Flexible Lag Definition for Experimental Variogram Calculation

Yupeng Li and Miguel Cuba

The inference of the experimental variogram in geostatistics commonly relies on the method-of-moments approach. Ideally, the available data should be on a regular pattern for stable experimental variogram calculation. However, in practice, data is sampled on an irregular pattern. Hence some binning of the variogram data pairs is required to obtain reliable estimates of the experimental variogram. This grouping of the variogram data pairs depends traditionally on parameters such as main anisotropic directions and lag vectors with tolerances. In the traditional estimation of the experimental variogram, the lag vectors with their tolerances do not rely on the configuration of the variogram data pairs in the variogram cloud but on a segment of it that is predefined along a specified direction. A new methodology to calculate the experimental variogram is proposed in this paper, where the lag vectors and their tolerances are decided from the variogram cloud. The proposed experimental variogram estimation results appear more reasonable, since less noise is introduced during the definition of the lag vectors. It yields a well-founded, practicable, and easy-to-automate methodology for irregularly sampled data. Results of a comparison study with the traditional approach are encouraging.

1. Introduction

The conventional methodology (Matheron, 1962) to obtain the variogram model of a domain considers two main steps 1) the estimation of the experimental variogram and 2) the fitting of the experimental variogram using a licit model. The latter is used as one input parameter to build a geostatistical model of the domain.

The experimental variogram is calculated based on the variogram cloud. In case the available data is sampled over a regular pattern, the lag vectors in the variogram cloud are easy to identify. In practice, the available data is usually sampled over a sparse patter. Hence lag vectors plus tolerance parameters have to be used to estimate the experimental variogram of the domain. In this approach, the lag vectors and their tolerances are defined and adjusted during the estimation process. They are used to classify groups of variogram data pairs from the variogram cloud. These classified groups are used as supporting information of the experimental variogram estimates at the specified lag vectors. In the conventional approach, the lag vectors are regularly defined along a specified direction resulting in almost regular bins to classify variogram data pairs.

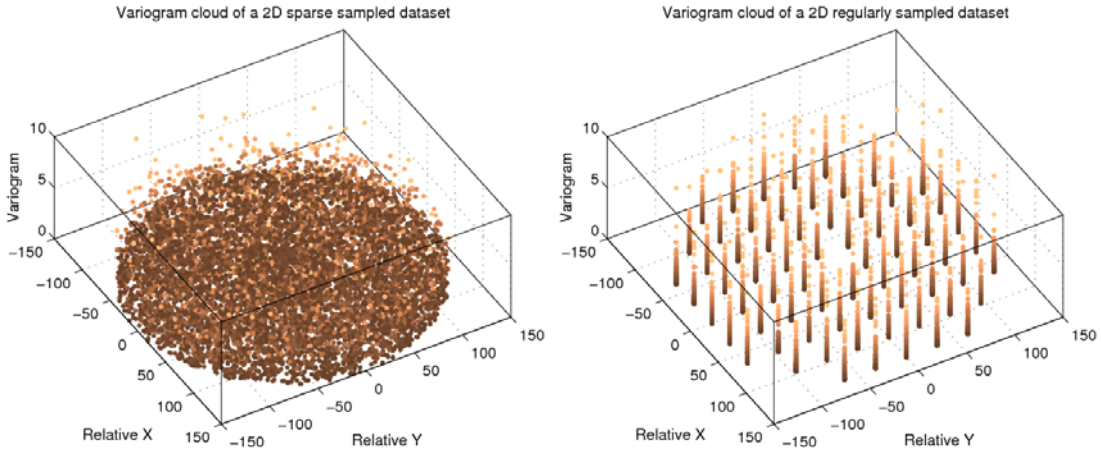

Figure 1: Variogram cloud of sparse (left) and regularly (right) sampled datasets.

Denser groups of variogram data pairs are spatially distributed over the variogram cloud depending on the sampling pattern of the available data. It is in those regions where the experimental variogram is better informed in terms of supporting information. The simplification of the conventional approach introduces bias in the estimation of the experimental variogram, since it considers a very strict configuration of bins over the variogram cloud. Accordingly, relevant information about the spatial correlation is lost in the process.

In this paper a new approach to estimate experimental variograms is proposed that accounts for the denser groups of variogram data pairs in the variogram cloud. It consists of identifying these denser regions to maximize the amount of supporting information for each of the lag vectors while minimizing the range of

tolerances to make the estimation more accurate. The proposed approach considers the use of clustering techniques on the variogram cloud to automate the estimation process.

In the next section, the basic theory about the experimental variograms is presented. After that, the methodology to implement the proposed approach is discussed in details and illustrated with a calculation example using the Jura dataset. Also, the improvement comparing with the tradition approach is done with a synthetic dataset which is simulated from a known variogram model. In the conclusions section, advantages and disadvantages of the proposed approach are presented.

2. Experimental semivariogram

In spatial statistics, the theoretical variogram $2\gamma(\mathbf{h})$ is a function that describes the degree of spatial dependence of an intrinsic random function Z whose increments are second-order stationary and is the expected squared increment of the values between locations $\mathbf{u}$ and $\mathbf{u}+\mathbf{h}$ as:

$$2\gamma(\mathbf{h}) = Var\{Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\}$$

Where, $2\gamma(\mathbf{h})$ is given the name of variogram function. The function of practical interest is $\gamma(\mathbf{h})$ which is named as semi-variogram, since it is one-half of the variogram function. Usually, only the function $\gamma(\mathbf{h})$ is used in implementing kriging, the prefix semi- is regularly dropped, and the function $\gamma(\mathbf{h})$ is interchangeably called variogram and the semi-variogram in the geostatistical literature. From now on, in this paper, the variogram will be mainly referred to as semi-variogram.

The above expression is a useful abstraction, but not easy to apply to observed values $z(\mathbf{u})$. To infer the experimental variogram $\hat{\gamma}(\mathbf{h})$ from the sampled locations, a natural estimator based on the method-of-moments (Matheron, 1962) is:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_i) - z(\mathbf{u}_i \pm \mathbf{h})\right]^2 \quad (0)$$

Where $N(\mathbf{h}) = \{\mathbf{u}_i - \mathbf{u}_j = \pm\mathbf{h}; i, j = 1, \ldots, n\}$ is the number of data pairs separated by $\mathbf{h}$. The experimental variogram is calculated by averaging one half of the squared differences of the values over all pairs of observations separated by the vector $\mathbf{h}$.

In practice, the available data is usually sampled over a sparse pattern; the variogram in estimator is usually smoothed as (Cressie N. A., 1993):

$$\hat{\gamma}(\mathbf{h}_\ell) = \frac{1}{2n(\mathbf{h}_\ell)}\sum_{j=1}^{n(\mathbf{h}_\ell)}[z^j(\mathbf{u}_i) - z^j(\mathbf{u}_i \pm \mathbf{h}_\ell)]^2$$
$$(\mathbf{u}_i, \mathbf{u}_i \pm \mathbf{h}_\ell) \in n(\mathbf{h}_\ell); \mathbf{h}_\ell \in T(\mathbf{h}_\ell \pm \boldsymbol{\varepsilon})\}; \ell = 1, \ldots, K$$

Where $n(\mathbf{h}_\ell)$ is the number of pairs in the lag region which is centered on lag vector $\mathbf{h}_\ell$ and $T(\mathbf{h}_\ell)$ is all the possible distances that is in the specified distance region around lag vector $\mathbf{h}_\ell$ with some predefined tolerance $\boldsymbol{\varepsilon}$. While $z^j(\mathbf{u}_i) - z^j(\mathbf{u}_i \pm \mathbf{h}_\ell)$ is the j^{th} pair of the observed values in this region. In order to get a stable experimental variogram $\hat{\gamma}(\mathbf{h})$, the tolerance to define the distance region has to be as small as possible to retain spatial resolution, yet large enough to obtain a reliable number distinct pairs of data pairs which should at least be 30 as recommend by Journel and Huijbregt (1978, p. 194)

Extensive research has been focused on variograms. For example, there are some researches on uncertainty estimation by the method of moments (B.P. Marchant and R.M. Lark, 2004). Also, some researches focused on variogram interpretation and model fitting (Pelletier, 2004; E. Gringartn & C.V. Deutsch, 2001). Some researchers are tried to find a more reliable experimental variogram calculation with the assumption that the lag parameters are predefined. The improvements of the experimental variogram are confined to more robust average calculation. Rosenblatt (1985) proposed a moving-windows methodology to build the estimator. All these research may improve the estimation of the experimental variogram somehow. However, as Armstrong (1984)

pointed out that the poor choice of distance classes may lead to a non-robust estimation. In the next sections a methodology to group variogram data pairs in the variogram cloud that leads to an improvement in the estimation of the experimental variogram is presented.

3. Methodology of the improved experimental variogram calculation

The proposed approach would work in the variogram cloud space. The first step is calculating the variogram cloud from the data set. The variogram cloud is a plot of the dissimilarity between any two observations as a function of their separation in geographical space (Chauvert, 1982), that is, a plot of the separation vector $\mathbf{h} = \mathbf{u}_i - \mathbf{u}_j$ of all the data pairs against their corresponding halved squared of the increment $0.5\left[z(\mathbf{u}_i) - z(\mathbf{u}_i - \mathbf{u}_j)\right]^2$.

Conventionally, once the variogram cloud is calculated, the experimental variogram is estimated along a selected direction by defining parameters such as number of lags and lag distance that establish a set of lag vectors. To ensure that enough data pairs are used for each lag vector, the use of tolerance parameters is implemented as expressed in equation (3) and illustrated in Figure 2.

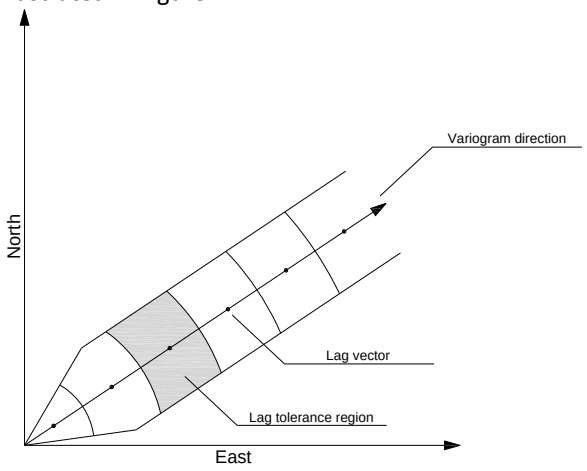

Figure 2: Diagram of conventional binning of the variogram cloud in experimental variogram calculation (2D case)

Each lag tolerance region will define a small partition of variogram data pairs from the variogram cloud. The variogram data pairs are classified according to the set of lag vectors established. The most critical part of calculating an experimental variogram relies on the definition of the lag tolerance regions. The variogram cloud of a sparse sampled data contains irregular denser regions of data pairs distributed along the variogram cloud. Using a very strict definition of lag tolerance regions might add bias in the estimation of the experimental variogram. A more appropriate estimation should account for the irregular locations of the denser regions in the variogram cloud as they provide more information than of the arbitrarily defined.

In the proposed approach, the experimental variogram estimation will be referred to as an optimal classification problem of the variogram cloud aiming to obtain accurate estimates. Thus, in second step is using a kind of partial optimization algorithm to classify the variogram cloud into different lag tolerance regions to obtain more accurate and representative estimates of each lag vector. In this paper, a self-organizing-map (SOM) algorithm which considers the implementation of neural networks is used to classify the input variogram cloud into optimized lag tolerance regions. Unlike the conventional approach, these lag tolerance regions are no longer defined by a regular increment of the lag separation vector.

The SOM model was first described as an artificial neural network algorithm by Teuvo Kohonen, and is sometimes called Kohonen-map (Kohonen, 2001). There are two constraints when implementing SOM in the variogram cloud classification. The first constraint is related to the number of data pairs $n(\mathbf{h}_l)$ and the amount of supporting data pairs $n(\mathbf{h}_l)$ should be as much as possible. The second constraint is the distance increment between each data pair. In each lag tolerance region, the difference of the increment for all the clustered data pairs should be as small as possible. That is the distribution of $T(\mathbf{h}_\ell)$ should have a very small variance. Any other optimal clustering algorithms that can satisfy these two constraints should be fine in the implementation of the proposed approach. Applying these two constraints to the variogram cloud, the data pairs would be assigned to some clusters obtained from the SOM, as shown in Figure 3.

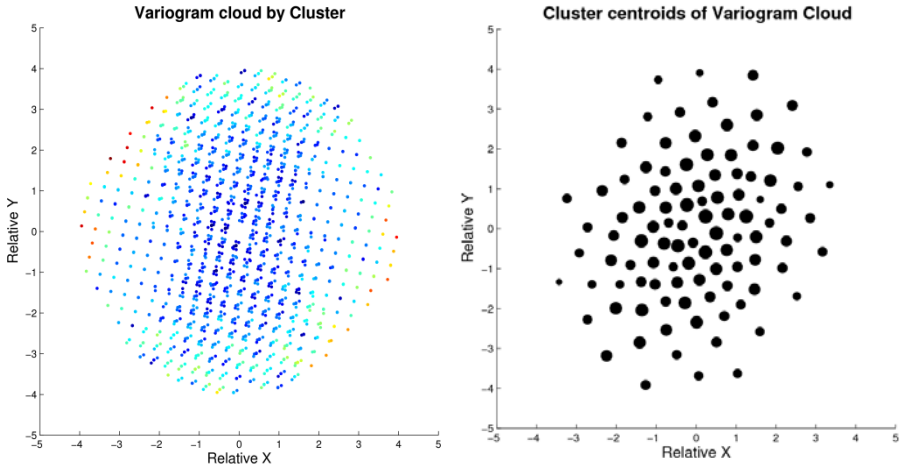

Figure 3: Variogram cloud (left) and center of cluster of the variogram cloud (right), the diameter of the circle is proportional to the number of data pairs.

The third step would be pick up one direction and calculate the experimental variogram. In the Jura data set, the experimental variogram along the 45NW would be recognized as the major continuous direction as shown in Figure 4. In the conventional experimental variogram approach, and given the all the tolerance and band width, the experimental variogram along the 45NW is also plotted in Figure 4.

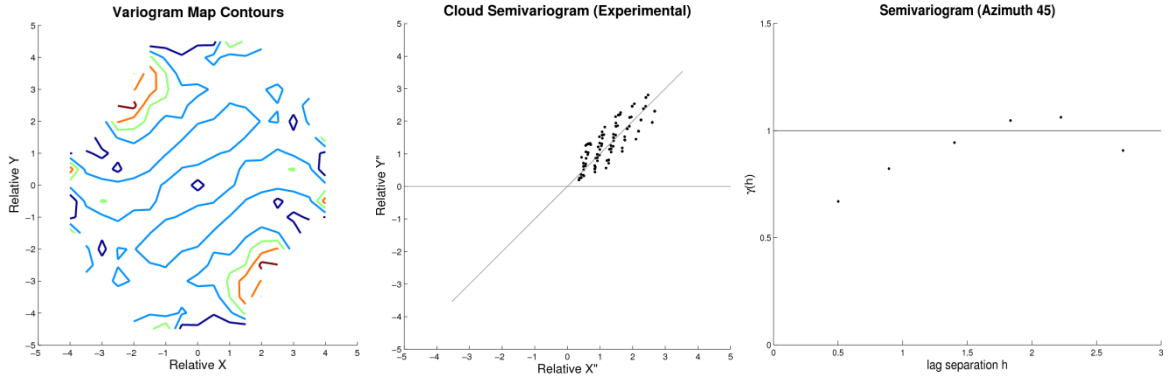

Figure 4: The variogram map contours (Left); the variogram cloud picked along 45NW and the predefined band width(Middle);the experimental variogram calculated from traditional approach

While using the proposed variogram, given the direction, direction tolerance and band width, the experimental variogram would be calculated without taking the lag number and lag tolerance into consideration. Instead, it will looking the clustered data pairs fall into the area defined by the direction, direction tolerance and band width as shown in Figure 5. The experimental variogram along 45NW is also plotted in Figure 5. Comparing with the traditional approach, the data pair in variogram cloud is divided into different lag regions by its mean, pair numbers and distance difference of these pair numbers not by the predefined lag number, lag distance and lag tolerance.

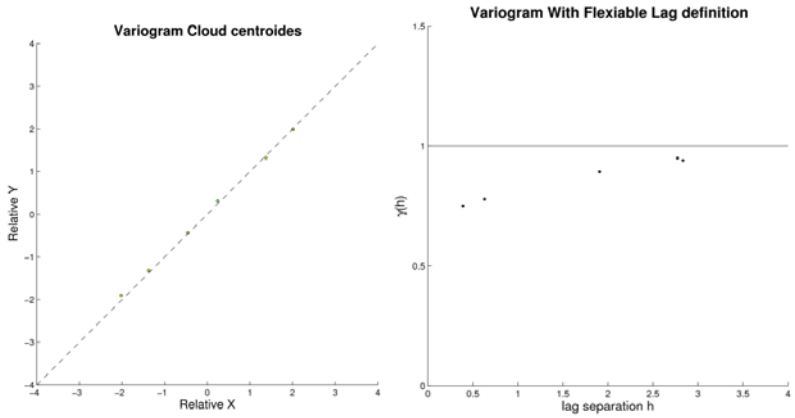

Figure 5: the entire clustered center fallen into the region defined by direction and band width (left) and the experimental variogram from the clustered centroides

4. Comparison with the known variogram model

In section 3, the proposed approach is illustrated with a spatial data set that the true distribution is unknown. In order to know the accuracy improvement, the proposed approach and the traditional approach will be used to one data set that is constructed with a known variogram model. Assuming the variogram model is:

$$\gamma(h) = 1 - \exp(-\frac{3h}{a})$$

Assuming the range a along 45 degree has the maximum range ($a_{\max}$) 100, and a minimum range ($a_{\min}$) 50.

A synthetic dataset that consists of 200 data points is sampled from an unconditional realization map simulated imposing this variogram model.

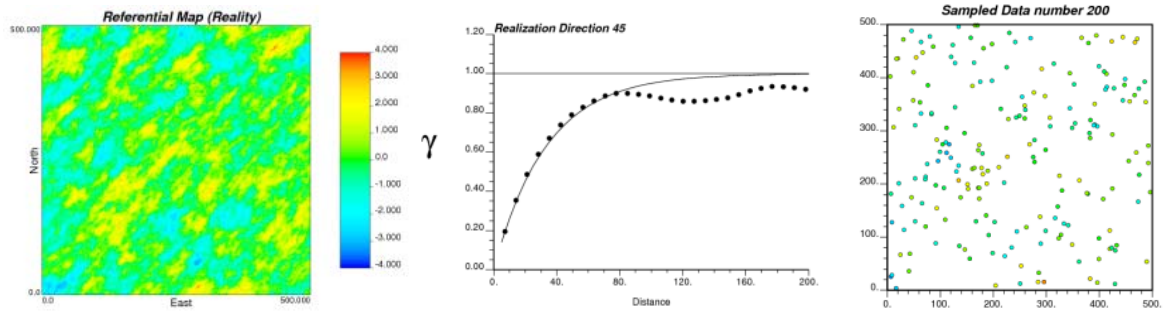

Figure 6: Left, the realization with the known variogram model. Middle: the used variogram model (solid line) and the experimental variogram calculated from the realization along 45°NW direction; Right: the randomly picked samples (Total number is 200)

The experimental variogram can be calculated from the proposed approach and the traditional approach using the randomly picked sampled data set and compared with the theoretical model. For example, along 50°NW, the experimental variogram from these two methods are listed in Table 1. The experimental variogram from both of them are also plotted together with the theoretical variogram model as plotted in Figure 7. Note, in the new proposed approach, each experimental variogram value is represented by the vector of the variogram cloud centers.

Table 1: the experimental variogram calculated from traditional and the proposed approach

Traditional Approach		Proposed approach		
distance	variogram	azimuth	Distance	variogram
30.57915	0.57475	46.735	21.237	0.48654
44.45322	0.74316	42.506	23.487	0.54924
60.03737	0.80341	34.136	43.855	0.82493
74.78219	0.80579	36.213	45.582	0.87652
89.76925	0.56642	52.963	62.873	0.64555
104.79942	0.70193	54.562	65.358	0.68857
120.25917	0.69833	31.239	70.929	0.84005
134.25806	0.76732	31.237	71.082	0.83126
149.74615	0.97956	57.783	84.729	0.82374
164.22216	0.66954	39.996	89.626	0.66504
179.66234	1.00833	56.94	89.923	0.791
194.1386	0.95888	40.279	89.97	0.66274
210.26956	0.68458	50.851	108.44	0.71418
224.99694	1.19715	50.611	110.69	0.69615
239.89279	1.49214	44.757	132.17	0.60863
254.94556	1.07555	44.466	132.93	0.66541

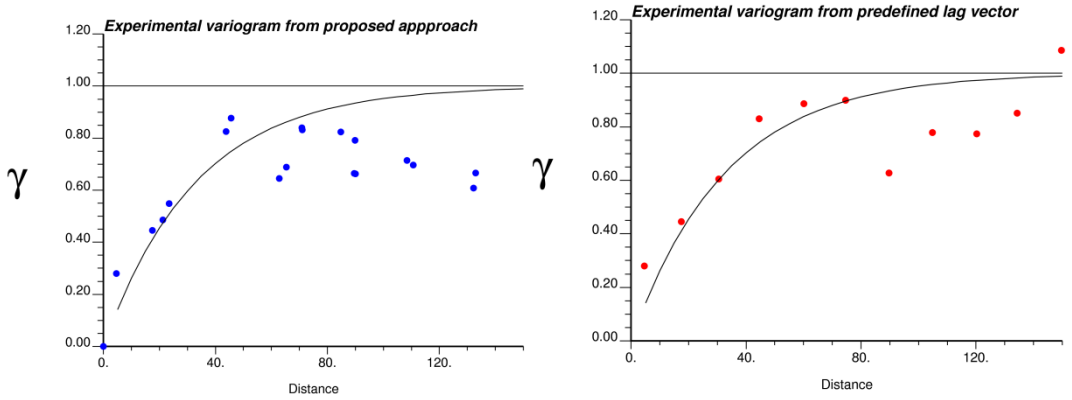

Figure 7: Experimental variogram from the proposed approach with flexible lag definition (Left) comparing with the traditional approach with predefined lag vector (right)

The same comparison is done for all the experimental variogram from 0°NW to 170°NW, along every 10 degree (with 15 degree of direction tolerance and 20 band width), the experimental variograms are calculated using the traditional and the proposed approach. For each experimental variogram point from both of them is compared with the theoretical model equation. The mean square error along each direction is listed in Table 2 and plotted in Figure 8. As it is shown the experimental variogram calculated from the proposed approach is more close the known variogram model.

Table 2 the mean square error of the experimental variogram from the theoretical variogram model

Main Azimuth	Traditional approach	Proposed approach	Improvments
0	0.0348	0.0415	-0.1932
10	0.0305	0.0226	0.2571
20	0.0322	0.0060	0.8128
30	0.0420	0.0146	0.6528
40	0.0514	0.0401	0.2204
50	0.0461	0.0414	0.1017
60	0.0354	0.0342	0.0334
70	0.0157	0.0314	-0.9994
80	0.0339	0.0184	0.4566
90	0.0293	0.0180	0.3844
100	0.0394	0.0302	0.2338
110	0.0441	0.0195	0.5579
120	0.0429	0.0305	0.2883
130	0.0514	0.0267	0.4809
140	0.0578	0.0251	0.5651
150	0.0375	0.0224	0.4032
160	0.0277	0.0225	0.1889
170	0.0422	0.0253	0.4013

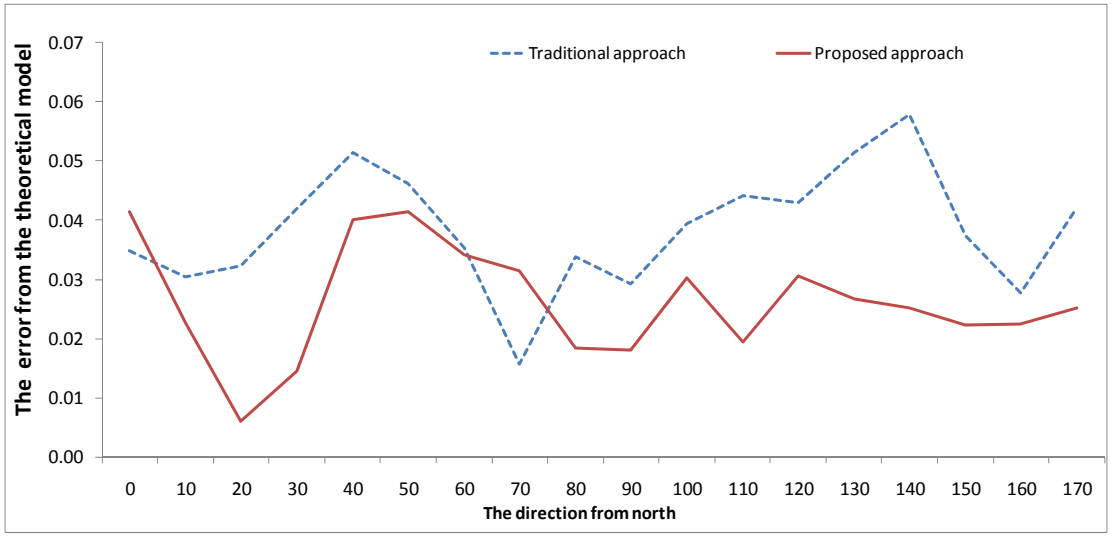

Figure 8: The comparison of experimental variogram calculated from traditional approach with the variogram model

Concluding remarks

Using the new proposed approach can improve the accuracy experimental variogram as considering the data pair configuration. The experimental variogram calculation lag regions are not defined with an equal lag distance and a tolerance around this lag distance but are optimized divided using some artificial algorithm. In this paper, the SOM cluster algorithm is used. Any other optimal algorithm could be used based on the constraints from the distance distribution and minimum number of each cluster.

After the cluster center of each lag region is found from the optimizing process, digging into the distance distribution of these data pairs would provide some spatial information. It would be possible using some cleaning process to these data pairs in order to obtain a more reliable mean from the clustered data pairs.

In the directional experimental variogram calculation, the traditional direction and direction tolerance is used in this research. In the coming research, the direction tolerance picking would be avoided if the entire

clusters are fitted at once. I would be possible using a positive definition surface to fit all the cluster centers. It will be the next step research topic.

Reference

E. Gringartn & C.V. Deutsch. (2001). Teacher's Aide Variogram interpretation and Modeling. *Mathematical Geology , 33* (4), 507-534.
Armstrong, M. (1984). Common problems seen in variograms. *Mathematical Geology , 16* (3), 305-313.
B.P. Marchant and R.M. Lark. (2004). Estimating variogram uncertainty. *Mathematical geology , 36* (8), 867.
Chauvert, P. (1982). AIME annual meeting. *AIME preprint*, (pp. semi-variogram estimatin using a simulated deposit). Dalllas.
Chilés, J. P., & Delfiner, P. (1999). *Geostatistics, Modeling Spatial Uncertainty.* New York: Wiley-Interscience Publication.
Cressie, N. A. (1993). *Statistics for spatial data.* John wiley & sons.
Cressie, N. (1985). When Are Relative Variograms Useful in Geostatistics? *Mathematical Geology* , 693-702.
Cressie, N., & Hawkins, D. (1980). Robust Estimation of the Variogram I. *Mathematical Geology* , 115-125.
Deutsch, C. V., & Journel, A. (1997). *GSLIB: Geostatistical Software Library and User's Guide.* New York: Oxford Press.
Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press.
Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics.* New York: The Blackburn Press.
Kohonen, T. (2001). *Self-organizing Maps.* Springer.
Matheron, G. (1962). *Traité de Géostatistique appliquée.* Paris : Editions Technip .
Pelletier, B. a. (2004). Fitting the Linear Model of Coregionalization by Generalized Least Squares. *Mathematical Geology , 36* (3), 323-343.
Wackernagel, H. (2003). *Multivariate Geostatistcs.* Berlin Heidelberg: Springer-Verlag.