ModF_RDA User's Guide

Version 1.0 "for Matlab users", June 2017

This documentation explains how to use the computer program **ModFRDA**, which implements two modified *F*-tests (in addition to the classical one) to assess the significance of the average R^2 (i.e., the average proportion of variance in the first set of variables that is reproducible by linear prediction from the variables in the second set) in a redundancy analysis (RDA) with spatial data; from Dutilleul and Pelletier (2017).

Reference:

Dutilleul, P. and Pelletier, B. 2017. A valid parametric test of significance for the average R^2 in redundancy analysis with spatial data. *Spatial Statistics* 19:21–41.

Installation

1- Download a copy of the zipped folder **ModF_RDA.zip** on your computer, and unzip it in a relevant place into a folder with the name of your choice. That folder contains Matlab files in .p format and two ASCII files (.txt) with datasets that can be used as examples when learning how to use **ModFRDA**, e.g. for the preparation of the input file.

2- The 'master file' being **ModFRDA**, the Matlab user must type a line of command that matches the line below, in syntax and structure, in the Command Window for his/her application:

ModFRDA('inputfilename.txt', ny, nx, standardization, 'outputfilename')

where

- 'inputfilename.txt' is the name of the ASCII file containing the ID number of the sampling location (first column), followed by the spatial coordinates in an appropriate unit (e.g. m), the (numerical) data for the criterion variables (to explain) first and then the predictor (explanatory) variables. The names of variables (with the exception of that of the ID number of the sampling location) must appear (as headings, not necessarily aligned with the data below) on the first row.
- ny is the number of criterion variables (denoted "*p*" in Dutilleul and Pelletier, 2017)
- nx is the number of predictor variables (or "*q*")
- standardization = 1 means the data for each variable will be centered to a zero mean and a variance of one; = 0, no standardization is requested.
- 'outputfilename' will be used as prefix in the names of the two output files, one with extension .txt (ASCII) and the other .xlsx (Microsoft Excel).

3- All the Matlab files contained in the folder **ModF_RDA** once unzipped are in .p format, and are files that the authors have created for this project. Of course, other Matlab functions

are called by their .p files, so the user needs to work with Matlab; the authors worked with Matlab R2015a for the preparation of their .p files.

4- The folder **ModF_RDA** contains two data files for 'training' (with same structure as the data files used as input for applications with the CRAD software, also posted on http://environmetricslab.mcgill.ca):

- Dataset_Ex1.txt: Example 1 in the *Spatial Statistics* article (n = 275);
- Dataset_Ex2.txt: Example 2 in the same article (n = 314).

5- Results are saved in two output files with different formats (see above), but with same content in terms of results of tests of significance for the average R^2. That is, each output file (in its own format) contains the results of a total of five tests: Miller's classical F-test; two Option 1 modified F-tests (with LMRs and Model-Free) and two Option 2 modified F-tests (with LMC and Model-Free).

6- For methodological aspects (e.g. fitting of a linear model of coregionalization by Estimated Generalized Least Squares, definition of distance classes), please see Dutilleul and Pelletier (2017) and the references therein, as well as the notes below.

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 some of the modified F-tests of Dutilleul and Pelletier (2017) is limited to two structures: a nugget effect and a spherical model.

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 the modified *F*-tests for the average R^2 in an RDA with spatial data (as in several other modified tests of significance of correlation coefficients with spatial data), second-order stationarity is assumed. In the presence of non-stationarity at first order (i.e., the mean is not constant over the sampling domain), the analysis should be performed on the residuals obtained after removing a drift estimated appropriately (e.g. by Estimated Generalized Least Squares).