Modt-LMC User's Guide

Version 1.0 "Standalone application", March 2011

This documentation explains how to use the computer program called "Modt-LMC", which implements modified *t*-tests for structural correlations in the linear model of coregionalization (LMC) from Dutilleul and Pelletier (2011).
Reference:
Dutilleul, P. and Pelletier, B. 2011. Tests of significance for structural correlations in the linear model of coregionalization. *Mathematical Geosciences* 43:819–846.

Installation

1- Following the instructions posted at http://environmetricslab.mcgill.ca, download **mcrinstaller.exe** from our ftp server and install it on your computer.

2- Retrieve a copy of the file **ModtLMC_standalone.zip** on your computer and unzip its content to a folder of your choice.

3- Launch the program by clicking on **ModtLMCui.exe**. Two windows should then appear: a user interface, to run the Modt-LMC program, and a command window. Note that the first time the program is launched, a folder containing encrypted Matlab files will be generated in your folder.

In the user interface, you are asked to perform the following tasks:

1- Data file: Write the name of a tab-delimited text file containing

- On the first row, the names (identifiers) of the spatial coordinates and *p* variables.

  Followed (starting on the second row) by

- In the first column, the indices from 1 to *N*, where *N* is the total number of sampling locations.
- In the second and third columns, the spatial coordinates of the *N* sampling locations in 2-D space.
- In the following *p* columns, the *N* observations for the variables from which pairs of variables will be selected.

2- Choice of variables: A unique set of variables or two different sets can be used for analysis. With a unique set, structural correlations are estimated and the significance of estimated structural correlations is assessed for each pair of variables among those chosen. With two different sets, structural correlations will be estimated only for the pairs involving one variable from each set. Use [Shift + right click] to choose a group of ‘contiguous’ variables in the list(s). Use [Ctrl + right click] to choose ‘non-adjacent’ variables from the list(s).

3- Box-Cox transformation: Check the box if you want to perform a Box-Cox transformation of your data. This procedure is aimed at improving the normal distribution assumption on the transformed data. This power transformation is defined by $y' = (y^{\lambda} - 1)/\lambda$ when $\lambda \neq 0$ and $y' = \ln(y)$ when $\lambda = 0$. In Modt-LMC, the coefficient λ is selected among values from -2.5 to 2.5, by steps of 0.1.
Reference:
Box, G.E.P and Cox, D.R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, series B* 26:211–243.

4- Standardization: Check the box if you want to standardize the variables to a zero mean and a variance of one. Note that when the Box-Cox transformation is used, variables are automatically standardized and this option is no longer available.

5- Threshold: The decision to include a structure for a variable in the LMC, or not, is based on information available in the experimental direct variogram of the variable, in the form of a percentage expressed relative to the total variance. This percentage corresponds to the smallest value of an estimated sill, at which the structure or basic variogram function (i.e., nugget effect, spherical model) is kept in the variogram modeling. When one structure is discarded for a given variable, the fitting of the LMC for the experimental variograms involving that variable is performed by using the other structure only. A value of “0” should be used when no threshold is desired.

6- Write a set of characters (see *Userdefinedprefix* below) that will be used to identify the three output files (*.txt) generated by the program and saved in the Matlab folder.

Output files

The content of the three output files can be described as follows.

*Userdefinedprefix*_results.txt: Comprises four sections of results for the modified *t*-tests performed on structural correlations estimated for the nugget effect and the spherical model, depending on whether two structures or only one structure were/was used in the LMC fitting. The last two sections will not appear if the corresponding situations are not encountered - this may be the case when no threshold is used. Each section of results comprises eight columns: the sill estimated for the experimental direct variogram of the first variable at structure *s* (Sill_direct1); the sill estimated for the experimental direct variogram of the second variable at the same structure (Sill_direct2); the sill estimated for the cross experimental variogram between the two variables at that structure (Sill_cross); the estimated range of the spherical model (Range); the estimated effective sample size

(M_s_hat); the estimated structural correlation (r_s); the observed value of the *t*-test statistic (t_obs); and the corresponding probability of significance (P_value).

*Userdefinedprefix*_variograms.txt: The first and second rows indicate the number of pairs of observations per distance class, and the mean distance value per class, respectively. The remaining rows are arranged as a series of three rows (triplets) corresponding to the two experimental direct variograms and the experimental cross variogram for each pair of variables involved in the analysis of structural correlations.

*Userdefinedprefix*_boxcox.txt: Comprises two columns with the coefficient λ used in the Box-Cox transformation, i.e., between –2.5 and 2.5, and an indicator of whether or not the variables were standardized to a zero mean and a variance of one, i.e., standardized = 1; non-standardized = 0.

Example

The dataset used by Pelletier *et al.* (2009) to examine the relationships between plant species diversity and soil and topographical variables provides an example for the application of Modt-LMC here (see results in Appendix). The dataset accompanies the computer program as a text file (*datatest.txt*). It comprises 339 sample points and eight variables. In order to satisfy the assumption of second-order stationarity, the dataset actually comprises residuals obtained after subtracting estimated drifts from the raw data. For the purpose of this example, structural correlations were analyzed using two different sets of variables: (1) two plant diversity variables – Trees & Herbaceous, and (2) six soil and topographical variables – pH, extractable magnesium (Mg), soil organic matter (SOM), total nitrogen (N), elevation (E_M), and fractal dimension of elevation (FD). Thus, the total number of pairs of variables was 12. A threshold of 5% was used for the determination of the number of structures (see above). Note that the estimated structural correlations of Pelletier *et al.* (2009) were obtained in Phase 2 of coregionalization analysis with a drift (CRAD), where coregionalization analysis is performed on all the variables together, whereas it is performed for two variables at a time in Modt-LMC.
Reference:

Pelletier, B., Dutilleul, P., Larocque, G., and Fyles, J.W. 2009. Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 2. Estimation of correlations and coefficients of determination. *Environmental and Ecological Statistics* 16:467–494.

Complementary note 1

This note in four parts is about the definition of distance classes in the computation of experimental variograms. First, the area covered by the sampling grid is estimated by convex hull, using the Matlab function “convhull”. Second, half the side length of the square with same area is used as maximum lag distance. Third, this maximum lag distance is divided by 12 to obtain the minimum lag distance, which is also used as the increment between distance classes. If there are less than 100 pairs of observations in at least one distance class, the maximum lag distance is divided by 11, 10, etc., until each distance class has at least 100 pairs of observations or the number of distance classes is four. Fourth and last, the mean distances of classes are used to plot experimental variograms and fit variogram models.

Complementary note 2

In coregionalization analysis, the level of uncertainty in the estimation of sills heavily depends on the number of sampling locations (i.e., the sample size) and the number of structures or basic variogram functions used to model experimental variograms in the LMC (Larocque *et al.*, 2007). For this reason, the LMC used in Modt-LMC is limited to only two structures: a nugget effect and a spherical model. A consequence of high uncertainty is the occurrence of estimated structural correlations (i.e., functions of estimated sills) that are very close to 1 (i.e., their largest value possible); in such a case, results should be interpreted with great caution. A high level of uncertainty in the estimation of structural correlations can also be reflected by an estimated effective sample size that is too low to allow the evaluation of the test statistic and its probability of significance. In this case, a NaN (Not a Number) value is written in the table of results. Datasets of 200 sampling locations or more are recommended for use of our software.
Reference:
Larocque, G., Dutilleul, P., Pelletier, B., and Fyles, J.W. 2007. Characterization and quantification of uncertainty in coregionalization analysis. *Mathematical Geology* 39:263–288.

Complementary note 3

In Modt-LMC, second-order stationarity is assumed. In the presence of non-stationarity at first order (i.e., the mean is not constant over the field), the analysis should be performed on the residuals obtained after removing a drift appropriately estimated (e.g. by Estimated Generalized Least Squares).

Appendix

Results for Nugget Effect (2 structures)

	Sill_direct1	Sill_direct2	Sill_cross	Range	M_s_hat	r_s	t_obs	P_value
Trees:pH	0.554	0.249	0.053	85.900	55.825	0.144	1.066	0.291
Trees:Mg	0.537	0.296	0.039	80.900	53.546	0.097	0.703	0.485
Trees:SOM	0.583	0.481	0.005	95.900	100.682	0.009	0.092	0.927
Trees:N	0.513	0.569	-0.078	73.400	71.866	-0.145	1.227	0.224
Trees:E_M	0.608	0.000	0.000	105.900	NaN	NaN	NaN	NaN
Trees:FD	0.576	0.941	0.080	93.400	154.905	0.109	1.354	0.178
Herbaceous:pH	0.462	0.274	0.083	93.400	54.373	0.233	1.731	0.089
Herbaceous:Mg	0.462	0.352	0.027	93.400	62.693	0.067	0.527	0.600
Herbaceous:SOM	0.470	0.500	0.002	103.400	90.116	0.005	0.048	0.962
Herbaceous:N	0.456	0.664	-0.106	90.900	93.690	-0.193	1.885	0.063
Herbaceous:E_M	0.471	0.000	0.000	105.900	NaN	NaN	NaN	NaN
Herbaceous:FD	0.473	0.986	-0.047	108.400	157.011	-0.068	0.850	0.396

Results for Spherical Model (2 structures)

	Sill_direct1	Sill_direct2	Sill_cross	Range	M_s_hat	r_s	t_obs	P_value
Trees:pH	0.421	0.751	0.008	85.900	37.266	0.014	0.085	0.933
Trees:Mg	0.432	0.752	0.206	80.900	37.175	0.361	2.296	0.028
Trees:SOM	0.400	0.542	0.162	95.900	29.045	0.348	1.929	0.064
Trees:N	0.446	0.493	0.195	73.400	28.582	0.417	2.364	0.026
Trees:E_M	0.382	1.020	-0.211	105.900	37.415	-0.338	2.139	0.039
Trees:FD	0.405	0.052	-0.068	93.400	6.200	-0.471	1.093	0.333
Herbaceous:pH	0.571	0.758	-0.036	93.400	41.275	-0.054	0.341	0.735
Herbaceous:Mg	0.571	0.738	0.098	93.400	39.605	0.152	0.940	0.353
Herbaceous:SOM	0.593	0.536	0.153	103.400	32.806	0.272	1.568	0.127
Herbaceous:N	0.571	0.410	0.206	90.900	28.948	0.426	2.447	0.021
Herbaceous:E_M	0.603	1.020	-0.030	105.900	45.914	-0.038	0.255	0.800
Herbaceous:FD	0.610	0.000	0.000	108.400	NaN	NaN	NaN	NaN

Results for Nugget Effect (1 structure)

	Sill_direct1	Sill_direct2	Sill_cross	Range	M_s_hat	r_s	t_obs	P_value
Herbaceous:FD	0.473	0.986	-0.047	NaN	322.551	-0.068	1.223	0.222

Results for Spherical Model (1 structure)

	Sill_direct1	Sill_direct2	Sill_cross	Range	M_s_hat	r_s	t_obs	P_value
Trees:E_M	0.382	1.020	-0.211	105.900	293.513	-0.338	6.136	0.000
Herbaceous:E_M	0.603	1.020	-0.030	105.900	293.513	-0.038	0.657	0.512