

Spatially-explicit generation of ecological data – Program SimSSD4 user's guide

Pierre Legendre
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec H3C 3J7, Canada

May 2005

E-mail: **Pierre.Legendre@umontreal.ca**

What does program SimSSD4 do?

The program SimSSD creates pairs of variables on a surface whose size is determined by the user. A variable observed at localities on a geographic surface is called a “surface” for short. The first surface corresponds to an explanatory variable, called “environmental variable”. The second surface is a response variable. Variable generation is done by a spatially-explicit process.

The statistical theory behind this program is described in the “SIMULATION STUDY: CANONICAL ANALYSIS VERSUS MANTEL TEST” section of the present paper. Fig. 1 illustrates the process with (a) the construction of an environmental surface consisting of a deterministic structure plus spatial autocorrelation in the environmental data plus normal error; (b) the construction of a response surface (e.g., a species) partly determined by the environmental surface described in (a) plus spatial autocorrelation in the species plus normal error; and (c) a sampling design applied to the two surfaces, simulating the action of sampling the species and simultaneously observing the environmental variable's values. The simulated surface contained $20 \times 20 = 400$ points in Fig. 1; it contained $100 \times 100 = 10000$ points in the simulation study reported in the present paper. The sampling design was a cross in Fig. 1; it was a regular grid with $10 \times 10 = 100$ points in the present study.

The program was first written as part of the realizations of the Working Group “Integrating the Statistical Modeling of Spatial Data in Ecology”¹. Results have been published, investigating the consequences of spatial structures for the design of ecological field surveys (Legendre et al. 2002) and field experiments (Legendre et al. 2004). The version presented here was redesigned to generate a large number of pairs of surfaces and write them out to output files. These files were then used as input for the calculation of the partial canonical analyses and partial Mantel tests reported in this paper.

The program was written in FORTRAN because it uses the subroutine SGSIM from the geostatistical software library GSLIB (Deutsch and Journel 1992), which is in FORTRAN. FORTRAN 77 was chosen because free-access compilers are available from the GNU PROJECT; see URL <http://www.gnu.org/>.

Parameter input file

The simulations produced by the program are controlled by a line of parameters included in a parameter file. The instructions for setting up the parameters for a run are the following. They are also found in the program listing.

nsim = number of simulations (negative number: debug is on)

nx, ny: size of the surface in the X and Y directions (discrete pixels)

¹ This working group was supported by the *National Center for Ecological Analysis and Synthesis* (NCEAS), a center funded by NSF (Grant #DEB-94-21535), the *University of California at Santa Barbara*, and the *State of California*.

deter = type of deterministic structure in resource variable

0 = no deterministic structure; only $N(0,1)$ error

1 = gradient from top (low values) to bottom (high values) + $N(0,1)$

2 = gradients in two directions + $N(0,1)$

3 = one big patch in center of the field + $N(0,1)$

4 = 4 waves + $N(0,1)$

5 = two zones separated by a sharp step + $N(0,1)$

height = maximum value of the deterministic structure

beta = transfer (or regression) parameter

Spatial autocorrelation (SA) in resource: parameters of the spherical variogram

nug1 = nugget value (usually 0); nug1 = -9 if no SA in resource

range11 = range value (0 if none) in the horizontal direction

range12 = range in vertical dir. (different from range11 if anisotropy)

Spatial autocorrelation (SA) in response: parameters of the spherical variogram

nug2 = nugget value (0 if none); nug2 = -9 if no SA in response

range21 = range value (0 if none) in the horizontal direction

range22 = range in vertical dir. (different from range11 if anisotropy)

varnor = variance of the normal error term for the response variables

isamp = sampling design:

0 = no sampling will take place (for problem 2 only)

1 = simple random sampling

2 = systematic (check first if compatible with grid size)

3 = aggregates of 5 sampling units in a systematic pattern

4 = horizontal transect with a single lag

5 = horizontal transect with 2 lags

6 = vertical transect with a single lag

7 = vertical transect with two lags

8 = field divided vertically in 2 strata; random within stratum

9 = cross

10 = 2 parallel vertical transects distributed evenly along X

11 = 3 parallel vertical transects distributed evenly along X

12 = 2 parallel vertical transects distributed at random along X

13 = 3 parallel vertical transects distributed at random along X

nsamp = number of sampling units, e.g. 25, 49, 50, 100, 101

nsp1 = number of generated species, not related to an environmental variable

nsp2 = number of generated species that are related to an environmental variable. “nsp2” is also the number of environmental variables to be generated.

Note to users: some of the sampling designs (“isamp”) require the number of samples (“nsamp”) to obey certain properties:

- systematic: ‘nsamp’ should be a square number.
- aggregated: ‘nsamp’ should be a multiple of 5, and the number of aggregates of 5 points should be a square.
- cross: ‘nsamp’ should be an odd number (only 1 point in the centre).
- transects with 2 lags: ‘nsamp’ should be a multiple of 2.
- 2 transects: ‘nsamp’ should be a multiple of 2.
- 3 transects: ‘nsamp’ should be a multiple of 3.

Example of a line of parameters

1000 100 100 2 5 0.50 -9 0 0 0 15 15 1 2 100 5 5

The parameters correspond to the following list, described above:

nsim, nx, ny, deter, height, beta, nug1, range11, range12, nug2, range21, range22, varnor, isamp, nsamp, nsp1, nsp2

The parameter line means that 1000 sets of simulated data files will be produced. The simulations will take place on a (100 × 100) grid. The environmental variable (resource) will contain a deterministic gradient running diagonally across the field (maximum height: 5). The environmental variable will be transferred to the species with a transfer parameter of 0.5. There will be no autocorrelation in the environmental variable, but there will be in the species (controlled by a variogram with a range of 15 pixels in both the X and Y directions). The normal error for the species (response variables) will have a variance of 1. The sampling design will be a systematic square grid (design #2) with 100 points. Each simulated data set will contain 5 species unrelated to the environmental variables plus 5 that are related to them.

Five different lines of parameters were used for the simulations reported in Table 1 of the paper:

1000 100 100 0 0 0.00 -9 0 0 0 15 15 1 2 100 5 5 ¹	Table 1A
1000 100 100 0 0 0.50 0 15 15 0 15 15 1 2 100 5 5	Table 1B
1000 100 100 0 0 0.25 0 15 15 0 15 15 1 2 100 5 5	Table 1C
1000 100 100 2 5 0.50 -9 0 0 0 15 15 1 2 100 5 5	Table 1D
1000 100 100 2 5 0.25 -9 0 0 0 15 15 1 2 100 5 5	Table 1E

¹ If “nsp1” = 10 and “nsp2” = 0 had been used, the same results in terms of species would have been obtained (10 species unrelated to the environmental variables), but no environmental variables would have been generated.

Output files

The output files are the following:

- **Coord.txt** contains the X and Y coordinates of the points selected by the sampling design.
- **Species.txt** contains successive tables of generated species data, either integers or 0-1. Each table has “nsamp” rows and “nsp1” + “nsp2” columns. If “Print simulation number and random multipliers” has been requested, each table is preceded by a line containing the following information: number of objects (i.e., number of rows of the generated data table), number of species, simulation number (from 1 to “nsim”), and list of the random multipliers for species. The random multipliers will be multiplied by the value of the “Standard deviation of the normal species multipliers? (e.g., 1, 1.5, 2)” before being applied to individual species during data generation.
- **Envir.txt** contains successive tables of generated environmental data (real numbers). Each table has “nsamp” rows and “nsp1” + “nsp2” columns. If “Print simulation number and random multipliers” has been requested, each table is preceded by a line containing the following information: number of objects (i.e., number of rows of the generated data table), number of environmental variables, and simulation number (from 1 to “nsim”).

Disclaimer

This program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers may use this program for scientific purposes, but the source code remains the property of Pierre Legendre. Users of the program may refer to the present user’s manual as follows:

Legendre, P. 2005. Spatially-explicit generation of ecological data – Program SimSSD4 user’s guide. In Supplement 1 of Legendre, P., D. Borcard and P. R. Peres-Neto (Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, 2005). Ecological Archives E000-000-E1. 7 pages.

Technical notes

The program is available in a variety of forms:

- FORTRAN source code for Mac OS X and Windows (file SimSSD4_g77.FOR), which can be compiled using the g77 FORTRAN compiler. The user may modify the Parameter statement at the beginning of the program, which fixes the maximum size of the simulation field.
- Compiled version for PowerPC processors for Macintosh OS X (file SimSSD4). The maximum size of the sampling field is 100 x 100 pixels.
- Compiled version for IBM compatible PC (file SimSSD4.exe). The maximum size of the sampling field is 100 x 100 pixels. The executable file is a Win32 “console” executable, not a DOS executable. Therefore it cannot run under plain DOS, nor in a DOS window under Windows 3.x. It can only run in Windows 95/98, Windows NT, or Windows XP consoles.

References

- Deutsch, C. V., and A. G. Journel. 1992. GSLIB – Geostatistical software library and user's guide. Oxford University Press, New York, NY.
- Legendre, P., M. R. T. Dale, M.-J. Fortin, J. Gurevitch, M. Hohn and D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* 25: 601-615.
- Legendre, P., M. R. T. Dale, M.-J. Fortin, P. Casgrain and J. Gurevitch. 2004. Effects of spatial structures on the results of field experiments. *Ecology* 85: 3202-3214.
- Legendre, P., D. Borcard and P. R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs* (in press in the "Concept & Synthesis" section).

Appendix 1 – Test run

In this example, the program is used to obtain 10 sets of simulated data tables, using parameters read from a file called "File_of_parameters.txt".

SimSSD: Simulation of Spatially Structured Data

NCEAS Working Group "Integrating the Statistical Modeling of Spatial Data in Ecology"

SimSSD4 for Legendre, Borcard, Peres-Neto simul.

(c) Pierre Legendre, 2004
 Departement de sciences biologiques
 Universite de Montreal

Name of input file with parameters?

File_of_parameters.txt

File of parameters: File_of_parameters.txt

Transformation of species data:

(0) Standardize only; no further transformation

(1) $y' = \text{int}(\exp(y))$

(2) $y' = \text{int}(\sqrt{\exp(y)})$

(3) $y' = 0$ or 1 (binary)

1

Standard deviation of the random normal
 species multipliers? (e.g., 1, 1.5, 2)

1.5

Output file for coordinates: Coord.txt

Output file for species data: Species.txt

Output file for environmental data: Envir.txt

Type a small positive integer (i.e. 1-10) to offset
 the random number generator for generation of data

5

Print simulation number and random multipliers?

(0) No - (1) Yes

1

Simulation No. 1

Simulation No. 2

Simulation No. 3

Simulation No. 4

5 simulations only were produced in this test run

End of the program.

Transformation of species data: four options are available in the program. Two of them, #1 and #3, have been used to generate the simulated data analyzed in the paper. They are further described in paragraph 3 of subsection “Data generation” of the “Simulation study: canonical analysis *versus* Mantel test” section of the paper. That paragraph also indicates that 1.5 was the multiplier for the standard deviation of the random normal deviates used to generate the species data.

Type a small positive integer (i.e. 1-10) to offset the random number generator for generation of data: different sets of generated species and environmental data can be obtained by using different values of this parameter and repeating the computer run using the same line of parameters.

Print simulation number and random multipliers? If requested, a line will be printed before each table of simulated data in files Species.DAT and Envir.DAT. That information also acts as a divider between successive data tables. See description in the next section.

Appendix 2 – Unix and Windows user’s notes

The Unix (MacOS X) and DOS versions of this program were built with g77, the GNU FORTRAN compiler.

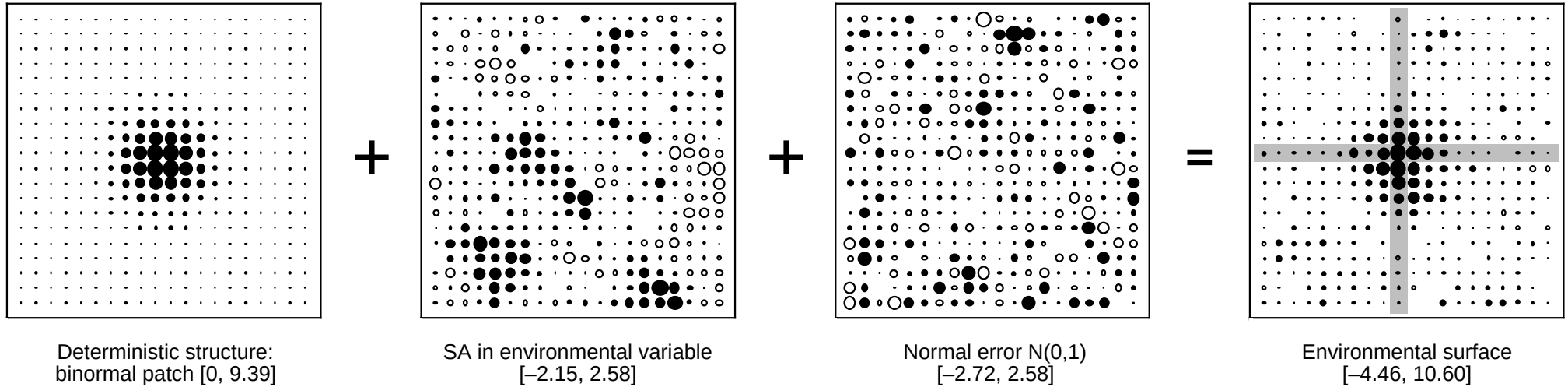
Click on the program’s icon to start it, then answer the questions that appear on the screen. You will be asked to provide the name of the file containing the data. In Windows, make sure the file name is DOS-compatible (maximum 8 characters, no space, etc.). If the input file has an extension, like “txt”, you have to type it when you provide the name of the input file, even if the extension does not appear in the file list.

In Unix and Windows, the files “Coord.DAT”, “Species.DAT”, and “Envir.DAT” produced by the program cannot be deleted by the FORTRAN program during a following run. If, after launching, the program ends abruptly and a message is displayed, such as:

In Mac OS X: open: 'new' file exists
 apparent state: unit 7 named Coord.DAT
 last format: list io
 lately reading direct formatted external IO
 Abort trap
 logout

In Windows: Abnormal program termination following which the console window shuts off,
 this means that a file called “Coord.DAT” already exists in the current directory. Rename or remove that file before running the program again.

Construction of the environmental surface



Construction of the response surface

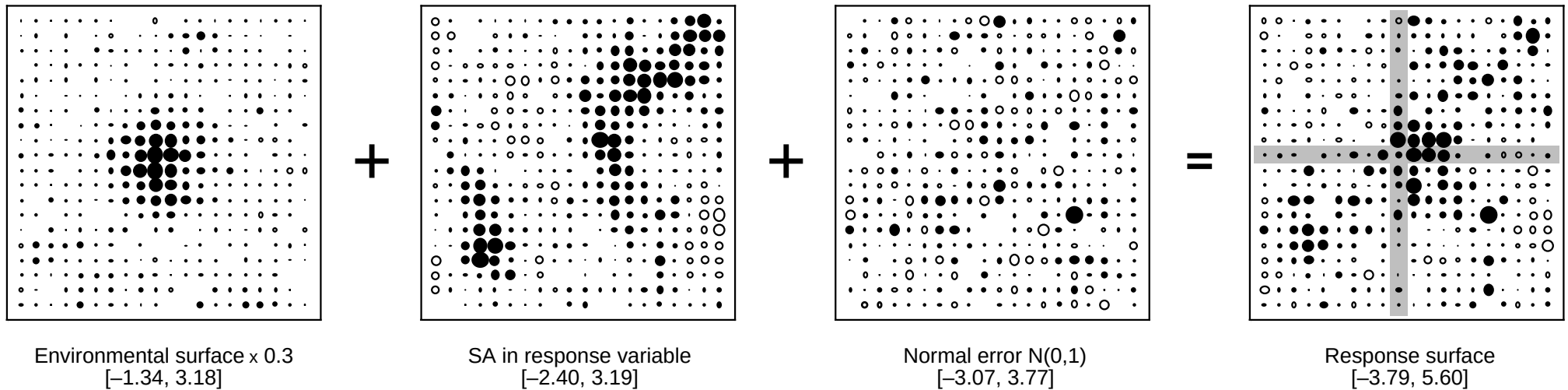


Fig. 1 (from Legendre et al. 2002). Construction of the environmental and response surfaces during the simulations. In this example, the field has 20 points wide and 20 points high. The environmental variable is the sum of a deterministic structure plus a spatially autocorrelated error component (SA) generated using a spherical variogram model with range of 5 in both directions. The response variable is the sum of three separate effects: the influence of the environmental variable multiplied by an effect-size parameter, spatial autocorrelation (SA) in the response variable generated using a spherical variogram model with range of 5 in both directions, and a spatially unstructured random error component $N(0,1)$. Bigger empty circles represent larger negative values; bigger black circles, larger positive values. The range of values in each graph is shown in brackets underneath. The sampling design is a cross in this example; the 39 sampled points are shown in grey.