Pongkor's management, in particular Ir. Rustaman, Ir. Joseph and Ir. Kristiawan, which has provided an opportunity to the author to obtain data and a wide range of existing facilities.

REFERENCES

1. L.G. Vendrusculo, P.S.G. Magalhães, S.R. Vieira and J.R.P. Carvalho, *Journal of Science Agricultural (Piracicaba, Braz.)* **5** (61), 100-107 (2004).
2. G. Matheron, *The theory of Regionalized Variables and its Applications*, Paris: Ecole Nationale des Mines de Paris (1971).
3. N. Cressie and D.M. Hawkins, *Journal of Association for Mathematical Geology*, **12** (2), 115-125 (1980).
4. R.M. Lark, *European of Journal Soil Sciences* **51**,137–157 (2000).
5. A.E. Annels, *Mineral deposit evaluation: A practical approach*, New York: Chapman & Hall (1991).

6. J.W. Barnes, *Ore and Minerals: Introducing Economic Geology*, Philadelphia: Open University Press (1988).
7. A.S. Kishné, E. Bringmark, L. Bringmark and A.A. Alriksson, *Journal of Environmental Monitoring and Assessment* **84**, 243–263 (2003).
8. D.S. Yates, D.D. Moore and D.S. Starnes, *The Practice of Statistics*, New York: Freeman (2003).
9. A.G. Journel and Ch.J. Huijbregts, *Mining Geostatistics*, London: Academic Press, 1978, 600 p.
10. D.D. Sarma, *Geostatistics with Applications in Earth Sciences*, Berlin: Spriger, 2009, pp. 124-138.
11. O. Schabenberger and C.A. Gotway, *Statistical Methods for Spatial Data Analysis*, New York: Chapman & Hall/CRC, 2005.
12. M.E. Hohn, *Geostatistics and Petroleum Geology*, New York: Van Nostrand Reinhold, 1988, pp. 107.
13. D.S. Maglione and A.M. Diblasi, *Exploring a valid model for the variogram of an isotropic spatial process*, New York: Springer-Verlag,
14. *Stoch. Environ. Res. Risk Assess*. **18**: 366-376 (2004).
15. W.G. Müller, *Journal of Statistics & Probability Letters* **43**, 93-98 (1999).
16. E. Gringarten and C.V. Deutsch, *Journal of Association for Mathematical Geology* **33** (4), 507-534 (2001).
17. R. Webster and M.A. Oliver, *Geostatistics for Environmental Scientists*, England: John Wiley & Sons Ltd, 2007, pp. 87-88.
18. R.S. Bivand, E.J. Pebesma and V.G. Rubio, *Applied Spatial Data Analysis with R*, New York: Springer Science+Business Media, LLC, 2008, pp. 215-216.
19. E.H. Isaaks and R.M. Srivastava, *Applied geostatistics*, New York: Oxford University Press, 1989, pp. 191.