Characterization of National Spatial Variation

Terence Fitz-Simons
U.S. Environmental Protection Agency, Research Triangle Park, NC 27711

Abstract

Spatial variability is an important quality of air pollutants for many areas of policy within the U.S. Environmental Protection Agency (EPA). Obviously, monitoring regulations depend heavily on knowledge of spatial variability. In addition, control strategies depend on this knowledge, which helps determine whether a local or regional program would be more effective. Action day programs and public information programs also benefit from this knowledge.

Traditionally, spatial variation has been depicted by isopleth maps, concentration maps, and box plots of various sites. Does this really give us useful knowledge about spatial variation? This paper explores a new way to examine spatial variability on a national scale and also presents an extension of this method in an attempt to characterize spatial variability in a useful way. The new methodology is presented along with its application using $PM_{2.5}$ and ozone data.

Introduction

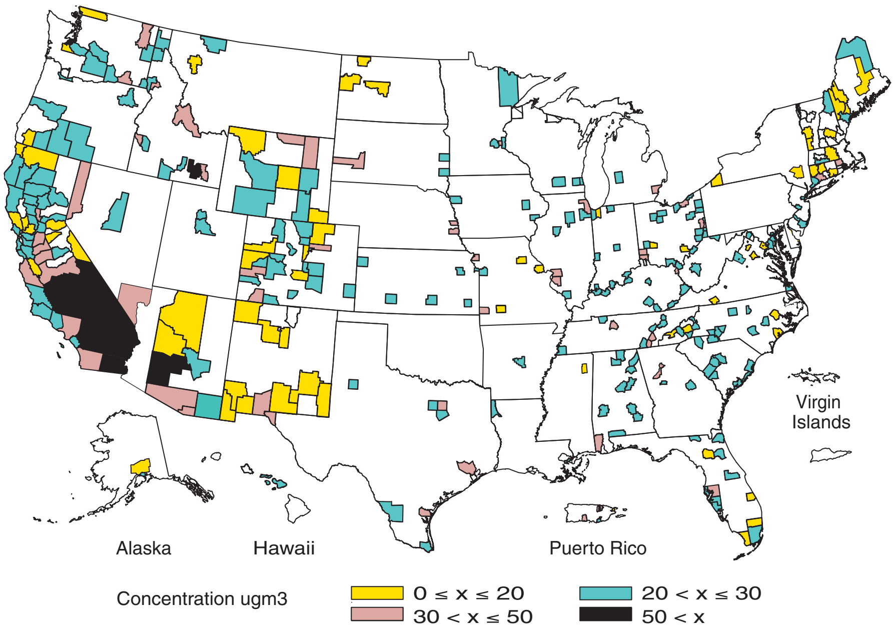
Spatial variability is a very important quality of air pollutants for many areas of EPA policy. Obviously, monitoring regulations and network design depend heavily on knowledge of spatial variability, as do implementation strategies and policies. Control strategies also depend heavily on this knowledge, which helps state and local agencies decide whether a local or regional program may be more effective. Action day programs and public information programs also depend on this information to facilitate decisions regarding how large of an area should be included in various alerts or information publications. Traditionally, spatial variation has been depicted by isopleth maps, concentration maps, and box plots of various sites. Each of these methods gives a crude idea of spatial variability. This paper explores a new way to visualize large-scale spatial variability and also presents an extension of this method in an attempt to characterize spatial variability in a useful way. The new methodology is presented along with its application using data from several pollutants nationwide.

Characterizing Spatial Variation

One of the first questions arising from almost any investigation of an air pollutant is, "What is the spatial and temporal variability or variation?" Very often, the spatial part of the question is answered with a map showing ranges of pollutant levels by county. These maps show where pollutant levels are higher and lower and, in general, where information is available or where monitoring sites are located (see Figure 1).

After the work of producing the map is done, the question is usually considered answered. However, this is a crude view of spatial variability. Looking at such a map, counties with

Figure 1. PM_{10} annual averages (county maximum).



higher values are easily spotted but it is hard to visualize how close adjoining counties are to others. Some analysts go a step farther and show a map of an estimated surface of pollutant levels. The latest and most popular way to do this is called kriging.[1] Kriging is a spatial interpolation technique developed for the mining industry in South Africa to predict ore reserves. With an interpolated surface, all the blank areas on the map are gone, and it is somewhat easier to see how pollutants may vary over space. Figure 2 provides an example of a kriged surface. Because the surface itself is smoothed by the process, kriging actually hides some of the spatial variation, which may or may not be a good result depending on the purpose of the analysis.

At the heart of kriging is a concept called a variogram, which is a representation of the statistical variance of the difference between two data points on a map as it relates to the distance between the two points on the map. Much like the mean, which is a measure of the center of a distribution of data, the variance is a measure of the spread of a distribution of data. In this case, the data are a series of measurements representing differences between two locations paired by time. Thus if d_i is the difference between two readings at two monitors at a given time i, then $d_i = x_{1i} - x_{2i}$. If x_1 and x_2 are both random variables from two locations, then the variance of the difference is $V(x_1 - x_2)$, or V(d). In fact, the variance of the difference is $V(d) = V(x_1) + V(x_2) - 2COV(x_1,x_2)$. This is the sum of the variances of the two random variables minus twice the covariance (a measure of how much the two random variables vary together). Basically, this says that the more the two random variables change together (they go up or down together but they do not necessarily change the same amount), the smaller the variance of the difference will be because the values at two different sites would be expected to vary together more if they are close together and vary more independently if they are far apart. This leads to the concept of the variogram, which, in this case, is the relationship between the variance of the differences and the distance between two sites (Figure 3). The dotted line in Figure 3 shows how the variance changes with the distance. At a distance of zero (0), there is still variation left that does not go away even if the sites are at the same location. This is called the nugget. Similarly, there is a point, called the sill, at which the variance levels out. The area between 0 and the sill is called the range. The range can be thought of as the region where there is a correlation between two sites. The region after the sill can be thought of as the distances at which sites appear to be independent of each other.

Figure 2. Example of a kriged surface.

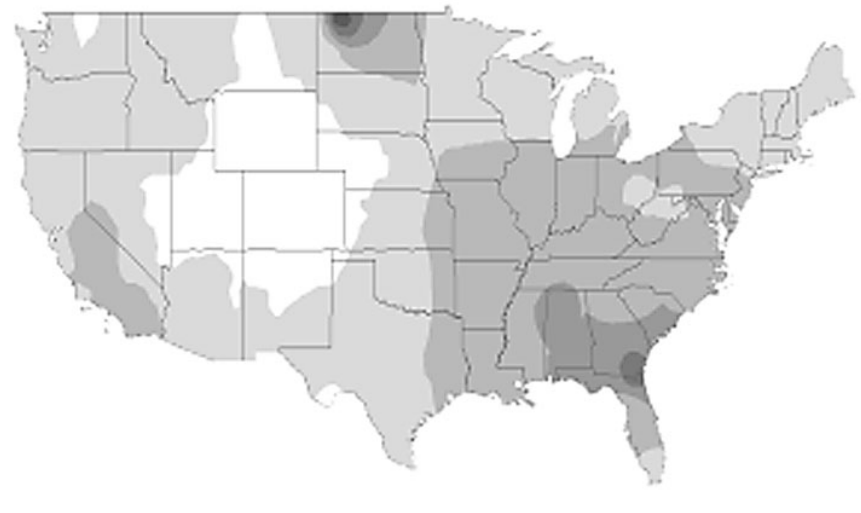


Figure 3. Schematic of a variogram.

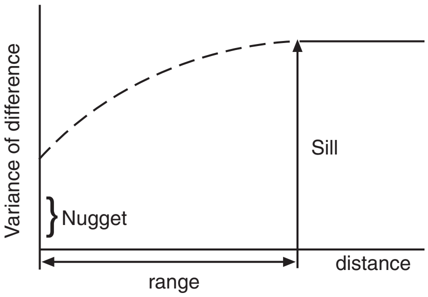


Figure 4 shows how $PM_{2.5}$ data can be used to plot the variance of the difference against distance. The difference in daily $PM_{2.5}$ values was calculated for various sites across the country. The variance of the differences was calculated, and the latitude and longitude of each site were used to calculate the distance between two sites. Each pair of sites then had a variance of the difference and a distance, which were plotted for all possible pairs of sites across the country.

Looking at the scatterplot, it is clear that there is no simple relationship between the variance of the difference and distance. A very dense cluster of points seems to center over 25 at 0 distance and then slowly increases as the distance increases. However, from a casual examination of the plot, enough points fall outside the dense cloud (in fact, many were cut off to actually see any trend at all by setting the maximum variance displayed to 500) to bring into question the assumption used in kriging, as shown in Figure 3, that the variance of the difference over distance can be described by a line.

The point of defining all these terms is to show that the variance of the differences between two measurements taken at the same time but at different locations is generally increasing because the covariance is decreasing over the distance. Because the correlation is covariance normalized by the variances, we can characterize the spatial dependence of data from two locations through the correlation. Because the variance of the difference generally increased, the covariance and, therefore, the correlation should decrease over distance. This raises the question, how does the correlation vary over distance? To answer this question, $PM_{2.5}$ data were used to calculate the correlation of daily $PM_{2.5}$ values between two sites, and the latitude and longitude were used to calculate the distance between two sites. Thus for each pair of sites, we have correlation and a distance. Looking at all the possible pairs of sites, scatterplots may be generated, such as the one in Figure 5. The values of the correlations are restricted to all values between -1 and 1, but the variance of the distance must be positive. These restrictions help provide a much more coherent picture. There is, again, a dense cloud that trends downward as the distance increases. Also, there are many points not in the dense cloud that fall beneath the trend. Again, these points are numerous enough to question the simplicity of the variogram used in kriging.

To simplify what is seen in this scatter plot, the data could be summarized by box plots of the data over 20-km intervals. This would result in Figure 6, which shows a much less confusing picture. The whiskers represent the maxima and minima of the intervals. The box represents the 75th and 25th percentiles, the plus sign (+) represents the mean, and the single line in the box represents the median or 50th percentile. Now a trend is much more apparent in the correlation than in the scatterplot. However, this display shows only how well the data "track" or follow a pattern. It does not show how well the data from different sites actually agree. In other words, the data from one site might track the data from another site very well but still have very different concentrations on average than data from the other site. Here we present a solution to this problem, a coefficient of perfect agreement, or CPA.

Figure 4. Variance of the difference vs. distance.

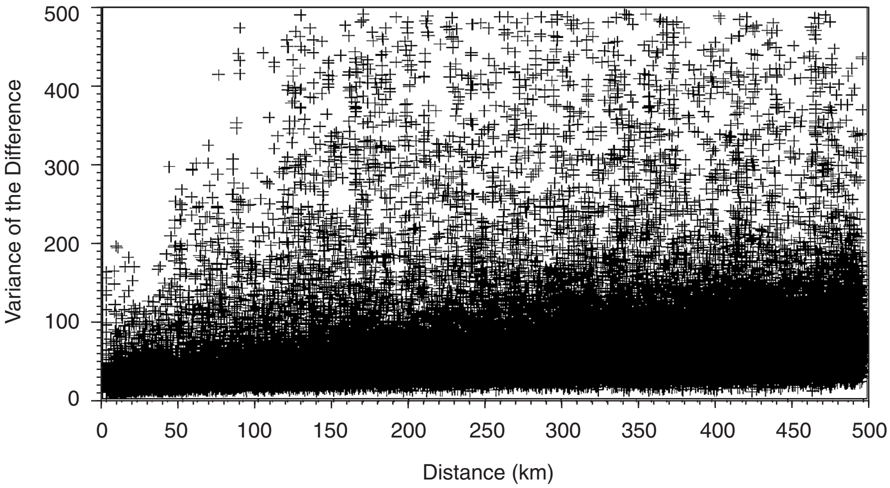


Figure 5. Correlation (r) vs. distance for $PM_{2.5}$.

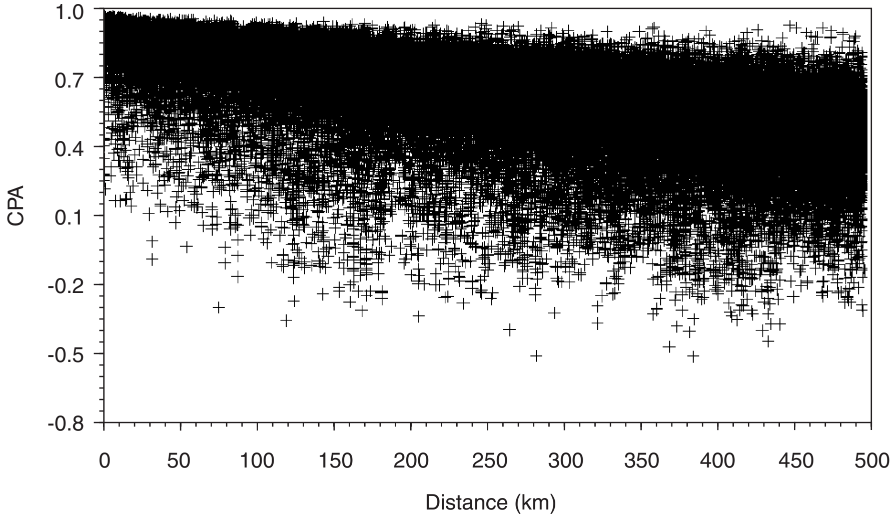
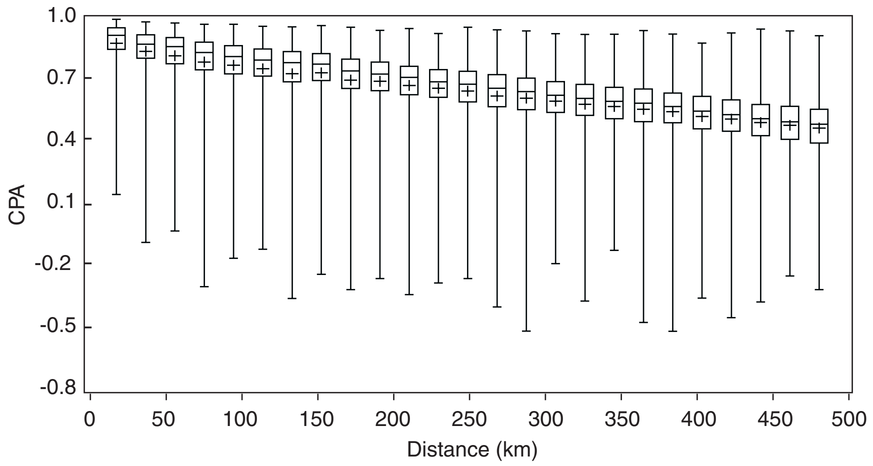


Figure 6. Box plot of correlation vs. distance.



The Coefficient of Perfect Agreement

The goal of formulating a CPA is to give a measure of agreement with many of the characteristics of the correlation coefficient.

The classical correlation coefficient is a measure of how well paired values track each other. The value 0 (zero) means they do not track each other at all, whereas a value of 1 means they track each other perfectly (all the points in a scatterplot would be on a straight line). A value of -1 also means perfect tracking, but the scatterplot line would have a downward or negative slope. The correlation coefficient is defined as shown in Equation A.

As stated earlier, the correlation coefficient has a nice feature in that, when the data from two sites agree in a perfectly linear fashion, then r is 1 (or -1). However, if the data agreed perfectly, the only line that mattered would be a line with a slope of 1 and an intercept of 0 (the line y = x). Therefore, the first characteristic we desire in a CPA is that the CPA = 1 when all points in a scatterplot fall on the line y = x, and the CPA = 0 if there is no systematic agreement. One way to create this would be to include a term in the denominator of the correlation coefficient as shown in Equation B.

Equation A

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\frac{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}{n}}$$

Equation B

$$CPA = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\frac{\sum (x-y)^2 + \sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}{n}}$$

Equation C

$$CPA = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\frac{\frac{\sum (x-y)^2}{n} + \sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}{n}}$$

Equation D

$$CPA = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\frac{\left(\sum (x-y)^2\right)n}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}} + \sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$

If there were no agreement, this term would become large and the CPA would become small (or close to 0). If there were perfect agreement, the term would be 0, and, because all the points would fall on a straight line, the rest of the equation (the correlation coefficient) would be 1, allowing the CPA to be 1. However, if the two data streams fell on a straight line that did not have a slope of 1 and an intercept of 0, then the

CPA would certainly not be 1 but less than 1 because y would not equal x everywhere. This seems to have all the characteristics desired in a CPA.

However, note that the $\sum(x - y)^2$ term will get larger and larger as the number of data points gets larger and larger, making the CPA get smaller and smaller. Unless there were a situation of perfect agreement, then such a CPA could be made to be arbitrarily small by taking larger and larger numbers of data points to compute the CPA. A further refinement would then be defined as shown in Equation C.

This solves the sample size problem, but there is one problem left. The correlation coefficient is a unitless or unit invariant quantity. This CPA is not, but it should be. Units have been reintroduced into the formula. Because a units conversion could result in a different CPA value, this is not a desirable trait for a coefficient. The added term is divided by the same divisor used to normalize the covariance to get the correlation resulting in Equation D.

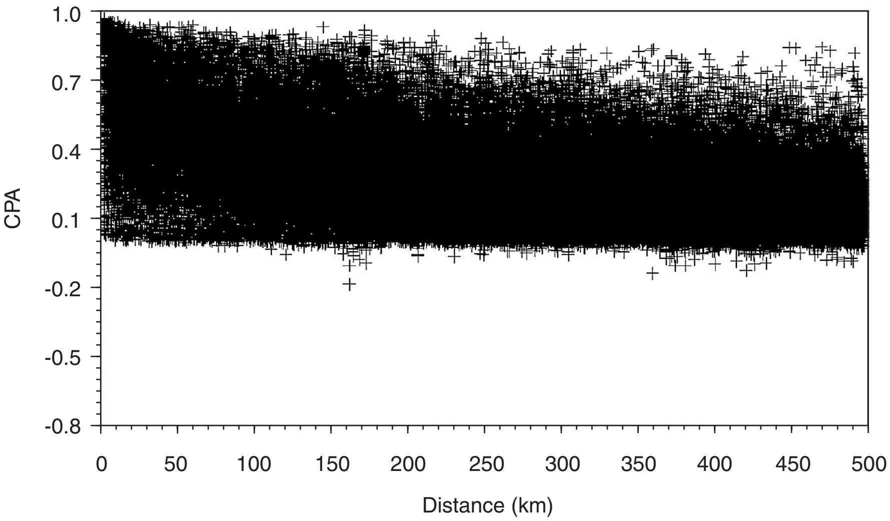
Now the CPA is unitless.

Monte Carlo studies of the CPA were performed by generating values from a straight line. In linear regression, $Y = a + bX + e$, where e has a normal distribution with a mean of 0 and a variance of σ^2. This last term is also called the variation about the line. Five hundred sets of values were generated with different slopes, intercepts, and variations about the line. Slopes ranged from 0 to 5, intercepts ranged from -10 to 10, and the variance about the line, σ^2, ranged from 0 to 100. In this case, whenever σ^2 is 0, then r is 1 (a perfect linear relationship). However, the CPA is equal to 1 only if a is 0, b is 1, and σ^2 is 0. The studies found the CPA to be relatively sensitive to the lack of perfect agreement when there was only a perfect linear relationship (when r is 1 and the CPA should be less than 1).

Application

Using the CPA instead of r, a new scatterplot can be constructed (Figure 7). Now the denser part of the distribution of points has a different trend. The trend dips quickly and then falls off gradually. If, as before, the data are displayed as box and whisker plots, the more pronounced trend in Figure 8 is revealed. This gives a national picture of the spatial variation of $PM_{2.5}$. The mean CPA starts off at around 0.6 and falls off rapidly out to about 150 km, then falls off gradually from there to about 0.2 at 500 km. The maximum and minimum of the coefficient (the whiskers on the box and whiskers plot) still vary almost across all possible values of the coefficient (perfect agreement, or 1, to no agreement at all, or 0) at any distance. Quantitatively, interpretation of this coefficient is difficult at best. Where it might be of most use is in comparisons with other pollutants.

Figure 7. CPA vs distance (km).



Comparison of Pollutants

Pollutants can be compared by following the previous steps used to produce Figure 8. The means in Figure 8 (the pluses [+]) can be joined by a line for several pollutants. This is where the usefulness of a CPA can be demonstrated. A comparison between pollutants could be made to help guide policy. For example, daily values of $PM_{2.5}$, daily values of PM_{10}, hourly values of CO (carbon monoxide), and hourly values of ozone were used to produce Figure 9. As can be seen from the plot, $PM_{2.5}$ has a mean CPA that is above ozone for most of the distances out to 500 km (at least until 450 km). This might suggest that if a regional control strategy is being pursued for the ozone problem in the United States, a regional strategy also makes sense for $PM_{2.5}$.

Figure 8. Coefficient of perfect agreement vs distance (km).

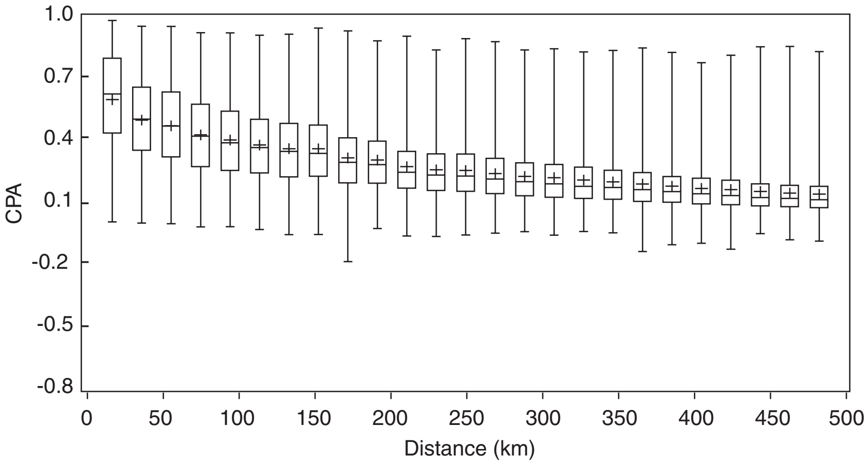
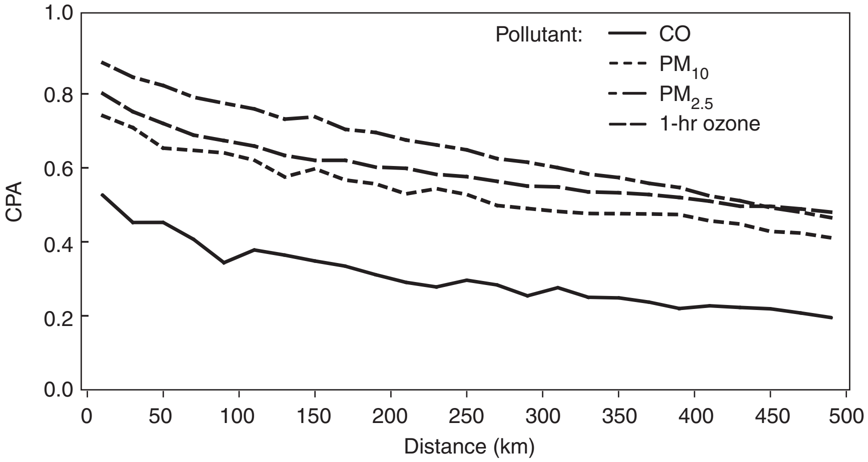


Figure 9. Comparison of mean CPA vs distance (km).



Conclusions

A CPA can be formulated that can be of some use in assessing spatial variation on a national scale. The statistical properties of the CPA used here are not known, and the CPA cannot be used to quantify this variability. However, it can be a useful comparative tool to visualize differences in national scale spatial variation among pollutants.

References

1. Matheron, G. Principles of Geostatistics. *Economic Geology*. **1963**, 58, 1246–1266.