

Spatial Correlation Analysis of Nitrogen Dioxide Concentrations in the Area of Milan, Italy

G. Dubois^{a1}, M. Saisana^b, A. Chaloulakou^c and N. Spyrellis^c

^aDept. of Mineralogy and Petrography, University of Lausanne, Switzerland (gregoire.dubois@usa.net)

^bIPSC-TERM Unit, Joint Research Centre of the EC, Ispra, Italy (michaela.saisana@jrc.it)

^c Dept. of Chemical Engineering, National Technical University of Athens, Greece (dchal@central.ntua.gr, nspyr@cemeng.ntua.gr)

Abstract: Monitoring atmospheric pollution in urban areas frequently involves mapping techniques that assist the researcher and/or the decision-maker to describe and quantify the pollution at locations where no measurements are available. The preparation of these pollution maps is a complex task, which is only feasible if a spatial correlation of the variable of interest is identified. Furthermore, the spatial correlation may not only change in time and space, but also according to the pollution levels. To illustrate this point, this paper investigates the fractal dimension of the spatial correlation of different levels of annual nitrogen dioxide concentrations [NO₂] in the greater area of Milan in Italy for the years 1997-1999. It is shown that levels below 20 ppb and levels higher than 32 ppb of annual [NO₂] present less correlation in space than mid levels (20-32 ppb) do. As a result, maps defining areas with high and low probability of exceeding a specific concentration threshold will have an uncertainty that is related to the different NO₂ annual levels. In the light of these results, the development of environmental policies related to the EC Directive target is briefly discussed. Rather than trying to define an optimal NO₂ sampling network, suggestions are made on how the information provided by the fractal analysis of the spatial correlation could be used to streamline the existing network.

Keywords: Spatial correlation; Fractal dimension; Probability mapping; Indicator semivariograms; NO₂

1. INTRODUCTION

In assessing air quality, knowledge of the spatial distribution of pollutants in a study area is often needed for successful monitoring design and risk analysis. During the past few decades, a large volume of literature has been published dealing with various aspects of the mapping of pollutants [e.g. Goovaerts and van Meirvenne, 2001; Hopkins et al., 1999] and the spatial correlation structure of environmental variables [e.g. Elsom, 1978]. Less work has been done on the analysis of the spatial correlation of air pollutants in the face of different concentration thresholds. The presence of a spatial correlation is not only a condition for the interpolation of the data in space in order to establish a map of pollution, but it also provides useful insights on the structure of the air quality patterns. At the same time, when an assessment of epidemiological consequences of environmental factors is needed, decision-makers will be more

interested in obtaining an estimation of the probability of the pollutant's concentration to exceed a certain threshold at unsampled locations than a pollutant's concentration *per se* at these locations. Indicator kriging (IK) has often been used to estimate such probability maps [see e.g. Journel, 1983; Bilonick 1988]. IK is a non-parametric geostatistical technique that is entirely distribution free but which requires nevertheless the analysis and the modelling of many correlation functions.

In this work, the spatial correlation of annual [NO₂] measured at several sites in northern Italy is analysed for several concentration thresholds throughout 1997-1999. The correlation structures for different thresholds and years are then confronted to each other with the help of the fractal dimension calculated from the semivariogram [Burrough, 1981; Bruno and Raspa, 1989]. A further discussion is then presented of how the

¹ Now at the European Patent Office, Den Haag, The Netherlands.

degree of the spatial correlation field for several concentration thresholds may influence the preparation of probability maps of $[\text{NO}_2]$. This can lead to an improvement of the existing network by identifying important and/or redundant monitoring stations on the basis of whether or not certain regulatory air quality thresholds are exceeded.

2. THE DATA SET

The study area is an almost 33 thousand km^2 region located in the northern Italy including the broader area of Milan and the cities of Bergamo and Brescia. Stations measuring NO_2 operate since 1983 and have been recording high concentrations of the pollutant. During the past three years 1997-1999, the annual $[\text{NO}_2]$ at most monitoring sites exceeded the limit value of $40 \mu\text{g}/\text{m}^3$ (21.3 ppb) set by the European Directive 1999/30/EC for the protection of human health.

Site selection for the present analysis was first based on data completeness and then on the variance of the $[\text{NO}_2]$ at close-distant monitoring sites. The selected stations had at least 75% valid hourly $[\text{NO}_2]$ for the summer (April-September) and winter (January-March and October-December) seasons of each year separately. This aggregation criterion for annual means is recommended by the EC in COM(2000) 613 final.

Furthermore, the h-scatterplots at short distances (less than 5 km) for each year were used to identify which stations should be excluded from further analysis. "h-scatterplots" are plots of the concentration $z(\mathbf{x})$ observed at a point \mathbf{x} , against the concentration $z(\mathbf{x}+\mathbf{h})$ measured at the location $\mathbf{x}+\mathbf{h}$, with \mathbf{h} a separation vector. Stations showing large differences in the annual $[\text{NO}_2]$ with the neighbouring stations were excluded from the analysis. These irregularities at some sites were due to influences from local sources (traffic or industrial). It was preferred to distinguish those stations that were close to highways or industries on a statistical basis rather than on a geographical one and eliminate those sites from the analysis. A detailed discussion of how the analysis of measurements at short distances can provide information on the expected variation between neighbouring samples can be found in Dubois and De Cort [2001].

From the stations that satisfied the two aforementioned criteria, we selected the 59 sites that operated through all three years. The fact that we used the common stations in the 1997-1999 was imposed with a view to analyse the degree of the spatial correlation in the study area over different years (assess the meteorological impact, if any) and for different air quality thresholds, and

not due to the different configuration of the network. Details for the number of the originally available NO_2 monitoring sites and the remaining sites after the application of each validity criterion are given in Table 1.

Table 1. Number of NO_2 monitoring stations in the study area for 1997-1999 and number of remaining stations after application of each validity criterion.

Year	Originally available stations	After the 75% validity criterion	After the h-scatterplots criterion	Common stations for 1997-1999
1997	122	110	89	59
1998	126	99	82	59
1999	134	89	73	59

The locations of the 59 finally selected stations for the present analysis are shown in Figure 1, where additionally the three most populated cities of the study area (Milan, Brescia and Bergamo) are depicted.

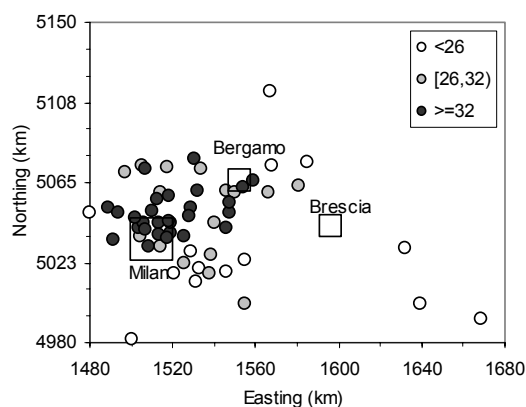


Figure 1. Locations of monitoring sites for NO_2 , distance units = km. White circles represent measurements with $[\text{NO}_2]$ (3-yr average) less than 26 ppb, grey circles between 26 and 32 ppb and black circles greater than 32 ppb. The three big cities of the study area, Milan, Brescia and Bergamo, are marked as squares.

The descriptive statistics of the annual $[\text{NO}_2]$ for the 59 selected stations are listed in Table 2, including the mean, median, minimum, maximum, standard deviation (SD) and skewness. The spatial annual average concentration and standard deviation for 1997 are 31.3 ± 8.4 ppb, while similar are the values for 1998 and 1999. The mean values are similar to the medians for all years, so the distributions of the annual $[\text{NO}_2]$ are not skewed. The low skewness and the low ratio of SD/mean (coefficient of variation) also indicate that there are no extreme values in the dataset.

Table 2. Descriptive spatial statistics of annual [NO₂] at the 59 selected sites for 1997-1999.

Year	mean (ppb)	median (ppb)	minimum (ppb)	maximum (ppb)	SD (ppb)	skew- ness
1997	31.3	32.8	12.4	49.5	8.4	-0.25
1998	30.9	32.0	11.2	47.5	8.2	-0.23
1999	29.0	29.8	12.0	47.2	7.3	-0.15

3. SPATIAL CORRELATION ANALYSIS

3.1 Indicator approach

The annual [NO₂] measured at the monitoring sites \mathbf{x} , $z(\mathbf{x})$, are converted into a binary data set of indicator codes, $I(\mathbf{x}, z_T)$, given a concentration threshold z_T :

$$I(\mathbf{x}, z_T) = \begin{cases} 1, & \text{if } z(\mathbf{x}) \geq z_T \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

That is, concentrations below the threshold are assigned a value of 0, whilst concentrations above the threshold are assigned a value of 1. Since the indicator variables have values that lie between 0 and 1, the krigged estimations will also have values between 0 and 1. Hence, the predictions can be interpreted as the probability of the variable exceeding the concentration threshold.

Twenty thresholds for annual [NO₂] have been used in the present analysis, ranging from 18 to 44 ppb. This procedure has been followed for each of the three years.

3.2 The semivariogram

Indicator codes can be used to describe spatial correlation structures. Such structures would not only justify the spatial interpolation of the data in order to generate maps of annual [NO₂] levels but it could also serve to optimise a sampling strategy [see i.e. Mc. Bratney et al., 1981]. Several functions exist to describe the spatial pattern of an environmental variable, the most frequently used being the semivariance:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [I(\mathbf{x}_i + \mathbf{h}, z_T) - I(\mathbf{x}_i, z_T)]^2 \quad (2)$$

where $\gamma(\mathbf{h})$ is the semivariance of the indicator codes and $N(\mathbf{h})$ is the number of pairs of $I(\mathbf{x}, z_T)$ for a separation distance (called the *lag distance*) \mathbf{h} . The plot of the semivariance versus $|\mathbf{h}|$ is called semivariogram. A condition for the existence of the semivariogram is that the difference between two samples depends only on \mathbf{h} and not on their locations. In other words, $\gamma(\mathbf{h})$ is expected to be invariant in space.

The value of the semivariogram at short distances, which should be at its minimum, is of particular interest. The *nugget*, for example, is the term used by geostatisticians to define the value of γ for $|\mathbf{h}|=0$. Although $\gamma|\mathbf{h}|=0$, in practice this is almost never the case because of micro-scale variations and/or measurement errors. Moreover, the nugget effect can only be derived from a model of the spatial correlation due to difficulty of taking two samples at exactly the same location and under the same experimental conditions. Ideally, the semivariance will increase with the distance (the correlation decreases between samples located farther apart) from a value equal to the nugget to a constant value called the *sill*. At distances shorter than the *range*, that is the distance corresponding to the sill, a correlation or covariance exists between two points in space. At larger distances the covariance is usually considered to be null, and the sill to correspond to the variance of the raw data. Further explanations about the definitions of these terms can be found in Isaaks & Srivastava [1989].

An initial structural analysis of the indicator sets revealed that the correlation structure of the variogram does not differ with the orientation for distances between samples smaller than 65 km. Therefore, only omnidirectional indicator semivariograms with a lag distance of 10 km were considered for the twenty thresholds. Figure 2 selectively shows the indicator semivariograms for a 32 ppb threshold for 1998 and 1999.

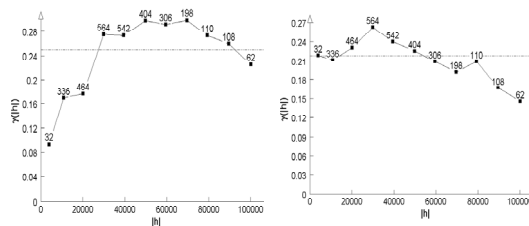


Figure 2. Omnidirectional semivariograms of indicator codes for the 32 ppb threshold in 1998 (left graph) and 1999 (right graph). Lag distance = 10,000 m. The theoretical variance is shown with a dashed line. The number of pairs of points used to calculate the variance at each lag are indicated.

The semivariogram for the measurements made in 1998 shows a low nugget effect and a range of approximately 50 km, while the semivariogram for the 1999 data does not reveal any spatial correlation. Consequently, the probabilities to exceed the 32 ppb threshold can not be estimated via interpolation for 1999. In such a case, one would have to rely on deterministic models to establish maps showing the probabilities to be above this threshold.

3.3 Fractal dimension of spatial correlation

In order to compare the spatial correlation of different semivariograms, one can use the ratio of the nugget to the sill after having fit a model to each semivariogram [Henebry, 1993]. Low ratios indicate strong spatial dependence and vice versa. However, the modelling of the semivariogram for each threshold is a long iterative process, during which the defined model is cross-validated; the parameters of the models are changed until the model does not produce any systematic errors. Less known is the use of the “fractal dimension” of the spatial correlation. The fractal dimension has the advantage that it does not require the modelling of the semivariogram since it can be calculated on the basis of the experimental semivariogram.

Mandelbrot [1977] introduced the term ‘fractal’ specifically for temporal or spatial phenomena that are continuous but not differentiable, and that exhibit partial correlations over many scales. Their dimensions need not be integers. A review of methods to estimate the fractal dimension of irregular surfaces is given in Haston [1996].

In this work, we applied the approach based on the fractal dimension D (the Hausdorff-Besicovitch statistic) of the spatial correlation [Burrough, 1981] estimated from the slope (Figure 3) of the log-log plot of the semivariogram as $|\mathbf{h}| \rightarrow 0$,

$$D = 2 - \frac{1}{2} \lim_{|\mathbf{h}| \rightarrow 0} \left(\frac{d \log \gamma(\mathbf{h})}{d \log |\mathbf{h}|} \right) \quad (3)$$

D varies between 1 (strong spatial correlation, slope = 2) and 2 (no correlation at all, slope = 0). The first six points from the semivariogram were used to define the log-log plots and calculate D for the various concentration thresholds for the years 1997-1999. Since the calculation of D is derived from the analysis of the semivariance, it is sensitive to the same parameters that affect the semivariogram (i.e. lag distance). To facilitate the comparisons, a lag distance of 10 km has been set by default for all the following calculations. One should also be aware of the fact that the calculation of D is based on regression analysis, and is therefore only estimation. Since a poor regression may introduce a bias in our conclusions, the correlation coefficients (r) associated to the respective estimation of D have also been taken

into account. When the correlation was found to be negative, which has no physical meaning in this context (see Figure 3 for the 35 ppb in 1999 for example), the value of D was set to 2 by convention.

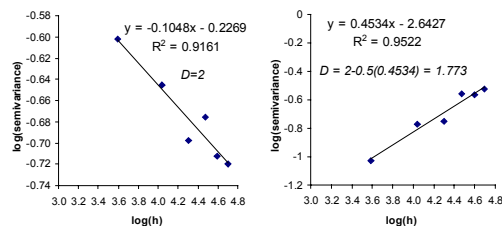


Figure 3. Calculation of the slope of log-log plot of the semivariograms of the indicator codes for 1998 at 32 ppb threshold (left) and for 1999 at 35 ppb (right). Lag distance = 10 km.

Figure 4 shows the values of D and the associated r 's calculated for the 20 indicator semivariograms of the NO_2 annual concentration thresholds for each of the three years. One can conclude that:

- D is highly dependent on the concentration threshold value, while it shows similar pattern over 1997 and 1998 and slightly different pattern in 1999. In a physical point of view, one would deduce that the meteorological conditions affecting the spatial distribution of NO_2 during the years 1997 and 1998 were, most probably, very similar, while 1999 shows different characteristics. Further studies are nevertheless required to assess properly the influence of meteorology on the spatial correlation structures of the analysed variable considering a longer timeseries of annual concentrations.
- The spatial correlation is at its maximum (i.e. minimum value of D) between 20-35 ppb for 1997, between 22-32 ppb for 1998 and between 20-31 ppb for 1999. Outside these boundaries, the spatial correlation becomes low and/or difficult to identify (low r associated with values of D close to 2).
- The range 20-31 ppb for which the spatial correlation is stronger over all years coincide with the range [minimum+SD, mean] of the annual $[\text{NO}_2]$ as given in Table 2.

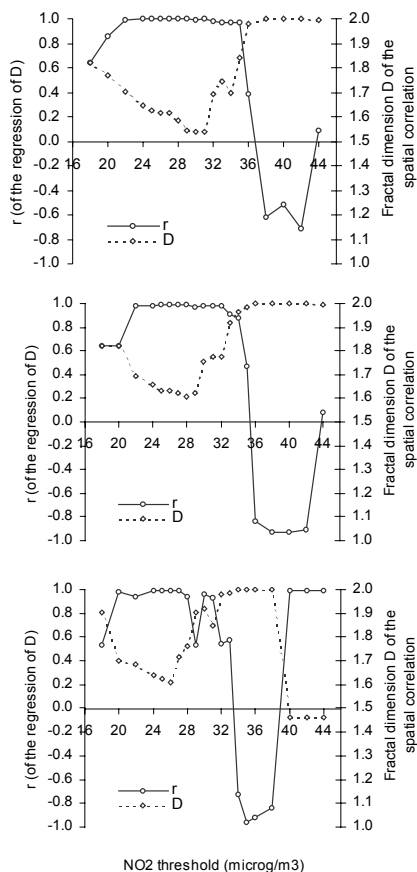


Figure 4. Fractal dimension D of the semivariograms for 1997 (top), 1998 (middle) and 1999 (bottom). Corresponding correlation coefficient r (solid line, left axis) when calculating D . Lag distance = 10 km.

In the light of these results, one can conclude that the spatial correlation of the probabilities to be above a specific threshold is not only fractal, but that it has a wide range of dimensionality. Identifying the thresholds for which the fractal dimension is low beforehand could be of enormous practical value because one could tailor sampling strategies to particular concentration levels, thereby improving the efficiency of expensive field investigations and the resulting interpolations. As a follow-up of this analysis, environmental policies related to air quality monitoring network design are discussed below.

4. DELINEATION OF POLLUTED AREAS

Knowledge of the cumulative probability to exceed several specific thresholds will help the decision-maker to classify a site x as polluted or not according to several criteria. A straightforward approach consists in the classification into “risk areas” of all the locations where the probability of

exceeding a concentration threshold z_T is greater than a critical probability threshold p_c :

$$x \text{ is polluted if } \text{prob} \{z(x) \geq z_T\} \geq p_c \quad (4)$$

In this work, cross-validation was used to investigate how the choice of p_c influences the proportion of locations that are misclassified. Cross-validation consists in the removing of one sampling measurement at a time and performing its estimation by means of an interpolation function using all of the remaining samples to estimate the probability of exceeding a given threshold at that site. This procedure is repeated until every sample has been, in turn, removed.

Most stations in the study area exceed the annual NO_2 limit value at $40 \mu\text{g}/\text{m}^3$ (21.3 ppb). However, the recommendations of the EC Directive regarding the minimum required number of monitoring sites in a zone are related to the lower ($26 \mu\text{g}/\text{m}^3$) and upper ($32 \mu\text{g}/\text{m}^3$) assessment thresholds, which are exceeded in nearly all stations. Therefore, we have selected, for illustrative purposes, two thresholds higher than the limit value: one at 26 ppb and one at 32 ppb.

At a first stage the probabilities of $[\text{NO}_2]$ to be lower than these two thresholds were estimated by cross-validation and were then confronted with the measured annual $[\text{NO}_2]$ at the respective monitoring stations. In total, 177 cases ($59 \text{ sites} \times 3 \text{ years}$) have been used. As shown in Figure 5, the smallest misclassification rate for the measured annual concentrations at the sites is reached at $p_c=0.73$ (23.4%) for the 26 ppb threshold and at $p_c=0.72$ for the 32 ppb threshold (25.4% misclassified cases). As misclassified cases are considered both the false negatives (i.e. locations where the actual measurements exceed z_T but the estimated probability of exceeding this threshold is less than the corresponding p_c) and the false positives.

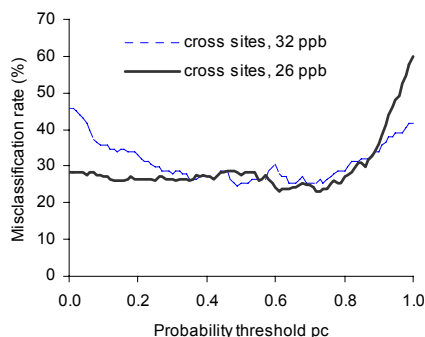


Figure 5. Impact of the probability threshold p_c on the proportion of the cross-validated sites that are wrongly classified as exceeding the 26 ppb (solid line) or the 32 ppb selected thresholds (dashed line)

The critical probability threshold, for which the minimum misclassification rate is reached at the cross-validation set, could be used as a guide to identify which of the existing stations (not included in the initial analysis) are important or redundant in verifying whether or not the two selected annual NO₂ assessment thresholds are being exceeded in the study area. This can be achieved by first estimating via indicator kriging (IK) the probability of exceeding a concentration threshold at the locations of the existing monitoring sites that were not included in the set of 59 sites. The locations where the estimated probability is greater than the p_c value, but the measured concentration is lower than the concentration threshold z_T and vice versa (Equations 5 and 6), are important for identifying whether this concentration threshold met.

$$\text{prob} \{z(\mathbf{x}) \geq z_T\} \geq p_c \text{ and } z(\mathbf{x}) < z_T \quad (5)$$

$$\text{prob} \{z(\mathbf{x}) \geq z_T\} < p_c \text{ and } z(\mathbf{x}) \geq z_T \quad (6)$$

The existing stations for which:

$$\text{prob} \{z(\mathbf{x}) \geq z_T\} \geq p_c \text{ and } z(\mathbf{x}) \geq z_T \quad (7)$$

$$\text{prob} \{z(\mathbf{x}) \geq z_T\} < p_c \text{ and } z(\mathbf{x}) < z_T \quad (8)$$

are redundant, because the exceedance or not of the regulatory concentration thresholds can be determined via indicator kriging.

5. CONCLUSIONS

This paper has described the variation of the spatial correlation for different annual [NO₂] thresholds, in terms of the fractal dimension of the semivariogram, in N. Italy. The fractal dimension provides useful information to decision-makers, which would help them not only to prepare reliable probability maps of NO₂ concentrations, but also to tailor sampling strategies to particular concentration levels, thereby improving the efficiency of expensive field investigations and the resulting interpolations. Finally, suggestions have been made of how the fractal analysis could be used as a guide to streamline the existing network by identifying important and/or redundant monitoring stations on the basis of whether or not certain regulatory thresholds of annual NO₂ concentrations are met.

6. REFERENCES

Bilonick, R. A., Monthly hydrogen ion deposition maps for the Northeastern U.S. from July 1982 to September 1984. *Atmospheric Environment*, 22(9), 1909-1924, 1988.

Bruno, R. and Raspa G., Geostatistical characterization of fractal models of surfaces.

In *Geostatistics* (pp. 77-89), M. Armstrong (Ed.). Dordrecht: Kluwer Academic, 1989.

Burrough, P.A., Fractal dimensions of landscapes and other environmental variables, *Nature*, 294, 240-242, 1981.

Dubois, G., and De Cort, M., Mapping ¹³⁷Cs: data validation methods and data interpretation. *Journal of Environmental Radioactivity*, 53(3), 271-289, 2001.

Elsom, D.M., Spatial correlation analysis of air pollution data in an urban area, *Atmospheric Environment*, 12, 1103-1107, 1978.

European Commission, Amended proposal for a Directive of the European Parliament and of the Council relating to ozone in ambient air, COM(2000) 613 final.

European Commission, Council Directive 1999/30/EC relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air, *Official Journal of the European Communities* L 163/41, 26.6.1999.

Goovaerts, P. and van Meirvenne M., Accounting for measurement and interpolation errors in soil contaminant mapping and decision-making, *Annual Conference of the International Association for Mathematical Geology*, Cancun, Mexico, September, 2001.

Haston, M.B., Shear strength testing and fractal analysis of rock discontinuities, M.S Thesis, University of Tennessee, Knoxville, 1996.

Henebry, G. M., Detecting change in grasslands using measures of spatial dependence with Landsat TM data. *Remote Sensing of Environment* 46, 223-234, 1993.

Hopkins, L. P., Ensor, K. B. and Rifai, H. S., Empirical evaluation of ambient ozone interpolation procedures to support exposure models. *J. Air & Waste Management Association*, 49, 839-846, 1999.

Isaaks, E. H., & Srivastava, R. M., *An introduction to applied geostatistics*. Oxford University Press. 1989.

Journel, A. G., Non parametric estimation of spatial distributions. *Mathematical Geology*, 15(3), 445-468, 1983.

Mandelbrot, B., *Fractals, Form, Chance and Dimension*, Freeman, San Francisco, 1977.

McBratney, A. B., Webster, R., & Burgess, T. M., The design of optimal sampling schemes for local estimation and mapping of regionalized variables. I. Theory and method. *Computers and Geosciences*, 7(4), 331-334, 1981.