Analysis of Multivariate Ecological Data

School on Recent Advances in Analysis of Multivariate Ecological Data

24-28 October 2016

The Abdus Salam International Centre for Theoretical Physics

Prof. Pierre Legendre
Dr. Daniel Borcard
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre Ville
Montréal QC
H3C 3J7 Canada
pierre.legendre@umontreal.ca
daniel.borcard@umontreal.ca

Day 4

Beta diversity

→ see course material by Pierre Legendre

Day 4

Origin of spatial structures

Spatial analysis of multivariate ecological data

1. Introduction

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.

The environment can be considered as primarily structured by broad-scale physical processes that generate gradients and/or patchy structures separated by discontinuities.

These structures induce similar responses in biological systems.

Spatial analysis of multivariate ecological data

1. **Introduction**

1.1 Conceptual importance

Spatial heterogeneity is *functional* in ecosystems, and 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.

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

Spatial analysis of multivariate ecological data

1. Introduction

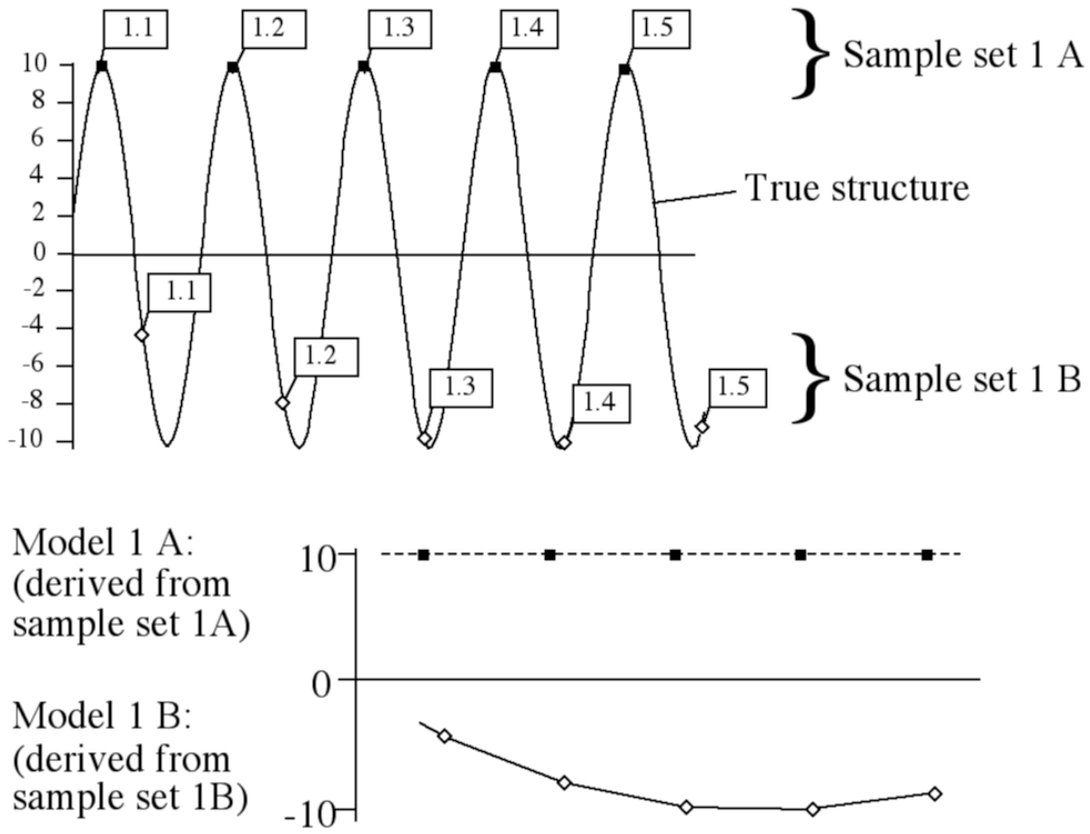
1.2 Importance in sampling strategy

Sampling strategy strongly influences the perception of the spatial structure of the sampled population or community.

Example: **systematic sampling**

Spatial analysis of multivariate ecological data

1.2 Importance in sampling strategy



Spatial analysis of multivariate ecological data

1. Introduction

1.3 Importance in statistics

Spatial structure in data → data points not independent!

Nonindependence of observations often leads to nonindependence of residuals of a regression model.

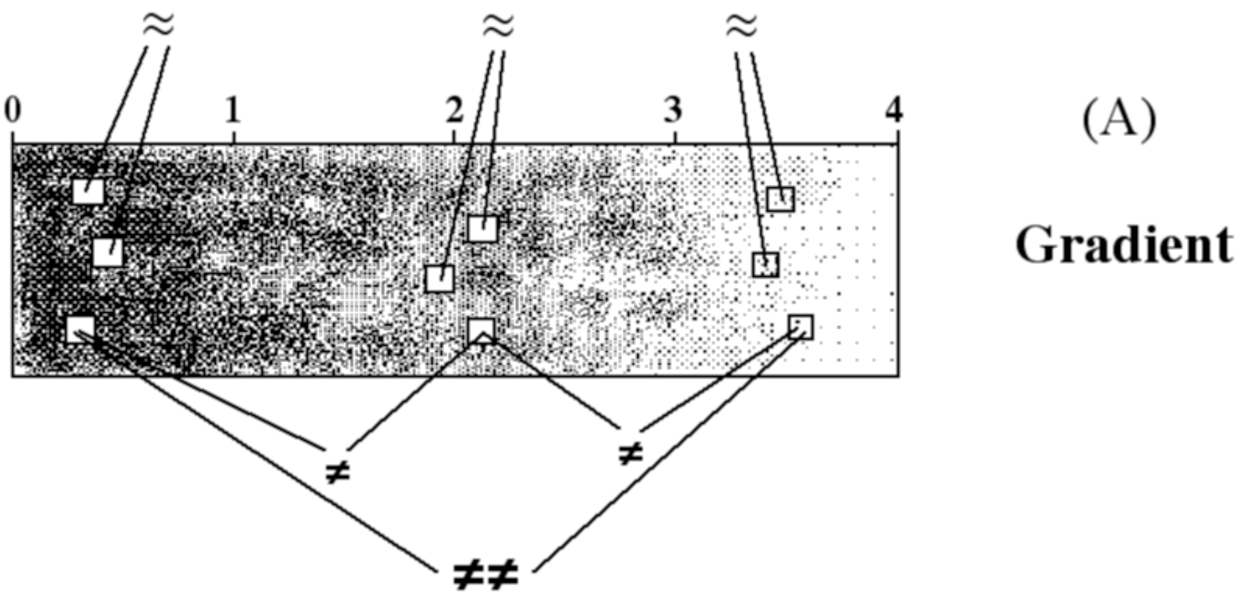
Most fundamental assumption of classical statistics violated!

Spatial structure in data → data **autocorrelated**

Spatial analysis of multivariate ecological data

1. Introduction

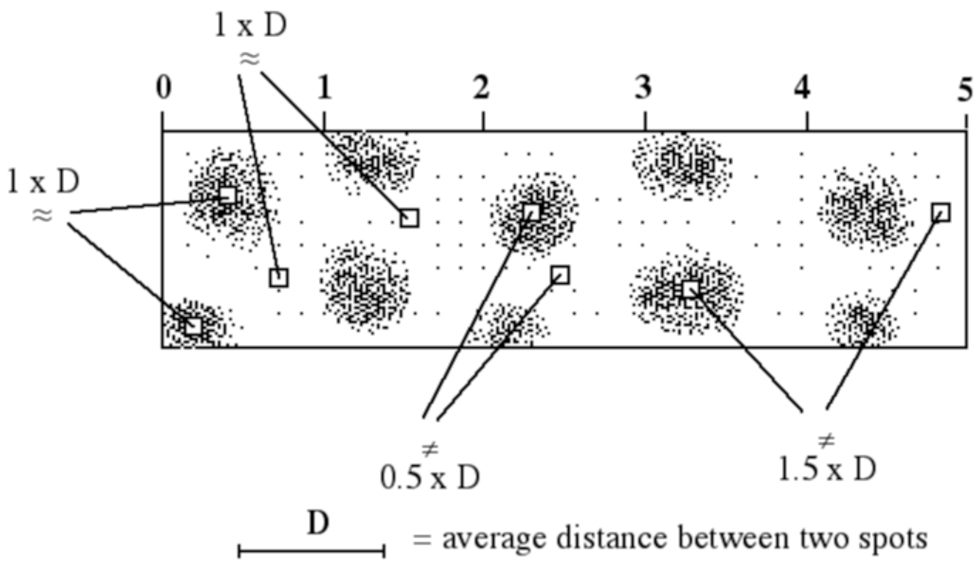
1.3 Importance in *statistics*



Spatial analysis of multivariate ecological data

1. Introduction

1.3 Importance in statistics



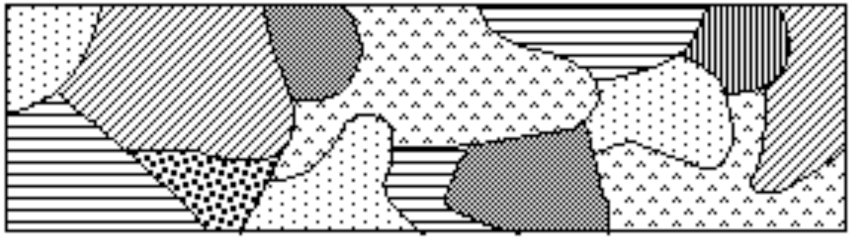
(B)

Patches

Spatial analysis of multivariate ecological data

1. Introduction

1.3 Importance in statistics



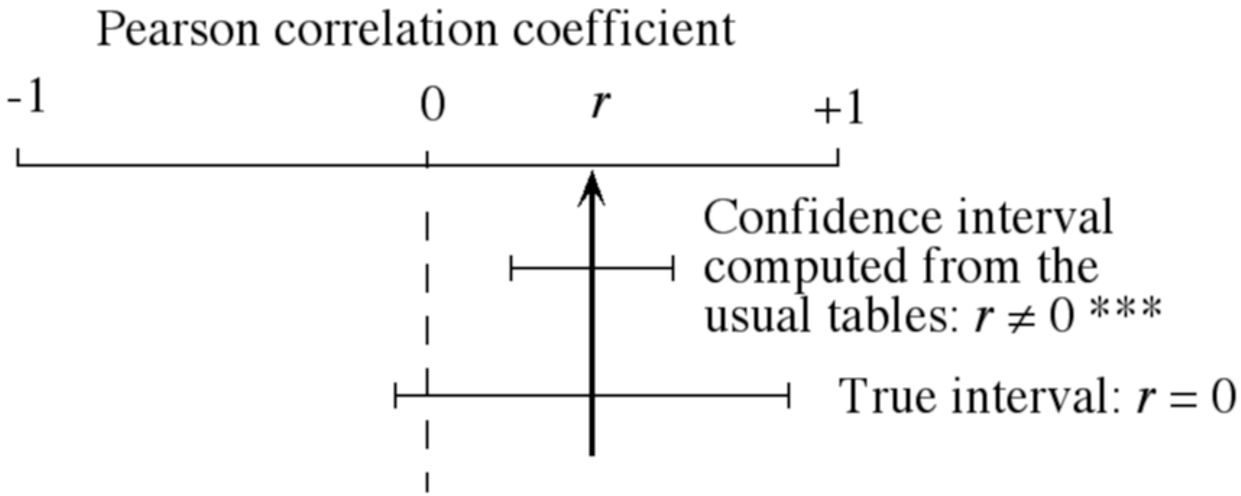
(C)

Mosaic

Spatial analysis of multivariate ecological data

1.3 Importance in statistics

Consequence (example): underestimation of the confidence interval around a Pearson *r* correlation coefficient in the presence of autocorrelation.



Spatial analysis of multivariate ecological data

1.4 Importance in data interpretation and modelling

Spatial structuration that is shared by response and explanatory variables can induce **spurious correlations**, leading uncorrect causal models to be accepted.

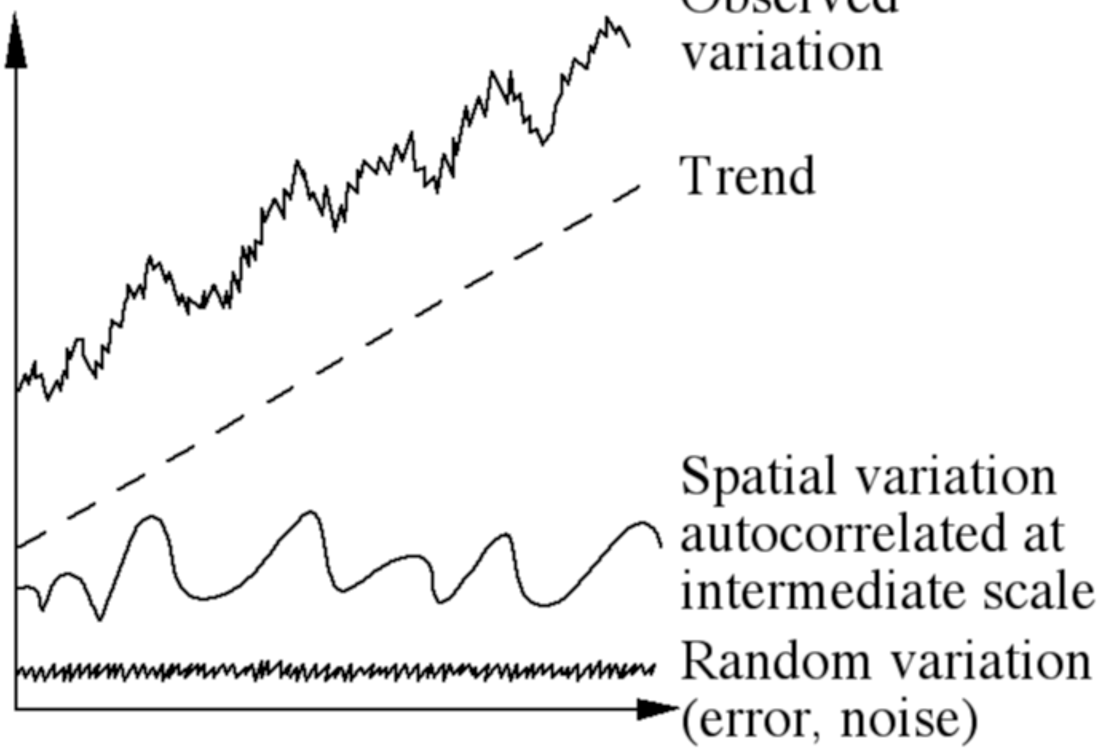
Proper handling of spatial descriptors allows to **explain the data variation in a more detailed way**, by disciminating between environmental, spatial, and mixed relationships.

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**.

Spatial analysis of multivariate ecological data

3. Modeling spatial structures

3.1 Introduction: the 3 components of spatial structure



Mantel correlogram

Mantel correlogram

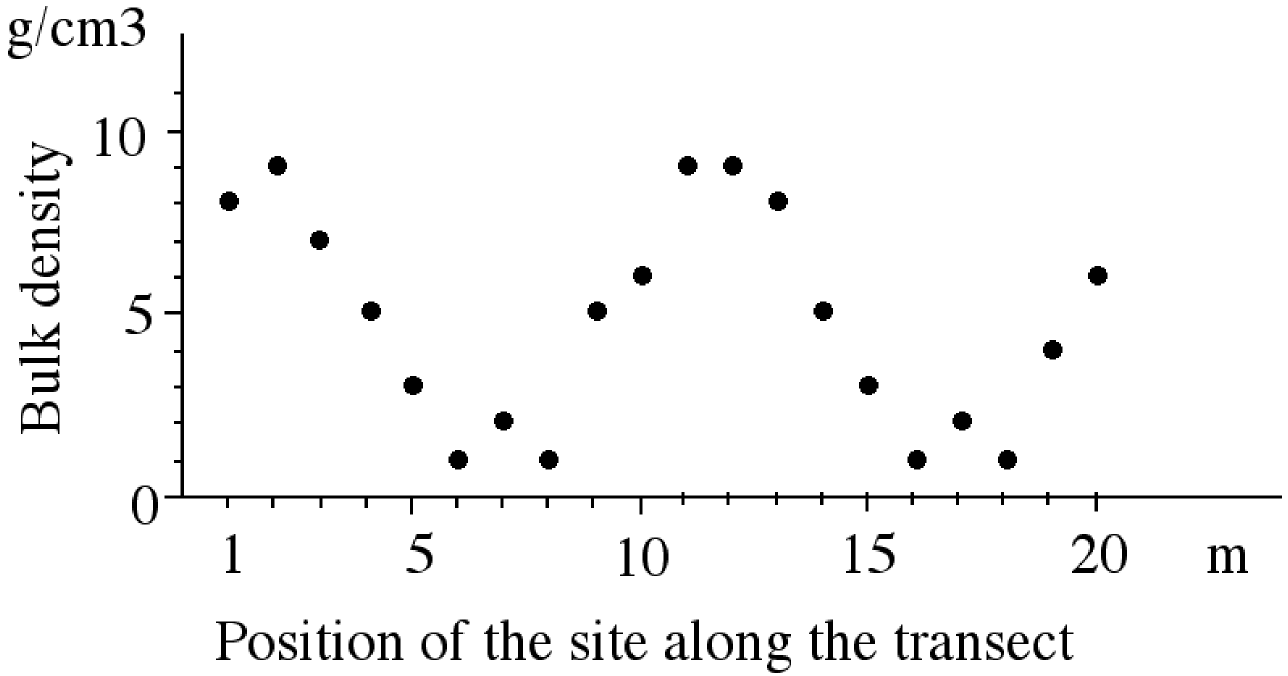
→A **correlogram** is a graphical representation of spatial or temporal correlations between sites, computed for a range of classes of geographical distances.

→These spatial or temporal correlations measure how much the sites resemble their neighbours of increasing spatial (or temporal) distances.

… how does it work?

Correlogram: an intuitive example

1. The data – one dimension



Substratum density along a transect (fictitious data)

Correlogram: an intuitive example

2. Correlation of progressively shifted series

```
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 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...)
```

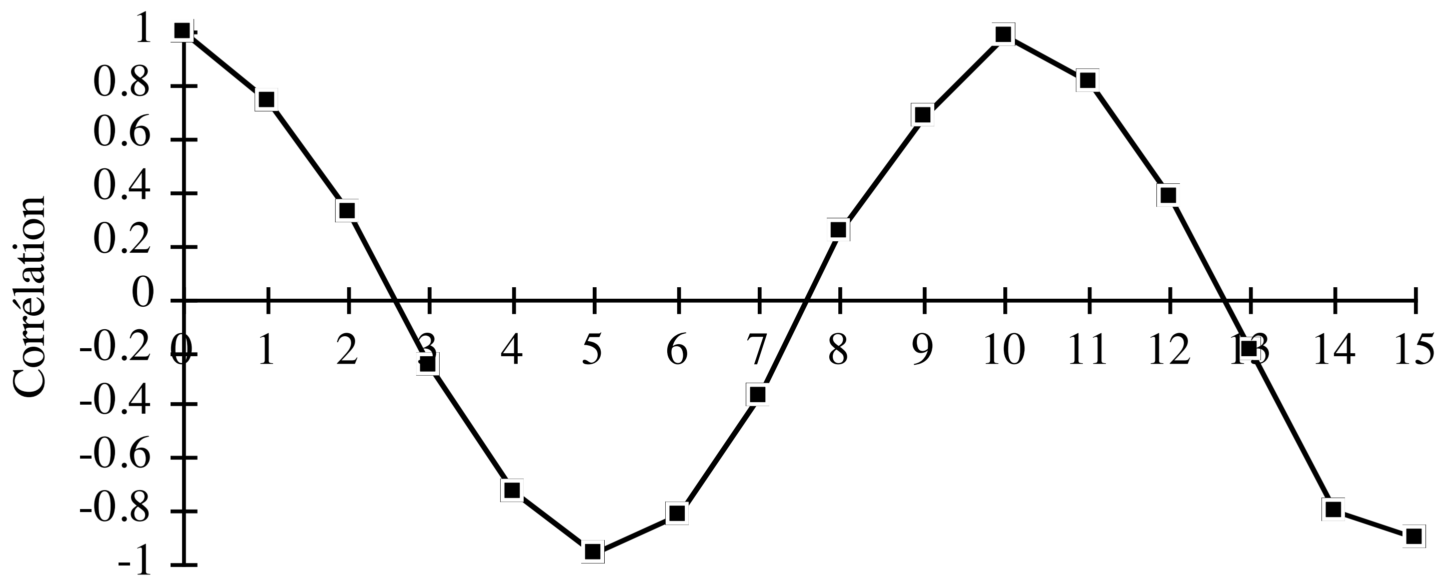
$r = -0.79$

```
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

Correlogram: an intuitive example

3. Correlogram of the fictitious data series



Lag = shift of the series with respect to itself

Correlogram: an intuitive example

4. In two dimensions

1. Matrix of Euclidean (geographical) distances among all pairs of sites.
2. The distances are grouped in classes

	2	3	4	5	6
1	0.23	2.82	1.65	0.89	1.23
2		1.45	2.44	0.32	1.87
3			3.56	0.09	2.11
4				2.70	1.15
5					1.34

Matrix of Euclidean distances

->

2	3	4	5	6
1	3	2	1	2
	2	3	1	2
		4	1	3
			3	2
				2

Matrix of distance classes

Correlogram: an intuitive example

5. The measure of univariate spatial autocorrelation

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

1. Moran's *I*

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

Close to Pearson's r correlation

2. Geary's *c*

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

Close to a measure of distance

Mantel correlogram

For **multivariate data**, the solution is to compute a **Mantel correlogram**.

… but first, let us see what a Mantel test is.

Mantel correlogram

1. Mantel test: matrix correlation

Linear correlation between similarity or dissimilarity matrices.

Formally, the hypotheses of the Mantel test can be stated as follows:

H_0: the dissimilarities (or similarities) among objects in matrix **Y** are not (linearly) correlated with the corresponding dissimilarities in matrix **X**.

H_1: the dissimilarities among objects in matrix **Y** are linearly correlated to the dissimilarities in **X**.

Mantel correlogram

1. Mantel test: matrix correlation

The original **Mantel *z* statistic**, i.e. the measure used to evaluate the resemblance between the two matrices, is:

$$z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij} y_{ij}$$

where *i* and *j* are row and column indices of the similarity or dissimilarity matrices.

Mantel correlogram

1. Mantel test: matrix correlation

Example:

	2	3	4
1	0.25	0.43	0.55
2		0.17	0.39
3			0.66

Dissim. among vegetal communities

	2	3	4
1	1	0	1
2		0	1
3			0

First geogr. distance class

$z = (0.25 \times 1) + (0.43 \times 0) + (0.55 \times 1) + (0.17 \times 0) + (0.39 \times 1) + (0.66 \times 0) = \mathbf{1.19}$

Mantel correlogram

1. Mantel test: matrix correlation

Example:

This value (1.19) is the "true" (observed) value.

It is compared to a reference distribution obtained by randomly **permuting** (99 or 999 or 9999 times) **the rows and corresponding columns** of one of the two (dis)similarity matrices.

Beware: the values of the dissimilarity matrices cannot be permuted completely at random. The permutation scheme is actually **equivalent to permuting the raw data** and recomputing the dissimilarities.

Mantel correlogram

1. Mantel test: matrix correlation

Standardized Mantel *r* statistic: same formula as that of Pearson's *r* correlation coefficient:

$$r_M = \frac{1}{d-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{x_{ij}-\bar{x}}{s_x}\right)\left(\frac{y_{ij}-\bar{y}}{s_y}\right)$$

where *i* and *j* are as above, *x*-bar, *y*-bar, s_x and s_y are the means and standard deviations of the dissimilarity values of each matrix, and $d = [n(n–1)/2)$ is the number of dissimilarity or similarity measures in one of the upper triangular matrices.

Mantel correlogram

Words of caution (Mantel test and distance approach)

Legendre et al. (2005) :

(1) The variance of a community composition table is a measure of beta diversity.

(2) The variance of a dissimilarity matrix among sites is **neither** the variance of the community composition table **nor** a measure of beta diversity; hence, partitioning on distance matrices **should not** be used to study the variation in community composition among sites.

Mantel correlogram

Words of caution (Mantel test and distance approach)

Legendre et al. (2005) :

(3) In all of the simulations, partitioning on dissimilarity matrices underestimated the amount of variation in community composition explained by the raw-data approach.

(4) The tests of significance in the distance approach had less power than the tests of canonical ordination.

Mantel correlogram

Words of caution (Mantel test and distance approach)

Legendre et al. (2005) :

Hence, the proper statistical procedure for partitioning the spatial variation of community composition data among environmental and spatial components (or any other components), and for testing hypotheses about the origin and maintenance of variation in community composition among sites, is **variation partitioning based on canonical ordination (RDA)**.

Mantel correlogram

Words of caution (Mantel test and distance approach)

Legendre et al. (2005) :

Whenever possible, use statistical procedures based on tables of raw data, such as correlation, regression, or canonical analysis. **Save the Mantel test and derived forms to test hypotheses that can <u>only</u> be formulated in terms of distances.**

Mantel correlogram

There is one application where the Mantel statistic is valid and useful, however: the **Mantel correlogram.**

Mantel correlogram

Basically, one computes a standardized **Mantel statistic** $\boldsymbol{r_M}$ (analogous to a Pearson's *r* coefficient) between a dissimilarity matrix among sites and a matrix where pairs of sites belonging to the same distance class receive value 0 and the other pairs, value 1.

	2	3	4
1	0.25	0.43	0.55
2		0.17	0.39
3			0.66

Dist. among vegetal communities

	2	3	4
1	1	0	1
2		0	1
3			0

First geogr. distance class

Mantel correlogram

The process is repeated for **each distance class**.

Each r_M value can be tested for by **permutations**.

The **expectation** of the Mantel statistic for no spatial correlation is $\boldsymbol{r_M = 0}$.

Mantel correlogram

In **R**, a Mantel correlogram can be computed by function **`mantel.correlog()`** {vegan}.

The only data necessary are a response dissimilarity matrix and either the geographical coordinates of the sites or a matrix of geographical distances among sites.

If the data contain a linear trend, they must be **detrended** before the computation of a Mantel correlogram.

Mantel correlogram

```
# Hellinger distance
mite.h <- decostand(mite, "hel")
# Detrending
mite.h.det <- resid(lm(as.matrix(mite.h) ~.,
as.data.frame(mite.xy)))
# Correlogram on detrended mite data
mite.mantel.c <-
mantel.correlog(dist(mite.h.det), dist(mite.xy))
mite.mantel.c
```

Mantel correlogram

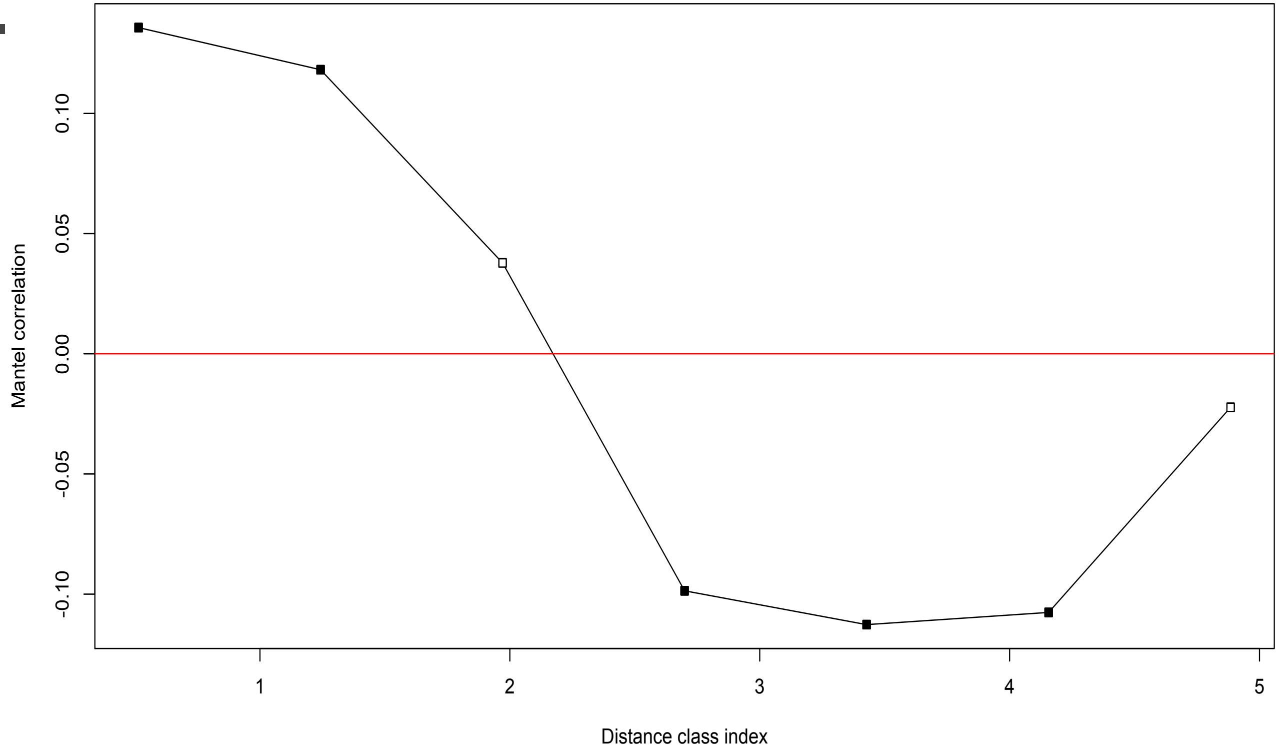
```
Mantel Correlogram Analysis

Call:

mantel.correlog(D.eco = mite.h.D1, D.geo = dist(mite.xy), nperm = 999)

        class.index     n.dist Mantel.cor Pr(Mantel) Pr(corrected)
D.cl.1     0.514182 358.000000   0.135713      0.001         0.001 ***
D.cl.2     1.242546 650.000000   0.118174      0.001         0.002 **
D.cl.3     1.970910 796.000000   0.037820      0.067         0.067 .
D.cl.4     2.699274 696.000000  -0.098605      0.001         0.004 **
D.cl.5     3.427638 500.000000  -0.112682      0.001         0.005 **
D.cl.6     4.156002 468.000000  -0.107603      0.001         0.006 **
D.cl.7     4.884366 364.000000  -0.022264      0.112         0.134
D.cl.8     5.612730 326.000000         NA         NA            NA
D.cl.9     6.341094 260.000000         NA         NA            NA
D.cl.10    7.069458 184.000000         NA         NA            NA
D.cl.11    7.797822 130.000000         NA         NA            NA
D.cl.12    8.526186  66.000000         NA         NA            NA
D.cl.13    9.254550  32.000000         NA         NA            NA
```

Mantel correlogram



Mantel correlogram of the Hellinger-transformed and detrended oribatid mite species data. Black squares indicate **significant** multivariate spatial correlation **after Holm correction for multiple testing**. The abscissa is labelled in metres since this is the unit of the data used to construct the distance classes.

Mantel correlogram

Correction for multiple testing

In a correlogram, a statistical test is performed for each distance class.

If no correction is made, each test has its own α rejection level (for instance $\alpha = 0.05$).

Running several tests increases the probability that **at least one of them will reject H_0 when it is true** (type I error).

Cure: correct the level of rejection.

Mantel correlogram

Correction for multiple testing

Bonferroni correction: divide the global threshold level by the number of simultaneous tests.

In our case, **globally**, the correlogram will be declared significant if **at least one of the 7 autocorrelation value** is significant at the corrected α level of 0.05/7 = **0.00714**.

One can verify that, with this correction, the global chances of accepting H_0 when it is true is equal to $(1 – 0.00714)^7 \approx 0.95$.

Mantel correlogram

Correction for multiple testing

The Bonferroni correction is very conservative, however, and can lead to type II error (failing to reject 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**.

Mantel correlogram

Correction for multiple testing

Holm correction:

1.Run all the k tests.

2. Order the probabilities in increasing order.

3. Correct each α significance level by dividing it by (1 + the number of remaining tests) in the ordered series.

Mantel correlogram

Correction for multiple testing

Example:

6 fictitious autocorrelation probability values:

	0.001 -	0.014 -	0.153 -	0.03 -	0.007 -	0.001
No correction	***	*		*	**	***
Bonferroni	**				*	**

Holm corr: order	0.001 =	0.001 <	0.007 <	0.014 <	0.03 <	0.153
Thresh.	0.0083	0.01	0.0125	0.0167	0.025	0.05
	***	**	**	*		

Bonferroni corrected *p*-value: $0.05/6 = 0.0083$