Torocor 1.0

a program to assess the association between spatially autocorrelated variables using torus-translation tests

written by Olivier HARDY

User's manual

Address for correspondence:

Evolutionary Biology and Ecology, CP160/12

Université Libre de Bruxelles

50 Av. F. Roosevelt

B-1050 Brussels, Belgium

e-mail: ohardy@ulb.ac.be

Last update: 30 June 2009

PURL to download TOROCOR: <http://purl.oclc.org/net/torocor>

Purpose

Torocor is designed (1) to characterise the spatial autocorrelation of quantitative and/or qualitative variables and (2) to test the significance of the association between variables, notably using torus-translation randomizations. The latter procedure removes the bias of classical tests applied on spatially autocorrelated variables where samples cannot be considered as independent (classical tests tend to be liberal, i.e. rejecting too often the null hypothesis that there is no association between variables).

The data required are geographically located sample points where a set of quantitative and/or qualitative variables are defined. In addition, to apply torus-translation tests, each sample must be positioned on one or several rectangular grids (otherwise only spatial autocorrelation analysis and association tests using potentially biased complete randomization tests can be performed). Missing data are allowed but should occur at low frequency.

What does *Torocor* do?

1. Characterizing the spatial structure of each variable

The spatial autocorrelation of quantitative variables is described using Moran's I statistic. First consider the spatial autocorrelation between two samples, i and j, at a variable x:

$$I_{ij} = \frac{(x_i - \bar{x})(x_j - \bar{x})}{Var(x)} + 1/(n-1)$$

where x_i is the value of variable x for sample i; $\bar{x}$ and $Var(x)$ are, respectively, the mean and variance of variable x estimated from the whole data set; n is the total sample size. The second term is a sample bias correction, ensuring that the average I_{ij} over all existing i, j pairs is equal to zero. Moran's I statistic is then the average I_{ij} over all pairs separated by a given distance interval:

$$I(d) = \sum_{i,j} \delta(i,j,d) I_{ij} \Big/ \sum_{i,j} \delta(i,j,d)$$

where the sums are taken over all possible pairs of samples (i, j); d denotes a distance interval (e.g. 100 meters to 200 meters); $\delta(i,j,d)$ is an indicator variable taking the value 1 when the spatial distance separating samples i and j is included into interval d and 0 otherwise.

An equivalent of Moran's I for qualitative variables can be defined as:

$$I'(d) = \frac{P(d) - \overline{P}}{1 - \overline{P}}$$

where $P(d)$ is the probability that two samples separated by a distance interval d share the same state at the variable, $\overline{P}$ being the probability that two random samples share the same state.

$I(d) = 1$ when pairs of samples separated by d always share the same value or the same state at the variable considered. When a variable is spatially autocorrelated, $I(d)$ or $I'(d)$ is expected to be positive at short distance, decreasing with increasing distances and eventually reaching negative values. Without spatial autocorrelation, $I(d) = 0$ for any d (except for local fluctuations due to limited sample size). A plot of $I(d)$ according to distance d is called an

autocorrelogram. The steeper the slope of the autocorrelogram, the stronger the spatial structure of the variable.

To quantify the magnitude of the spatial autocorrelation by a single statistic, I_{ij} is regressed on the distance between samples, d_{ij}, or its logarithm, $\ln(d_{ij})$, providing the regression slopes, b_{lin} or b_{log}. In option, these regression slopes can be assessed for a restricted range of distances (*mindist* to *maxdist*), that is considering only *i-j* pairs such that $mindist \leq d_{ij} \leq maxdist$.

Torocor also allows to assess the spatial autocorrelation of variables for particular pairs of samples chosen according to a qualitative variable. For example, if samples are vegetation plots and a variable indicates the type of habitat of each plot, the spatial autocorrelation of another variable, such as the abundance of a particular species, could be assessed only for pairs of samples occurring in the same habitat, or only for pairs of samples occurring in distinct habitats. This allows to control for a possible habitat effect on the spatial autocorrelation of the species abundance. In addition, comparison of within-habitat and between-habitat correlograms could show whether there is an habitat effect while controlling for the distance between samples, i.e. if the two correlograms are equivalent, one can conclude that there is no habitat effect.

2. Characterizing the association between variables

The association between variables is quantified as following:

(1) for two quantitative variables, by Pearson's correlation coefficient

(2) for two qualitative variables, by the Khi-square statistic derived from a contingency table

(3) between quantitative and qualitative variables, by the intra-class correlation coefficient of the quantitative variable classified by the qualitative variable (as in an ANOVA).

3. Testing the spatial structure of each variable

The spatial autocorrelation of each variable is tested using **complete randomizations**, whereby the values of a variable are randomly shuffled among all samples. $I(d)$, as well as the regression slopes b_{lin} and b_{log}, are recomputed for many randomized data sets to assess their distributions under the null hypothesis that there is no spatial structure. P-vales are estimated by comparing the observed statistics with their respective distributions, and 95% envelopes can be constructed.

Torus-translation randomizations (see below) can also be applied to test whether the spatial structure is affected by this randomization procedure, which is designed to keep this structure as intact as possible.

4. Testing the association between variables

Complete randomizations can also be used to test the association between variables if at least one variable shows no spatial autocorrelation. However, complete randomizations usually result in liberal tests (i.e. tests providing too many false positive) for testing the association between two spatially autocorrelated variables, as is the case of classical tests assuming that samples are independent. To account for the spatial structures of the variables, the goal of

torus-translation randomizations is to break down the association between variables while keeping their respective spatial autocorrelation patterns intact, as far as possible.

Torus-translations require that samples are located on one or several rectangular grid(s), or one or several transect(s), where each position of the grid(s) contains a sample. A torus-translation randomization consists in translating all the samples within each grid by a random number of steps in each direction. Because all samples move in parallel, their spatial relationships, hence the spatial structure of the variables, are preserved. When samples are translated beyond one border of the grid, their are re-introduced through the opposite border, as if the grid was inscribed on the surface of a torus. If samples are actually located on a transect, torus-translation randomizations somehow assume that the transect is inscribed in a circle. In addition to random translations, each grid can be turned upside down or can be rotated by 180° (and transects can be inverted). The statistic characterizing the association between two variables is computed after many such torus-translation randomizations applied on one of the variables, so that its distribution under the null hypothesis can be assessed.

Torus-translation randomization is not perfect because, unless the samples were actually located on a circle or on a torus, which is very unlikely, border effects disturb to some extent the original spatial structure. Such disturbance is usually inconsequential except when a macro-scale spatial structure occurs, such as an overall gradient throughout the grid. In extreme cases it may limit the validity of torus-translations for testing the association between variables. To investigate the impact of torus-translation randomization on the spatial structure of a variable, $I(d)$ can be tested using such randomization, non-significant values meaning that torus-translations preserve the spatial autocorrelation pattern.

5. Computing average local values of quantitative variables

For some types of variable, such as one describing the occurrence of a rare event per location, it can be interesting to compute local averages over all samples located within a given radius around a position, because the presence of such event in nearby locations can be an interesting information, even if the event was not observed on the location itself. *Torocor* allows the creation of such new variables by duplicating all quantitative variables found in the data set. One must provide the threshold distance defining the radius of the circle within which local average are computed. Note that the resulting variables are strongly spatially autocorrelated.

Data file

The data file is tab-delimited text file that can be prepared using a spreadsheet (e.g., Excel) and saved as text file (with tab-delimited columns).

The **first line** is a title line that will be copied in the result file

The **second line** contains four numbers separated by a TAB:

- the number of samples
- the number of grids or transects (0 or several)
- the number of variables (quantitative or qualitative)
- the symbol(s) assigned to missing data (optional)

The **third line** starts by the number of distance intervals followed the maximal distance defining each interval (all values are separated by a TAB)

The **fourth line** contains the labels of the columns, including the names of the variables. To force a variable to be considered as a **qualitative** one, its name must be preceded by the symbol **$** (otherwise, the program will automatically detect qualitative variables as the ones for which at least one non-numeric value is given).

The **following lines** contains the data, with one line per sample, in the following order :

- the name of the sample
- the geographic X position of the sample (e.g. UTM longitude)
- the geographic Y position of the sample (e.g. UTM latitude)
- if the number of grids > 1: the name of the grid on which the sample is located
- if the number of grids > 0: an integer giving the X position on the grid
- if the number of grids > 0: an integer giving the Y position on the grid
- for each variable, its value (numeric for quantitative variables, alphanumeric for qualitative variables) for the given sample (or the symbols defining a missing data)

Example:

Example of data set
20 2 3 missing
5 2 4 8 20 60

Sample	Long	Lat	Grid	X	Y	$Var1	Var2	Var3
s1	1.7	8.2	g1	1	1	4	84	0.91
s2	4.1	9.9	g1	1	2	A	95	0.3
s3	7.8	11.9	g1	1	3	Bleu	81	0.96
s4	11.6	14.1	g1	1	4	Bleu	64	0.16
s5	5.2	6.1	g1	2	1	A	64	0.07
s6	9.7	8.4	g1	2	2	A	47	0.16
s7	12	10.4	g1	2	3	4	95	0.55
s8	15.3	11.7	g1	2	4	Bleu	35	0.11
gh2	34.3	16.7	g2	1	1	Bleu	80	0.85
gh5	37.3	18.7	g2	1	2	A	97	0.92
gh7	41.2	20.8	g2	1	3	A	72	0.26
gh8	36.5	16	g2	2	1	Bleu	73	missing
gh9	39.4	17.8	g2	2	2	Bleu	11	0.7
gh10	42.1	19.8	g2	2	3	A	85	0.55
gh11	38.1	14.6	g2	3	1	missing	59	0.92
gh12	41	16.7	g2	3	2	4	23	0.46
gh13	44.4	19	g2	3	3	4	46	0.33

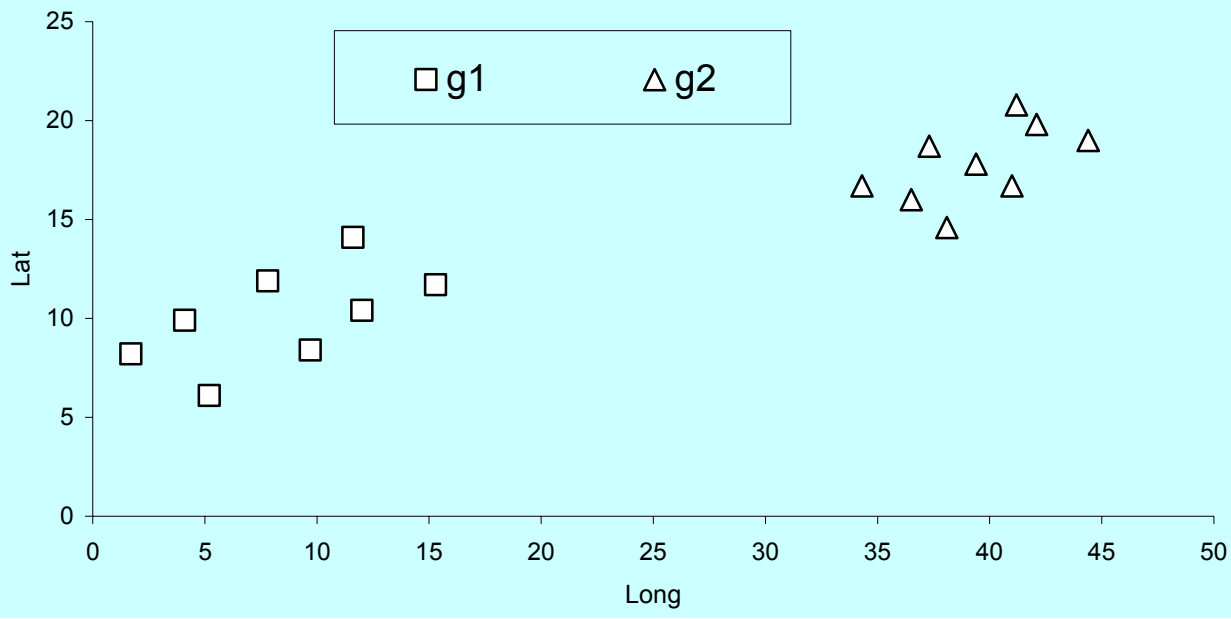

Figure: Spatial positions of the samples given in the example.

Notes on data file:

- Missing values can only be attributed to variables, they are not allowed for X, Y geographical or grid positions.
- The data set must be complete, empty cells are not allowed.
- If there is no grid system, there is no column for grid names and grid X, Y positions. If there is only one grid, there is no column for grid names.
- The position on the grid in each direction must be given by an integer, starting from 1 to the number of grid units in that direction. When grids are defined, the program will check that there is exactly one sample per grid position.
- If samples are located along uni-dimensional transects rather than on grids, the Y grid position must be given systematically the value 1.
- Ideally, the geographical positions of the samples should fit into one or several rectangular grid(s) or transect(s) but some level of spatial distortion can be acceptable (cf. figure), as long as the torus-translation randomization procedure maintains the observed level of spatial autocorrelation. For example, if samples are located along a sinuous trajectory (e.g. along a river) and with approximately a constant distance between adjacent samples, samples could be treated as if occurring regularly along a linear transect.
- Spatial autocorrelation is always computed according to the actual geographic positions of the samples; the grid positioning system serves only to perform torus-translation randomizations.

Running the program

The program runs on PC with Windows 9x or later versions, but has no fancy windowing features. It is written in C language using functions conforming to ANSI C standard (except for one console I/O function).

You can launch the program by double-clicking on its icon (the data file must then reside in the same folder as the program file) or by drag-and-drop of the data file icon on the program icon, in which case the data file is automatically recognized.

Once the program is launched, you are requested to enter the names of the data and results files. If you just press *RETURN* to these questions, the default names "in.txt" and/or "out.txt" will be considered. If a file with the same name as the results file already exists in the folder, the program will ask if you wish to: erase the existing file first (enter '*e*'), add results to the end of this file (enter '*a*' or simply press *RETURN*), or change the name of the results file (enter the new name).

Then the program displays basic characteristics of the data, that you must check for consistency, and asks to select one of three types of analyses to perform:

1- Spatial autocorrelation analysis
2- Correlation between variables
3- Computing local means for quantitative variables

Additional choices can be requested, after which the program perform the tasks asked and write down all results in a single file. The latter is a text file (with TAB delimited pieces of information) that should be opened with a spreadsheet (e.g. Excel).

Options for spatial autocorrelation analysis

1. You can specify the name of a qualitative variable of the data set to conduct analyses for particular pairs of samples classified by this variable (i.e. within or between states defined by the variable). Specifically, Moran's *I* will be computed considering pairs of samples (*i*, *j*) belonging to the same, or different, states following the selected variable. In that case, you are asked to select between:

    1- same state (*i* and *j* belong to the same state)
    2- distinct class (*i* and *j* belong to the different state)
    3- the same given state (*i* and *j* belong to a given state)
    4- two given states (*i* belongs to one given state, *j* to another given state, or the reverse)

    If you select option 3 or 4, you will be asked to specify the state(s) for selecting samples.

2. You can define a restricted distance interval for computing regression slopes.

3. The number of randomizations to test Moran's *I* statistics is asked.

    Enter 0 if you do not want any test but just the correlograms

Complete randomizations are used but if you enter a negative number and have defined grids, torus-translations will also be performed, which is useful to check whether the latter preserve the spatial structure of the variable.

4. If you have provided a number of randomizations for tests, you can ask to obtain the details of the randomization tests, in which case *Torocor* will give, for each variable, the average and the lower and upper limits of the 95% envelopes of $I(d)$, b_{lin} and b_{log} values obtained after randomization.

Options for association between variables

1. The number of randomizations to test the statistics describing the association between variables is asked.

   Enter 0 if you do not want any test but just the statistics

   Torus-translation randomizations are used, at least if grids are defined in the data set, but if you enter a negative number or have not defined grids, complete randomizations will be used. The results of such complete randomizations are reliable only if at least one of the two variables being compared is not spatially autocorrelated.

2. You can ask to obtain the details of the randomization tests, in which case *Torocor* will give, for each association statistic between variables, the following characteristics of their distribution after randomization: the number of distinct values, the average, the standard deviation, the lower and upper limits of the 95% envelopes, the P-values associated with the bilateral and the two unilateral tests. P-values for unilateral tests correspond to the proportion of values obtained after randomization being lower (for one test), or higher (for the other test), than the observed value prior to randomization.

Options for computing local means for quantitative variables

You will be asked to enter the distance threshold within which local averages are computed. For each sample, the average value of each quantitative variable will be computed over all surrounding samples (including the focal sample) occurring at a distance inferior or equal to the threshold distance. Results are written in a single file called “newvar.txt”. The new variables of interest, which are given the name of the original variables followed by “(d<=[threshold distance])”, can be copied from this file to create a new data file and perform further analyses.

Result file

The result file is a text file with TAB delimited pieces of information that can be opened with a worksheet program (e.g. Excel).

The result file first contains basic information about the data set: the number of samples, the number of grids, the total number of variables (among which the number of qualitative variables), the code indicating missing data, and the range of distances separating the samples.

Then basic statistics describe each quantitative variable (mean, standard deviation, range, proportion of missing data) and then each qualitative variable (number of states, proportion of samples of each state).

Spatial autocorrelation analysis

First, the types of pairs of samples analyzed are specified following the option chosen.

Second, characteristics are provided for each distance interval (one per column): the maximal distance defining the interval, the number of (selected) pairs of samples falling into the interval, the mean distance between samples for pairs belonging to the interval. Then, for each variable (successive line), Moran's *I* statistics are given for each interval, followed by the regression slopes according to the linear distance (b_{lin}) and the logarithm of the distance (b_{log}).

If randomization tests were asked, the same table is shown afterwards but with statistically significant values marked by * ($P<0.05$ and at least 99 randomizations), ** ($P<0.01$ and at least 499 randomizations), or *** ($P<0.001$ and at least 4999 randomizations). The type of randomization (complete or torus-translation) is also indicated. Bilateral tests are used for Moran's *I* values, whereas unilateral tests are used for regression slopes (because they are always expected to be negative under spatial autocorrelation).

If one had asked for a detailed report, an additional table of Moran's *I* statistics (and b_{lin}, b_{log}) is given for each variable, including the observed values, the mean values after randomization, and the lower and upper limits of the 95% confidence envelope.

Association between variables

The characterization of the association between variables depend on the types of variables being compared. First, a matrix of Pearson's correlation coefficients between quantitative variables is given. Second, a matrix of Khi-square (χ^2) statistics of independence between qualitative variables is given. Third, the intra-class correlation coefficient of each quantitative variable classified by each qualitative variable is given.

If randomization tests were asked, the same tables are shown afterwards but with statistically significant values marked by * ($P<0.05$ and at least 99 randomizations), ** ($P<0.01$ and at least 499 randomizations), or *** ($P<0.001$ and at least 4999 randomizations) and preceded by + or – to indicate whether the observed value was higher or smaller than the mean value after randomization. The type of randomization (complete or torus-translation) is also indicated. Bilateral tests are used for Pearson's correlation coefficients between quantitative variables, whereas unilateral tests are used for the χ^2 statistics and the intra-class correlation coefficients.

If one had asked for a detailed report, for each pair of variables are given: the observed value of the association statistic, characteristics of the distribution of the statistic after

randomization (number of distinct values, average, standard deviation, lower and upper limits of the 95% envelope), and the P-values of the associated test (one bilateral, two unilateral tests).

Duplication of quantitative variables with local means

Local averages are given for each sample and each quantitative variable in a file called "newvar.txt". The new variables are given the name of the original variables followed by "(d<=[threshold distance])". They can be copied from this file to create a new data file and perform further analyses.

Notes

Present limitations:

- Maximal number of samples: limited by available memory only
- Maximal number of variables: 1000
- Maximal number of states per quantitative variable: 1000
- Maximal number of randomizations: 100000
- Maximal number of distance intervals: 100
- Maximal number of characters for names (sample, variable, states, files): 100
- Names: no blanks allowed
- Maximal number of characters per line in the data file: 10000

Computation time:

Spatial autocorrelation analyses can be time consuming when the number of samples is large (>1000). This can be limiting for randomization tests. Hence one can try first few randomizations (e.g. 99), and increase the number afterwards if computation time is not problematic. *Torocor* displays the number of randomizations done while computing.

Testing the association between variables is usually fast, except when the number of variables is high (>200). Here again, a low number of randomizations can be asked first to evaluate computation time. Nevertheless, with many variables, there will be many pairwise tests so that one should consider only highly significant values (low P-values), which require a large number of randomizations.

Interpreting spatial autocorrelation tests:

For spatial autocorrelation, there are as many Moran's *I* values tested as distance intervals given. Thus, under the null hypothesis, the chance that at least one value lies outside the 95% confidence envelope is much higher than 5% and there is a risk of rejecting the null hypothesis too often (liberal test). It is therefore advisable to conclude for significant spatial autocorrelation only if the regression slope is statistically significant, choosing the regression with distance or ln(distance) depending on whether correlograms appear more linear with distance or ln(distance), respectively.

Interpreting association tests between variables:

With many variables, there are many pairwise association tests. Hence, several tests may be significant just by chance so that one should be careful in interpreting the results, for example by being more restrictive regarding the P values before concluding that an association is really significant. Nevertheless, I do not advise to apply strictly a Bonferroni correction, whereby a test is considered significant only if the P value is less than the nominal P value (e.g. 0.05) divided by the total number of tests, firstly because the procedure is very conservative and many actual correlations might be missed, second because the minimal number of randomizations required to get a test significant at a low P-value is about 5 / (P value), which can become too high to be performed.

Interpreting association tests between variables using several grids:

When several grids are defined, torus-translation randomizations occurs only within each grid. Hence, torus-translation tests can detect only association between variables occurring within grids. If two variables are associated because their values differ markedly among grids in a correlated way, this will not be tested. Nevertheless, such situation can be detected because the distribution of the correlation coefficient after torus-translation randomization does not centre on zero.