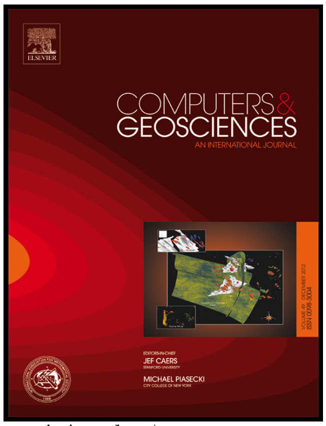

Variable lag variography using k-means clustering

I.K. Kapageridis

PII: S0098-3004(15)00087-4
DOI: http://dx.doi.org/10.1016/j.cageo.2015.04.004
Reference: CAGEO3528

To appear in: *Computers and Geosciences*

Received date: 1 September 2014
Revised date: 29 December 2014
Accepted date: 10 April 2015

Cite this article as: I.K. Kapageridis, Variable lag variography using k-means clustering, *Computers and Geosciences,* http://dx.doi.org/10.1016/j.cageo.2015.04.004

Variable Lag Variography Using k-means Clustering

I.K. Kapageridis

Laboratory of Mining Information Technology and GIS Applications, Department of Environmental Engineering, Technological Educational Institute of Western Macedonia, Greece, Email: ioannis.kapageridis@gmail.com

Abstract

Experimental variography in three dimensions based on drillhole data and current modelling software requires the selection of particular directions (azimuth and plunge) and a basic lag distance. Variogram points are then calculated on distances which are multiples of that basic lag. As samples rarely follow a regular grid, directional and distance tolerances are applied in order to have sufficient pairs to calculate reliable variogram points. This process is adequate when drillholes follow a drilling pattern (even if not an exactly regular grid) but can be time consuming and hard when the drilling pattern is irregular or when drillhole orientations vary considerably. Having all variogram points being calculated on multiples of a fixed lag, and the same tolerance being applied throughout the range of distances used, can be very restrictive and a reason for considerable time wasting or even failure to calculate an interpretable experimental variogram. The method discussed in this paper is using k-means clustering of sample pairs based on pair separation distance leading to a number of clusters each representing a different variogram point. This way, lag parameters are adjusted automatically to match the spatial distribution of sample locations and the resulting variogram is improved. Case studies are provided showing the benefits of this method over current fixed-lag experimental variogram calculation techniques.

keywords: experimental variogram, k-means clustering, variogram modelling

Introduction

Drilling patterns, in mineral exploration programmes in particular, very rarely follow a strictly regular pattern. In some cases, this is due to practical issues causing a deviation from a constantly spaced pattern, while in other cases, it is the geometry of the targeted orebody envelope that requires a more flexible pattern to be followed. As exploration takes place in stages, in-fill drilling to increase the level of confidence in certain areas, also causes local changes in sample spacing. Drilling from underground workings is, in most cases, irregular and leads to extreme variation of sample spacing.

Regardless of the reason and the degree of irregularity, when the sample spacing is not reasonably constant, the variography practitioner can face a very difficult task in finding a set of lag parameters that work for all variogram points in a particular direction. In most cases, parameters that work for the variogram points at smaller separation distances will not work for the points at larger separation distances and vice versa. The problem is further exaggerated by the way the lag parameters are applied by geostatistical software. Depending on how dynamic is the application of the parameters, it is possible to reach, after a considerable amount of time and effort, a set of parameters that works reasonably well for most variogram points. However, some of the geostatistical software is not dynamic at all in the application of lag and direction parameters. Setting of these parameters and running the computation of experimental variogram values are independent and take place separately, in which case, it can be almost impossible to find a set of parameters that will produce an interpretable experimental variogram.

The problem of grouping pairs of samples that fall in a particular direction seems to be quite suitable for solving using a clustering algorithm like the k-means. The main concept is that pairs are selected to belong to a particular direction according to the vector they define and then they are grouped according to a criterion like the separation distance. This way, the user does not have to spend time and effort in finding appropriate lags and tolerances that work in that particular direction. A separate clustering run will have to be performed for each direction considered. The directions themselves could be chosen using a similar approach to better match the most sampled directions, but this is something that most geostatistics practitioners would probably like to keep control of, and computationally it would require a lot more time to perform.

Current State of Variogram Calculation

The information provided in this section is based on the experience and knowledge of the author and does not necessarily cover every geostatistical software package and method available to the geostatistics practitioner. However, the author believes that most of the available packages follow, in some way, one of the paradigms discussed here.

Variogram Points Based on Multiples of a Basic Lag and an Absolute Lag Tolerance

Most geostatistical software packages follow this concept. As shown in Figure 1 (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. 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.

The search area is split into multiple search windows which are defined using a basic lag and a lag tolerance. Each search window is centred on a distance along the direction vector derived by a multiple of the basic lag plus some lag offset. The extents of the search window are controlled by the lag tolerance, which, in most cases, is fixed to an absolute distance value and does not change with distance. For example, if the basic lag is 50m and the tolerance is 15m, then the search window at the sixth variogram point will be centred at 6x50 = 300m and start at 285m and stop at 315m (for a zero lag offset). The sample pairs that fall within the search area are checked against the search windows and they get grouped into different variogram points according to the corresponding separation distance.

As a percentage of the separation distance, the tolerance decreases with every multiple of the basic lag, leading to an ever decreasing number of pairs found at higher distances. In the previous example, for the first variogram point, 15m is 30% of the 50m window distance, 15% of the 100m, 10% of the 150m, and so on. Of course, this is not the only reason for the number of pairs to decrease with distance - it will happen inevitably as we reach the maximum distance between drillholes. However, it is probably the only reason or factor that can be controlled using a different approach of searching for pairs, i.e. a different way of defining the search windows. It should be noted that for variogram smoothing purposes, some of the packages allow overlapping of the search windows, in which case, some sample pairs are used in more than one variogram points.

Some of the main examples of geostatistical packages (the list is far from comlete) that follow the approach described above in two or three dimensions are Isatis (Bleines et al, 2004), GSLIB (Deutsch et al, 1992) including the implementation for standard directional variography in Vulcan 3D software (Maptek Pty Ltd), GEO-EAS (Englund et al, 1991), SGeMS (Remy et al, 2011), and VarioWin (Pannatier, 1996). In some of these software packages, a file containing all pairs of samples and their separation vector parameters (distance, azimuth, plunge) is formed before the variogram points are calculated (called a pair comparison file).

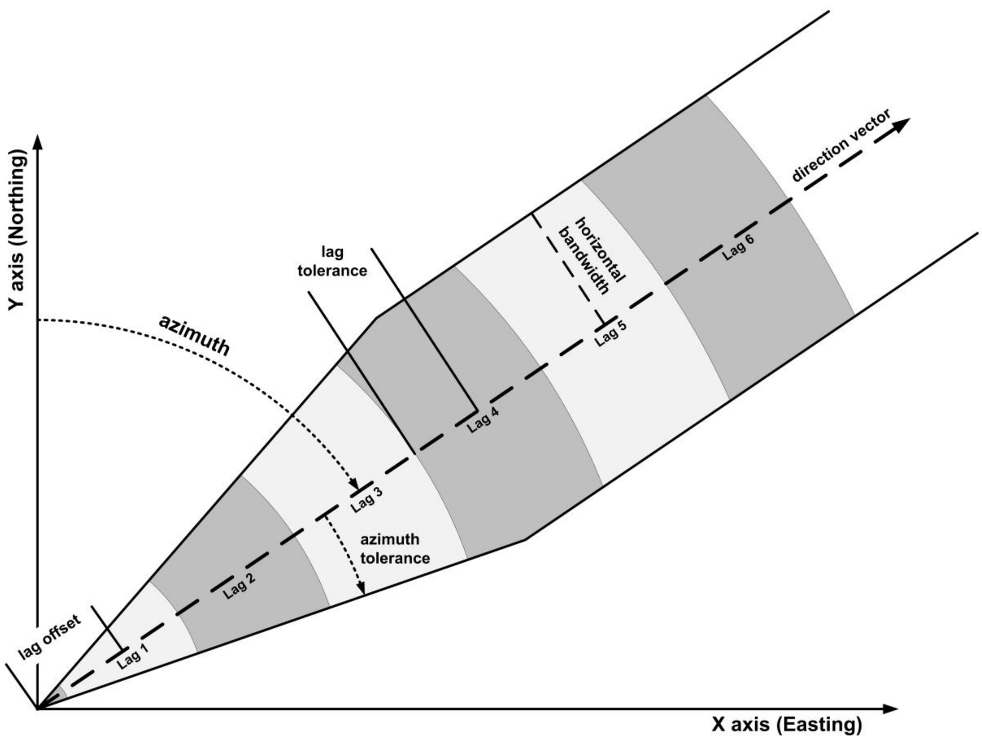

Figure 1: Standard sample pair selection in two dimensions, used in most current geostatistical software.

Variogram Points Based on 3D Block Search Windows

This method is based on orthogonal blocks forming a 3D model centred on the origin of variography polar coordinates – the model always has an odd number of blocks along X, Y and Z. The centroid of each block together with the origin defines a different vector with its own azimuth and plunge. The block extents control the lag, azimuth and plunge tolerance but in a way quite different to the technique described in the previous section. Each of the sample pairs is checked against each block, with one sample at the model origin. If the second sample of the pair falls is nearer the centroid of a particular block, the pair is used to calculate the variogram value for that block (Figure 2). Once all pairs are assigned to particular blocks, the block variogram values are calculated and stored in the model. The end result is a three-dimensional variogram map that can be displayed using a number of different methods (contours, slices, shells, etc.) in two or three dimensions. In addition to the variogram value (different variogram types are available such as standard semivariogram, general relative, pairwise relative, etc.), some other useful information is calculated and store in each block, including the number of pairs, average distance, average head and tail values of the pairs.

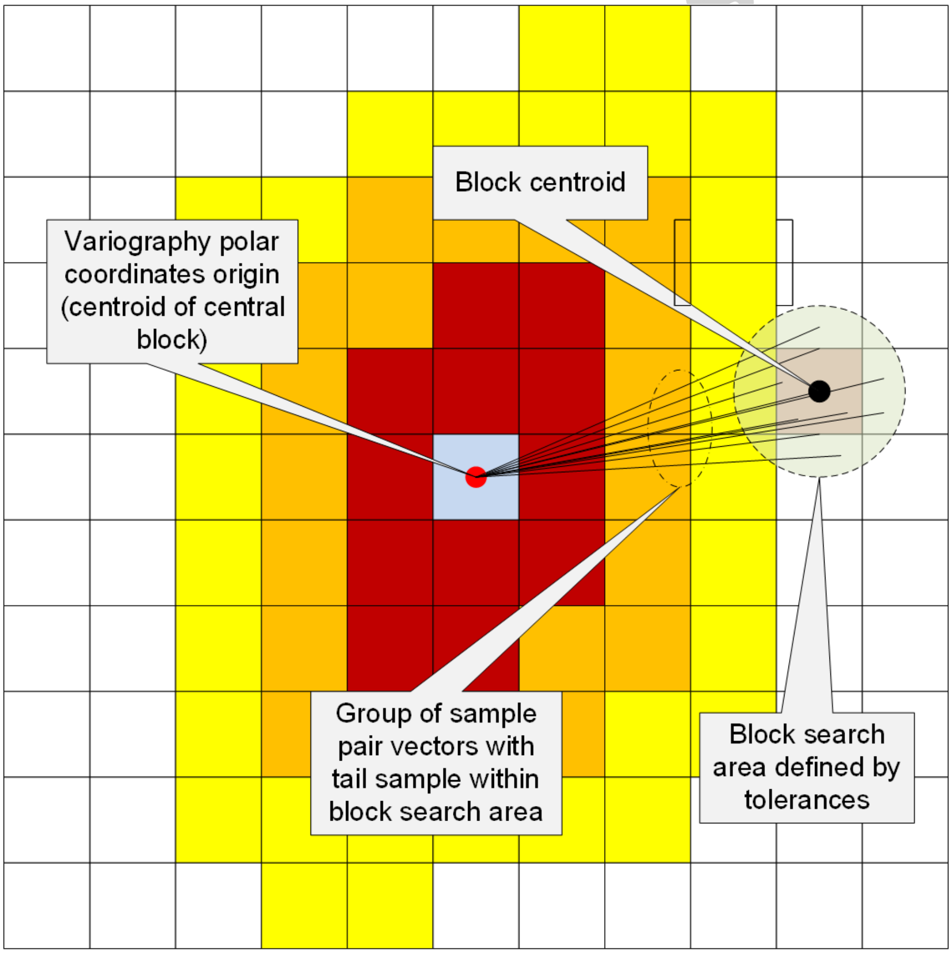

Figure 2: Simple 2D representation of the way sample pairs are selected for the calculation of a particular block variogram value in 3D block variography. Blocks are coloured by variogram value and can potentially reveal the orientation of anisotropy.

The quality of the produced variogram map is controlled by the block sizes and tolerances along X, Y, and Z relative to the sample spacing. Block sizes work similarly to the lag sizes in the previous technique. The tolerances along X, Y, and Z work relative to the block centroids (blocks can be allowed to overlap when checked against sample pairs). As an approach, it is more suitable to investigating the existence and orientation of geometrical (ellipsoidal) anisotropy, and finding the directions that work better with the available sampling pattern, rather than forming the basis for variogram modelling. This technique, called cube variography, is available in the more recent versions of Vulcan 3D (Maptek Pty Ltd).

As a technique, it is still not particularly dynamic in its application, as the block model has to be calculated first using a parameter file that needs to be modified and called again if a different block setup is necessary. However, it is an improvement in the visual aspect of checking the effects of different block sizes (i.e. the effect of different lag sizes) in different directions, potentially leading to a better lag setup in directional variography. The orientation of the anisotropy ellipsoid can be easily determined and the number of directional variograms to calculate can be potentially reduced. At its current form, this technique can work as a preparatory step before the common directional variography described in the previous section.

Variable Lag Experimental Variogram Calculation Based on k-means Clustering

Concept

The techniques described in the previous sections require a time consuming trial and error procedure to be followed in order to reach a lag setup that will produce a reasonably interpretable experimental variogram. The information to reach a good lag setup is already available in the sample pairs for any particular direction, even for irregular sampling patterns, but it is probably too much for the practitioner to handle. Figure 3 shows a histogram of sample pairs based on separation distance for the data of the first case study. Some separation distances present much higher pair frequencies than others, and, together with some very low frequency distances, they can define a group of pairs that could produce a reliable experimental variogram point with sufficient number of pairs. For example, such a point could be considered between 250 and 300 meters around the peak at 280. Other points can be identified in a similar manner. The end result would be a set of experimental variogram points defined at variable separation distances, not multiples of a basic lag, and with varying distance tolerances – a concept called variable lag variography (VLV) from this point on (Figure 4). These points would have sufficient pairs to be considered reliable even at higher separation distances.

Following such logic in a manual way, by examining a histogram like the one in Figure 3, could be beneficial but would still require a fair amount of time and work. It would be better if these groups of pairs can be identified by an automated procedure, as unsupervised as possible. The author has chosen k-means clustering for its speed and simplicity. It is not necessarily the best method for this problem and still requires some minimum input by the user. There are many similar clustering methods and variations, and it is one of the aims for future work to identify one that is more appropriate and will produce the most optimum results. However, k-means proved sufficient to demonstrate the validity of the VLV concept. IBM SPSS Statistics, the software package used for clustering in this study, provides two more clustering methods, the TwoStep Cluster Analysis, and the Hierarchical Cluster Analysis (IBM SPSS Statistics Base 20, 2011).

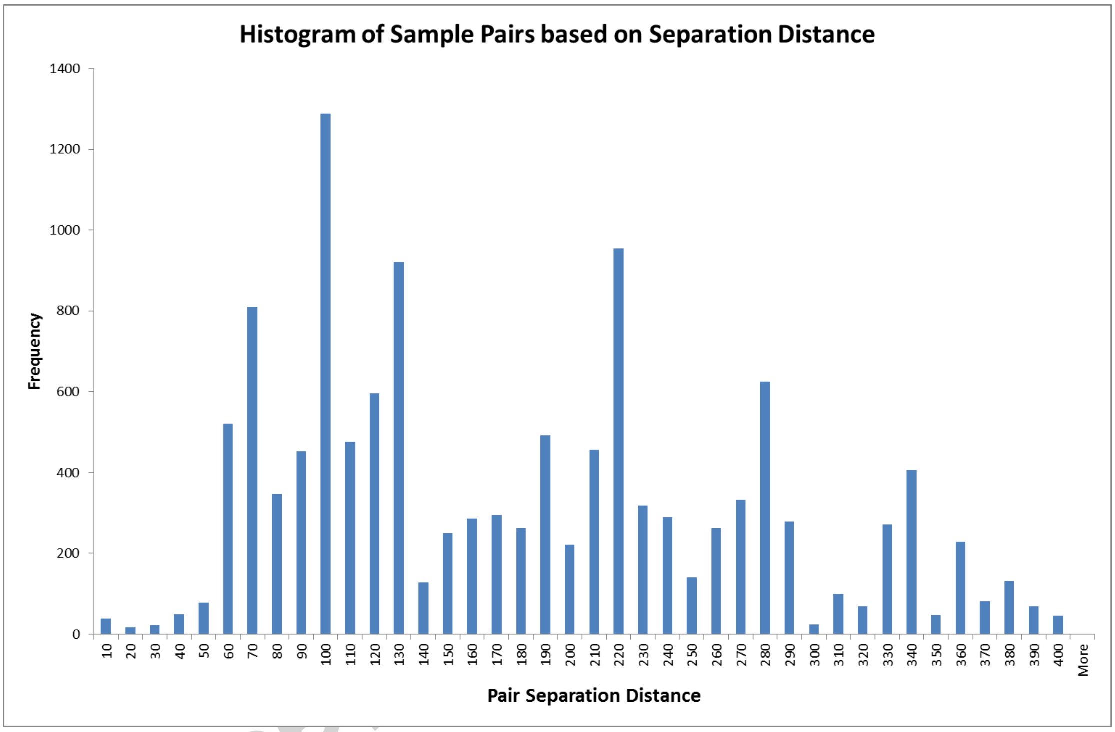

Figure 3: Histogram of sample pairs for a particular direction based on separation distance in the case of the underground drilling pattern of case study 1.

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, not a multiple of a standard distance.
- 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 4 shows how these parameters define search windows in the case of VLV in two dimensions.

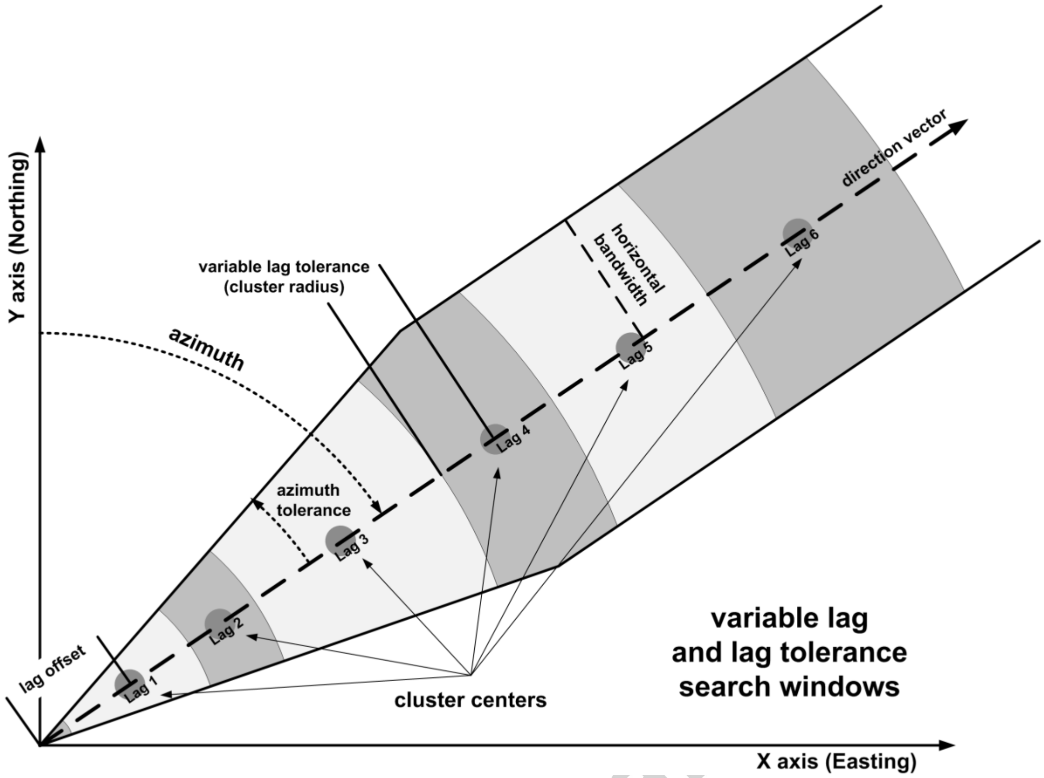

Figure 4: Proposed variable lag sample pair selection based on k-means clustering of pairs.

k-means Clustering

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 as a technique for pulse-code modulation. Forgy published essentially the same method (Forgy, 1965), which is why it is sometimes referred to as Lloyd-Forgy. A more efficient version was proposed and published 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.

The implementation of the k-means clustering algorithm in SPSS (called QUICK CLUSTER) can handle large numbers of cases. It attempts to identify relatively homogeneous groups of cases based on selected characteristics. As with most k-means clustering algorithms, it requires that the number of clusters is specified a priori. The initial cluster centres can be manually selected if required. There are two methods available for classifying cases, either updating cluster centres iteratively or classifying only. Information such as cluster membership, distance information, and final cluster centres can be stored after clustering. Optionally, a variable can be specified whose values will be used to label case-wise output. The first iteration of the algorithm involves three steps (as described in IBM SPSS Statistics 20 Algorithms, 2011):

Step 1 - Initial Cluster Centre Selection

Selection of the initial cluster centres involves a single pass of the data. The values of the first NC (number of requested clusters) cases are selected as cluster centres, and the remaining cases are reprocessed as follows:

a) If $\min_i d(x_k,M_i) > d_{mn}$ and $d(x_k,M_m) > d(x_k,M_n)$, then x_k replaces M_n. If $\min_i d(x_k,M_i) > d_{mn}$ and $d(x_k,M_m) < d(x_k,M_n)$, the x_k replaces M_m; that is, if the distance between x_k and its closest cluster mean is greater than the distance between the two closest means (M_m and M_n), then x_k replaces either M_m or M_n, whichever is closer to x_k.

b) If x_k does not replace a cluster mean in (a), a second test is made:
   - Let M_q be the closest cluster mean to x_k.
   - Let M_p be the second closest cluster mean to x_k.
   - If $d(x_k,M_p) > \min_i d(M_q,M_i)$, then $M_q = x_k$;
     That is, if x_k is further from the second closest cluster's centre than the closest cluster's centre is from any other cluster's centre, replace the closest cluster's centre with x_k.

where, NC is the number of requested clusters, M_i the mean of the ith cluster, x_k the vector of the kth observation, $d(x_i,x_j)$ is the Euclidean distance between vectors x_i and x_j, and d_{mn} is the distance between the two closest means ($\min_{i,j} d(M_i,M_j)$). After one pass through the data, the initial means of all NC clusters are set.

Step 2 – Initial Cluster Centres Updating

Starting with the first case, each case in turn is assigned to the nearest cluster, and the cluster means are updated. The initial cluster centre is included in this mean. The updated cluster means are considered as the classification cluster centres.

Step 3: Cases Assigning to Nearest Cluster

The third pass through the data assigns each case to the nearest cluster, where distance from a cluster is the Euclidean distance between that case and the (updated) classification centres. Final cluster means are then calculated as the average values of clustering variables for cases assigned to each cluster. Final cluster means do not contain classification centres.

When the number of iterations is greater than one, the final cluster means in step 3 are set to the classification cluster means in the end of step 2, and QUICK CLUSTER repeats step 3 again. The algorithm stops when either the maximum number of iterations is reached or the maximum change of cluster centres in two successive iterations is smaller than ε (the convergence criterion) times the minimum distance among the initial cluster centres.

Application of k-means Clustering to Sample Pairs Grouping for Variography

As it was mentioned before, in order to test the proposed methodology, two software packages were used: Vulcan 3D, a mine planning package, and IBM SPSS Statistics, a package used for statistical analysis. Vulcan was used to provide the samples database environment and general variography tools for displaying and modelling. Vulcan's Isis database module provides all the necessary functions for manipulating a drillhole or other sample database, while the Envisage graphical environment provides advanced 3D tools for graphically displaying samples. Vulcan's geostatistical functionality is based on GSLIB (Deutsch et al, 1992). IBM SPSS Statistics was used to provide the k-means clustering algorithm.

A script written by the author in Perl and utilising Vulcan's Lava Perl modules was developed to take care of all the sample pairs preparation work and calling SPSS for clustering. Each direction is processed separately, i.e. the script works in one direction at a time. Lava modules give access to all Vulcan database and model structures as well as the graphical environment through a Perl script. The script allows the user to select the samples database and required sample location and grade fields, as well as the directional parameters for the searching. Currently, it allows the application of an azimuth and plunge with separate tolerances and a bandwidth that is applied in both (Figure 5).

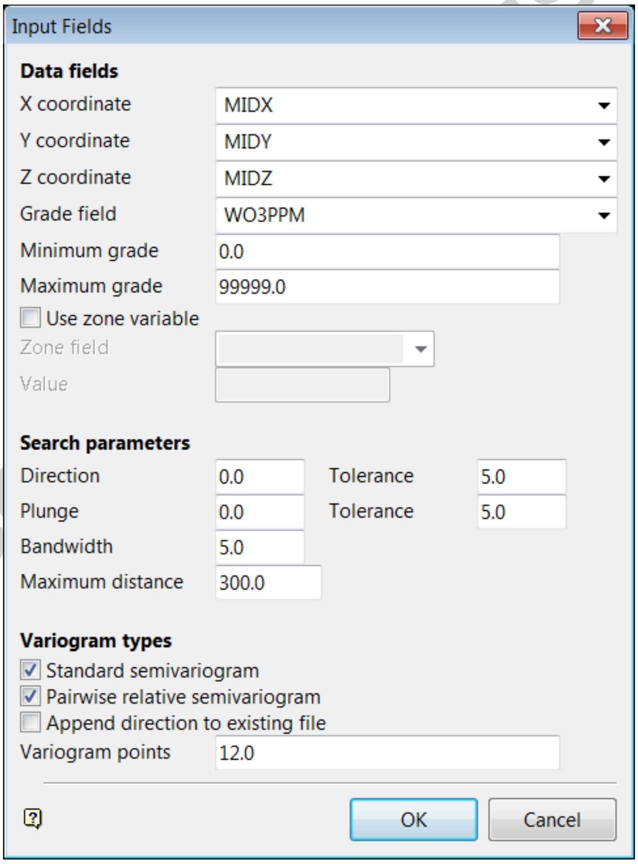

Figure 5: Specification panel from the script responsible for generating the pairs file for a particular direction and direction tolerances and running SPSS for clustering with k-means.

The script goes through the following steps when it runs:

1. User selects the variogram direction and directional tolerances to apply (horizontal, vertical, and related bandwidth). A maximum separation distance can also be applied to speed up the pair formation process.
2. User also selects the required number of variogram points – this controls the number of clusters that will be used by the k-means algorithm.
3. Composites database is scanned and composites pairs are formed and stored to a file (Table 1). The 3D separation distance, squared difference of composites grades (semi), and squared difference divided by pair mean (pairwise) are also stored.
4. An SPSS syntax file is generated referencing the pairs file.
5. SPSS is called using the syntax file and k-means clustering is performed. The Lava script waits for SPSS to complete the clustering process before it continues. The steps performed by SPSS are the following:
   i. Opens the pairs file.
   ii. Executes k-means clustering (QUICK CLUSTER) based on pair separation distance. Two new columns are added to the pairs data table – the resulting cluster number for each pair (QCL_1), and the distance from the cluster centre (QCL_2).
   iii. Aggregates the resulting table based on cluster numbers (QCL_1).
   iv. Calculates average separation distance, maximum distance from cluster centre (QCL_2_max), sum of squared differences (semi), sum of pairwise squared differences and number of pairs (N_BREAK) for each cluster.
   v. Sorts the aggregated table by average separation distance in ascending order (Table 2).
   vi. Saves the aggregated table to a text file.
6. The saved table from SPSS is read by the Lava script and is converted to a Vulcan compatible variogram file that can be opened and displayed in Envisage.

Table 1: Part of a pairs data table after clustering with k-means in SPSS (from case study 1, azimuth 90^{o}, plunge 20^{o}).

pair	sample1	sample2	distance	hordev	verdev	semi	pairwise
P0	H-314-16.622	H-316-11.79	6.963741379	0.095184207	0.358188481	1161.650889	0.471699959
P1	H-314-16.622	H-316-10.8	7.551314654	0.725616456	0.129110504	321960.9171	3.218301598
P2	H-315-5.539	H-314-2.778	3.194344221	0.441519213	0.079610374	185731.6932	2.380098251
P3	H-315-5.539	H-314-1.789	3.982700968	0.153695571	0.239766397	3180251.622	3.483169808
P4	H-315-5.539	H-316-1.89	4.498334581	0.486475155	0.003702789	8043661.738	3.662662936
P5	H-315-5.539	H-314-0.8	4.846072843	0.761050168	0.372952977	34481440.97	3.831501105
P6	H-315-5.539	H-316-0.9	5.056332663	0.347141651	0.247642419	295832122.5	3.941245206
P7	SD-08-43	H-15-52.247	6.678770995	0.301930063	1.11746493	9846435.237	3.643351118
P8	SD-08-43	H-15-53.226	6.050389822	0.223973326	0.492179109	10400625	3.652345679
P9	SD-08-43	H-15-54.205	5.523861783	0.781942854	0.116150433	21137751.05	3.751232685
P10	M-21-19.7	H-112-51.75	45.89995157	2.255489159	0.611567557	1745181.029	3.191502971
P11	M-21-19.7	H-112-52.757	46.57318349	1.510416662	0.821022109	1745181.029	3.191502971
P12	M-21-19.7	H-112-53.764	47.25828136	0.761767907	1.026430026	1745181.029	3.191502971

P13	M-21-19.7	H-112-54.771	47.95348424	0.011819937	1.227332402	1745181.029	3.191502971
P14	M-21-19.7	H-112-55.779	48.66086482	0.7403636	1.425085903	4041768.472	3.438658745
P15	M-21-19.7	H-112-56.786	49.37828845	1.494646484	1.618298387	9740971.829	3.624044192
P16	M-21-19.7	H-112-57.793	50.10516554	2.252006088	1.80931579	9740971.829	3.624044192
P17	SD-08-42	H-15-54.205	5.46253659	0.493360014	0.837717765	20908497	3.671776097
P18	SD-08-44	H-15-52.247	6.677113448	0.05490167	0.150186911	9609398.609	3.474857844
P19	SD-08-44	H-15-53.226	6.15910586	0.506035101	0.467754473	10156969	3.487799746
P20	SD-08-44	H-15-51.268	7.291150869	0.580861371	0.782508272	10384403.13	3.492891538

Table 2: Aggregates table sorted by cluster average separation distance in SPSS (from case study 1, azimuth 90^{o}, plunge 20^{o}).

QCL_1	distance_mean	QCL_2_max	semi_sum	pairwise_sum	N_BREAK	semivariogram	pairwise
3	8.966	8.368	4.93E+10	1755.058	2382	2.07E+07	0.737
2	19.581	5.389	6.50E+10	2308.253	2595	2.50E+07	0.890
4	30.360	8.892	4.04E+10	1515.350	1446	2.80E+07	1.048
1	48.192	8.909	6.94E+10	2572.557	2420	2.87E+07	1.063
6	63.216	13.971	5.64E+10	1825.052	1865	3.03E+07	0.979
5	91.347	14.005	1.43E+09	105.705	222	6.46E+06	0.476

Case Studies

Two case studies using data from real deposits are presented in this paper, selected from a number of examples used to test the VLV approach. Due to the sensitivity of the data, the discussion on its origin and characteristics is kept at a minimum.

Experimental variograms were calculated using standard fixed lag variography in Maptek Vulcan 3D, and using variable lag variography with the developed script and IBM SPSS Statistics. The comparison was made on two variography modes, semivariograms and pairwise relative variograms as these are the only modes currently supported by the script. An effort was made to calculate experimental variogram points using both approaches up to the same maximum distance and for the same number of points to make the comparison easier and more objective.

A small number of directions were selected in each case to calculate variograms. There are minor differences in the application of azimuth/plunge tolerances and bandwidths in the two approaches compared, as the script does not necessarily replicate the way Vulcan 3D forms and selects pairs from particular directions. This has some effect to the difference in the number of pairs reported by the two approaches.

The variogram cloud for each direction considered was constructed to gain some understanding for the produced experimental variogram points. The variogram cloud is commonly used as a diagnostic tool in geostatistics and can help detect the presence of outlier points affecting the produced values in the calculation of experimental variograms (Chauvet, 1982, Isaaks et al. 1989, Cressie, 1991). It is essentially a scatter plot of sample pair squared differences against their separation distance (Figures 8 and 15). Other forms of variogram clouds have been proposed and used, such as the square-root differences cloud (Cressie, 1991). The variogram cloud can be used to detect (Plonner, 1999):

- Global outliers which are clearly away from the main group of the data.
- Local outliers which are more difficult to trace but differ from their neighbouring values and result in high squared differences for small distances.
- Small areas of non-stationarity in cases were a cluster of points presents a larger variability than surrounding points.

Case Study 1 – Tungsten Deposit

Data for the first case study come from a tungsten deposit contained within a number of tabular, bedding-conformable skarn horizons. Data includes both underground and surface drilling. The mountainous terrain of the area and the geometry of the underground workings resulted in a fairly irregular sampling pattern. Drillhole samples from one of the tabular horizons were used to produce equal length (1m) composites of WO_3 values. A total of 5,466 composites were produced and formed the basis for the first case study. Figure 6 shows the spatial distribution of these composites in plan view.

A number of directions were selected to calculate the experimental variogram in semivariogram and pairwise relative mode. These are shown in Figure 6. At 90° and 135° azimuth, two different plunges were tried, 20° and -20°, and 0° and 10° respectively.

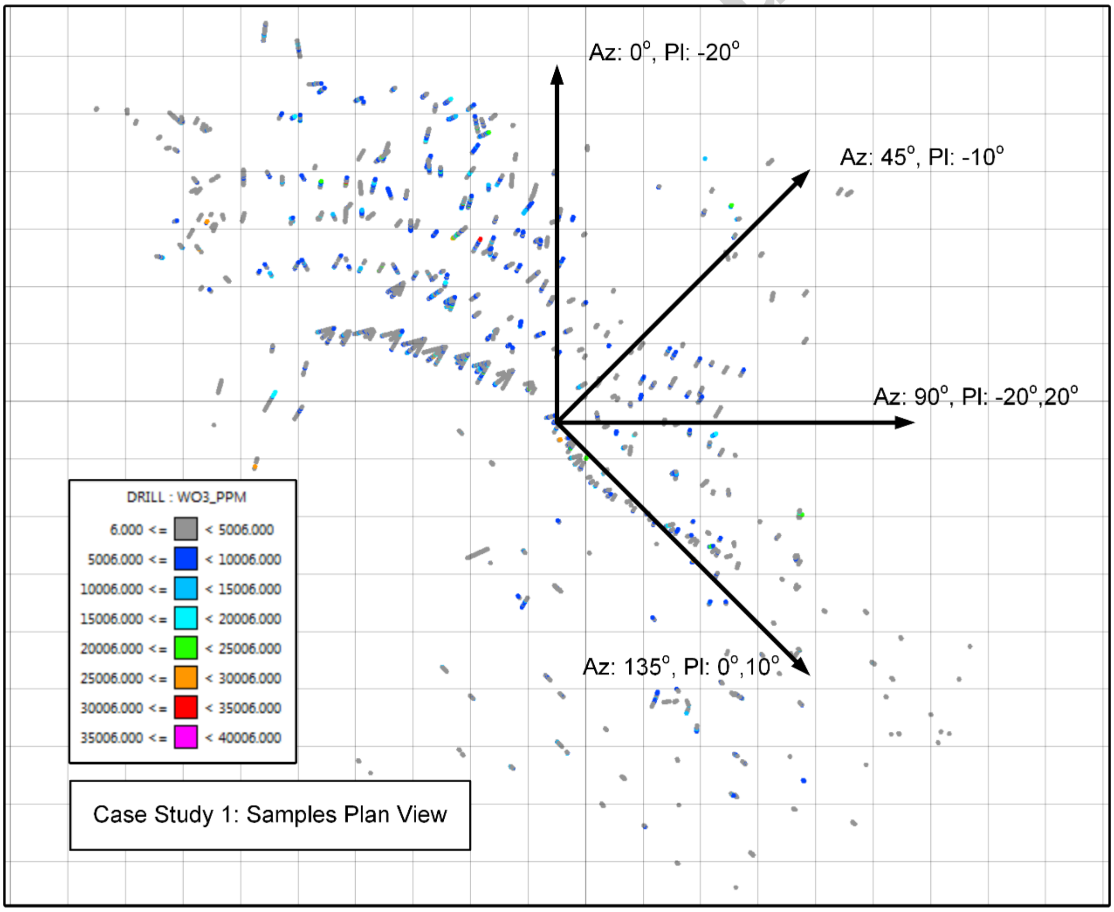

Figure 6: Plan view of underground drillhole sample composites from a particular zone of a tungsten deposit. Irregularity of the pattern is mostly due to drilling following underground openings and drillholes fanning out from almost the same collar location.

For each of the directions, a histogram of selected pairs was produced based on the separation distance to help decide the number of clusters for k-means clustering, i.e. the number of variogram points. The same number of points was used (more or less) in standard fixed lag variography to make comparison of the two approaches easier and more conclusive. The histograms were examined visually in order to establish groups of pairs around frequency peaks. In some cases this was possible while in others not quite. For example, in the middle right histogram (Azimuth 90^{o}, Plunge -20^{o}) of Figure 7, the number of peaks is approximately eight. Thus, during k-means clustering, the number of required clusters was set to eight. However, other histograms were not as clear and presented a much more continuous distribution of pairs along the separation distance bins. In those cases, the number of clusters was set according to the number of drillholes along the particular direction up to the maximum distance considered. Clearly, a direct improvement to the VLV approach would be a different clustering algorithm that can set the number of clusters automatically using some criteria (not necessarily just the pair separation distance).

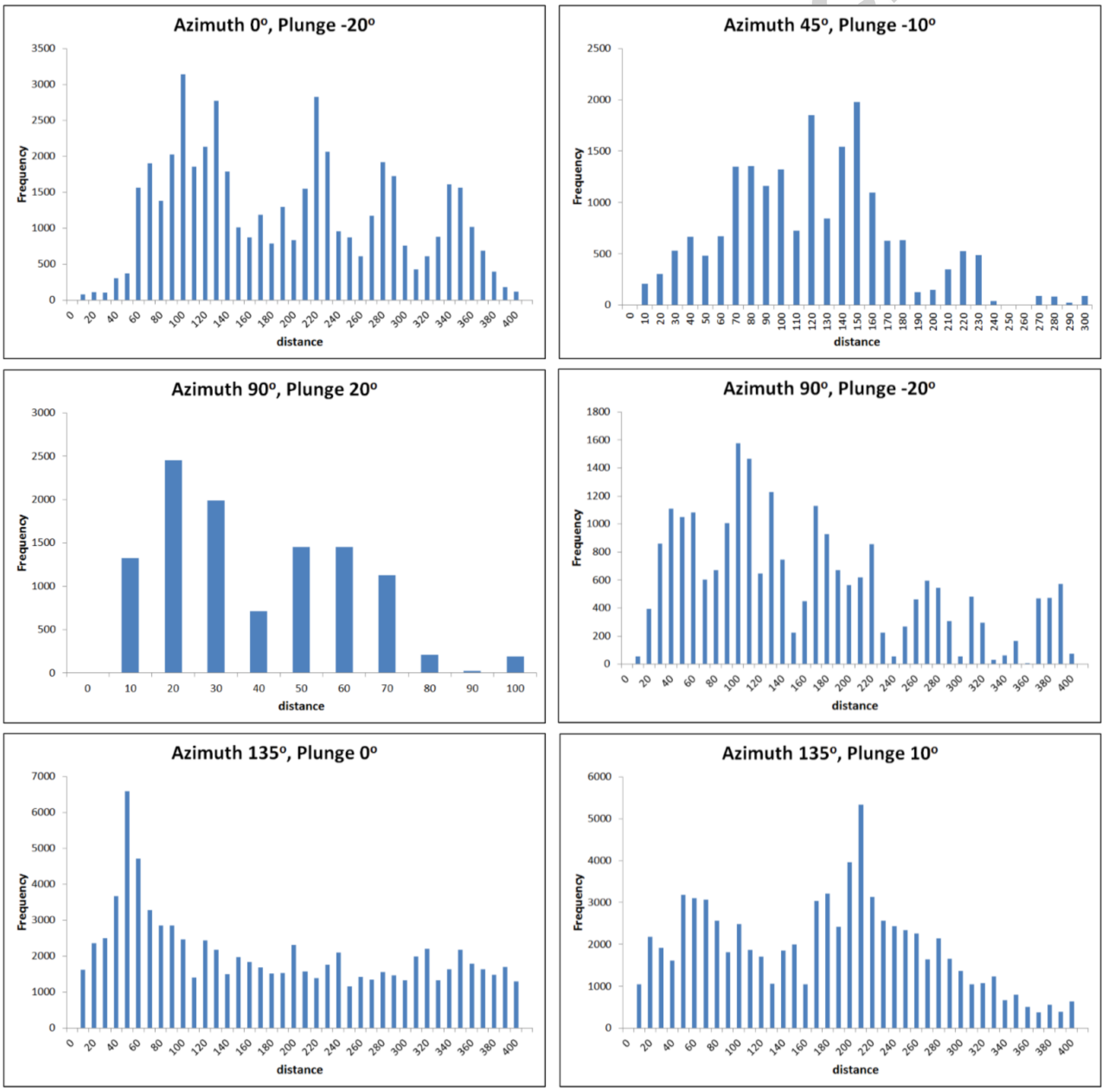

Figure 7: Histograms of pairs based on separation distance for each of the six directions considered in case study 1.

The variogram clouds presented in Figure 8 show the presence of a considerable amount of outliers affecting most distances in each direction considered in this study. The pattern of outliers is mostly uniform in all directions, with the exception of the largest distances of 45°/-10° and 90°/20° where they are absent. This was reflected in the last couple of points of the corresponding fixed lag variogram graphs (Figure 9) and the last point of the corresponding variable lag variogram graphs (Figure 10).

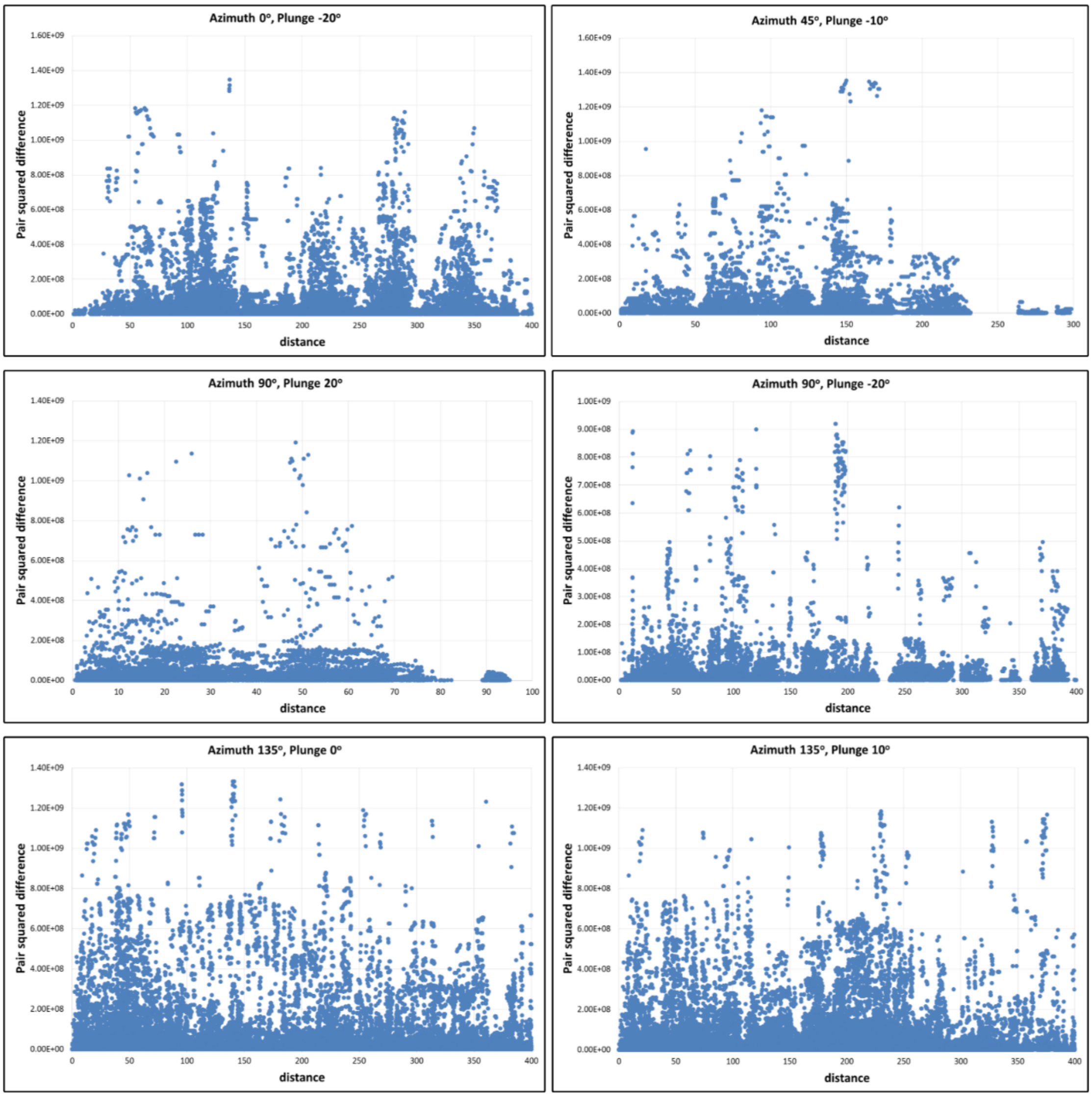

Figure 8: Variogram clouds showing the distribution of pair squared difference values along distance for the six directions considered in case study 1.

Standard fixed lag variography and VLV were performed in all six directions. The lag setups used by the two approaches for each direction are summarised in Table 3. The number of pairs found for each point is also included in the table. In most cases, VLV achieves a much more balanced distribution of pairs along the various points, even at higher separation distances.

Table 3: Variable lag setup defined by k-means clustering and fixed lag setup defined manually for each of the four directions of case study 1.

	Variable Lag				Fixed Lag			
	Point	Lag	Lag Tolerance	Number of Pairs	Point	Lag	Lag Tolerance	Number of Pairs
Azimuth 0, Plunge -20	1	57.73	56.67	5262	1	50	20	7495
	2	93.70	17.98	7702	2	100	20	20132
	3	128.34	23.98	7914	3	150	20	10096
	4	176.31	23.98	4585	4	200	20	13693
	5	220.41	28.15	8274	5	250	20	8022
	6	276.74	28.13	6384	6	300	20	7586
	7	330.01	26.58	4363	7	350	20	10706
	8	361.65	38.34	3027	8	400	20	3607
Azimuth 45, Plunge -10	1	29.29	28.29	2211	1	50	20	2577
	2	71.97	21.32	4607	2	100	20	5179
	3	109.69	20.17	4651	3	150	20	5318
	4	150.18	31.45	5932	4	200	20	1171
	5	213.09	31.43	1623	5	250	20	133
	6	278.82	19.90	288	6	300	20	126
Azimuth 45, Plunge -10	1	8.97	8.37	2382	1	20	10	4227
	2	19.58	5.39	2595	2	40	10	1108
	3	30.36	8.89	1446	3	60	10	1668
	4	48.19	8.91	2420	4	80	10	246
	5	63.22	13.97	1865	5	100	10	202
	6	91.35	14.01	222	6	120	10	0
Azimuth 90, Plunge -20	1	39.78	37.93	4834	1	50	20	3077
	2	90.89	25.55	4802	2	100	20	4974
	3	125.06	22.41	3007	3	150	20	2733
	4	169.95	22.43	3193	4	200	20	2904
	5	207.97	18.99	2319	5	250	20	1460
	6	264.20	27.31	2197	6	300	20	1185
	7	310.20	31.72	958	7	350	20	766
	8	373.64	31.51	1748	8	400	20	692
Azimuth 135, Plunge 0	1	24.26	24.10	10470	1	50	20	14352
	2	57.04	23.74	17363	2	100	20	9186
	3	104.54	26.54	11232	3	150	20	7008
	4	157.62	28.10	9312	4	200	20	6811
	5	213.85	28.10	9926	5	250	20	6043
	6	268.69	27.42	7321	6	300	20	6998
	7	319.53	25.61	9051	7	350	20	7255
	8	370.75	29.22	9025	8	400	20	4477
Azimuth 135, Plunge 10	1	22.48	22.33	7161	1	50	20	9040
	2	60.46	22.26	11820	2	100	20	7982
	3	105.00	28.42	9247	3	150	20	8073
	4	161.86	28.42	11463	4	200	20	15059
	5	205.83	23.51	16326	5	250	20	8841
	6	252.85	26.04	10796	6	300	20	5254
	7	304.95	31.46	7146	7	350	20	2392
	8	367.95	32.05	3388	8	400	20	3307

Figures 9 and 10 present the standard semivariogram in each direction produced by fixed lag variography and VLV respectively. It must be noted that lag tolerances in the case of fixed lag variography were set after some testing while VLV lag tolerances were set automatically by the clustering process. Figures 11 and 12 present the pairwise relative variogram in the same manner. In both variogram modes and in most cases, the points produced by VLV define a smoother, easier to interpret, graph.

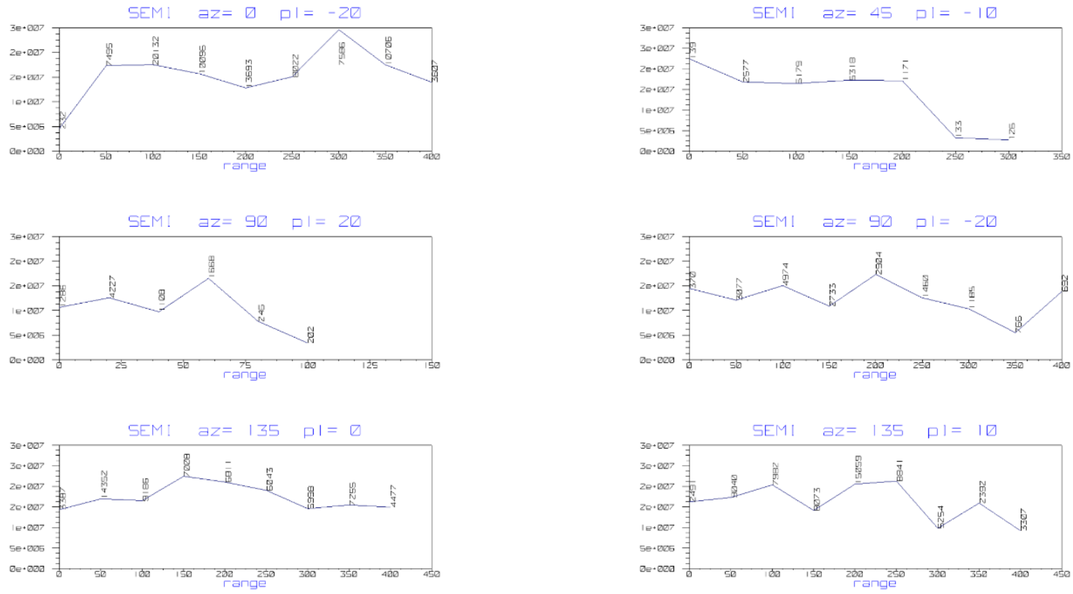

Figure 9: Standard semivariograms produced using fixed lag variography for case study 1.

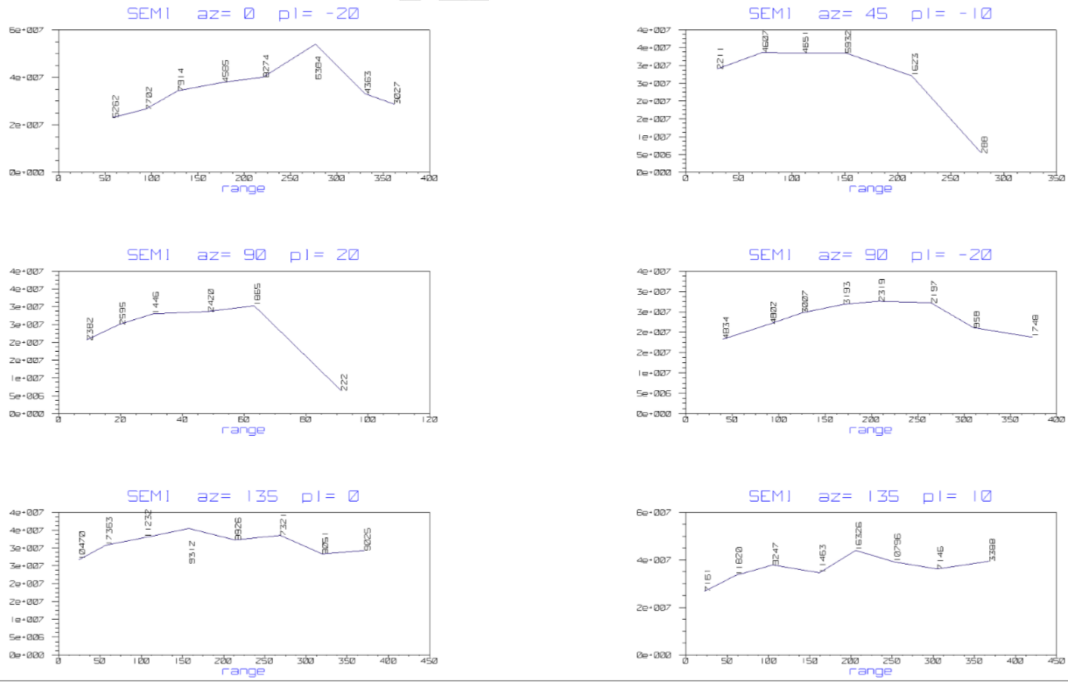

Figure 10: Standard semivariograms produced using variable lag variography for case study 1.

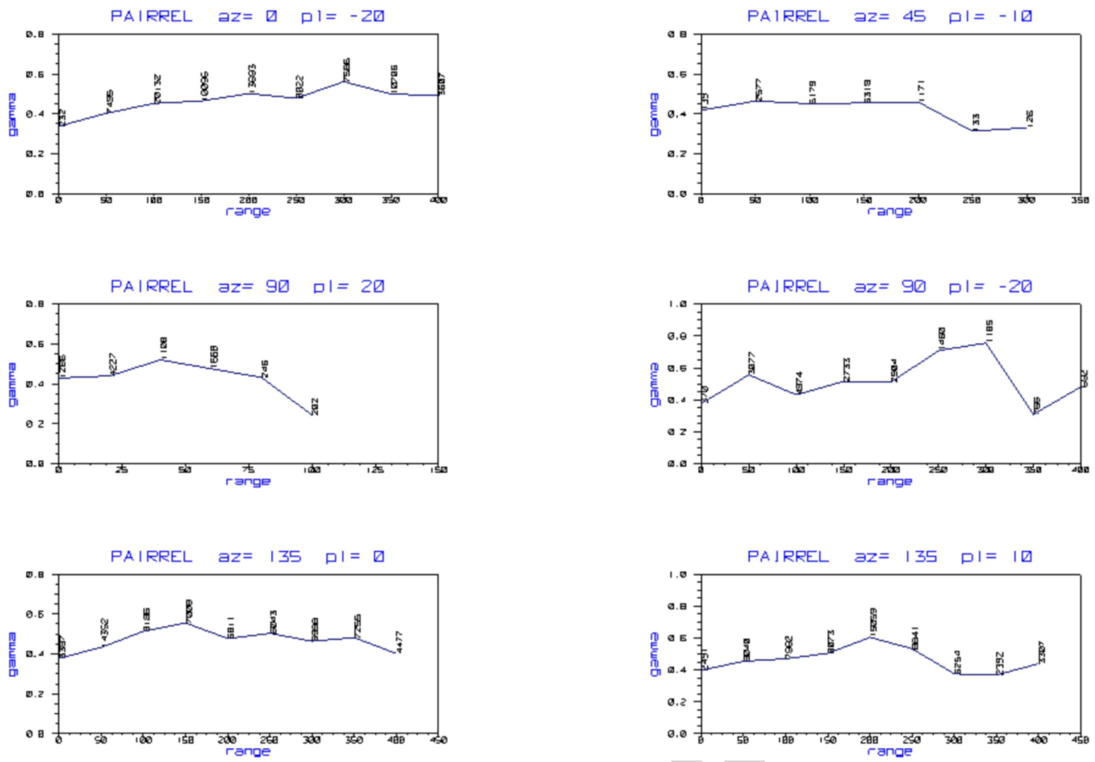

Figure 11: Pairwise relative variograms produced using fixed lag variography for case study 1.

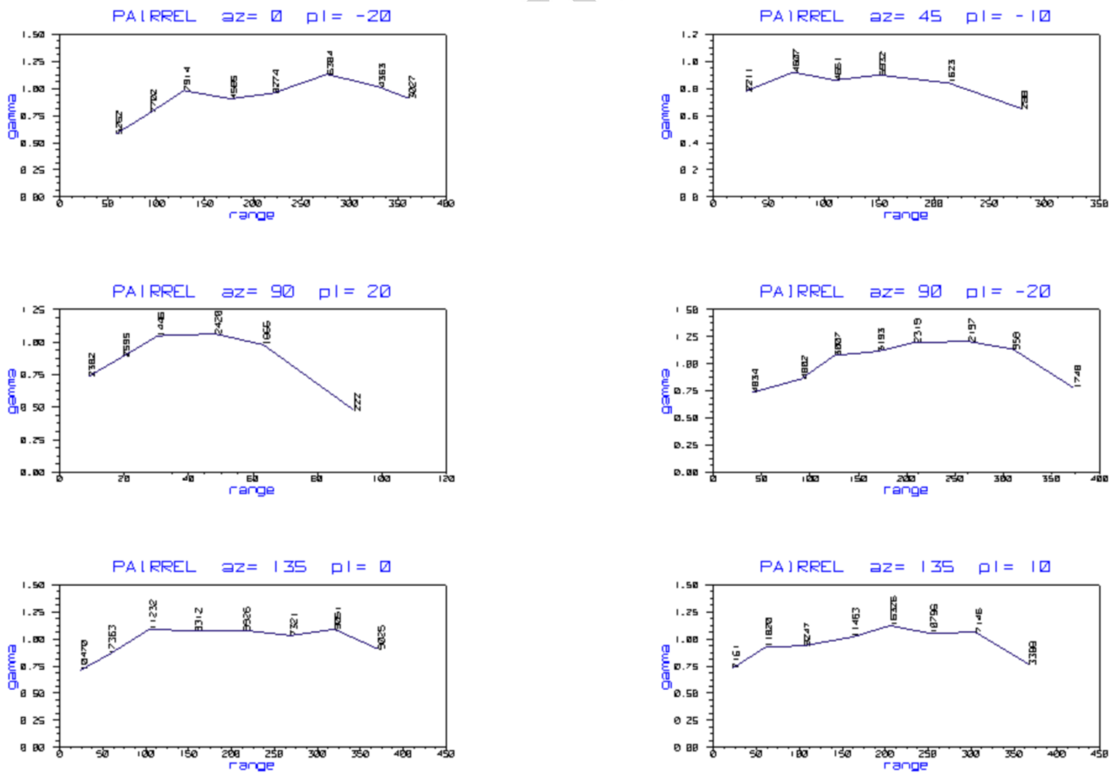

Figure 12: Pairwise relative variograms produced using variable lag variography tools for case study 1.

Running the Perl script to perform VLV required considerably more time than running normal variography tools in Vulcan (a few minutes compared to a few seconds). This is due to the fact that Perl is an interpreted language, and pairs were written to and read from an ASCII file to save memory.

Case Study 2 – Silver Vein Deposit

Data for the second case study include 573 composites of approximately 1m length from a near vertical silver vein. Two separate drilling campaigns (original exploration plus infill drilling) resulted in some irregularity of the sampling pattern – considerably less though compared to the first case study. Figure 12 shows a side view of the data and the selected directions for experimental variogram calculation.

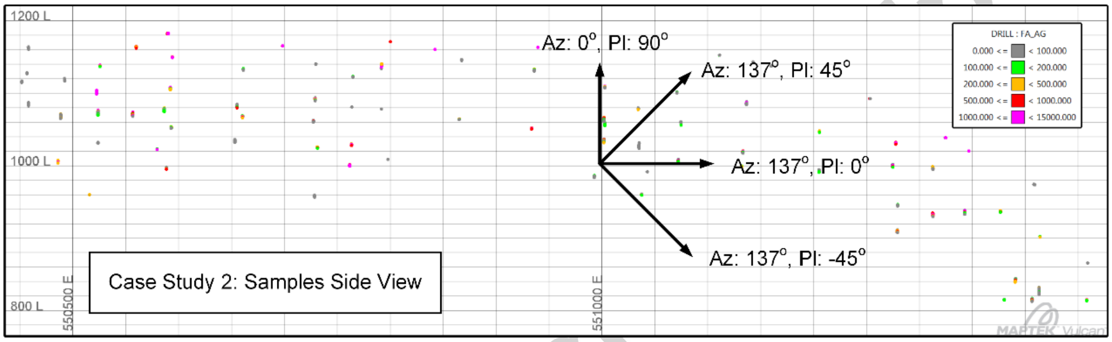

Figure 13: Side view of samples from a vertical silver vein showing a fair amount of irregularity in spacing, partly due to infill drilling.

Histograms were constructed with the pairs according to separation distance as in case study 1. These reflected the much more regular sampling pattern (Figure 14). Particularly the one along the strike direction (azimuth 137°, plunge 0°) presents exactly the original drillhole spacing (50m) through the corresponding peaks. Thus, in the second case study, it was much easier to decide the number of required clusters or variogram points to calculate.

The lag setup used in fixed lag variography and the setup configured with VLV are shown in Table 4. The differences in the way directional tolerances are applied in the two approaches were quite evident in the number of pairs reported. VLV still produced a slightly more balanced distribution of pairs on the different variogram points, but not to the extent shown in the first case study as the sampling pattern is much more regular this time and it is much easier for a fixed lag setup to follow it.

The variogram clouds for the four directions considered are shown in Figure 15. In this case study, the distribution of outliers is not as uniform across the range of distances (with the exception of direction 137°/0° which follows the drilling pattern more closely). This means that there could be room for improvement of the semivariogram graphs by trimming of outliers at appropriate levels.

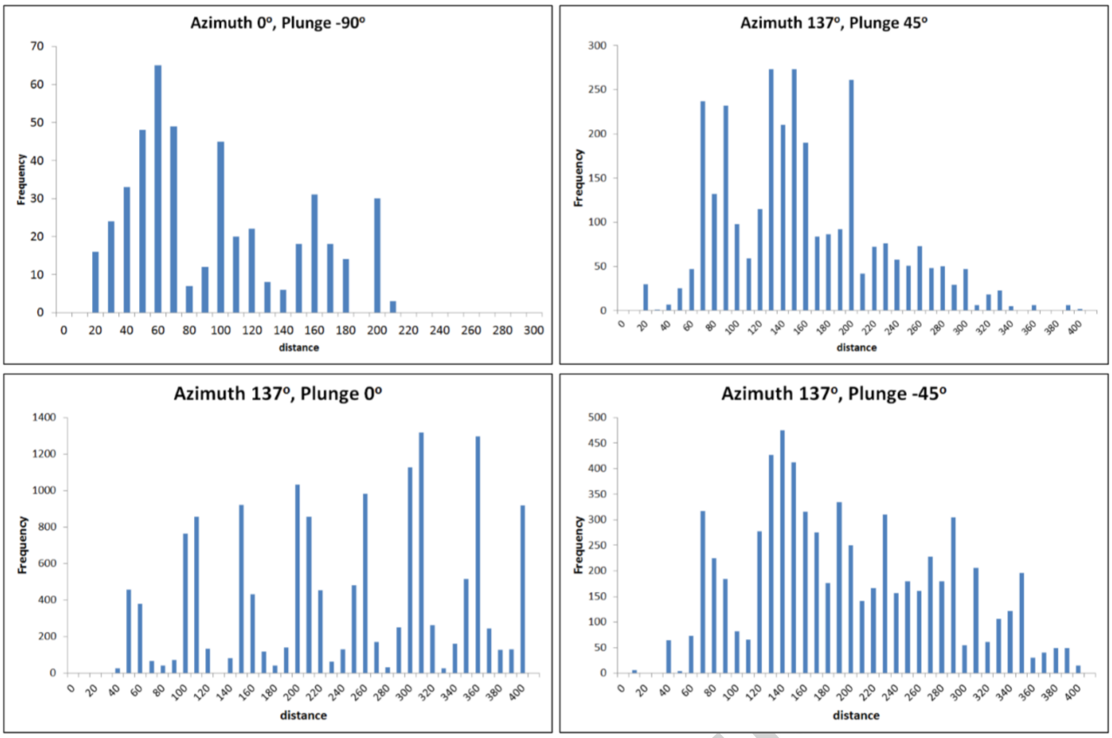

Figure 14: Histograms of pairs based on separation distance for each of the four directions considered in case study 2.

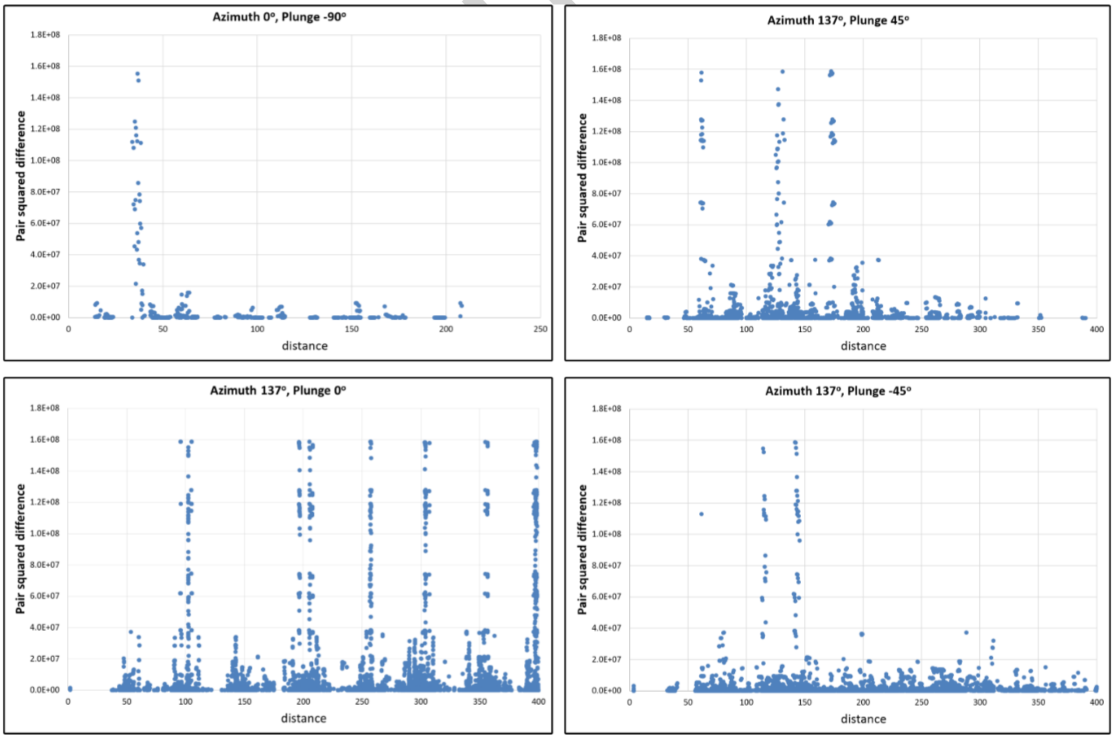

Figure 15: Variogram clouds showing the distribution of pair squared difference values along distance for the four directions considered in case study 2.

Table 4: Variable lag setup defined by k-means clustering and fixed lag setup defined manually for each of the four directions of case study 2.

	Variable Lag				Fixed Lag			
	Point	Lag	Lag Tolerance	Number of Pairs	Point	Lag	Lag Tolerance	Number of Pairs
Azimuth 0, Plunge -90	1	28.12	13.47	70	1	50	15	1303
	2	56.72	23.33	156	2	100	15	658
	3	103.61	28.14	105	3	150	15	865
	4	169.33	30.07	104	4	200	15	310
Azimuth 137, Plunge 45	1	36.31	21.33	82	1	50	20	290
	2	78.45	28.70	772	2	100	20	545
	3	136.14	28.75	1091	3	150	20	845
	4	190.22	28.42	617	4	200	20	496
	5	247.18	30.99	359	5	250	20	188
	6	309.48	81.06	143	6	300	20	116
					7	350	20	19
					8	400	20	25
Azimuth 137, Plunge 0	1	51.22	49.31	941	1	50	20	824
	2	99.86	23.91	1862	2	100	20	1743
	3	149.26	26.03	1598	3	150	20	1472
	4	201.66	24.33	2536	4	200	20	2352
	5	251.72	23.26	1798	5	250	20	1606
	6	300.28	24.30	2991	6	300	20	2876
	7	352.94	24.47	2294	7	350	20	2219
	8	394.03	20.42	1111	8	400	20	2635
Azimuth 137, Plunge -45	1	70.52	66.98	968	1	50	20	437
	2	134.33	31.82	1923	2	100	20	575
	3	180.65	25.21	1132	3	150	20	1462
	4	231.15	26.24	1038	4	200	20	799
	5	283.99	31.32	1029	5	250	20	709
	6	346.71	53.28	630	6	300	20	665
					7	350	20	421
					8	400	20	185

Figures 16 and 17 show the standard semivariogram produced by both approaches. The benefit of using VLV over fixed lag variography is evident as the produced points define a very clear structure. In pairwise relative mode (Figures 18 and 19) the improvement is smaller but still significant. The time required to run VLV was considerably less compared to the first case study as the number of composites and possible pairs was much smaller. The actual clustering process in IBM SPSS Statistics required a couple of seconds in both case studies for each of the directions.

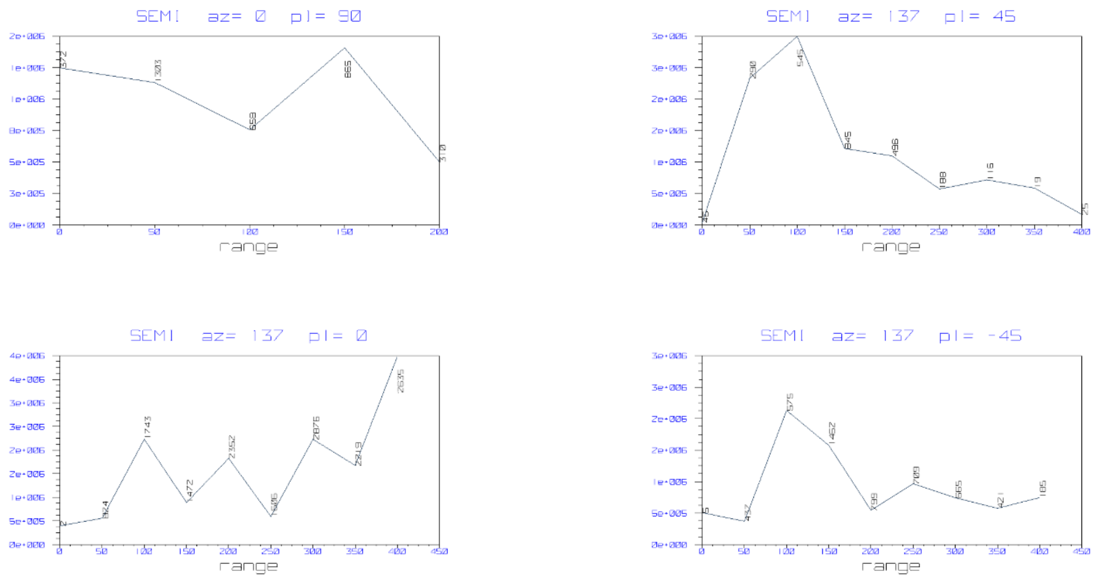

Figure 16: Standard semivariograms produced using fixed lag variography for case study 2.

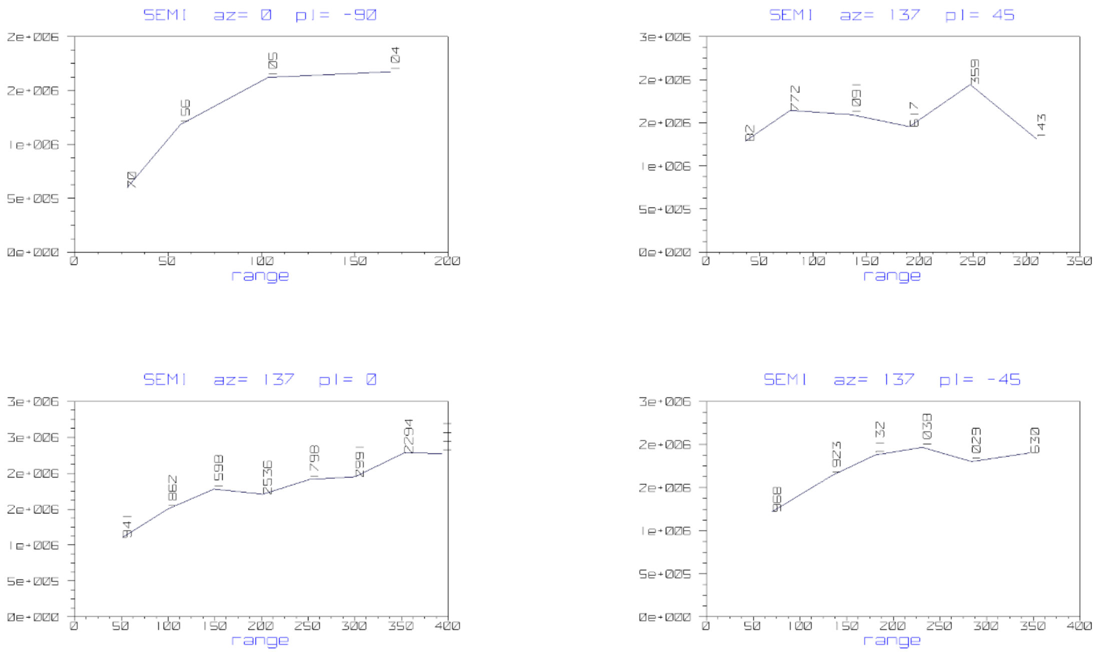

Figure 17: Standard semivariograms produced using variable lag variography for case study 2.

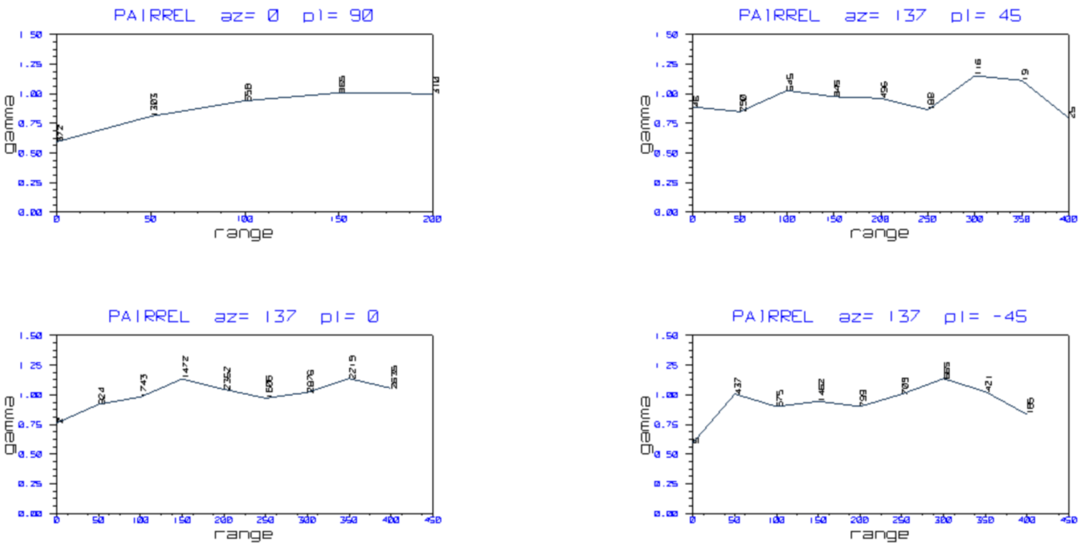

Figure 18: Pairwise relative variograms produced using fixed lag variography for case study 2.

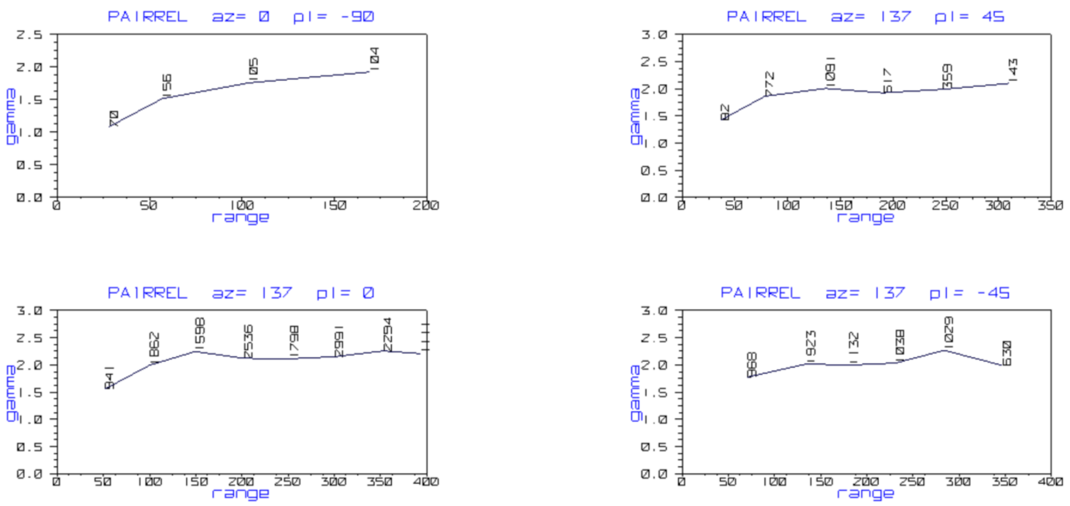

Figure 19: Pairwise relative variograms produced using variable lag variography tools for case study 2.

Conclusions

This paper presented an alternative method for calculating experimental variograms based on variable lags defined through a clustering process. The development of this method was motivated by the excessive time and effort required in setting up lags and tolerances for variography in cases of irregular sampling patterns. K-means clustering was considered as the algorithm for clustering sample pairs based on separation distance. A script was developed that runs through a mine planning package, which creates the pairs, runs the clustering process through a statistical software package and produces the experimental variogram in two variography modes (semivariogram and pairwise relative). The method was tested on data from a number of real deposits, two of which are presented as case studies in this paper. Normal variography was also performed using standard geostatistical tools in order to evaluate the benefits of the variable lag variography (VLV) concept. An effort was made to keep all variography parameters (other than lag and lag tolerance) the same to make the comparison easier and more effective.

The results from calculating experimental variograms using both approaches have shown that VLV can relieve the practitioner from the trouble of finding an appropriate lag setup and at the same time produce experimental variograms which are smoother and easier to interpret in cases of irregular sampling patterns. More testing is required with other datasets to have a better understanding of the effects of using VLV.

Some input is still required to VLV as the k-means algorithm does not define the number of clusters automatically. It is one of the aims for future work to develop a method to find the optimum number of clusters (or variogram points) based on a given set of pairs. Currently, VLV is using separation distance as the sole criterion for clustering. Other criteria are considered, such as the squared difference of the pair sample grades. The effect of such criteria needs to be evaluated.

The script that was developed to perform VLV will also be rewritten to speed up the pair formation process. An implementation of the k-means algorithm will also be included in the scrip so that an external statistical package is no longer required, if such development does not have a negative effect on the speed of clustering.

References

Bleines, C. J., Deraisme, F., Geffory, N., Jeannee, S., Perseval, F., Rambert, D., Renard, O., Torres, and Touffait, Y. (2013) Isatis Beginner's Guide, Geovariances & Ecole Des Mines De Paris.

Chauvet, P. (1982). The Variogram Cloud. In Proceedings of the 17th APCOM Symposium. Colorado School of Mines, Golden, CO, pages 757-764.

Cressie, N. A. C. (1991). Statistics for Spatial Data. New York: Wiley & Sons.

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

Englund, E., and Sparks, A. (1991). GEO-EAS 1.2.1 User's Guide, US Environmental Protection Agency & Computer Sciences Corporation.

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.

IBM SPSS Statistics 20 Algorithms (2011), IBM Corporation.

IBM SPSS Statistics 20 Base (2011), IBM Corporation.

Isaaks, E. H. and Srivastava, R. M. (1989). An Introduction to Applied Geostatistics. New York: Oxford University Press.

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.

Pannatier, Y. (1996). VARIOWIN: Software for Spatial Data Analysis in 2D. Springer-Verlag, New York.

Ploner, A. (1999). The Use of the Variogram Cloud in Geostatistical Modelling, Environmetrics, 10, 413-437: John Wiley & Sons.

Remy, N., Boucher, A., and Wu, J. (2011). Applied Geostatistics with SGeMS – A User's Guide. Cambridge.