CRAD User's Guide

Version 2.0 "for stand-alone application", February 2011

This documentation explains how to use the computer program called "CRAD", which implements the method of coregionalization analysis with a drift presented in the following two articles.
References:
Pelletier, B., Dutilleul, P., Larocque, G., and Fyles, J.W. 2009a. Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 1. Estimation of drift and random components. *Environmental and Ecological Statistics* 16:439–466.
Pelletier, B., Dutilleul, P., Larocque, G., and Fyles, J.W. 2009b. 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.

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 **CRAD_standalone.zip** on your computer and unzip itscontent to a folder of your choice.

3- Launch the program by clicking on **CRADui.exe**. Two windows should then appear: a user interface, to run the CRAD 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 provide/perform the following inputs/tasks:

1- A tab-delimited text file[1] 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.

[1] If creating the text file from an Excel spreadsheet, make sure that there are no empty lines or extra spaces.

- In the fourth and following columns, the corresponding observations for the *p* variables. For redundancy analysis (RDA), the dependent variables must precede the independent variables.

Note: After you have selected the data file, the total number of variables (as counted by the program) will appear.

2- A set of characters (prefix) that will be used to identify the output files (*.txt) generated by the program and saved in the Matlab folder.

3- Indicate whether the data should be submitted to a Box-Cox transformation or not. 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 CRAD, the coefficient λ is selected among values from -2.5 to 2.5, by steps of 0.1.

Note: A standardization of the data to a zero mean and a variance of one is always performed *after* the application of the Box-Cox transformation.

4- Indicate whether the data should be standardized to a zero mean and a variance or one, or not. If you have selected "Yes" for the Box-Cox transformation, this box will disappear.

5- Indicate your choice of drift estimation procedure:

- Global procedure using a polynomial of order 3;
- Local procedure using a moving window and a polynomial of order 1 (i.e., a plane) inside the window.

6- Give a value for the threshold (i.e., a percentage expressed relative to the total variance of each variable). Below this value, a scale component (i.e., large-scale spatial, small-scale spatial, small-scale non-spatial) will be considered missing. In Phase 1, a variable with a pseudo-variance below that percentage is considered to have no large-scale spatial component. In Phase 2, the 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.

7- Indicate whether you want to restrict the data analysis to Phase 1 of CRAD (i.e., the univariate phase).

8- If you choose to perform Phase 2 of CRAD, you need to select the type of multivariate analysis that you want to be performed:

- Principal Component Analysis (PCA);

- Redundancy Analysis (RDA).

9- If you choose RDA, then you need to specify the number of dependent and independent variables.

Output files

The results of the data analysis are included in text files saved in the Matlab folder. The output files whose content is described below can be identified by the set of characters (prefix) defined by the user in the program interface.

*outputprefix*_Boxcox.txt: Comprises the coefficients obtained for and used in the Box-Cox transformation of each of the *p* variables.

*outputprefix*_Estimated_drifts.txt: The first two columns comprise the geographical coordinates. The third and following columns comprise the values of estimated drifts for the *p* variables.

*outputprefix*_Estimated_residuals.txt: The first two columns comprise the geographical coordinates. The third and following columns comprise the values of residuals for the *p* variables.

*outputprefix*_Direct_variograms.txt: On the first and second rows are indicated the number of pairs of observations per distance class and the mean distance value per class, respectively. On the following rows are reported the ordinates of experimental variograms of the *p* variables.

*outputprefix*_Results_Phase1.txt: Comprises the results of Phase 1 (univariate phase) of CRAD. They include the nugget effect (column 1), the estimated sill of the spherical model (column 2), the pseudo-variance of the estimated drift (column 3), the estimated range of the spherical model (column 4), and the size of the window used in the local drift estimation procedure (column 5) for each of the *p* variables (rows).

The following output files are generated only if Phase 2 of CRAD is performed.

*outputprefix*_All_variograms.txt: On the first and second rows are indicated the number of pairs of observations per distance class and the mean distance value per class, respectively. On the following rows are reported the ordinates of the $p(p+1)/2$ direct and cross experimental variograms. The names of variables for which experimental direct and cross variograms were computed are given in the second and third columns.

*outputprefix*_Correlations_total: Comprises the p x p correlation matrix computed from the raw data.

*outputprefix*_Correlations_nugget: Comprises the p x p matrix of structural correlations estimated for the nugget effect.

*outputprefix*_Correlations_spherical: Comprises the p x p matrix of structural correlations estimated for the spherical-model.

*outputprefix*_Correlations_drift: Comprises the p x p matrix of pseudo correlations computed from the drift estimates.

*outputprefix*_Sills_nugget: Comprises the p x p matrix of sill estimates (i.e. the estimated coregionalization matrix) for the nugget-effect structure.

*outputprefix*_Sills_spherical: Comprises the p x p matrix of sill estimates (i.e. the estimated coregionalization matrix) for the spherical-model structure.

*outputprefix*_Pseudovar_drift: Comprises the p x p pseudo variance-covariance matrix computed from the drift estimates.

*outputprefix*_Structural_variance: Comprises the respective amounts of variation associated with the nugget-effect structure (column 1), the spherical-model structure (column 2) and the drift (column 3) for each of the p variables (first p rows) and for all the dependent variables and all the independent variables (last two rows).

*outputprefix*_ScoresY_total: In PCA, it comprises the scores of the p variables on biplot axes (i.e., the principal components). In RDA, it comprises the scores of the dependent variables on biplot axes. The analysis is performed on the ‘total variables’ in the terminology of Pelletier et al. (2009a, 2009b) (i.e., before decomposition into scale components).

*outputprefix*_ScoresY_nugget: In PCA, it comprises the scores of the p variables on biplot axes (i.e., the principal components). In RDA, it comprises the scores of the dependent variables on biplot axes. The analysis is performed for the nugget-effect structure.

*outputprefix*_ScoresY_spherical: In PCA, it comprises the scores of the p variables on biplot axes (i.e., the principal components). In RDA, it comprises the scores of the dependent variables on biplot axes. The analysis is performed for the spherical-model structure.

*outputprefix*_ScoresY_drift: In PCA, it comprises the scores of the *p* variables on biplot axes (i.e., the principal components). In RDA, it comprises the scores of the dependent variables on biplot axes. The analysis is performed on the estimated drifts.

The following output files are generated only if an RDA is performed.

*outputprefix*_R2.txt: Comprises the coefficients of determination for the classical regression/RDA (column 1) and regionalized regressions/RDAs for the nugget-effect structure (column 2), the spherical-model structure (column 3) and the drift (column 4). Results are presented for each dependent variable separately and for the dependent variables all together (last row).

*outputprefix*_ScoresX_total: In an RDA, it comprises the scores of the independent variables on biplot axes. The analysis is performed on the 'total variables'.

*outputprefix*_ScoresX_nugget: In an RDA, it comprises the scores of the independent variables on biplot axes. The analysis is performed for the nugget-effect structure.

*outputprefix*_ScoresX_spherical: In an RDA, it comprises the scores of the independent variables on biplot axes. The analysis is performed for the spherical-model structure.

*outputprefix*_ScoresX_drift: In an RDA, it comprises the scores of the independent variables on biplot axes. The analysis is performed on the estimated drifts.

Complementary note

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.