zt: a software tool for simple and partial Mantel tests

Eric Bonnet* and Yves Van de Peer

Department of Plant Systems Biology
Flanders Interuniversity Institute for Biotechnology (VIB)
Ghent University
KL Ledeganckstraat 35
B-9000 Ghent
Belgium
***corresponding author <eric.bonnet@gengenp.rug.ac.be>**

1. Introduction

Different methods of data analysis (e.g. clustering and ordination) are based on distance matrices. In some cases, researchers may wish to compare several distance matrices with one another in order to test a hypothesis concerning a possible relationship between these matrices. However, this is not always self-evident. Usually, values in distance matrices are, in some way, correlated and therefore the usual assumption of independence between objects is violated in the classical tests approach. Furthermore, often, spurious correlations can be observed when comparing two distances matrices. A classic example is the comparison between genetic and environmental distances. Colonies that are in close proximity of each other tend to have similar environments and therefore there will be a positive correlation between environmental and geographical distances. Such colonies will also be more likely to exchange migrants so that genetic distances will be positively correlated with spatial distances. The consequence is that an observed positive association between genetic and environmental distances may be simply due to spatial effects. The most widely used method to account for distance correlations is a procedure known as the Mantel test (Mantel, 1967; Mantel and Valand, 1970 following the pioneering work of Daniels, 1944 ; Daniels and Kendall 1947). The simple Mantel test considers two matrices while an extension known as the partial Mantel test considers three matrices. These tools are widely used in different fields of research such as population genetics, ecology, anthropology, psychometrics and sociology.

Since the Mantel test proceeds from distance (dissimilarity) matrices, it can be applied to variables of different logical types (e.g. categorical, rank, interval-scale...). This is especially interesting in research areas such as ecology that often use categorical variables. Since dissimilarity D is the equivalent of the inverse of similarity S (D = 1 – S), using similarity instead of dissimilarity has no qualitative effect on the analysis and only the sign of the coefficient will change.

In the Mantel test, the null hypothesis is that distances in a matrix A are independent of the distances, for the same objects, in another matrix B. In other words, we are testing the hypothesis that the process that has generated the data is or is not the same in the two sets.

Then, testing of the null hypothesis is done by a randomization procedure in which the original value of the statistic is compared with the distribution found by randomly reallocating the order of the elements in one of the matrices.

1.1 Simple Mantel test

The statistic used for the measure of the correlation between the matrices is the classical Pearson correlation coefficient:

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\frac{(A_{ij} - \overline{A})}{s_A} \right] \left[\frac{(B_{ij} - \overline{B})}{s_B} \right] \quad [1],$$

where N is the number of elements in the lower or upper triangular part of the matrix, $\overline{A}$ is mean for A elements and s_A is the standard deviation of A elements.

Note that if matrices A and B are normalized:

$$a_{ij} = \frac{A_{ij} - \overline{A}}{s_A} \; ; \; b_{ij} = \frac{B_{ij} - \overline{B}}{s_B},$$

we then have:

$$\overline{a} = 0 \; ; \; s_a = 1; \; \overline{b} = 0; s_b = 1,$$

which simplifies equation [1] as:

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{ij}$$

This coefficient measures the linear correlation and hence is subject to the same statistical assumptions. Consequently, if non-linear relationships between matrices exist, they will be degraded or lost.

The testing procedure for the simple Mantel test goes as follows:

Assume two symmetric dissimilarity matrices A and B of size n x n. The rows and columns correspond to the same objects. The first step is to compute the Pearson correlation coefficient between the corresponding elements of the lower (or upper)-triangular part of the matrices.

1. Compute the reference value r_{AB} using eq [1].
2. Permute randomly rows and the corresponding columns of one of the matrices, creating a new matrix A'.
3. Compute the $r_{A'B}$ statistic between matrix A' and matrix B using equation [1].
4. Repeat steps 2 and 3 a great number of times (>5000). This will constitute the reference distribution under the null hypothesis. The number of repeats determine the overall precision of the test (≈ 1000 for $\alpha = 0.05$; ≈ 5000 for $\alpha = 0.01$; ≈ 10000 for greater precision (Manly 1997).
5. For a one-tailed test involving the upper tail of the distribution, the p value is equal to the proportion of values $r_{\hat{A}'B.C}$ greater than or equal to $r_{AB.C}$. Symmetrically, the p value for the lower tail is the proportion of values $r_{\hat{A}'B.C}$ smaller than or equal to $r_{AB.C}$.

1.2 Partial Mantel test

The partial Mantel test involves three matrices. The goal is to test the correlation between matrices A and B while controlling the effect of a third matrix C, in order to remove spurious correlations. Different authors have suggested different possibilities to do this. Legendre (2000) simulated the properties of different forms of partial Mantel test and concluded that the method of permutation of the residuals of a null model can be used in most of the cases while the method of permutation of raw values (Smouse *et al.* 1986) is more suitable if a small sample size (n < 20) is combined with highly skewed data and the presence of outliers.

Permutation of the residuals of a null model was originally proposed by Freedman and Lane (1983) and further developed by Anderson and Legendre (1999). The principle is the following:
Given the multiple regression equation:

$$y = b_0 + b_1 x + b_2 z + e_{x,z},$$

where y is the dependent variable, x is a covariable and z is the explanatory variable of interest. The null hypothesis is that:

$$b_2 = 0$$

If we consider a null model where H_0 is true, then the regression equation can be rewritten as:

$$y = b_0 + b_1 x + e_x$$

So, all variation of y not explained by x is contained in e. Residuals are exchangeable among observations if they are independent.

The complete procedure is then as follows (according to Anderson and Legendre, 1999):

The reference statistic used is the well-known partial correlation coefficient:

$$r_{AB.C} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{1 - r_{AC}^2}\sqrt{1 - r_{BC}^2}} \quad [2],$$

where r is the simple Mantel statistic and A, B and C are the reference matrices in the study.

Note that if there's no link between C and matrices A and B:

$$r_{AC} = 0 \; ; \; r_{BC} = 0,$$

we have the Pearson correlation coefficient between A and B:

$$r_{AB.C} = r_{AB}$$

1. Compute the residuals Â from the simple linear regression of distances in A over the distances in C.
2. Compute r_{AB}, r_{AC} and r_{BC} and calculate the reference value $r_{AB.C}$ using equation [2].
3. Permute Â randomly using the same procedure as in simple Mantel test (see above), obtaining Â'.
4. Compute $r_{Â'B}$ and $r_{Â'C}$ and with r_{BC} compute the partial correlation statistic $r_{Â'B.C}$ using equation [2].
5. Repeat steps 2 and 3 a great number of times (>5000). This will constitute the reference distribution under the null hypothesis.
6. For a one-tailed test involving the upper tail the p value is equal to the proportion of values $r_{Â'B.C}$ greater than or equal to $r_{AB.C}$. Symmetrically, the p value for the lower tail is the proportion of values $r_{Â'B.C}$ smaller than or equal to $r_{AB.C}$.

For the method of permutation of the raw values, the algorithm is exactly the same as above except that no regression is done, and raw values of matrix A are used in the test.

2. Code & algorithm description

2.1 Licence

2.2 Language

The zt program has been written in the C programming language, both for huge matrix management and speed of computation. As the software is ANSI - C compliant (Kernighan and Ritchie, 1988), compilation can be done without modifications with any ANSI compliant compiler. Successful compilations and tests were done for Solaris and Linux with GNU gcc and for Windows with Borland bcc32.

2.3 Memory

Since dynamic memory allocation is used, the size of the matrices is only dependent on the available memory. In zt, only the lower half matrix elements are loaded into memory, without diagonal elements and without any labels.

2.4 Matrix permutation

Instead of randomly rearranging the elements in the matrix, only the labels of the columns and the corresponding rows are permuted. Suppose for example that we have the following 3 x 3 matrix:

	1	2	3
1	a11	a12	a13
2	a21	a22	a23
3	a31	a32	a33

The values of interest (lower triangular matrix) are in black. Note that these are the values that will be effectively used for computations.

The initial order of the labels is {1,2,3}. After random permutation, the order will be for example{3,1,2}. Elements in the matrix will thus be rearranged according to the new order.

	3	1	2
3	a33	a31	a32
1	a13	a11	a12
2	a23	a21	a22

Note that due to the randomization, elements of the upper half matrix are now in the lower triangle. But as the matrix is symmetric, value a13 = a31, and so we can compute even without upper values.

Two methods can be used for the randomization of labels. The first one will be called permutation and it involves the enumeration of all possible permutations sets for n elements. The second one will be called randomization and involves the sampling of random sets of all possible permutations sets for n elements**.**

The total number of permutations of a vector of n elements is given by n!. This number grows exponentially with increasing values of n (see below). Thus, the permutation procedure is applicable only for small values of n.

Matrix size n	n !
5	120
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	479001600

Furthermore, for small values of n, the permutation procedure is better than randomization. For example, for a 6 x 6 matrix there will be 720 possible permutations sets. Thus with the option of 1000 randomizations there will be surely some repetition for some sets and a bias in the p value. The software zt automatically selects the permutation procedure for matrices for which the size is smaller than 8 x 8.

For the complete enumeration of all permutations for a set of n elements, a possible procedure is the generation of sets in lexicographic order.
Consider a set of five data values labeled 1 to 5, so that the initial order is 12345. This is the smallest possible number that can be formed with these digits. The next largest one is 12354, which is found by permuting values 4 and 5. So we will have:

12345 12354 12435 12453 12534 12543 54321

The algorithm used in zt software was kindly provided by Glenn C Rhoads with some minor modifications by the author (see http://remus.rutgers.edu/~rhoads/Code/code.html).

The randomization procedure chooses random sets from all the orders possible. The algorithm used in zt software is a modified version of Knuth (1981):

1. set j = 1.
2. generate a random number U uniformly distributed between 0 and 1.
3. set k = jU+1, so that k is an integer between j and n.
4. exchange j and k.
5. set j = j + 1 ; if j < n return to step 2 otherwise stop.

2.5. Performance
Simple Mantel tests were run on a single Sun Ultrasparc 450 Mhz processor with a number of randomizations of 10000.

Size of matrices	**Computation time**
10 x 10	< 1sec.
100 x 100	11 sec.
1000 x 1000	26 min. 24 sec.

3. Syntax and case study

zt is a command line program, with text output. Thus it can be easily include in scripts and batch procedures.

3.1 Syntax and options

Simple Mantel test:
```
zt -s <file1> <file2> <number of randomizations>
```

Partial Mantel test:
```
zt -p <file1> <file2> <file3> <number of randomizations>
```

Complete path to data files should be given according to the syntax of the operating system used:
```
/statistics/data/foo.dat (Unix-like systems)
c:\statistics\data\bar.dat (Windows)
```

Options
-s simple Mantel test (mandatory)
-p partial Mantel test (mandatory)
-r partial Mantel test with raw option (optional)
-e force exact permutations procedure (optional)
-l print licence terms
-h display help

-s and –p option are mutually exclusive.

For the partial Mantel test, the default method is the permutation of the residuals of a null model. In case the option –r is chosen, the permutation of the raw values will be used.

The –e option will force the program to use the exact permutation set for a given size of matrices. Note that for matrix size < 8 this option will be automatically selected. The maximum size allowed for this option is 12 x 12. Of course with this option, the number of randomizations do not have to be indicated.

Options can be combined to the same 'word'. For example '–pre' or '–p –r –e' both mean a partial Mantel test with permutation of the raw values and exact enumeration of all possible permutations.

-h option display some basic help.

Matrices format
Input matrices are in text ASCII format. They contains only the numeric values of the lower half matrix without diagonal values separated by spaces. The first number is the size of the matrix.

Example:
```
8
0.3
0.14 0.5
0.23 0.5 0.54
0.3 0.4 0.5 0.61
-0.04 0.04 0.11 0.03 0.15
0.02 0.09 0.14 -0.16 0.11 0.14
```

```
-0.09 -0.06 0.05 -0.16 0.03 -0.06 0.36
```

Note that you can provide all values in a single column:

```
8
0.3
0.14
0.5
0.23
0.5
0.54
...
0.36
```

3.2. A case study using 'zt' on earwigs (Manly, 1997)

Two datasets will be used, both are taken from Manly, 1997.

3.2.1 Distribution of earwigs species across continents.
Earwigs species may have evolved in the northern hemisphere and subsequently spread into the southern continents or, alternatively, they may have evolved throughout the southern proto continent of Gondwanaland, 150 millions years ago.
If the first hypothesis is correct, then similarities between species in different part of the world should reflect their present distances. If not, then southern continents should contain species that are more similar.
For the example being considered, rows and columns will be eight different areas in the world, i.e. Europe and Asia, Africa, Madagascar, the Orient, Australia, New Zealand, South America and North America.
- assoc.txt is the matrix of the species coefficient of similarities across continents.
- gond.txt is the matrix of distances between areas in term of "steps" required to go from one to another, based on positions of the areas in Gondwanaland.
- pres.txt is the matrix of distances between areas at present time.

We will use a simple Mantel test to test the correlation of species similarities with (1) distances at present time and (2) distances in Gondwanaland. As the size of the matrices is relatively small, the exact permutation method will be used.

3.2.1.1 Simple Mantel test between similarity coefficients and geographical distance at present time with exact permutation method.
Command:

```
zt –se assoc.txt pres.txt
```

Output:

```
File A:                 assoc.txt
File B:                 pres.txt
Size of matrices:       8 x 8
Number of iterations:   40320
Options:                simple exact

Randomizing...

r =                     -0.216964
p =                     0.178323 (one-tailed)
```

Conclusion:
The correlation coefficient is –0.22 with an associated probability of 0.18. The correlation is a typical value obtained by chance and there is no real evidence of a relationship between species distances and the present day distances between continents.

3.2.1.2 Simple Mantel test between similarity coefficients and geographical distance at Gondwanaland time with exact permutation method.
Command:
zt –se assoc.txt gond.txt

Output:

```
File A:                 assoc.txt
File B:                 gond.txt
Size of matrices:       8 x 8
Number of iterations:   40320
Options:                simple exact

Randomizing...

r =                     -0.605379
p =                     0.001587 (one-tailed)
```

Conclusion:
The observed correlation is –0.6 with an associated p value of 0.0016. Therefore, there is strong evidence that earwig species evolved in Gondwanaland before it broke up in different continents.

3.2.2 Colonies of *Euphydryas editha*
The original work was done by McKechnie *et al.* (1975).
The problem here is to determine if genetic distances between 21 colonies of the butterfly *Euphydryas editha* are correlated with environmental distances taking into account (1) the fact that colonies that are geographically close can be expected to have similar environments and (2) that colonies that are geographically close can be genetically relatively similar because of past migration.
The analysis will begin by testing the association between the matrices with simple Mantel tests and then to perform partial Mantel tests between genetic (gene.txt), environmental (env.txt) and geographical distances (geo.txt).

3.2.2.1 Simple Mantel test for genetic and environmental distances with 10000 randomizations.
Command:

```
zt –s gene.txt env.txt 10000
```

Output:

```
File A:                 gene.txt
File B:                 env.txt
Size of matrices:       21 x 21
Number of iterations:   10000
Options:                simple

Randomizing...

r =                     0.291544
p =                     0.006599 (one-tailed)
```

Conclusion:
The correlation value of 0.29 between genetic and environmental distances is significantly different from zero with a p value smaller than 0.01. Note that command line options are resumed on the line "Options" in the results.

3.2.2.2 Simple Mantel test for genetic and geographic al distances with 10000 randomizations.
Command:

```
zt –s gene.txt geo.txt 10000
```

Output:

```
File A:                 gene.txt
File B:                 geo.txt
Size of matrices:       21 x 21
Number of iterations:   10000
Options:                simple

Randomizing...

r =                     0.489822
p =                     0.000100 (one-tailed)
```

Conclusion:
The correlation value of 0.49 is significantly different from zero with a p value smaller than 0.001.

3.2.2.3 Simple Mantel test for environmental and geographical distances with 10000 randomizations.
Command:

```
zt –s env.txt geo.txt 10000
```

Output:

```
File A:                 env.txt
File B:                 geo.txt
Size of matrices:       21 x 21
Number of iterations:   10000
Options:                simple

Randomizing...

r =                     0.038256
p =                     0.367163 (one-tailed)
```

Conclusion: The correlation between geographical and environmental distances is not significantly different from zero.

3.2.2.4 Partial Mantel test between genetic and environmental distances while controlling the effect for geographical distances with 100000 randomizations (for a valid p value).
Command:

```
zt –p gene.txt env.txt geo.txt 100000
```

Output:

```
File A:                 gene.txt
File B:                 env.txt
File C:                 geo.txt
```

```
Size of matrices:       21 x 21
Number of iterations:   100000
Options:                partial residuals

Randomizing...

r =                     0.313143
p =                     0.000830 (one-tailed)
```

Conclusion:
There is a significant correlation of 0.31 between genetic and environmental distances while controlling effect for geographical distances.

3.2.2.5 Partial Mantel test between genetic and geographical distances while controlling the effect of environmental distances with 100000 randomizations.
Command:

```
zt –p gene.txt geo.txt env.txt 100000
```

Output:

```
File A:                 gene.txt
File B:                 geo.txt
File C:                 env.txt
Size of matrices:       21 x 21
Number of iterations:   100000
Options:                partial residuals

Randomizing...

r =                     0.500774
p =                     0.000030 (one-tailed)
```

Conclusion:
This test shows that there is a significant correlation of 0.5 between genetic and geographical distances, independently from environmental distances.

3.2.2.6 Overall conclusion
Genetic distances are significantly correlated to both environmental and geographical distances. There is no link between geographical and environmental distances.

4. Acknowledgments

I would like to thank Daniel Petit and Jean-François Lenain from the University of Limoges who introduced me to the Mantel test in the late 90's. Thanks also to Philippe Casgrain from the University of Montreal for his comments on previous versions of the program.

5. References

Anderson, M. J. and Legendre, P. (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Statistis. Comput. Simulation* 62: 271-303.

Daniels, H. E. (1944) The relation between measure of correlation in the universe of sample permutations. *Biometrika* Vol. 33 No. 2 : 129-135.

Daniels, H. E. and Kendall, M. G. (1947) The significance of rank correlations where parental correlation exists. *Biometrika* Vol. 34 No 3-4 : 197-208.

Freedman, D. and Lane, D. (1983) A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.* 1: 292-298.

Kernighan, B. and Ritchie, D. (1988, 1978) *The C Programming language, 2nd edition*, Bell Telephone laboratories, Incorporated.

Knuth, D.E. (1981) *The art of computer programming. Volume 2, Semi-numerical algorithms*, Addison-Wesley, Reading, Massachussets.

Legendre, P. (2000) Comparison of permutation methods for the partial correlation and partial mantel tests. *J. Statist. Comput. Simul.* 67: 37-73.

Manly, B.J.F. (1997, 1991) *Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd edition.* London: Chapman and Hall.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

Mantel, N. and Valand, R. S. (1970) A technique of nonparametric multivariate analysis. *Biometrics* 26: 547-558.

McKechnie, S. W., Ehrlich, P. R. and White, R. R. (1975) Population genetics of *Euphydryas* butterflies. I. Genetic variation and the neutral hypothesis. *Genetics* 81: 571-94.

Smouse, P.E., Long, J.C. and Sokal, R. R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* 35 : 727-32.