Journal of Environmental Sciences Vol.13, No.4, pp.453—458, 2001

ISSN 1001—0742
CN11—2629/X

Article ID: 1001-0742(2001)04-0453-06 **CLC number**: X11, X14 **Document code**: A

Utility of semivariogram for spatial variation of soil nutrients and the robust analysis of semivariogram

GUO Xu-dong[1,2], FU Bo-jie[1*], MA Ke-ming[1], CHEN Li-ding[1]

(1. Department of Systems Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China. E-mail: land@mail.rcees.ac.cn; 2. Open Laboratory of Land Use, China Institute of Land Surveying and Planning, Ministry of Land and Resources, Beijing 100029, China)

Abstract: The spatial variation of soil nutrients in topsoil (0—20cm) was analyzed using semivariogram in the Zunhua County of Hebei Province, China. The effect on semivariogram with randomly deleted data and kriged estimates using various reduced sample sizes was also analyzed. The semivariograms of available N, total N, available P, organic matter were best described by a spherical model, except for available K, which best fitted a complex structure of exponential model and linear with sill model. The ratio of nugget to total sample variance ranged from 34.4% to 68.4%, indicating the spatial correlation of tested soil nutrients on a large scale was moderately dependent. Among five soil nutrients, available nitrogen and available phosphorus had the shortest spatial correlation range (5 km and 5.5 km), available K had the longest range (25.5 km), whereas total nitrogen and organic matter had intermediate spatial correlation range (14.5 km and 8.5 km). The semivariograms of available N, total N, available P, and organic matter were insensitive to a 50%—60% reduction in original sampling density, while for available K, it is up to 70%. The estimated spatial distributions of total N by kriging, under various reduced sample sizes, all correlated significantly ($P = 0.001$) with those obtained from original data. The results showed that the semivariogram was a relatively robust tool when used in a large region and sufficient spatial variation information could be retained regardless of a higher deletion proportion of the original data. The original sample data could be reduced by kriging and the estimates showed no loss of spatial information, however, the results may be unreliable unless a clearly identified semivariogram model could be obtained. The results may provide useful information for determining the appropriate sampling densities for these scales of soil survey.

Keywords: semivariogram; robust; soil nutrients; spatial variation

Introduction

Geostatistics has proven useful for characterizing and mapping the spatial variation of soil properties (Webster, 1985; Trangmar, 1985; Goovaerts, 1999). Most previous geostatistical studies focused on the data of a relatively small spatial scale (Chien, 1997). The application of geostatistics techniques to large areas can produce a large sample size. Too much data will bring an extra burden for soil sampling and analysis, however; too little a sample size will not fully express the spatial character of the soil properties. So, it is necessary to determine an appropriate soil sampling density for soil spatial analysis of large regions.

There is some detailed information about spatial variation of soil nutrients (Cahn, 1994; Meirvenne, 1989). Some researchers have also reported that the semivariogram is a relatively robust tool for spatial analysis when used for large regions (Meisel, 1998; Chien, 1997; Chang, 1998). In this paper, we evaluate the spatial variation character of soil nutrients in the Zunhua County of North China, discuss the influence on semivariograms with a random data deletion, and compare the resulting estimates obtained by kriging under various reduced sample sizes. The results can contribute to local agricultural decisions and also may provide useful information for determining appropriate sampling densities at these scales of soil survey.

1 Materials and methods

1.1 Study area

The Zunhua County (39°55′—40°22′N, 117°34′—118°14′E), located at Hebei Province of China, was chosen as the study area. The county covers an area of 1520 km^2 consisting of alluvial and diluvial plains (36%) in the middle with hilly land (64%) surrounded. The elevation of the plain area ranges from 20—80m and 90% of hilly land is less than 300m. In addition, a narrow low mountain about 2—3 km wide and 200m high runs from west to east across the central part of the county. The area has a continental climate. Many seasonal rivers such as River Lihe, River Linhe etc. flow from NE to SW and converge to form the Yuqiao Reservoir. Land use types include crop, forest and grassland. The soil types are made up of brown soil, cinnamon soil and meadow soil, primarily of sandy loam, light loam, medium loam and heavy loam texture.

1.2 Laboratory analysis

The total nitrogen of soil was determined by the semi-micro Kjeldahl method after wet digestion with $H_2SO_4 + HClO_4$. Available nitrogen was determined by a micro-diffusion technique after alkaline hydrolysis. Available phosphorus was extracted with 0.5 mol/L $NaHCO_3$ solution (pH 8.5). Phosphate-P in solution was determined colorimetrically by the formation of the blue phosphomolybdate complex following reduction with ascorbic acid. Organic matter was determined by the oil bath-$K_2Cr_2O_7$ titration method. Available K is determined by the 1 mol/L NH_4OAc-flame photometry method.

1.3 Geostatistical approach

The basic theory of geostatistics have been well documented (Matheron, 1963; Clark, 1979; Journal, 1978; Rossi, 1992; Burgess, 1980). Assuming the intrinsic stationarity of the data, the degree of spatial dependency can be measured by calculating the semivariance (Journal, 1978).

Foundation item: The National Natural Science Foundation of China (No.49831020); * Corresponding author

$$r(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2, \quad (1)$$

where $r(h)$ is the semivariance; h is the lag distance separating pairs of points; $N(h)$ is the number of pairs separated by the lag distance h; $Z(x)$ is the value of the regionalized variable at location x; and $Z(x + h)$ is the value at the location $x + h$. It should be noted that semivariograms are typically calculated only to one-half of the maximum distance between the points (Rossi, 1992; Webster, 1985).

The map of sampling sites was digitized and used for geostatistical analysis according to the data of soil nutrients in topsoil (0—20 cm) obtained from the Second National Soil Survey of China (Fig.1). The examined soil nutrients included available N, total N, available P, available K, and organic matter. The sample size was 1059, the minimum distance between samples was 0.21 kilometers and the maximum distance was 56 kilometers. The sample density was about one sample every 1.4 km^2. The valid sample size of available N, total N, available P, available K and organic matter was 1025, 1044, 1034, 1035, 1030, respectively, subtracting the missing values and some outliers. The Kolmogorov-Smirnov test found that both original and log-transformed data were not normally distributed, but the log-transformed data showed a good symmetry and were used in the following analysis.

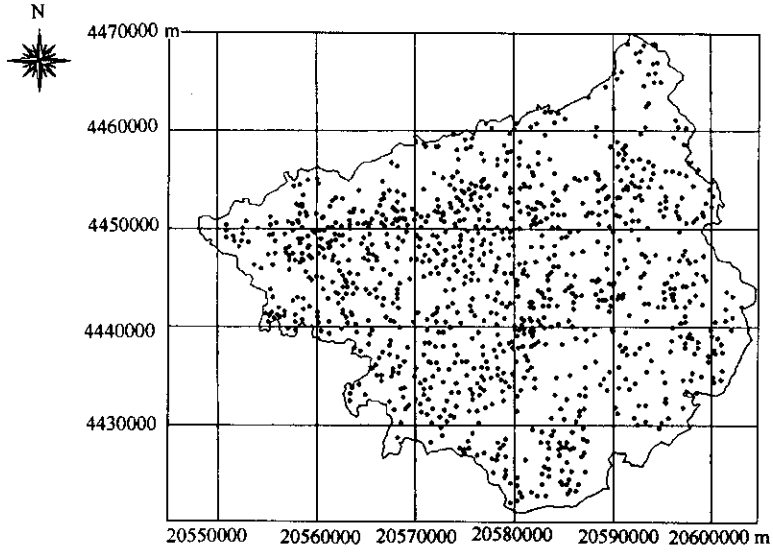

Fig.1 The sample distribution pattern of soil nutrients

Because the sample was distributed irregularly, the distance class was used to determine the lag. The lag used to calculate semivariance represented the middle value of a distance class. The lag space was 0.5 km, thus for distance 0, 0.5, 1, 1.5, ···, 28 km represented the distance interval of 0—0.25 km, 0.25—0.75 km, 0.75—1.25 km, ···, 27.75—28.25 km, respectively (ITC, 1998). Because the number of the pairs of points (< 8) were not sufficient between the distances of 0—0.25 km, the first semivariance was determined by distance 0.5 km.

Appropriate models to describe the semivariogram were selected through a least-square best-fit procedure, selecting the model with the highest coefficient of determination (R^2) and lowest nugget variance.

The semivariance obtained from the original data (denoted by X) and those obtained with reduced sample data (denoted by Y) were correlated by linear regression ($Y = X$). The Pearson correlation coefficients were used to indicate the degree of spatial variation information maintained.

Total N, for example, was used to detect the results estimated by ordinary kriging (O.K.) for various sample sizes. Similarly, the correlation coefficients of total N between the kriged results estimated from the original data (denoted by X) and those obtained from various combinations of reduced sample data (denoted by Y) were used to indicate the degree of spatial distribution maintained with reduced sample size. The mean square error (MSE) was used to compare the accuracy of estimates.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} | z(x_i) - z^*(x_i) |^2, \quad (2)$$

where $z(x_i)$ and $z^*(x_i)$ represented true values (original data) and estimates (reduced data), respectively. 6007 estimates were generated by interpolation the 500 × 500m node. When sample size was reduced by 30%—60%, the nearest 16 to 24 neighbor values were used, while the nearest 8 to 16 neighbor values were used when the sample size was reduced by 70%—80%. The maximum radius was 14 km. The semivariance and O.K. were performed using ILWIS (ITC, 1998). The Kolmogorov-Smirnov test and correlation coefficient were performed using SPSS7.0.

2 Results and discussion

2.1 Descriptive statistics

Among these five soil nutrients, available P had the highest $C.V.$ of 66.1%. Total N, available N, organic matter had $C.V.$ values that ranged from 23.7% to 27.3%, while available K had intermediate $C.V.$ values of 46% (Table 1). The highest variation of available P may be attributed to uneven application of fertilizer P. Very little fertilizer K was used in

Zunhua County. The variation of available K could be primarily caused by the topography, soil types, and land use types (Fu, 2000).

Table 1 Descriptive statistics of soil nutrients

	Sample size	Mean	*S.D.*	*C.V.*, %	Median	Minimum	Maximum
Available nitrogen, mg/kg	1025	66.48	16.47	24.8	65.00	28.00	149.62
Total nitrogen, %	1044	0.072	0.017	23.7	0.070	0.034	0.191
Available potassium, mg/kg	1035	87.95	40.43	46.0	77.70	11.40	266.69
Available phosphorus, mg/kg	1034	20.82	13.78	66.1	17.60	1.15	91.60
Organic matter, %	1030	1.16	0.32	27.3	1.13	0.41	3.29

2.2 The spatial variation of soil nutrients

Fig.2 is the model-fitted semivariograms of the soil nutrients. Table 2 summarized the corresponding parameters of these semivariograms based on the best-fitted model. The semivariograms of total nitrogen, available nitrogen and available phosphorus were best described by a spherical model, except for available K, which best fitted the nested structure of exponential model and linear with sill model.

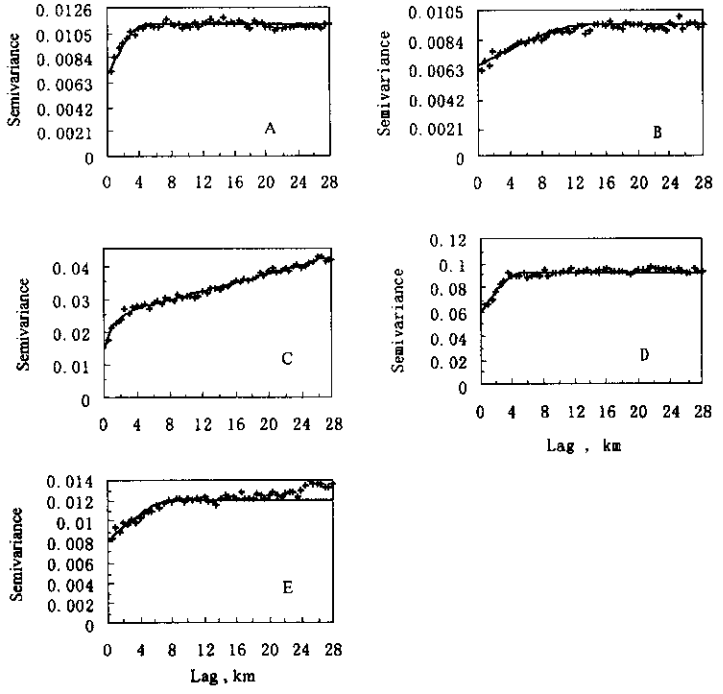

Fig.2 Bested-fitted and experimental semivariograms for (A) available N, (B) total N, (C) available K, (D) available P, (E) organic matter (+ experimental value)

Table 2 The best-fitted semivariogram models of soil nutrients and corresponding parameters

	Model	Nugget	Sill	Nugget/sill	Rang, km	R^2
Available nitrogen	Spherical	0.0069	0.0112	61.6%	5	0.90
Total nitrogen	Spherical	0.0065	0.0095	68.4%	14.5	0.93
Available potassium	Exponential	0.013	0.023	34.4%	25.5	0.99
	Linear with sill	0.0015	0.019			
Available phosphorus	Spherical	0.059	0.092	64.1%	5.5	0.88
Organic matter	Spherical	0.008	0.012	66.7%	8.5	0.69

The nugget effect, representing the undetectable experimental error and field variation within the minimum sampling spacing, ranged from 34.4% to 68.4%, indicating the soil nutrients exhibited moderately dependency (Cambardella, 1994).

The range of spatial correlation of these soil nutrients varied from 5 km to 25.5 km. The shortest range is 5 km for available N while the longest range was 25.5 km for available K. The range of available P was 5.5 km, quite similar with available N. Total N and organic matter had the intermediate ranges of 14 km and 8.5 km, respectively. The lowest range of available N and P could be attributed to the fact that the two soil nutrients were influenced more by soil management practices such as fertilizer application than were the rest of the soil nutrients.

2.3 The effect on semivariograms under various sample sizes

The original data were reduced by randomly deleting sample points. The proportion of deleted sample sizes to original data ranged from 30% to 80%. Table 3 is the sample sizes of soil nutrients under various deletion proportions. Fig.3 is the semivariograms of soil nutrients under various deleted sample sizes. Table 4 is the correlation coefficient of semivariance between original data and reduced sample sets. When sample sizes of available N and total N were reduced by 30%, the shapes of semivariograms (Figs. 3A and 3E) had little change, the range, nugget and sill were almost the same as those of the original data. The same consequence occurred to available P (Fig. 3I) and organic matter (Fig. 3P) with a reduction of 30%—40% sample data and available K (Fig.3M) with a 30%—50% reduction. When original data were reduced by 40%—50%, the semivariogram of available N began to show a little fluctuation (Fig. 3B), total N appeared a small valley (Fig. 3F), however, the range, nugget and sill can be identified clearly in the maps. The semivariograms of available P and organic matter (Figs. 3J and 3Q) also showed a slight fluctuation when the sample sizes were reduced by 50% and 50%—60%, respectively. We also can easily obtain the range, nugget and sill from their semivariograms. When the sample size were reduced by 60%—70%, the semivariograms of available N, total N, available P, organic matter, (Figs. 3C, 3G, 3K and 3R) began to show a large fluctuation, there were many abrupt local peaks and valleys in the map, that obscured the true range and sill. So under this deletion proportion the results of semivariogram analysis of these four nutrients can be seriously affected. But the semivariogram of available K (Fig. 3N) still could be best described by lineal with sill model even if the sample size was reduced by 70%. When the reduced sample size is over 80%, semivariograms of soil nutrients became effectively unreadable (Figs. 3D, 3H, 3L, 3O and 3S), it was hardly possible to draw the sill, nugget and range from such semivariograms, and the results obtained were likely to be meaningless.

Table 3 Sample sizes of soil nutrients under various deletion proportions

Deletion proportion, %	Sample size						Sample densities*, km²/one point
	Total	Available N	Total N	Available K	Available P	Organic matter	
30	730	706	719	710	708	712	2
40	628	608	618	609	610	613	2
50	529	508	518	512	516	517	3
60	429	404	413	407	412	412	4
70	316	290	296	291	294	296	5
80	218	190	195	190	193	196	7

including missing values

Table 4 The correlation coefficient of semivariance of soil nutrients between original and various reduced sample sets

	30%	40%	50%	60%	70%	80%
Available N	0.97	0.91	0.86	0.81	0.66	0.49
Total N	0.97	0.94	0.93	0.88	0.79	0.33#
Available K	0.99	0.97	0.97	0.93	0.87	0.86
Available P	0.95	0.93	0.89	0.78	0.68	0.77
Organic matter	0.97	0.95	0.90	0.86	0.78	0.64

Note: All correlation is significant at 0.001 (2-tailed) except #, significant at the 0.05 (2-tailed)

An obvious alteration to the shape of semivariogram of available K, as we expected, was that the exponential model that primarily occurred in relatively short ranges (< 3.5 km) did not show in the map when sample size was reduced by 60%—70%. Under this deletion proportion, the semivariogram only displayed the linear with sill model, which reflected the spatial variation of available K in large ranges (3.5—25.5 km). The reason was that decreasing sampling density practically increased the sampling intervals, the variation over short intervals was diminished. It also indicated that several processes had different contributions to the variation of available K at different scales. Some management practices may be the primary factors influencing the spatial variation of available K in relatively small ranges, while the variation in large ranges may be attributed to topography and land use types. The semivariograms were insensitive to randomly reduced sample data, which may be attributed to the fractal behaviours of soil properties. The fractals embody the idea of "self-similarity", that is, the manner in which variation at one scale is repeated another (Burrough, 1983). Reducing sample size was the process that spatial variation of soil nutrients changed from small scales to large scales, since the variation could be repeated, the semivariograms would not be changed much. However, the fractal behaviors of soil nutrients was limited in a range of scales, the shapes of semivariograms were seriously affected at the distances beyond "self-similarity" scale.

The degree of influence on semivariograms under various reduced sample size was different among tested soil nutrients. Available K was the most insensitive to the reduction of sample size. When the original data were reduced by 70%, the semivariogram of available K still could be identified clearly though the variation in relatively small ranges was ignored (Fig. 3N). Even under the deletion proportion of 80%, the correlation coefficient was up to 0.86 (Table 4). The value was higher than that of other soil nutrients and indicated that more spatial variation information could be maintained. The slightest effect degree on semivariogram of available K could be attributed to its longest spatial correlation range, which usually was used to represent the "self-similarity" scale. For the other four soil nutrients, the effect degree on semivariograms was little different despite their different ranges. However, the trend was found that the more spatial variation information of soil nutrients with longer distances could be maintained based on the correlation coefficients. For example, total N and organic matter had a longer correlation distance, the correlation coefficient, indicating that the retaining degree of spatial variation information, was approximately 0.8 with a reduction of 70%, while that for available N and P with shorter ranges was less than 0.7 (Table 4).

The results represented here were consistent with those reported by Meisel and Turner (Meisel, 1998). They evaluated the influence of the amount and spatial distribution of absent data on semivariogram results and interpretation using computer simulation. The semivariograms were found to be insensitive to missing or deleted data, when the size of the deleted blocks of data was small (i.e., less than 10% of the maximum lag distance, or less than 0.25% of the map). The original sample density in this study is about one point every 1.4 km², which is less than 0.25% of the area (4 km²). This result is encouraging, since it identifies semivariograms as a relatively robust tool used in large regions where soil sampling and analysis

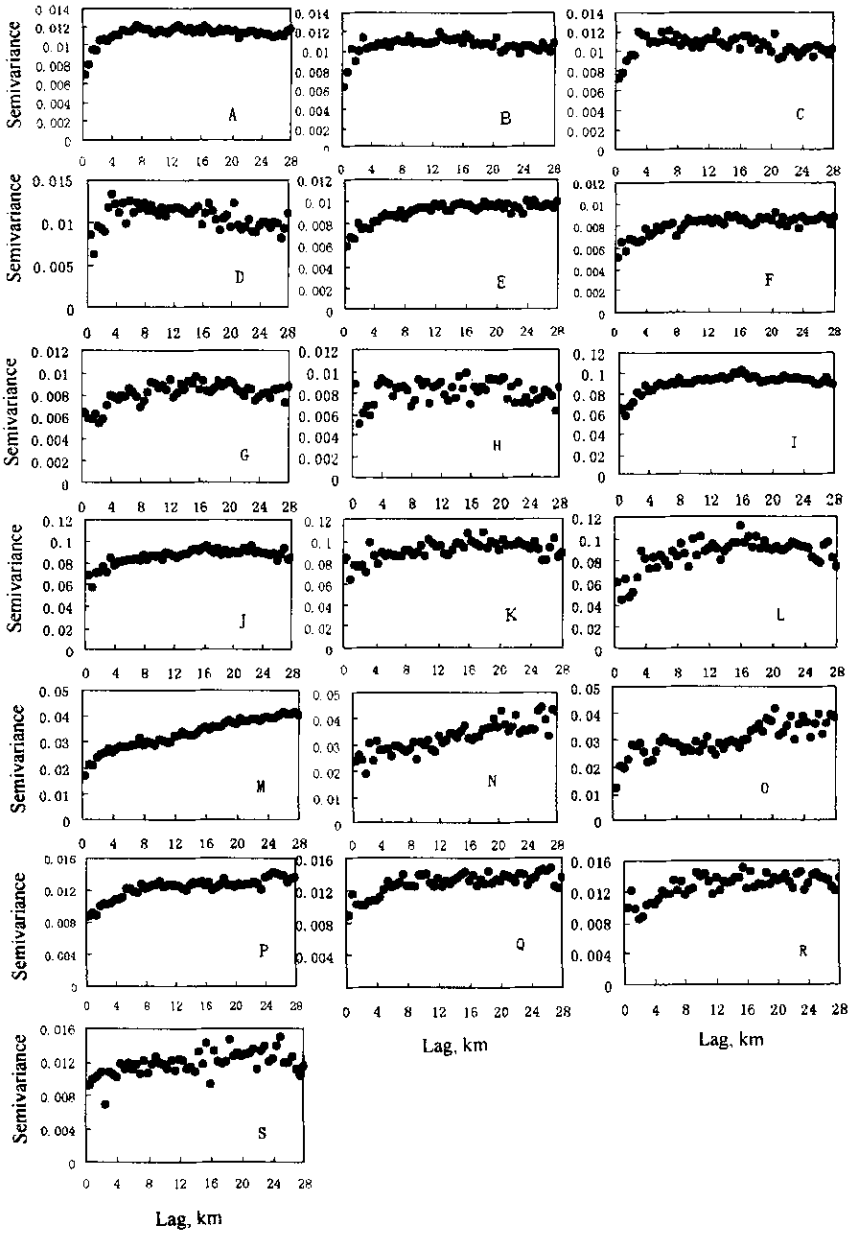

Fig. 3 The semivarograms of soil nutrients under various reduced sample sizes
(A)AN*, 30%; (B)AN, 40%—50%; (C)AN, 60%—70%; (D)AN, 80%; (E)TN*, 30%; (F)TN, 40%—50%, (G)TN, 60%—70%; (H)TN, 80%; (I) AP*, 30%—40%; (J)AP, 50%; (K)AP, 60%—70%; (L)AP, 80%; (M)AK*, 30%—50%; (N)AK, 60%—70%; (O)AK, 80%; (P)OM*, 30%—40%; (Q)OM, 50%—60%; (R)OM, 70%; (S)OM, 80%; * AN: avail able N; TN: total N; AP, available P; SK, available K; OM, organic matter

are time-and labor consuming. Thus, relatively fewer samples are enough to fully express the spatial variation of soil properties on the large scale.

2.4 Effect on the estimates by kriging under various sampling densities

The *MSE*, correlation coefficient of total N between the original and reduced sample sets are listed in Table 5. The correlation coefficient between estimates derived from a reduction of 80% sample data and estimates derived from the original data is still up to 0.65. It indicated that the estimates by kriging would retain much of spatial distribution information regardless of a higher proportion of deleted data. The results were in agreement with those reported by Chien *et al*. (Chien, 1997) and Chang *et al*. (Chang, 1998). Chien *et al*. (1997) observed when the original sampling density was reduced by nearly half, the overall spatial information of the sum of exchangeable bases (SEB) estimated by kriging and cokriging could still be maintained in comparison with the results obtained from the original data. Chang *et al*. (1998) also drew an analogous conclusion for an area of low agricultural land, which has been allowed to flood through tidal inundation. They also found the

estimates obtained by cokringing were more accurate than estimates by kriging. But the detailed information about semivariograms with reduced sample density in their studies was not described.

Table 5 The *MSE*, correlation coefficients of estimates by kriging of total N between the original and various reduced sample sets

Total N	30%	40%	50%	60%	70%	80%
Person correlation	0.90	0.86	0.88	0.87	0.77	0.65
MSE	4.55E-04	6.02E-04	5.26E-04	5.55E-04	1.14E-03	1.46E-03

Note: all significantly correlated at 0.001 level(2-tailed)

In this study, the results obtained from a reduction of 50% sample data were quite similar with those from a reduction of 60% if only from the correlation coefficient and *MSE* (Table 5). But the semivariograms in this study displayed many local valleys and peaks when the sample densities were reduced by 60% or higher. It was difficult to identify the range and sill from such semivariograms. Kriging relies directly on a semivariogram of the sampled data to derive the model for interpolation, a model that may give spurious results if the semivariogram itself is problematic. So the results of total N by kriging in the study with a reduction of 60% sample data or higher may be unreliable regardless of its higher correlation coefficient.

3 Conclusions

The best-fitted semivariogram model of available N, total N, available P, organic matter was a spherical model, except for available K, which was best described by the nested structure of exponential model and linear with sill model. Among five soil nutrients, available nitrogen and available phosphorus had the shortest spatial correlation range (5 km and 5.5 km), available K had the longest range (25.5 km), whereas total nitrogen and organic matter had an intermediate spatial correlation range (14.5 km and 8.5 km). The ratio of nugget to total sample variance ranged from 34.4% to 68.4%, indicating that the spatial correlation of tested soil nutrients on the large scale was moderately dependent. The semivariogram analysis for available N, total N, available P and organic matter in the study can be practicable and reliable when the original sample size is reduced by 50%—60%, while for available K, it is up to 70%. Under this deletion proportion, the appropriate sampling density for available N, total N, available P and organic matter is about one point every 4 km^2, and that for available K is about one point every 5 km^2. The estimated spatial distributions of total N by kriging, under various reduced sample sizes, all correlated significantly ($p = 0.001$) with those obtained from the original data. Our results suggested that the semivariogram was a relatively robust tool when used to large areas, the sufficient spatial variation information could be retained regardless of a randomly higher deletion proportion of original data. While maintaining a clearly identified semivariogram, the estimates by kriging showed almost no loss of spatial distribution information and were reliable even when the original sample density was reduced by 50%.

References:

Burgess, Webster R, 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging[J]. J of Soil Science, 31: 315—331.

Burrough P A, 1983. Multiscale sources of spatial variability in soil variation[J]. Journal of Soil Science, 34:577—579.

Cahn M D, Hummel J W, Brouer B H, 1994. Spatial analysis of soil fertility for site-specific crop management[J]. Soil Sci Soc Am J, 58: 1240—1248.

Cambardella C A, Moorman T B, Novak J M *et al*., 1994. Field-scale variability of soil properties in central low soils[J]. Soil Sci Soc Am J, 58:1501—1511.

Chang Y H, Scrimshaw M D, Emmerson R H C *et al*., 1998. Geostatistaical analysis of sampling uncertainty at the tollesbury managed retreat site in blackwater estuary, Essex, UK: kringing and cokriging approach to minimize sampling density[J]. The Science of the Total Environment, 221: 43—57.

Chien Y J, Dar-Yuan Lee, Horng-Yuh Guo *et al*., 1997. Geostatistical analysis of soil properties of mid-west Taiwan soils[J]. Soil Science, 162(4): 291—298.

Clark Isobel, 1979. Practical geostatistics[M]. London: Applied Science Publishers LTD.

Fu B J, Chen L D, Ma K M *et al*., 2000. The relationship between land use and soil conditions in the hilly area of loess plateau in northern Shaanxi, China[J]. Catena, 39: 69—78.

Goovaerts P, 1999. Geostatistics in soil science: state-of-the-art and perspectives[J]. Geoderm, 89: 1—45.

ITC, 1998. ILWIS 2.2 for Windows[M]. The Netherlands.

Journal A, Huijbregis Ch, 1978. Mining geostatistics[M]. London, UK: Academic Press.

Matheron G, 1963. Principles of geostatistics[J]. Economic Geology, 58:1246-1266.

Meirvenne M Van, Hofman G, 1989. Spatial variability of soil nitrate nitrogen after potatoes and its change during winter [J]. Plant and Soil, 120: 103—110.

Meisel J E, Turner A G, 1998. Scale detection in real and artificial landscape using semivariance analysis[J]. Landscape Ecology, 13: 347—362.

Rossi R E, Mulla D J, Journel A G *et al*., 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence[J]. Ecological Monographs, 62:277—314.

Trangmar B B, Yost R S, Uehara G, 1985. Application of geostatistics to spatial studies of soil properties[J]. Advanced Agronomy, 38: 44—94.

Webster R, 1985. Quantitative spatial analysis of soil in the field[J]. Advance in Soil Science, 3:1—70.

(Received for review August 21,2000. Accepted December 24,2000)