6. Spatial analysis of multivariate ecological data

6.1 Introduction

6.1.1 Conceptual importance

Ecological models have long assumed, for simplicity, that biological organisms and their controlling variables are distributed in nature in a random or uniform way. This assumption is actually quite remote from reality: field biologists know from experience that neither the variables they use to describe the environment nor the living beings themselves are distributed uniformly or at random. The environment can be considered as primarily structured by broad-scale physical processes (geomorphology on land, currents and winds in fluids), that generate gradients and/or patchy structures separated by discontinuities (interfaces). These structures induce similar responses in biological systems. Furthermore, even in zones that appear homogeneous at a given spatial scale, finer-scale contagious abiotic or biotic processes take place, generating more spatial structuring through reproduction and death, predator-prey interactions, food availability, parasitism, and so on.

Thus one can see that spatial heterogeneity is *functional* in ecosystems. It is not the result of some random, noise-generating process. Therefore, it is important to study it for its own sake. Ecosystems without spatial structuring would be unlikely to function. Imagine the consequences: large-scale homogeneity would cut down on diversity of habitats, feeders would not be found close to their food, mates would be located at random throughout the landscape, newborns would be spread around instead of remaining in favorable environments... Irrealistic as it may seem, this view is still present in several of our theories and models describing population and community functioning.

Spatial organization of ecosystems has thus to be incorporated in theories, otherwise these will be suboptimal. In general terms, more

and more theories admit that the elements of an ecosystem that are close to one another in space or time are more likely to be influenced by the same generating processes. Such is the case, for instance, for the theories of competition, succession, evolution and adaptation (historic autocorrelation), maintenance of species diversity, parasitism, population genetics, population growth, predator-prey interactions, and social behaviour.

6.1.2 Importance in sampling strategy

The very fact that every living community is spatially structured has its consequences on the sampling strategies. One should be aware that the sampling strategy strongly influences the perception of the spatial structure of the sampled population or community. For instance, in a site where the variables to be sampled are structured in more or less regular patches, systematic sampling could lead to completely altered estimations of the spatial structure if the intersample distance is larger than one half of the inter-patch distance (Figure 35, see below).

This example may seem trivial especially to botanists....but botanists have a major advantage in that they see what they sample. This is not the case in, say, ecology of aquatic or soil organisms. When one samples a soil community following a systematic pattern, for example, it is quite possible that a part of the sampled species distributions will be estimated correctly, whereas some other will be totally misinterpreted. Thus, when one aims to study the spatial distribution of organisms, a random-based sampling strategy seems preferable, in that it allows various and unrelated inter-sample distances to be sampled. Even when mapping is planned, it is always possible to estimate a regular grid of values on the basis of a randomly placed set of measurements.

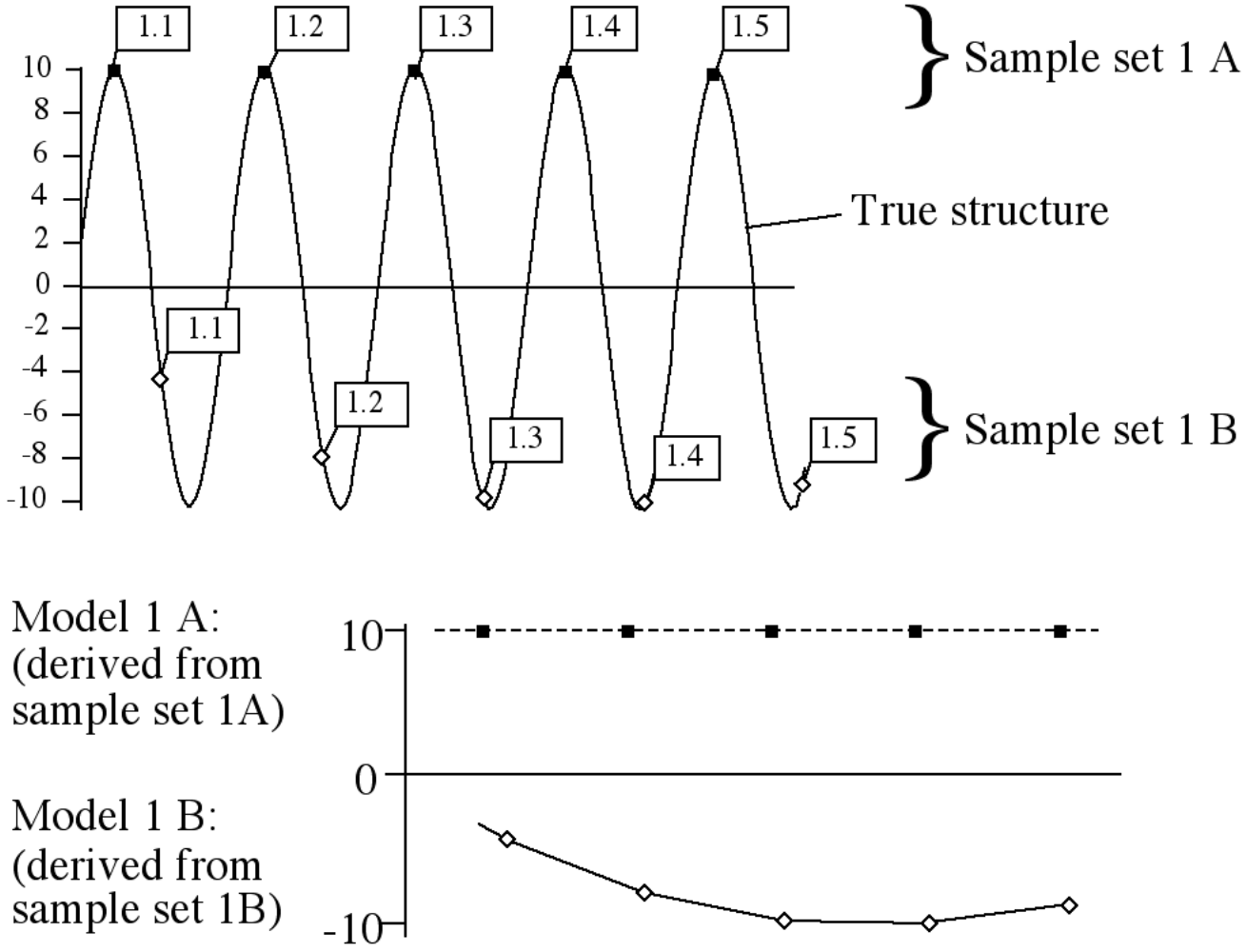

Figure 35 - The danger of systematic sampling

6.1.3 Importance in statistics

When the variables to be sampled are spatially structured, one of the most fundamental assumptions of the classical statistics, that is, the assumption of independance of the observations (or, more precisely, of residuals), is violated. In a situation where no spatial structure is present, the resemblance patterns between all pairs of sample units are independant of the geographical distances between the sample units. In other words, one cannot predict the value of a variable (or the species composition of a community sample unit) on the basis of a few other, neighbouring units. On the contrary, when a spatial structure exists

and one knows its general shape (gradient, patches...), one can predict at least roughly the content of a sample unit on the basis of the other ones on the basis of their locations. Such a sample set (or its variables) is said to be **spatially autocorrelated** (Figure 36).

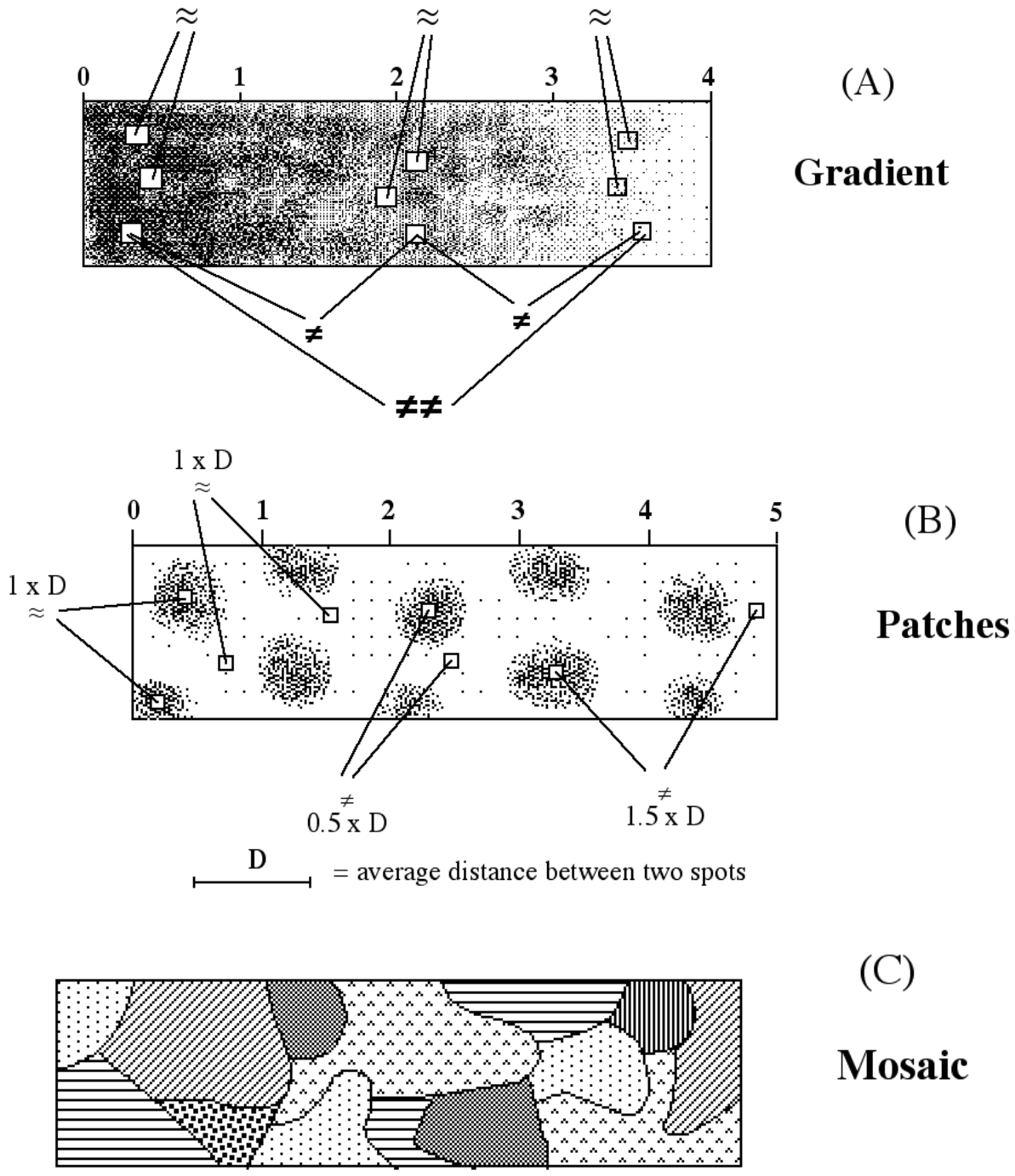

Figure 36 - Three types of spatial structures.

Due to the violation of the assumption of independance of the observations, it is not possible to perform standard statistical tests of hypotheses on spatially autocorrelated data. In most of the natural cases, the data are positively autocorrelated at short distances, which means that any two close sample units resemble more each other than predicted by a random (uncorrelated) structure. In such cases, for instance, the classical statistical procedures estimate a confidence interval around a Pearson correlation coefficient narrower than it is in reality, so that one declares too often that the coefficient is different from zero (inflated type I error, Figure 37):

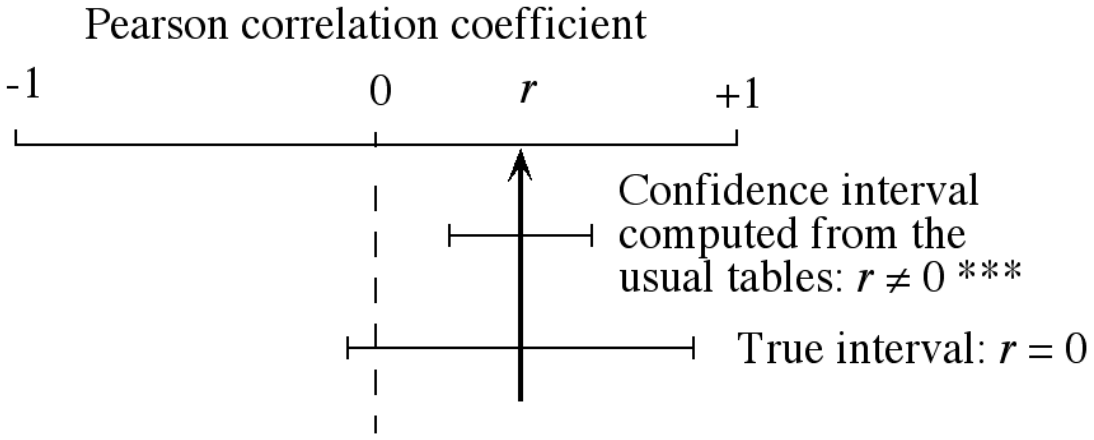

Figure 37 - Underestimation of the confidence interval around a Pearson *r* correlation coefficient in the presence of autocorrelation.

One could understand this from the point of view of the degrees of freedom: normally one counts one degree of freedom for each independant observation, and this allows one to choose the appropriate statistical distribution for the given test. Now, if the observations are not independant but rather autocorrelated, each new observation does not bring a full degree of freedom, but only a fraction. It follows that the total actual number of degrees of freedom is smaller than estimated by the classical procedures. Thus the consequence: for a given total variance, the smaller the number of d.f., the broader the actual confidence interval.

In limited cases there is now possible to correct for autocorrelation when estimating the number of d.f. But this is by no means an easy task.

6.1.4 Importance in data interpretation and modelling

As said previously, the spatial structuration of the living communities and their controlling environmental factors is functional. Thus, it is important to include this structuration into the theories, into the data analyses, and into the models.

Spatial structure in the data has many consequences on the interpretation of analytical results. For instance, spatial structuration that is shared by response and explanatory variables can induce spurious correlations, leading uncorrect causal models to be accepted. Partial analyses may allow to avoid this pitfall. On the other hand, proper handling of spatial descriptors allows to explain the data variation in a more detailed way, by disciminating between environmental, spatial, and mixed relationships.

So, by taking the spatial structure into account when analyzing multivariate data sets, it is often possible to elucidate more ecological relationships, avoid misinterpretations, and explain more data variation. This insight allows one to build more realistic models.

Last but not least, spatial structures can be mapped. Mapping is not only a nice tool (or toy!) for illustration, but also a powerful way of exploring data structure and generating new hypotheses, especially when the data contain a significant amount of spatial structure that is not related with the measured environmental variables.

This chapter addresses the topics of some measures of autocorrelation, the components of spatial structure, and some techniques of spatial modeling.

6. 2 The measure of spatial autocorrelation

The introduction above makes clear that there are multiple reasons to test in a unambiguous way if the data are autocorrelated. One can run such tests either to verify that there is *no* spatial structure, and use parametric tests afterwards, or, on the contrary, to confirm the presence of a spatial structure to be able to study it in more detail.

Such tests are built on **coefficients of spatial autocorrelation**, that have the double advantage to test for spatial structures and provide a simple description of it. Here we will first expose the tests for univariate data (Moran's *I* and Geary's *c*), and then the Mantel correlogram for multivariate data.

6.2.1 One singe variable: intuitive introduction

6.2.1.1 One spatial dimension

Imagine that you have collected a series of measures of bulk density of a soil along a transect. You can construct a graph associating the position of the sites and the measures of density (Figure 38):

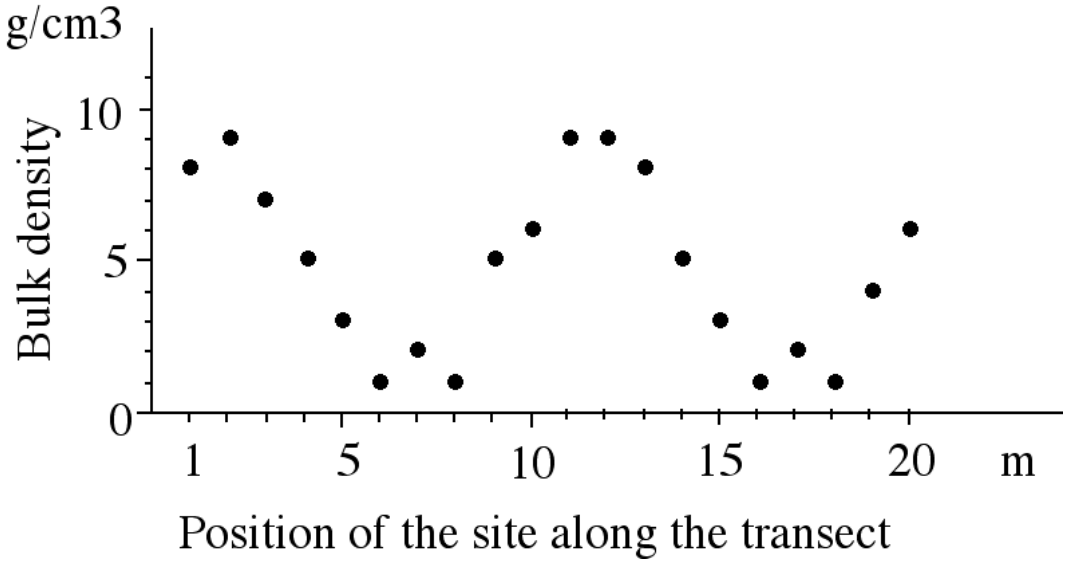

Figure 38 - Substratum density along a transect (fictitious data)

In these data, one can clearly see a periodic structure. In such a case, one can predict at least approximately the value at one site on the basis of the values nearby. The series is thus **autocorrelated**. An easy way of studying such a series is to correlate it with itself (*auto-correlating it!*) several times, introducing a shift (Table XII):

Table XII - Correlation of a data series with itself: auto-correlation. In this example, for demonstration purposes only, Pearson's *r* correlation coefficient is used for simplicity.

Lag 0 : 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6	Pearson's *r between the 2 series* = 1
Lag 1 :(8)9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4(6)	r = 0.75
Lag 2 :(8 9)7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2 1(4 6)	r = 0.33
Lag 3 :(8 9 7)5 3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1 2(1 4 6)	r = – 0.25
Lag 4 :(8 9 7 5)3 1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3 1(2 1 4 6)	r = – 0.73
Lag 5 :(8 9 7 5 3)1 2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5 3(1 2 1 4 6)	r = – 0.95
Lag 6 :(8 9 7 5 3 1)2 1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8 5(3 1 2 1 4 6)	r = – 0.81
Lag 7 :(8 9 7 5 3 1 2)1 5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9 8(5 3 1 2 1 4 6)	r = – 0.36
Lag 8 :(8 9 7 5 3 1 2 1)5 6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9 9(8 5 3 1...)	r = 0.25
Lag 9 :(8 9 7 5 3 1 2 1 5)6 9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6 9(9 8 5 3...)	r = 0.69
Lag 10:(8 9 7 5 3 1 2 1 5 6)9 9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5 6(9 9 8 5...)	r = 0.99
Lag 11:(8 9 7 5 3 1 2 1 5 6 9)9 8 5 3 1 2 1 4 6 8 9 7 5 3 1 2 1 5(6 9 9 8...)	r = 0.82

```
Lag 12:(8 9 7 5 3 1 2 1 5 6 9 9)8 5 3 1 2 1 4 6
                                8 9 7 5 3 1 2 1(5 6 9 9...)      r = 0.38

Lag 13:(8 9 7 5 3 1 2 1 5 6 9 9 8)5 3 1 2 1 4 6
                                  8 9 7 5 3 1 2(1 5 6 9...)     r = – 0.19

Lag 14:(8 9 7 5 3 1 2 1 5 6 9 9 8 5)3 1 2 1 4 6
                                    8 9 7 5 3 1(2 1 5 6...)     r = – 0.79

Lag 15:(8 9 7 5 3 1 2 1 5 6 9 9 8 5 3)1 2 1 4 6
                                      8 9 7 5 3(1 2 1 5...)     r = – 0.89
```

... and we stop at lag 15, because there are not enough pairs of values left.

The result of such an analysis is generally presented in the form of a **correlogram**, where the abscissa represents the lag, and the ordinate are the correlation values (Figure 39):

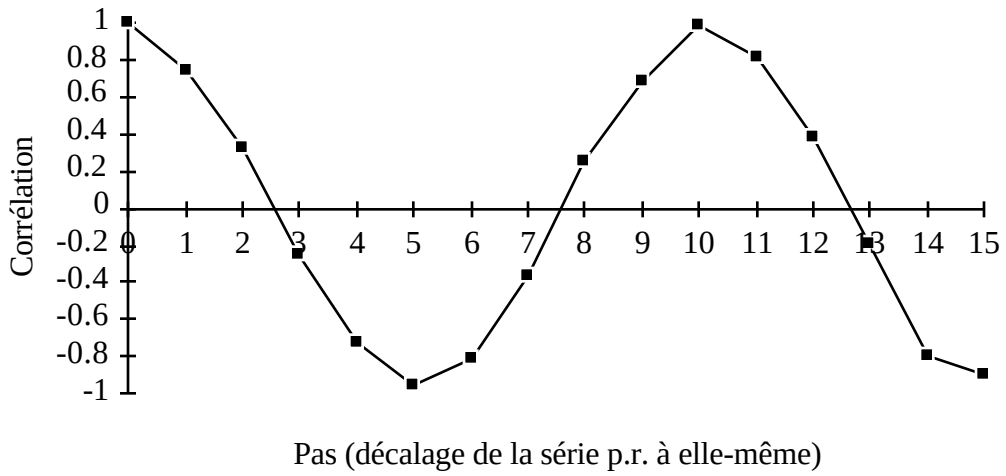

Figure 39 - Correlogram of the fictitious data of Fig. 38. **Beware**: the lag ("pas") does not represent spatial coordinates, but reciprocal distances among sites! For instance, at lag 5, the $r = –0.95$ means that any pair of sites whose members are 5 units apart (for instance sites 3 and 8, or 10 and 15...) probably have very different values of density: one of the sites has a high value, the other a low one.

6.2.1.2 Two spatial dimensions (surface)

The next step of our reasoning is to extend it to a surface, in the case of an isotropic process (the spatial pattern is more or less identical in all directions). In this case, one can no more organize the data in a linear series. Rather, one computes a matrix of Euclidean (geographical) distances among all pairs of sites, and then the distances are grouped in classes. For instance, all distances shorter than 1 meter belong to class 1, the distances between 1 and 2 meters are put in class 2, and so on. Table XIII gives an example:

Table XIII - Construction of a matrix of classes of distance

	2	3	4	5	6		2	3	4	5	6
1	0.23	2.82	1.65	0.89	1.23		1	3	2	1	2
2		1.45	2.44	0.32	1.87			2	3	1	2
3			3.56	0.09	2.11	->			4	1	3
4				2.70	1.15					3	2
5					1.34						2
	Matrix of Euclidean distances						Matrix of distance classes				

On this basis (but with more than 6 sites!) one could compute autocorrelation indices for the 4 classes of geographical distances. For instance, for class 1 the value would be computed on the basis of pairs 1-2, 1-5, 2-5 and 3-5. And so on. For n objects, one always has $n(n-1)/2$ distances, that should be grouped into an appropriate number of classes: neither too high (too few values in each class) nor too low (to avoid the analysis to be too coarse). Sturge's rule is often used to decide how many classes are appropriate:

Nb of classes = $1 + 3.3 \log_{10}(m)$ (rounded to the nearest integer)

where m is, in this case, the number of distances in the upper triangular matrix of distances (excluding the diagonal).

6.2.2 Indices of spatial autocorrelation: Moran's I *and Geary's* c

In practice, the two mostly used indices are Moran's *I*, which behaves approximately like a Pearson correlation coefficient, and Geary's *c*, which is a sort of distance.

Moran's *I* is computed as follows (for distance class *d*):

$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \bar{y})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i$$

and Geary's *c*:

$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i$$

y_h and y_i are the values of the variable in objects *h* and *i*; *n* is the number of objects; the weights *w* are equal to 1 when pair (*h*,*i*) belongs to distance class *d* (the one for which the index is computed) and 0 otherwise. *W* is the sum of all w_{hi}, i.e. the number of pairs of objects belonging to distance class *d*.

Moran's *I* generally varies between –1 and +1, although values beyond these limits are not impossible. Positive autocorrelation yields positive values of *I*, and negative autocorrelation produces negative values. Moran's *I* expected value in the absence of autocorrelation is equal to $E(I) = -(n-1)^{-1}$ i.e., close to 0 when *n* is large.

Geary's *c* varies from 0 to some unspecified value larger than 1. Positive autocorrelation translates into values from 0 to 1, while negative autocorrelation yields values larger than 1. Geary's *c* expectation in the absence of autocorrelation is equal to $E(c) = 1$.

Example

The 8 x 8 grid below (Table XIV and Figure 40) is a fictitious data set where the value of a variable *z* has been measured on a regular sampling design. One can see a gradient in the data.

Table XIV - Fictitious, spatially referenced sample set. 64 sites.

11	10	9	7	7	6	4	2
9	11	7	6	5	3	2	2
10	9	6	10	8	4	5	3
8	9	7	5	4	3	3	2
7	6	5	6	4	4	3	2
5	5	5	4	5	3	1	3
5	4	3	2	3	3	2	2
3	4	2	2	1	3	1	1

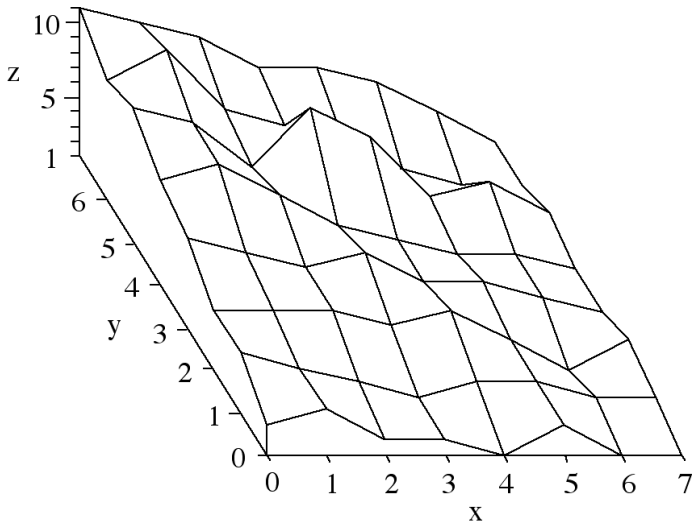

Figure 40 - Grid map of the data given in Table IX.

Each site is characterized by its x - y spatial coordinates and the value of the measured variable, *z*. In this example, let us state that the horizontal and vertical intersite distance is equal to 1 m. The computation of a correlogram involves 4 steps:

1. Computation of a matrix of Euclidean distances among sites.
2. Transformation of these distances into classes of distances.
3. Computation of Moran's *I* or Geary's *c* for all distance classes.
4. Drawing of the correlograms.

The hypotheses of the tests run on each distance class can be worded as follows:

H_0: there is no spatial autocorrelation. The values of variable *z* are spatially independent. Each Moran's *I* or Geary's *c* value is close to its expectation.

H_1: there is significant spatial autocorrelation in the data. At least one autocorrelation value is significant at the Bonferroni-corrected level of significance (see below).

The result below (with 6 equidistant classes) has been obtained using the R package for multivariate data analysis by Pierre Legendre, Philippe Casgrain and Alain Vaudor. This package, not to be confused with the R language, is available for Macintosh (Classic environment) from Pierre Legendre's web site:

http://www.bio.umontreal.ca/legendre/

```
Classes équidistantes

    Classe  Limite sup.   Fréq.
         1   1.64992       210
         2   3.29983       556
         3   4.94975       442
         4   6.59967       560
         5   8.24958       218
         6   9.89950        30
```

```
FICHIER DE DONNEES grad.8x8.r
Option du mouvement:   Matrice SIMIL
Note: les probabilités sont plus significatives près de zéro
      les probabilités sont à plus ou moins;   0.00100
```

H0:		I = 0	I = 0		C = 1	C = 1	
H1:		I > 0	I < 0		C < 1	C > 1	
Dist.,	I(Moran),	p(H0),	p(H0),	C(Geary),	p(H0),	p(H0),	Card.*
1	0.6723	0.000		0.2366	0.000		420
2	0.3995	0.000		0.4601	0.000		1112
3	0.0225	0.177		0.8660	0.008		884
4	-0.3678		0.000	1.4245		0.000	1120
5	-0.7674		0.000	2.0580		0.000	436
6	-1.0667		0.000	2.7131		0.000	60
Total							4032

***** Card. means "cardinality", i.e. the number of pairs of observations in each distance class, in a *square* distance matrix, diagonal excluded.

The following correlograms can be drawn from the results above:

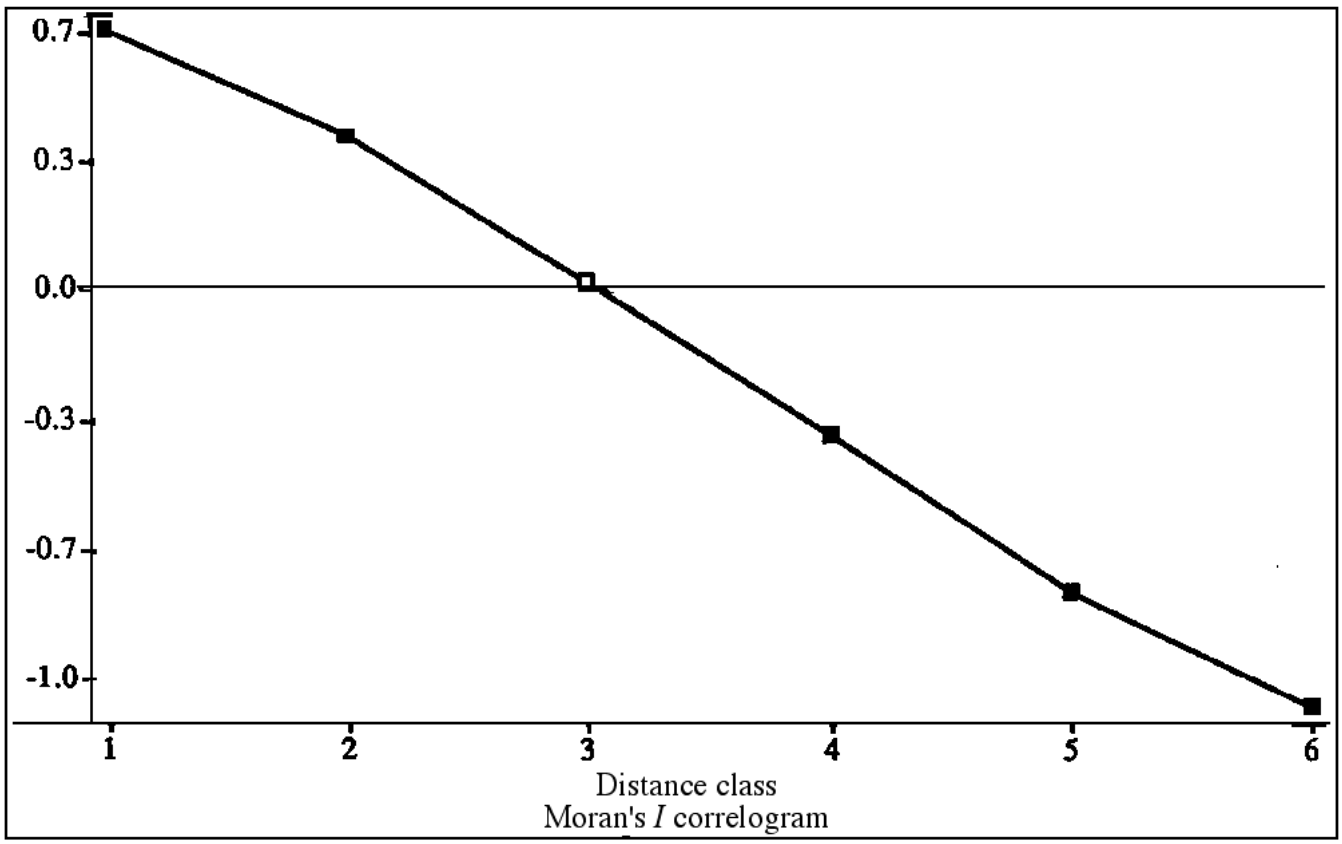

Figure 41 - Moran's *I* correlogram, data from Table XIV.

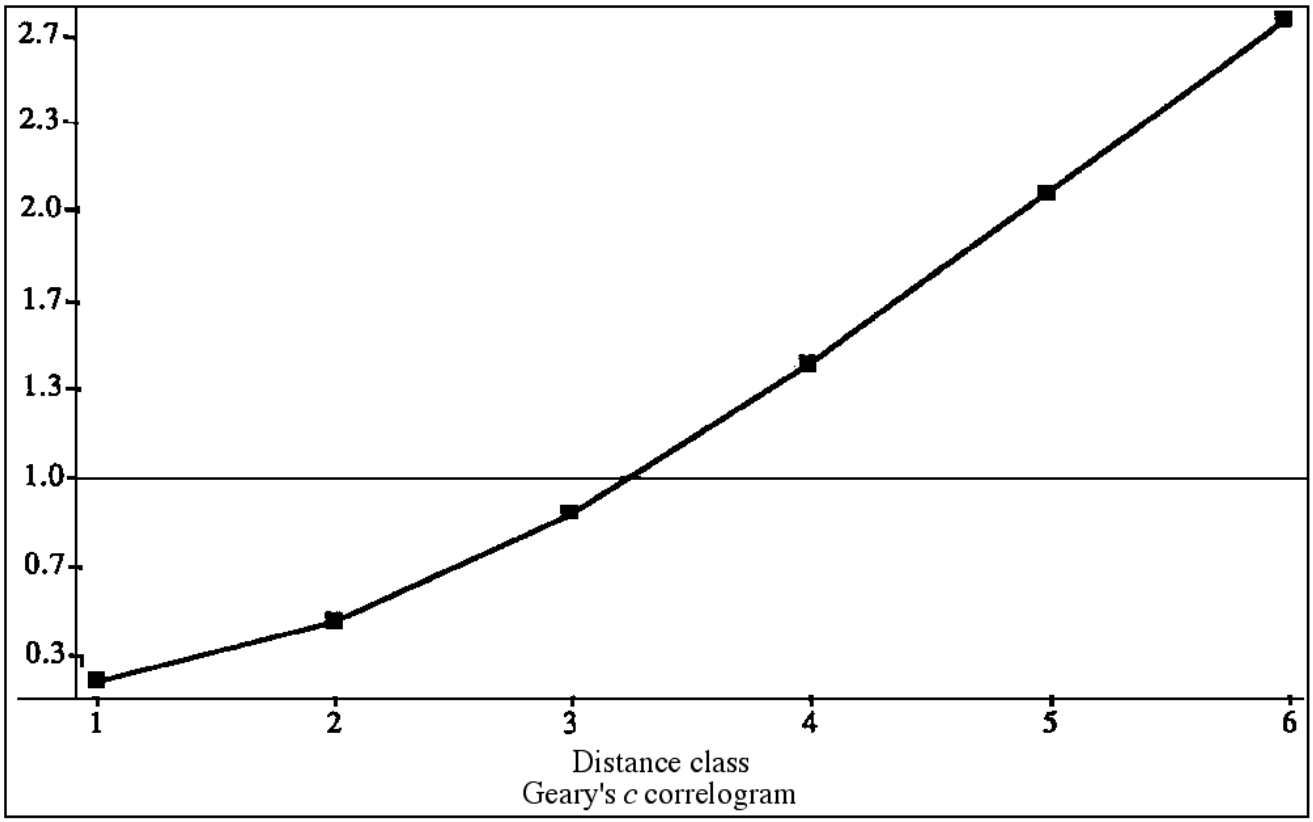

Figure 42 - Geary's *c* correlogram, data from Table XIV.

In these graphs, the black squares are significant values (at an uncorrected α level of 0.05, but see below!) and white squares non significant ones. The output file gives the exact probabilities.

The opposite aspect of the curves illustrates well the fact that Moran's *I* behaves like a correlation (positive values mean positive autocorrelation) while Geary's *c* behaves like a distance (lowest value for highest positive autocorrelation).

6.2.3 Correction for multiple testing

One problem remains, related to the α probability level of rejection of the null hypothesis. Such a threshold is defined for **one** test. In our case, since we computed six autocorrelation values (one for each distance class), **six** tests have been performed simultaneously. In such

a case, the type I error (rejecting H_0 when it is true) is no more 0.05 (if this was the selected level). For one test, a threshold of 5% means that, on average, among 100 tests run on completely random variables, 5 will (wrongly) reject H_0 and 95 will not. Cumulating 6 tests at the 5% level means that the chances of rejecting H_0 at least once is equal to $1 - 0.95^6 = 0.265$ instead of 0.05 ! The most drastic (conservative) remedy to this situation is called the **Bonferroni correction**, which consists in **dividing the global threshold level by the number of simultaneous tests.** In our case, for each correlogram, this means that, **globally**, the correlogram will be declared significant if at least one autocorrelation value is significant at the corrected α level of 0.05/6 = 0.0083. One can verify that, with this correction, the global chances of accepting H_0 when it is true is equal to $(1 - 0.0083)^6 \approx 0.95$.

This correction is very conservative, however, and can lead to type II error (accepting H_0 when it is false) when the tests involved in the multiple comparison are not independent, i.e. when they address a series of related questions and data. Several alternatives have been proposed in the literature. An interesting one is the **Holm correction**. It consists in (1) running all the *k* tests, (2) ordering the probabilities in increasing order, and (3) correcting each α significance level by dividing it by (1 + the number of remaining tests) in the ordered series. In this way, the first level is divided by *k*, the second one by $k - 1$, and so on. The procedure stops when a non-significant value is encountered.

6.2.4 Multidimensional data: the Mantel correlogram

If one wants to explore the autocorrelation structure of a multivariate data set (for instance a matrix of species abundances), on can resort to the Mantel correlogram. This technique is based on the Mantel statistic. Each class of the matrix of distance classes is represented by a binary (0 - 1) matrix where the pairs of objects belonging to that class receive the value 1 and the others 0. The tests are permutational (as in

the case of the Mantel test). If the standardized Mantel statistic is used, the values are ranged between -1 and +1 and the interpretation of the Mantel correlogram is similar to that of the Moran correlogram, bearing in mind that the expectation of the Mantel statistic under H_0 is strictly 0.

6.2.5 Remarks

1. Many other structure functions are available to describe and model spatial structures. A very important one is the **semi-variogram**, which uses a measure of variance among values of sites belonging to various distance classes to build a model that can be used either to describe the spatial structure (as in the case of the correlogram) or to model it, especially for mapping and prediction purposes (as in the case of kriging).

2. Unless otherwise specified, these methods consider that the spatial structures are isotropic (same in all directions). But anisotropy can be addressed in several ways, for instance by modifying the matrix of distance classes.

3. When statistical tests are run to identify spatial structures, they require that the condition of **second-order stationarity** be satisfied. This condition states that the expected value (mean) and spatial covariance (numerator of Moran's I) of the variable are the same all over the study area, and the variance (denominator) is finite. A relaxed form of stationarity hypothesis, the **intrinsic assumption**, states that the differences $(y_h - y_i)$ for any distance must have zero mean and constant and finite variance over the study area.