Econ. Environ. Geol.
Vol. 28, No. 3, p. 271-277, 1995.

Variogram Analysis for Spatial Similarity Measures : A Case Study using Geochemical Data Sets in the Taebaek Area

Kiwon Lee* and Byung-Doo Kwon*

ABSTRACT : The geological information analysis based on spatial statistical techniques have been studied in relation to mineral exploration. The applicability of outlier detection using moving-window statistics and directional cross-variography analysis have been verified by using geochemical data sets surveyed in the Taebaek area for mineral exploration. The directional variogram analysis has been basically known as a geostatistical method for spatial continuity measures. In this study, the application of this proposed method was extended to measure spatial correlation or similarity problems between two geochemical elements. For the appraisal of the usefulness of this scheme, five kinds of variogram functions were computed for original data and revised data, obtained by removing outliers detected by moving-window statistics and the results were compared. It is concluded that these advanced spatial statistical methods at the interpretation stage of spatial similarity provide us with valuable quantitative results as decision-supporting information for regional mineral exploration task.

INTRODUCTION

Recently, spatial statistics has taken on a renewed interest along with various developments and applications of GISs (Geographic Information Systems) to geological problems. Furthermore, the spatial analyses to explore mineral deposits based on geostatistics have been studied as the viewpoint of mathematical geology (Clark, 1982; Krige and Magri, 1982; Myers *et al.*, 1982; McArthur, 1988; Pan, 1993). Especially, when handling geochemical data sets for applications of GISs to mineral exploration in a large region, the following queries are raised:

(1) How do we detect outliers in each data set, if any ?
(2) How much spatially correlated between geochemical elements ?

In conjunction with these questions, unclassified outliers may severely affect to interpretation of a decision-supporting layer generated by one of GISs' modules for data integration or by other spatial reasoning techniques. Therefore, Haslett *et al.* (1990) discussed the practical importance of EDA (Exploratory Data Analysis) for geological problems concerning geochemical data sets because most of them are intrinsically or spatially correlated.

As for the data representation and the subsequent interpretation stages, it should be significantly considered to find statistical inter-relationship among geochemical elements in a certain study area, particularly when spatial integration task is performed. Normally, the correlation coefficient matrix method is a basic and useful statistical method to represent the relationship among multivariables located at same sampling site: however, it does not provide enough information to recognize the spatial similarity between geochemical elements surveyed at the decision-making stage.

The main objectives of this study are to investigate the applicability of spatial geostatistics to detect outliers at each data set and to check spatial similarity or variability between control variable and target variable using variogram analysis. By the application of CV (Cross Variography), one of variogram analyses, spatial continuity measures can be used to find spatial interrelationship between two variables. While, the variogram studies on univariate problems have been partially used for geological application on mineral deposit (McArthur, 1988), but there are few applications of CV for GISs' aspect. In order to verify the application of these approaches to actual geological problems, a case study using geochemical data sets such as Cu, Pb, Zn, and Ag (Jin *et al.*, 1983) with regards to the Homyeong geologic map (1 : 50,000 scale : 1962) located in the Taebaek area was performed and discussed.

The moving-window statistics (Issaks and Srivatava, 1989; Murray and Baker,1991), one of EDA approaches, was first used to detect outliers, and then the CV methodologies, one of the various application of spatial continuity measures to geological problems, were carried

*Dept. of Earth Science, Seoul National University, Seoul 151-742, Korea

out to check spatial similarity of the data sets.

EDA and CV

The general steps for geostatistics emphasized on measures of spatial continuity can be outlined: descriptive statistics, EDA, variogram analysis and statistical modelling. Descriptive statistics refers to a general process to obtain fundamental statistical parameters such as mean, skewness, kurtosis, variance, and so forth. Though this process is based on histogram analysis without consideration of any spatial properties of a given data sets, the descriptive statistics related to central tendency provides useful information for further geostatistical applications.

EDA is conceptually to explore spatial data themselves; therefore, one of main purposes using this scheme is to detect outliers existed within a data set, if any. The term 'outlier' refers to the abnormally large or small values in a given spatial data set. Sometimes, outliers are not distinguished from true anomalous data indicating existing ore deposits. Therefore, to confirm whether those estimated outliers are meaningless values or not in each data set, some independently supporting data or information are necessary. If a value is estimated as an outlier after proper application of EDA, it is normally ignored during further statistical processes. Meanwhile, moving-window statistics (Issaks and Srivastava, 1989; Murray and Baker, 1991) is known to be more effective than the original EDA scheme using the two-way table methodology proposed by Tukey(1977), especially when handling geochemical element data sets (Haining, 1990). As for EDA using moving-window statistics, summary statistics of a variable with local window that makes up a study area are first determined. In general, the summary statistics are composed of local window mean and standard deviation. These two parameters are also used to check whether a variable contains proportional effect, which represents fluctuated local average and local variability, or not. Furthermore, the outlier can be easily detected by the scatterplot using these two parameters, if it is possible to assume normal distribution or other well-known statistical distribution.

As the viewpoint of practical aspects, it is known that the proper choice of moving-window size depends on the coefficient of variation obtained by descriptive statistics. If the coefficient of variations is greater than a critical value, which is normally taken to be unit number, many data per size of local window are needed.

CV (Cross Variography) was initially motivated by variogram analysis in order to measure spatial continuity; furthermore, this concept can be utilized as a direct indicator to find the spatial similarity and variability in bivariate problems.

Suppose that h_{ij} is a vector lag distance with respect to (i,j) position on spatial grid system of the base map. Then, variogram analysis based on spatial statistics offers quantitative spatial relationship between a value of control variable at a certain position A and that of target variable at a shifted position $A+h_{ij}$. Therefore, CV is defined as a study for bivariate spatial statistics associated with directional aspects. Using this concept, quantitative measures related to spatial relationship can be obtained by the following equations (1)~(5). The control variable and the target variable involved in each equation are denoted by u and v, respectively. While, N(h) is the total number of lag distance along the given direction from (i,j) position. γ_{uv}, $\gamma_{s_{uv}}$, and M_{uv} at equations (1)~(5) refer cross semivariogram, cross standardized semivariogram, cross covariance function, cross correlogram or cross coefficient correlation function, and cross semimadogram, respectively. m_{u-h} and σ^2_{u-h} of equation (6) and (8) are the mean value and the standard deviation value of the control variable at locations that are $-h$ away from target variable data locations, respectively.

While, m_{v-h} and σ^2_{v+h} in equations (7) and (9) are the mean value and the standard deviation value of the target variable at locations that are $+h$ away from control variable data locations, respectively.

$$\gamma_{uv}(h)=\frac{1}{2N(h)}\sum_{(i,j)\,|\,h_{ij}=h}(u_i-u_j)\cdot(v_i-v_j) \tag{1}$$

$$\gamma_{s_{uv}}(h)=\frac{\gamma_{uv}(h)}{\sqrt{\sigma_{v-h}\cdot\sigma_{v-h}}\cdot\sqrt{\sigma_{u-h}\cdot\sigma_{u-h}}} \tag{2}$$

$$C_{uv}(h)=\frac{1}{N(h)}\sum_{(i,j)\,|\,h_{ij}=h}u_i\cdot v_j-m_{u-h}\cdot m_{v+h} \tag{3}$$

$$\rho_{uv}(h)=\frac{C_{uv}(h)}{\sigma_{u-h}\cdot\sigma_{v+h}} \tag{4}$$

$$M_{uv}(h)=\frac{1}{2N(h)}\sum_{(i,j)\,|\,h_{ij}=h}\sqrt{|u_i-u_j|}\cdot\sqrt{|v_i-v_j|} \tag{5}$$

$$m_{u-h}=\frac{1}{N(h)}\sum_{(i,j)\,|\,h_{ij}=h}u_i \tag{6}$$

$$m_{v-h}=\frac{1}{N(h)}\sum_{(i,j)\,|\,h_{ij}=h}v_j \tag{7}$$

$$\sigma^2_{u-h}=\frac{1}{N(h)}\sum_{(i,j)\,|\,h_{ij}=h}u_i^2-m^2_{u-h} \tag{8}$$

$$\sigma^2_{v+h}=\frac{1}{N(h)}\sum_{(i,j)\,|\,h_{ij}=h}v_j^2-m^2_{v+h} \tag{9}$$

While, the symmetric properties of these cross vario graphy functions are as follows:

$$\gamma_{uv}(h) = \gamma_{vu}(h) = \gamma_{uv}(-h)$$

$$\gamma_{suv}(h) = \gamma_{svu}(h) = \gamma_{suv}(-h)$$

$$C_{uv}(h) = C_{vu}(h) \neq C_{uv}(-h) \quad (10)$$

$$\rho_{uv}(h) = \rho_{vu}(h) \neq \rho_{uv}(-h)$$

$$M_{uv}(h) = M_{vu}(h) = M_{uv}(-h)$$

By derivation for relationships between equations (1) and (3), though it is not presented in this paper, it can be proven that the greater the cross semivariogram value is, the smaller the cross covariance function and the cross correlogram is. Therefore, it is implied that cross semivariogram is regarded as an indicator of spatial variability which means the spread of scatterness, and cross correlogram and cross covariance function are that of spatial similarity between control variable and target variable chosen. Although γ_{uv}, γ_{suv}, C_{uv}, ρ_{uv}, and M_{uv} are formulated as different expressions, their physical meanings are closely related with each other. Therefore, the combined interpretations of these parameters may provide more confirmative information on spatial conti-

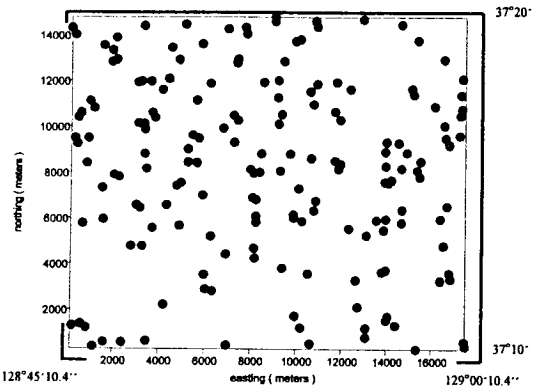

Fig. 1. Geochemical sampling sites within the Homyeong geologic map (scale 1:50,000) in the Taebaek area (Jin *et al.* 1983).

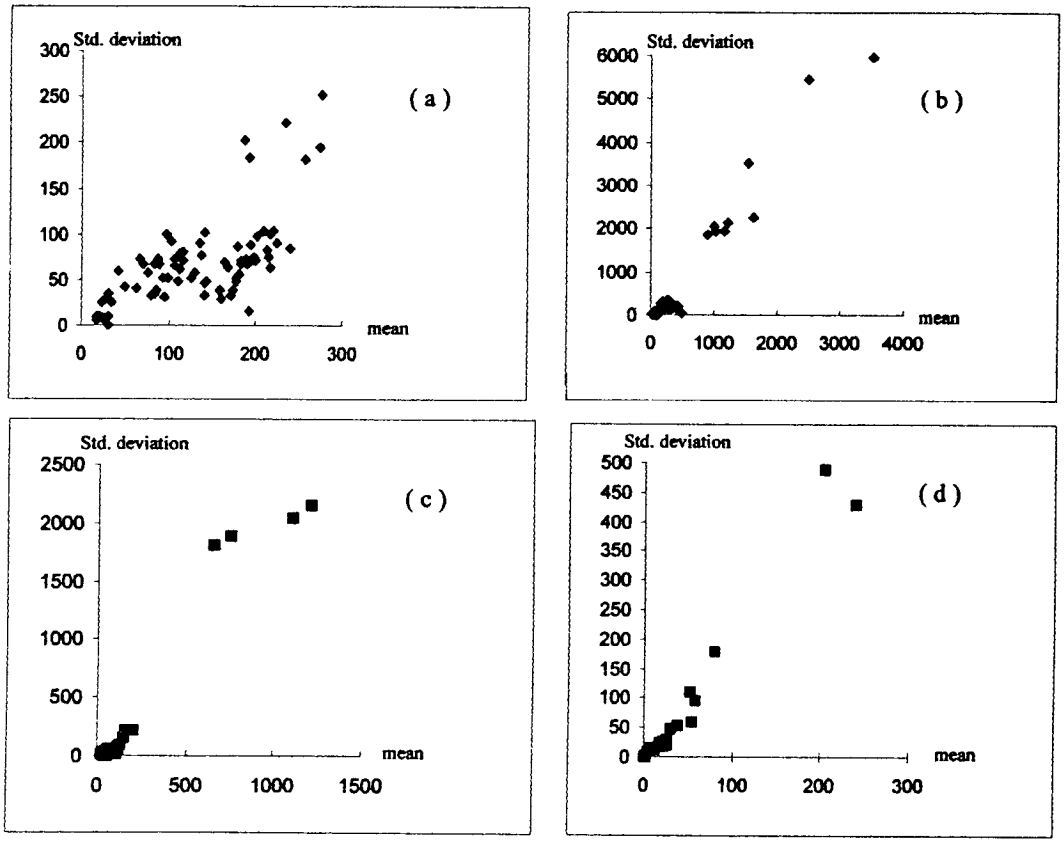

Fig. 2. Scatterplot of local window mean values vs. local window standard deviation values:(a) Cu, (b) Pb, (c) Zn, and (d) Ag.

nuity. While, variogram surfaces provide an effective way to visualize the anisotropy of phenomenon under consideration to identify preferential direction, along which five kinds of directional cross variograms mentioned above should be calculated. Among practical parameters required to calculate the directional CV for evaluating similarity or variability, two parameters should be carefully chosen: direction for increment of lag spacing and angular tolerance concerned with direction angle. For real applications of directional CV, angular tolerance of 90 degree is perferred because it is equivalent to the averaging variograms calculated in all directions and provide less subjective or biased results. In this terminology, variogram means any one selected among five functions of equations (1)~(5). This type of CV is referred to as an omni-directional variogram.

CASE STUDY

As a case study to test the validity of the these proposed scheme in real geological problem, four geochemical elements data sets such as copper, lead, zinc, and argentum concentrations, which were obtained from widely distributed 175 stream rock samples of the same latitudinal and longitudinal ranges with the Homyeong geologic base map (1:50,000 scale) in the Taebaek area (Fig. 1) (Jin *et al.*, 1983), were analyzed and the results were evaluated.

Fig. 2 shows the scatterplots between local window standard deviation vs. local window mean. From this process, 6, 9, 4, and 2 outliers were detected at original Cu, Pb, Zn, and Ag data sets, respectively. Related to these outliers estimated, spatial features of concentration for each element after removed outliers should be reasonably matched with polymetallic ore deposits within the study area, compared to those of previous one to which these outliers severely controlled. While, each data set was interpolated from random distribution to a common size of 25×25 grid to directly compare spatial characteristics of data before and after EDA. After interpolation, same condition for comparative study was prepared for direct analysis between original and revised data sets at common locations. The descriptive statistics and correlation matrix of four variables before and after EDA were presented in Tables 1 and 2.

The Pearson's product-moment correlation coefficient showing linear relationship between two variables was computed for this matrix. Conventionally, the correlation matrix has been considered as an useful one for multivariate geochemical data analysis to check statistical relationship with respect to each data set (Singh *et al.*, 1989). The correlation coefficient varies between −1.0 and 1.0; a positive value of this parameter indicates the variables are directly related, whereas a negative value implies an inverse relationship and values close to zero indicate a lack of linear relationship between two data sets composed of variables. As seen in Table 2, there appear no negative relationships among these variables before and after EDA, but main contents are remarkably changed. As a direct comparison between original data and revised data, the shift of computed correlation coefficient is quite different. In original data sets, Zn-Ag show the strong positive relationship showing the coefficient value of 0.6861 and Cu-Pb show relatively poor correlated (0.1549). However, with the revised data, the maximun and minimum values of the correlation coefficients appeared as Pb-Zn (0.6392) and Cu-Ag(0.0845),

Table 1. Descriptive statistics of Cu, Pb, Zn and Ag: (a) before moving-window statistics and (b) after moving-window statistics.

(a)

	Cu	Pb	Zn	Ag
Mean	120.43	368.28	126.29	44.64
Min	0.27	14.81	10.02	0.10
Max	521.55	8164.34	2328.87	3173.19
Skewness	0.90	7.12	6.47	12.08
Kurtosis	1.55	74.83	52.14	168.32
Std. Error	3.28	22.20	7.78	8.14
Std. deviation	81.95	554.94	194.46	203.52

(b)

	Cu	Pb	Zn	Ag
Mean	110.19	133.52	69.30	11.37
Min	0.54	7.89	10.02	0.10
Max	280.88	632.04	306.14	490.22
Skewness	1.17	1.50	2.54	12.40
Kurtosis	−0.97	2.14	10.47	211.89
Std. Error	2.63	3.97	1.31	1.01
Std. deviation	65.83	99.23	32.83	25.32

Table 2. Cross correlation matrix of Cu-Pb-Zn-Ag: (a) before moving-window statistics and (b) after moving-window statistics.

(a)

	Cu	Pb	Zn	Ag
Cu	1.0000			
Pb	0.1546	1.0000		
Zn	0.3297	0.3986	1.0000	
Ag	0.3049	0.3455	0.6861	1.0000

(b)

	Cu	Pb	Zn	Ag
Cu	1.0000			
Pb	0.4146	1.0000		
Zn	0.2624	0.6392	1.0000	
Ag	0.0845	0.3884	0.3404	1.0000

respectively. An experiment-based analysis with real samples is necessary to investigate these differences, but this is beyond the main scopes of this study.

The correlation coefficient discussed here is somewhat different from cross correlation function, one of the CV functions, but cross correlation coefficient function by complete overlap with no shift is equivalent to the component of this matrix. Therefore, cross correlogram can be interpreted as the form of spatially extended correlation.

As for spatial continuity measures, five variogram functions such as γ_{uv}, $\gamma_{s_{uv}}$, C_{uv}, ρ_{uv}, and M_{uv} were calculated for the original data and the revised data, and the results were directly compared. In this analysis, unit lag to compute the directional CV is chosen as approximately 660 meter. With regards as sampling space, this size is thought to be reasonable, and the lag increment direction and angular tolerance are 0° and 90°, respectively. The angular tolerance of 90°, so-called omni-directional CV, means the averaging variograms are calculated in all direction. Therefore, the omni-directional CV calculation was performed to analyze the unbiased results along specific trending in this study. For the purpose of comparison, Cu was taken as a control variable, and remaining elements such as Pb, Zn, and Ag were taken as target variables. Fig. 3(a)~(e) show the results of computed γ_{uv}, $\gamma_{s_{uv}}$, C_{uv}, ρ_{uv}, and M_{uv} using these geochemical elements. As can be seen in Fig. 3(a), variability of Cu-Pb appeared relatively high; however, this result was not coincident with those of cross standardized semivariogram function and cross semimadogram. There are several reasons to explain these phenomena, and nugget effect is regarded as one among them. Myers *et al.* (1982) noted that the discontinuity at the origin may be caused by irregular mineralization due to small high-grade accumulation; consequently, nugget effect of spatial distribution of geochemical element always happens for both major and trace elements in mining application. In such a case, some difficulties are arisen in integrated interpretations of γ_{uv}, $\gamma_{s_{uv}}$, C_{uv}, ρ_{uv}, and M_{uv}. However, as similarity measures using these data sets, Cu-Zn was shown to be correlated within the regime covered 10 lags, i.e. intermediate scale of the overall study area.

The results of omni-directional CV studies using revised geochemical elements obtained by removing some estimated outliers were presented in Fig. 4(a)~(e). Compared to previous ones, the results of cross semivariogram, cross standardized semivariogram, and cross semimadogram appear to be in accord with each other very well. From cross semimadograms of Fig. 3(e) and Fig. 4(e), it is noted that relationships of Cu-Pb, Cu-Zn, and Cu-Ag show similar trends. While, spatially high similarity between Cu and Pb is more apparent than those of Cu-Zn and Cu-Ag, especially within intermediate area within 10 lags. The gradual decrease trend of the similarity of Cu-Pb with increment of lag distance

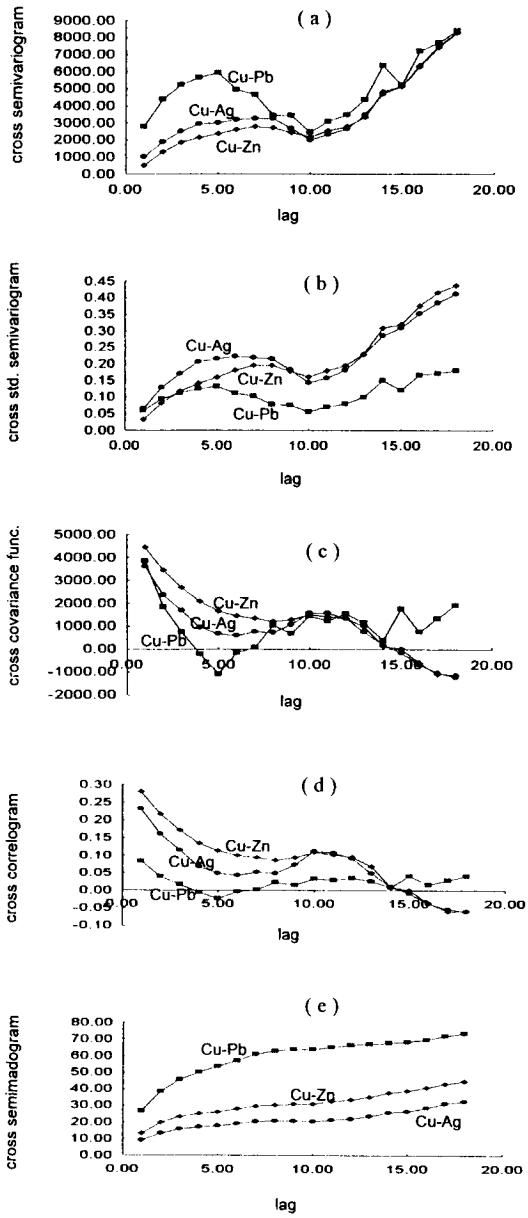

Fig. 3. An example of omni-directional cross variography along with azimuth of 0° from easting direction before moving-window statistics: (a) cross semivariogram, (b) cross standardized semivariogram, (c) cross covariance function, (d) cross correlogram, and (e) cross semimadogram.

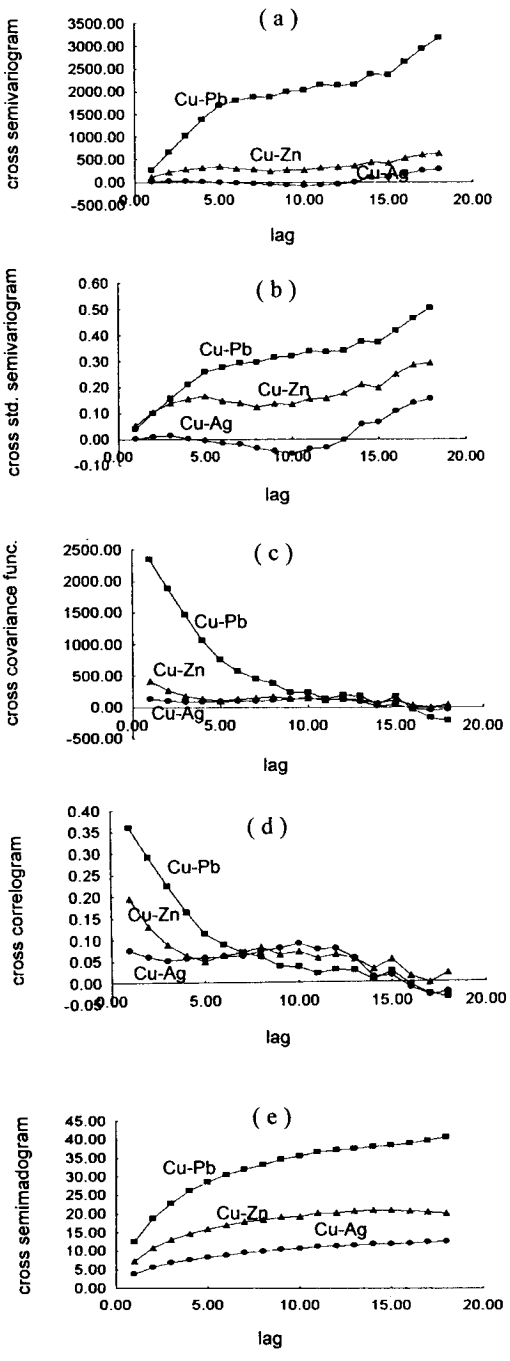

Fig. 4. An example of omni-directional cross variography along with azimuth of 0° from easting direction after moving-window statistics: (a) cross semivariogram, (b) cross standardized semivariogram, (c) cross covariance function, (d) cross correlogram, and (e) cross semimadogram.

can be interpreted due to the gradual increase of the scatterness of Cu-Pb as shown in Fig. 4(a), (b), and (e).

Conclusively, as long as mineral application task is performed for the target of these four elements, it is likely that the relation of Cu-Pb with ore deposits is more prominent than those of Cu-Zn and Cu-Ag.

CONCLUDING REMARKS

1. In this study, the well known moving-window statistics method was used to detect outliers, which is especially regarded as a kind of EDA to handle geochemical data. Currently, it has been known that one of the important problems in relation to applications of this method is concerned with determination of local window size; however, there are not established criteria, except rather a subjective one based on the coefficient of variation. Tentatively, the proper window size can be decided with consideration of distribution of sampling sites and descriptive statistics because of compensation effect of overlapping window.

2. Compared to a conventional correlation matrix method within data sets, spatial correlation functions computed by the directional CV should be interpreted on the basis of concepts of similarity and variability with various scales of the whole studied area.

3. The appearence of spatial features before and after removing outliers at each data set are remarkably changed; therefore, different results are given with the original data set and the revised data set. The results from omni-directional CV ignored outliers by the moving-window technique show reasonable consistency of γ_{uv}, $\gamma_{s\,uv}$, C_{uv}, ρ_{uv}, and M_{uv}, and are thought to be preferred at real geological problems. The scheme proposed in this article may provide an useful decision-supporting information, with help of statistical clustering, for regional mineral exploration for undiscovered base metal ore deposits as well as discovered ones. Therefore, these schemes are regarded as one of new methodologies for the geological applications of GISs handling spatial data sets.

REFERENCES

Clark, I. (1982) Practical geostatistics. Applied Science Publishers Ltd., London, 129p.

Haining, R. (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, 409p.

Haslett, J., Wills, G. and Unwin, A. (1990) SPIDER-an interactive statistical tool for the analysis of spatially distributed data. Int. Jour. of Geographical Information Systems, v. 4, p. 285-296.

Krige, D. G. and Magri, E. J. (1982) Studies of the effects of outliers and data transformation on variogram estimate for a base metal and a gold ore body. Math. Geology, v.

14, p. 557-564.
McArthur, G. J. (1988) Using geology to control geostatistics in the Hellyer Deposit. Math. Geology, v.20, p. 343-366.
Murray, M. R. and Baker, D. E. (1991) MWINDOW: An interactive Fortran-77 program for calculating moving-window statistics. Computer and Geosciences, v. 17, p. 423-430.
Myers, D. E., Begovich, C. L., Butz, T. R., and Kane, V. E. (1982) Variogram models for regional groundwater geochemical data. Math. Geology, v. 14, p. 629-644.
Issaks, E. H. and Srivastava, M. R. (1989) An introduction to applied geostatistics. Oxford University Press, 561p.
Jin, M. S., Lee, J.S., Kim, S. J., and Lee, C. Y. (1983) Geochemical maps for Homyeong Sheet in the Taebaegsan mineralized belt, Korea Institute of Energy and Resources, 21p.
Pan, G. (1993) Regional favorability theory for information synthesis in mineral exploration. Math. Geology, v.25, p. 603-631.
Singh, V., Moon, W. M., and Fedikov, M. (1989) Investigation of airborne MEIS-II and MSS data for biochemical exploration of mineralized zones, Farley Lake, Manitoba. Canadian Jour. of Remote Sensing, v. 15, p. 122-133.
Tukey, J. W. (1977) Explorative data analysis. Reading, MA, Addison-Wesley.

Manuscript received 27, March 1995

공간적 상관도 측정을 위한 변이도 분석 : 태백지역의 지화학자료를 이용한 사례 연구

이기원 · 권병두

요 약 : 공간통계를 바탕으로 한 지질정보의 정량적 처리및 분석을 위한 여러 응용방법들이 최근에 광물탐사문제와 관련되어 연구되고 있다. 본 연구에서는 지화학자료와 관련된 이상점판단(outlier detection)과 방향성 상호 변이도 측정의 적용성을 검토하였고, 아울러 태백지역내의 광물탐사를 위한 지화학자료에 에 대한 사례연구를 수행하였다. 이상점판단방법으로는 이동창(moving window)통계법이 이용되었다. 한편 상호 변이도는 공간적 연속성 측정을 위한 통계적 방법으로 알려져 있으나, 본 연구에서는 이 개념을 자료의 공간적 상관도 문제로 확장하였다. 한편 다섯가지 형태의 변이도 표현식을 이상점처리전후의 결과와 연관하여 비교하였다. 이러한 비교연구의 결과로 이 두가지 공간 통계법에 의한 자료처리과정및 분석방법은 실제의 결정판단단계에서 결정적인 영향을 미치는 것으로 나타났으며, 광역적 광물탐사에서 유용한 해석보조자료로 제공될 수 있을 것으로 생각된다.