15th International Congress of the Geological Society of Greece
Athens, 22-24 May, 2019 | Harokopio University of Athens, Greece
Bulletin of the Geological Society of Greece, Sp. Publ. **Ext. Abs. 00000**

Application of Variable Lag Variography on Directions Derived Using k-Means Clustering of Sample Pairs

I. Kapageridis, Technological Educational Institute of Western Macedonia, Koila, Kozani, Greece,
ioannis.kapageridis@gmail.com.

Abstract

Variography is an important step in any geostatistical resource estimation study. Calculation and modelling of experimental variograms using samples from irregular sampling patterns is a painful and time-consuming task. The more random the sampling pattern is, the more difficult it becomes to find directions and basic spacings (lags) that have a high enough number of sample pairs to produce reliable variogram points. Even after the application of directional and spacing tolerances, fixed lag variography can fail to produce interpretable experimental variograms that can be used to derive a model of the underlying structure. Finding directions that produce interpretable variograms can also be a fairly difficult task, even with the interactive and dynamic interfaces that most modern geostatistical packages provide. Variable Lag Variography (VLV) based on k-means clustering of sample pairs has been successfully used in the past to address the deficiencies of applying fixed lags and lag tolerances to irregularly spaced data. Sample pairs are grouped into clusters based on their sample distance, with the centre of each group representing a single variogram point lag, and the maximum difference in the group from this centre representing the particular variogram point lag tolerance. The study presented in this paper, takes the concept of variogram sample pair clustering one step further, using a k-means clustering process to derive the most populated directions in the sampling space. Sample pairs are first grouped into clusters based on their azimuth and plunge. Each cluster has a centre representing a particular azimuth and plunge (direction), and a radius representing azimuth and plunge tolerances. In these automatically produced directions, VLV is then applied to produce the experimental variogram, making maximum usage of the available sample pairs. This two-step automated approach can significantly reduce the time and effort required in producing the most interpretable experimental variograms from a set of irregularly spaced samples.

1. Standard Calculation of Experimental Variograms

Most geostatistical software packages follow this concept. As shown in Figure 1a (for a 2D case), for a particular direction chosen, a search area is defined using some direction tolerance which can be controlled both horizontally and vertically (Deutsch *et al.* 1992). Some packages use the same direction tolerance in both cases while others allow for separate tolerances to be applied for azimuth and plunge. These tolerances are allowed to expand the search area as the separation distance increases up to a maximum distance (bandwidth) from the direction vector. This way, the search area begins as a cone of circular or elliptical section (depending on whether the azimuth and plunge tolerances are different), and then becomes a cylinder of similar section to the cone, once the maximum distance from the direction vector is reached. Some packages allow for a separate maximum distance to be applied horizontally and vertically.

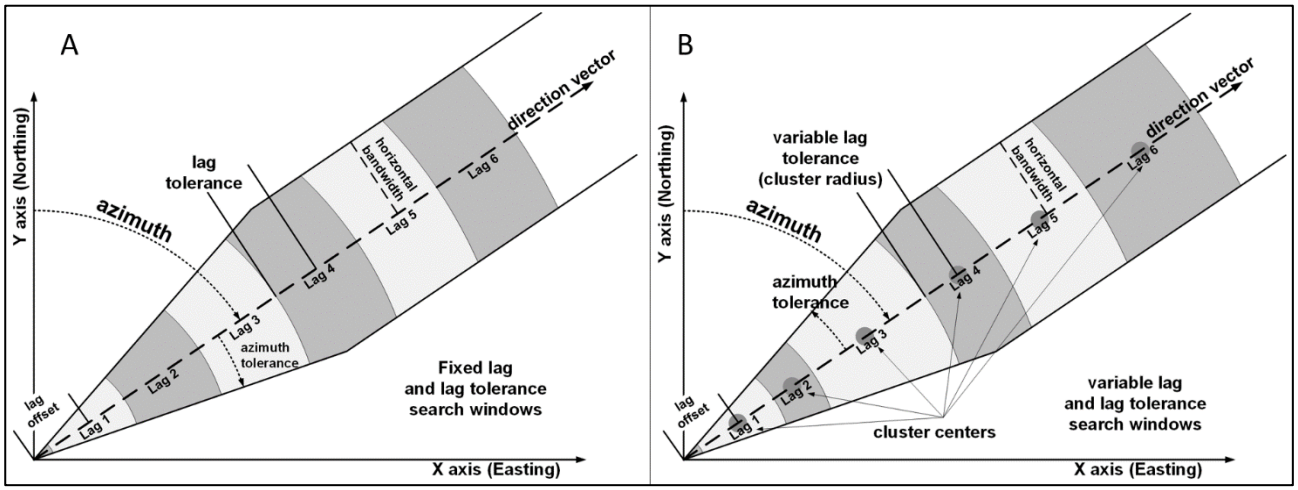


Figure 1. (A) Standard sample pair selection and (B) variable lag sample pair selection based on k-means clustering.

2. k-Means Clustering of Variogram Directions – Step 1

k-means clustering is a method originally used in signal processing, commonly used for cluster analysis in data mining. k-means clustering groups *n* observations into *k* clusters, with each observation assigned to the cluster with the nearest mean. The term "k-means" was introduced by MacQueen in 1967. The standard algorithm was first proposed by Lloyd in 1957, while Forgy published essentially the same method in 1965. A more efficient version was proposed by Hartigan

and Wong in 1975 and 1979. It is an iterative algorithm that is performed in steps. Before any iteration, the clusters are initially centred on an equal number of observations. These observations can be chosen using different methods. Iterations involve two steps. In the first step, each observation is assigned to the cluster whose mean yields the least within-cluster sum of squares. The second step involves the calculation of the new means to be the centroids of the observations in the new clusters. The algorithm converges when there is no change in the assignments. IBM SPSS Statistics was used to provide the k-means clustering functionality in this study.

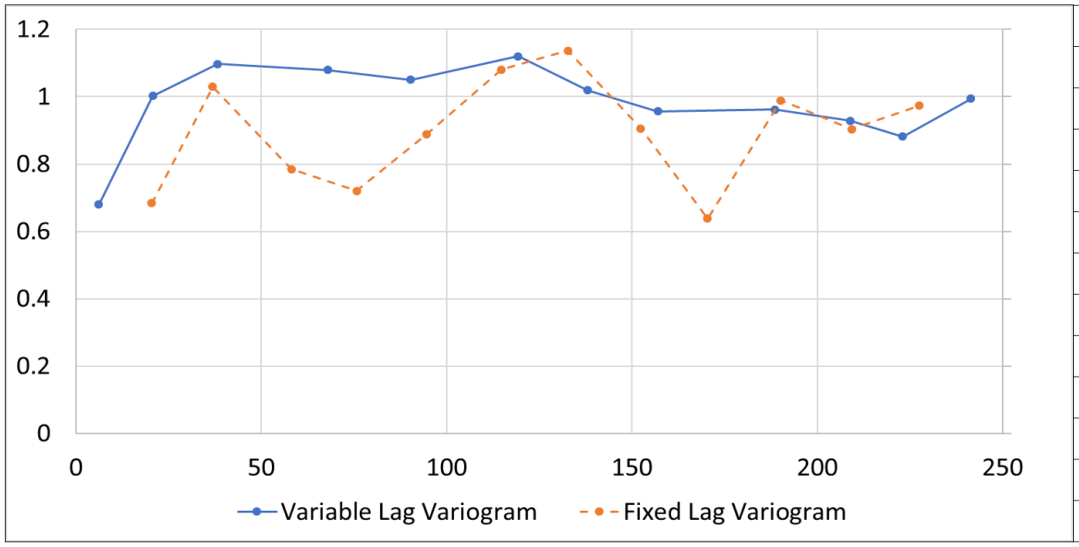
In this study, drillhole samples from an underground tungsten mine with a fairly irregular drilling pattern were used to form pairs for variography purposes. The separating distance, azimuth and plunge of each pair vector were calculated. Sample pairs were grouped into clusters using k-means based on their azimuth and plunge. The produced clusters represented the most populated (with sample pairs) directions in 3D space. The average azimuth and plunge of each cluster constituted the variogram direction, and the difference of the average and the maximum azimuth and plunge of each cluster were considered as the azimuth and plunge tolerances for each direction. The produced directions were not following a fixed angle step, while the tolerances of each direction were different and specific to that direction.

Variable Lag Variography – Step 2

VLV has been introduced by the author as a way to automatically adjust lag parameters to match the spatial distribution of sample locations and improve the resulting experimental variogram (Kapageridis, 2015). As shown in Figure 1b, sample pairs selected for a particular direction using standard criteria or the clustering process described in the previous section, are grouped into clusters based on the 3D distance between their samples. Variography parameters using k-means clustering in VLV are represented by the resulting clustering information:

- *Lag offset*: the average separation of the first cluster (first variogram point).
- *Lag*: the average separation of each cluster (each variogram point) - different for each variogram point.
- *Lag tolerance*: the maximum distance of the pairs classified in a specific cluster from that cluster's center - different for each variogram point, not a fixed value.
- *Pair count*: the number of pairs classified in each cluster.

Figure 2 shows an example of an experimental variogram calculated using the standard method based on fixed lag and lag tolerances (dashed), and an experimental variogram calculated using VLV (solid). The direction in which the variogram was calculated is the same in both cases and was one of the directions derived using the clustering method described in Step 1 (azimuth 54º and plunge -19º). The improvement in the interpretability of the variogram using VLV is clear.



Point	VLV	Fixed Lag
1	7694	294
2	5244	723
3	3172	1707
4	7697	3356
5	9707	3661
6	9633	3772
7	5457	2826
8	6034	3618
9	4084	873
10	3858	2326
11	3874	2263
12	3180	2065

Figure 2. Fixed and variable lag variogram calculated in the same direction, and corresponding numbers of sample pairs.

References

Deutsch, C.V., and Journel, A.G. 1992. GSLIB - Geostatistical Software Library and User's Guide, Oxford University Press.

Forgy, E.W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21: 768–769.

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hartigan, J. A.; Wong, M. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C 28 (1): 100–108.

Kapageridis, I. 2015. Variable Lag Variography Using k-means Clustering. Computers & Geosciences, Volume 85, Part B: 49-63.

Lloyd, S. P. 1982. Least square quantization in PCM. Bell Telephone Laboratories Paper. IEEE Transactions on Information Theory 28 (2): 129–137.

McQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297.