STATISTICAL TECHNIQUES FOR SPATIAL DATA ANALYSIS

Prachi Misra Sahoo
I.A.S.R.I., Library Avenue, New Delhi-110 012
rprachi@iasri.res.in

1. Introduction

Spatial data analysis aims at extracting implicit knowledge such as spatial relations and patterns that is not explicitly stored in spatial databases. It distinguishes itself from classical data analysis in that it associates with each object the attributes under consideration including both non-spatial and spatial attributes.

We distinguish three prevalent spatial data types, defined by the topology of the entity to which the recorded information refers. These are point, lines and area. Features having a specific location, but without extent in any direction are considered as points. A pair of coordinates represents a point. Village locations, industrial locations, cities *etc.* are the examples of the point data. Lines features consist of series of x, y coordinate pairs with discrete beginning and ending points. Features like rivers, road networks, represents lines. Features defined by a set of linked lines enclosing an area are known as polygons. Polygons are characterized by area and perimeter. Administrative boundaries, land use, soil map *etc*. are the polygon features.

Statistical analysis which deals with spatial data is termed as the science of Spatial statistics. Spatial statistics span many disciplines, with methods varying in relation to the specific research questions being addressed, whether predicting ore quality in mining, examining suspiciously high frequencies of disease events, or handling the vast data volumes being generated by GPS (global positioning system) and satellite remote sensing. A unique feature of spatial data is that geographical location provides a key shared either exactly or approximately between data sets of different origins. Census data can be overlayed over patient or customer data; environmental data can be integrated with disease frequencies; problems which hitherto did not admit ready empirical testing are becoming approachable It is an area of spatial analysis that has grown significantly in the last twenty years. It encompasses an impressive array of sophisticated methods and techniques for visualization, exploration and modeling of spatial data which are described here.

2. Descriptive Spatial Statistics

A set of descriptive spatial statistics has been developed (Table 1) that are areal or locational equivalents to the nonspatial measures.

Table 1: Nonspatial and Spatial Descriptive Statistics

Statistic	Central tendency	Absolute Dispersion	Relative Dispersion
Nonspatial	Mean	Standard Deviation	Coefficient of Variation
Spatial	Mean Center or Median Center or Euclidean Median	Standard Distance	Relative Distance

2.1 Spatial Measures of Central Tendency

Mean Center

The mean is an important measure of central tendency for a set of data. If this concept of central tendency is extended to locational point data in two dimensions (X and Y coordinates), the average location, called the mean centre, can be determined.

Consider the spatial distribution of points shown in Fig. 1. These points might represent any spatial distribution of interest, the only stipulation is that the phenomenon can be displayed graphically as a set of points in a two-dimensional coordinates system.

Once a coordinate system has been established and the coordinates of each point determined, the mean center can be calculated by separately averaging the X and Y coordinates, as follows:

$$\overline{X}_c = \frac{\Sigma X_i}{n} \quad \text{and} \quad \overline{Y}_c = \frac{\Sigma Y_i}{n}$$

where

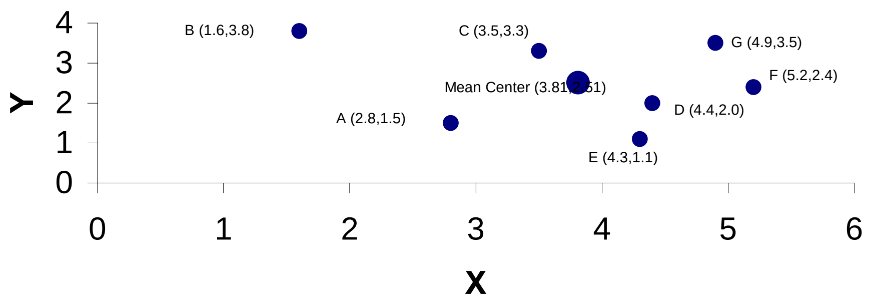
$\overline{\mathrm{X}}_{\mathrm{c}}$ = mean center of X, $\overline{\mathrm{Y}}_{\mathrm{c}}$ = mean center of Y

X_{i} = X coordinate of point i, Y_{i} = Y coordinate of point i

n = number of points in the distribution

For the point pattern shown in Fig. 1, the mean centre coordinates are
$\overline{\mathrm{X}}_{\mathrm{c}} = 3.81$ and $\overline{\mathrm{Y}}_{\mathrm{c}} = 2.51$.

Fig. 1: Graph of Locational Coordinates and Mean Center



The mean center may be considered the center of gravity of a point pattern or spatial distribution. In many geographic applications, it is appropriate to assign differential weights to points in a spatial distribution. The weights are analogous to frequencies in the analysis of grouped data (e.g., weighted mean).

$$\overline{X}_{wc} = \frac{\Sigma f_i X_i}{\Sigma f_i} \quad and \quad \overline{Y}_{wc} = \frac{\Sigma f_i Y_i}{\Sigma f_i}$$

$\overline{\mathrm{X}}_{\mathrm{wc}}$ = weighted mean center of X

$\overline{\mathrm{Y}}_{\mathrm{wc}}$ = weighted mean center of Y

f_{i} = frequency (weight) of point i

The mean center serves as a spatial analogue to the mean, in that it is the location that minimizes the sum of squared deviations of a set of points. Thus, the mean center has the same least squares property as the mean. The mean center $\left(\overline{\mathrm{X}}_{\mathrm{c}},\ \overline{\mathrm{Y}}_{\mathrm{c}}\right)$ minimizes:

$$\sum[\left(X_i - \overline{X}_c\right)^2 + \left(Y_i - \overline{Y}_c\right)^2]$$

In a location coordinate system, deviations such as $\left(\mathrm{X}_{\mathrm{i}} - \overline{\mathrm{X}}_{\mathrm{c}}\right)$ and $\left(\mathrm{Y}_{\mathrm{i}} - \overline{\mathrm{Y}}_{\mathrm{c}}\right)$ are, in fact, distances between points. One standard procedure for measuring distances is based on straight line or Euclidean distance. The Euclidean distance $\left(\mathrm{d}_{\mathrm{i}}\right)$ separating point i $\left(\mathrm{X}_{\mathrm{i}}\ \ \mathrm{Y}_{\mathrm{i}}\right)$ from the mean center $\left(\overline{\mathrm{X}}_{\mathrm{c}},\ \overline{\mathrm{Y}}_{\mathrm{c}}\right)$ is defined by the Pythagorean theorem as follows:

$$\mathrm{d}_{\mathrm{i}} = \sqrt{\left(\mathrm{X}_{\mathrm{i}} - \overline{\mathrm{X}}_{\mathrm{c}}\right)^2 + \left(\mathrm{Y}_{\mathrm{i}} - \overline{\mathrm{Y}}_{\mathrm{c}}\right)^2}$$

Thus, the mean center is the location that minimizes the sum of squared distances to all points. This characteristic makes the mean center an appropriate center of gravity for a two-dimensional point pattern, just as the mean is the center of gravity along a one-dimensional number line.

Euclidean Median

For many geographic applications, another measure of "center" is more useful. Often, it is more practical to determine the central location that minimizes the sum of unsquared, rather than squared, distances. This location, which minimizes the sum of Euclidean distances from all other points in a spatial distribution to that central location, is called the Euclidean median $\left(\mathrm{X}_{\mathrm{e}},\ \mathrm{Y}_{\mathrm{e}}\right)$ or median center. Mathematically, this location minimizes the sum:

$$\Sigma\sqrt{\left(X_i - X_e\right)^2 + \left(Y_i - Y_e\right)^2}$$

Determining coordinates of the Euclidean median is complex methodologically. A weighted Euclidean median is a logical extension of the simple (unweighted) Euclidean median. The coordinates of the weighted Euclidean median $\left(X_{we},\ Y_{we}\right)$ will minimize the expression

$$\sum f_i\sqrt{\left(X_i - X_{we}\right)^2 + \left(Y_{i_i} - Y_{we}\right)^2}$$

The weights or frequencies may represent population, sales volume, or any other feature appropriate to the spatial problem.

2.2 Spatial Measures of Dispersion

Standard Distance

As the mean center serves as a locational analogue to the mean, standard distance is the spatial equivalent of standard deviation. Standard distance measures the amount of absolute dispersion in a point pattern. After the locational coordinates of the mean center have been determined, the standard distance statistic incorporates the straight-line or Euclidean distance of each point from the mean center. Standard distance (S_D) is written as follows:

$$S_D = \sqrt{\frac{\Sigma\left(X_i - \overline{X}_c\right)^2 + \Sigma\left(Y_i - \overline{Y}_c\right)^2}{n}} \quad \text{or} \quad S_D = \sqrt{\left(\frac{\Sigma X_i^2}{n} - \overline{X}_c^2\right) + \left(\frac{\Sigma Y_i^2}{n} - \overline{Y}_c^2\right)}$$

Like standard deviation, standard distance is strongly influenced by extreme or peripheral locations. Because distances about the mean center are squared, "uncentered" or atypical points have a dominating impact on the magnitude of the standard distance. The standard distance is calculated in Table 2 and shown as the radius of a circle whose centre is the mean center in Fig.2.

Weighted standard distance is appropriate for those geographic applications requiring a weighted mean center. The definitional formula for weighted standard distance (S_{WD}) is:

$$S_{WD} = \sqrt{\frac{\Sigma f_i \left(X_i - \overline{X}_c\right)^2 + \Sigma f_i \left(Y_i - \overline{Y}_c\right)^2}{n}}$$

Table 2: Table for Calculating Standard Distance

	Locational Coordinates			
Point	X_i	Y_i	X_i^2	Y_i^2
A	2.8	1.5	7.84	2.25
B	1.6	3.8	2.56	14.44
C	3.5	3.3	12.25	10.89
D	4.4	2.0	19.36	4.00
E	4.3	1.1	18.49	1.21
F	5.2	2.4	27.04	5.76
G	4.9	3.5	24.01	12.25

$\overline{X}_c = 3.81$ *and* $\overline{Y}_c = 2.51$, $\overline{X}_c^2 = 14.52$ *and* $\overline{Y}_c^2 = 6.30$, therefore $S_D = 1.54$.

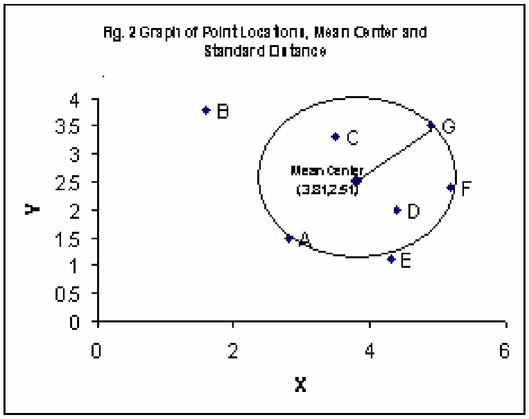


Fig. 2 Graph of Point Locations, Mean Center and Standard Distance

Relative Distance

The coefficient of variation (standard deviation divided by the mean) is the nonspatial measure of relative dispersion. A perfect spatial analogue to the coefficient of variation does not exist for measuring relative dispersion.

To derive a descriptive measure of relative spatial dispersion, the standard distance of a point pattern is divided by some measure of regional magnitude. One possible divisor is the radius (r_A) of a circle with the same area as the region being analyzed. A useful measure of relative dispersion, called relative distance (R_D), can now be defined:

$$R_D = \frac{S_D}{r_A}$$

This relative distance measure allows direct comparison of the dispersion of different point patterns from different areas, even if the areas are of varying sizes.

3. Spatial Association

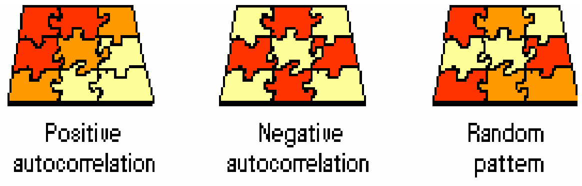
Spatial association enables to assess statistically the degree of spatial dependence in the data. Finding the degree of spatial association (correlation) among data representing related locations is fundamental to the statistical analysis of dependence and heterogeneity in spatial patterns. In 1960s the most challenging spatial question was: In an unbiased way, how is one to account for the correlation in spatially distributed variables? The next problem was the difficulty in dealing with unequally sized and irregularly shaped units.

3.1 Chi-Square Statistic

The Chi-Square statistic measures the strength of association between spatial distributions of two variables. For example relationship between wheat yield and precipitation or relation between two maps showing high and low yields and high and low precipitation.

3.2 Spatial Autocorrelation

Given a group of mutually exclusive units or individuals in a two dimensional plane, if the presence, absence or degree of a certain characteristic affects the presence, absence or degree of the same characteristic in neighbouring units, then the phenomenon is said to exhibit spatial autocorrelation (Cliff and Ord, 1973). Spatial autocorrelation tests whether or not the observed value of a variable at one locality is independent of values of that variable at neighbouring localities. A positive spatial autocorrelation refers to a map pattern where geographic features of similar value tend to cluster on a map, whereas a negative spatial autocorrelation indicates a map pattern in which geographic units of similar values scatter throughout the map. When no statistically significant spatial autocorrelation exists, the pattern of spatial distribution is considered to be random (Figure).



Classical Measure of Spatial Autocorrelation

- Moran's I
- Geary's C

Moran (1950) proposed the following measure to calculate the spatial autocorrelation (β);

$$\beta = \frac{N}{S} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Where, x_i is the observed value at location i, N is the number of locations and $S = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$, $(i \neq j)$. The weighting function w_{ij} is used to assign weights to every pair of locations in the study area, w_{ij} =1 , if i and j are neighbours and = 0, otherwise.

The range of Moran's autocorrelation varies from approximately -1 to +1. Positive sign represents positive spatial autocorrelation, while the converse is true for negative. Zero indicates no spatial autocorrelation. For calculation of weighing function one needs to identify whether the two locations are neighbours or not. This requires the criteria to decide about the definition of neighbours.

Geary's C

In this case the interaction is not the cross-product of the deviations from the mean, but the deviations in intensities of each observation location with one another. It is inversely related to Moran's I. It does not provide identical inference because it emphasizes the differences in values between pairs of observations, rather than the covariation between the pairs. Moran's I gives a more global indicator, whereas the Geary coefficient is more sensitive to differences in small neighborhoods.

$$C = \frac{[(N-1)[\sum_i \sum_j W_{ij}(X_i - X_j)^2]}{2(\sum_i \sum_j W_{ij}(X_i - \bar{X})^2}$$

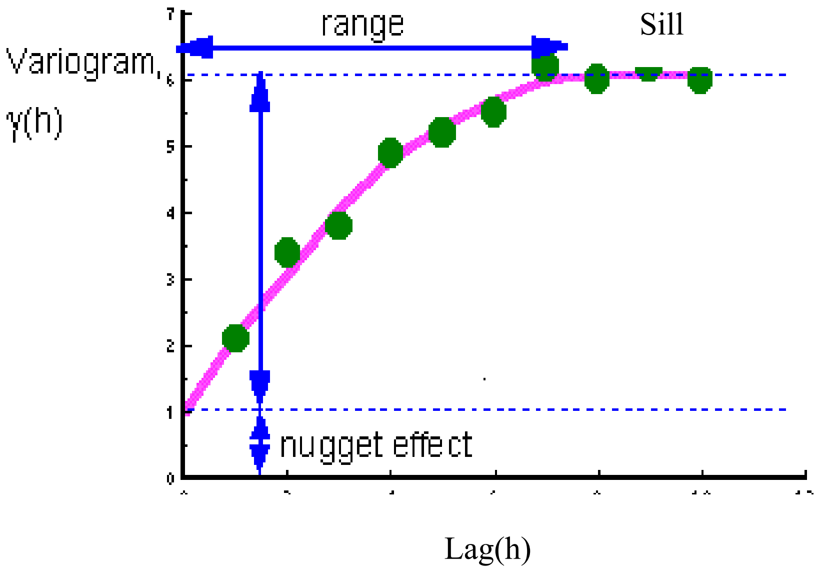
4. Spatial Interpolation

Spatial interpolation describes a process of using points with known values to estimate values at other points i.e it is the procedure of predicting the values of attributes at unsampled sites from measurements made at point locations within the same area or region. Interpolation is used to convert the data from point observations to the continuous fields so that the spatial patterns sampled by these measurements can be compared with the spatial patterns of other spatial entities. Spatial interpolation is thus a means of converting point data into surface data. It is a process of using points with known values to estimate values at other points forming the surface. For example while mapping precipitation if there is no weather reporting station within the grid cell, an estimate is based on nearby weather stations. The rationale behind interpolation is that, on average, values of the attribute are more likely to be similar at points close together than at those further apart.

The word "kriging" is synonymous with spatial interpolation. It is a method of interpolation which predicts unknown values from data observed at known locations. This method uses variogram to express the spatial variation , and it minimizes the error of predicted values which are estimated by spatial distribution of the predicted values. Kriging is also the method that is associated with the acronym B.L.U.E. (best linear unbiased estimator.) It is "linear" since the estimated values are weighted linear combinations of the available data. It is "unbiased" because the mean of error is 0. It is "best" since it aims at minimizing the variance of the errors. The difference of kriging and other linear estimation method is its aim of minimizing the error variance.

4.1 Semi-variogram

Semivariance is a measure of the degree of spatial dependence between samples. The magnitude of the semivariance between points depends on the distance between the points. A smaller distance yields a smaller semivariance and a larger distance results in a larger semivariance. The plot of the semivariances as a function of distance from a point is referred to as a semivariogram. The semivariance increases as the distance increases until at a certain distance away from a point the semivariance will equal the variance around the average value, and will therefore no longer increase, causing a flat region to occur on the semivariogram called a sill. From the point of interest to the distance where the flat region begins is termed the range or span of the regionalized variable. Within this range, locations are related to each other, and all known samples contained in this region, also referred to as the neighborhood, must be considered when estimating the unknown point of interest. Further for h zero, the value of semivariance should strictly be zero but due to several factors, such as sampling error or short scale variability, may cause sample values separated by extremely small distances to be quite dissimilar. This causes a discontinuity at the origin of the variogram. The vertical jump from the values of zero at the origin to the value of the variogram at extremely small separation distances is called the nugget effect. The figure below shows a general semivariogram.



4.2 Ordinary Kriging

Ordinary kriging gives both a prediction and a standard error of prediction at unsampled locations. The aim of kriging is to estimate the value of a random variable z at one or more unsampled points or over large blocks from more or less sparsed data say $z(x_1),..., z(x_N)$ at $x_1,... , x_N$. The data may be distributed in one, two or three dimensions, though applications in the agricultural sciences are usually two-dimensional. It assumes that the mean is unknown, we estimate Z at a point x_0 by $\hat{Z}(x_0)$, with the same support as the data, by

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i)$$

Where λ_i are weights. To ensure that the estimate is unbiased, the weights are made to sum to 1.

$$\sum_{i=1}^{N} \lambda_i = 1 \text{ and the expected error is } E[\hat{Z}(x_0) - Z(x_0)] = 0.$$

The estimation variance is

$$\text{var}[\hat{Z}(x_0)] = E[\{\hat{Z}(x_0) - Z(x_0)\}^2]$$

The kriging equations can be represented in matrix form as

$$A\lambda = b$$

where,

$$\mathbf{A} = \begin{bmatrix} \gamma(x_1,x_1) & \gamma(x_1,x_2) & \cdot\ \cdot\ \cdot & \gamma(x_1,x_N) & 1 \\ \gamma(x_2,x_1) & \gamma(x_2,x_2) & \cdot\ \cdot\ \cdot & \gamma(x_2,x_N) & 1 \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot & \cdot \\ \gamma(x_N,x_1) & \gamma(x_N,x_2) & \cdot\ \cdot\ \cdot & \gamma(x_N,x_N) & 1 \\ 1 & 1 & \cdot\ \cdot\ \cdot & 1 & 0 \end{bmatrix}$$

$$\boldsymbol{\lambda}' = [\lambda_1 \lambda_2 ...... \psi(x_0)] \qquad \text{and} \qquad \mathbf{b}' = [\gamma(x_1,x_0)\gamma(x_2,x_0)...\gamma(x_N,x_0)1]$$

Where, $\gamma(x_i, x_j)$ = semivariance of Z between the data points x_i and x_j and
$\gamma(x_i, x_0)$ = semivariance of Z between the data points x_i and x_0.

The weights and the Lagrange multiplier are obtained as $\lambda = A^{-1} b$
The kriging variance is given by

$$\hat{\sigma}^2(x_0) = b^t \lambda$$

Ordinary kriging is an exact interpolator in the sense that when equation given above is used, the interpolated values, or best local average, will coincide with the values at data points.

5. Spatial Sampling

Spatial sampling is that area of survey sampling which is concerned with sampling in two dimensions like the sampling of fields, groups of contiguous quadrats or other planar surface. Spatial sampling is difficult problem to deal with , since the idea is to select an unbiased sample, but finding independent observations are impossible. One approach to spatial sampling is through a population of MN units, usually points or quadrats, arranged in M rows and N columns. The sampling designs to choose mn units fall into three distinct types: designs in which the sample units are aligned in both the rows and column directions; designs in which the sample units are aligned in one direction only, say the rows, and unaligned in column directions; designs in which the sample units are unaligned in both the directions. For designs that have the sampling units aligned in both the directions, the number of sample elements in any row of the population will be 0 or n and the number of sampled elements in any column of the population will be 0 or m. For designs that have sampled units aligned in the rows and unaligned in column, the number of sample elements in any row of the population will be 0 or n and the number of sampled elements in any column will be at most m. Designs that have sample units unaligned in both the directions are characterized by having at most n sample elements in any row and at most m elements in any column of the population with the exception of simple random sampling without replacement of mn units from the MN in the population. Three traditional sampling designs have generally been applied for selection of the sampling units in different ways such as (a) simple random sample of row/columns, (b) a stratified sample of rows and for each selected row independent stratified sample of columns and (c) a systematic sample unaligned in both the directions.

A second approach to spatial sampling is in a more general population structure, where the spatial population is composed of a number of non-overlapping domains. Without imposing any more structure on the population, three sampling schemes can be considered: random sampling, stratified sampling and systematic sampling.

Dependent Areal Unit Sequential Technique (DUST) is a GIS based sequential technique characterized by variable inclusion probabilities at each step. The principle for sample selection for DUST is that the probability of selection of any unit increases as the distance from the areas already sampled increases. The steps for DUST includes estimation of spatial correlation coefficient (β) for the auxiliary variable x at various spatial lags and stationarity testing at various order spatial correlations followed by sample selection. The first unit is selected randomly out of N units. The subsequent units are selected by applying weight $W_n = \prod_{i=1}^{n-1}\left(1-\beta^{d_{in}}\right)$ for the units selected at the nth draw. β is the spatial correlation for the auxiliary character, n is the sample size and d_{in} is the distance between ith and nth units. Suitable estimators are used for estimation of population parameters.

6. Spatial Regression

Regression is often used in analysis of spatial data to obtain predictive relationships between variables. The assumption that the errors from the regression model are statistically independent will often not be plausible, due to spatial dependence in the sources of error. This is a problem for the regression analysis resulting in estimation of the standard deviation of the errors from the model is biased (downwards) which invalidates confidence limits on predictions made with the model, and which could lead to a false conclusion that the regression is statistically significant. While the estimates of the regression coefficient(s) are not necessarily biased they are not minimum-variance estimates when the errors are correlated.

Regression is used to estimate an equation for predicting a *dependent variable* from values of one or more *independent variables.*. The most useful applications of regression analysis are where the independent variable(s) can be rapidly collected at low unit cost by comparison to the dependent variable. A limited number of costly observations of the dependent variable may then be used to compute the regression equation, which is then applied to predict the dependent variable for all locations where the independent variables are measured. There are many examples of this application of regression analysis. Variables computed from digital elevation models have served as independent variables to predict soil properties, crop yields and air temperatures. Remote sensor data have been used as the independent variables to predict vegetation variables, water quality and forest resources. Regression has been used to predict soil salinity (measured directly by auger sampling and laboratory analysis) from measurements of electromagnetic induction.

7. References

Griffith,D.A. and Amrhein,C.G (1991). Statistical analysis for geographers. Prentice Hall, New Jersey.

McGrew,C. and Monroe,C.B (1993). Statistical problem solving in Geography. Brown Publishers.

Burrough, P.A. and McDonnell, R. A. (1998) Principles of Geographical Information Systems. Oxford University Press.

Cliff, A. D. and Ord, J. K. (1973) Spatial Autocorrelation. Pion London.

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

Journel, A. G. and Huijbregts, CH. J. (1981) Mining Geostatistics. Academic Press.