Samβada: User manual

Sylvie Stucki and Stéphane Joost[*]

November 22, 2019

Version 0.8.3

Contents

[*] Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland
Webpage: `lasig.epfl.ch/sambada`
Contact: sylvie.stucki@a3.epfl.ch, stephane.joost@epfl.ch

1 What is Samβada?

Samβada is an integrated software for landscape genomic analysis of large datasets. The key features are the study of local adaptation in relationship with environment and the measure of spatial autocorrelation in environmental and molecular datasets. When studying local adaptation, Samβada uses logistic regressions to estimate the probability that an individual carries a specific genetic marker given the habitat that characterises its sampling site. Multivariate analysis, i.e. including several predictors variables simultaneously in the models, enables the user to assess the effect of a combination of environmental variables or to include prior knowledge, e.g. the population structure, in the analysis. Regarding spatial statistics, Samβada measures spatial autocorrelation with Moran's I and local indicators of spatial association, in order to assess whether the observed data in each location depends on the values in the neighbouring locations. Underlying models are kept simple to put emphasis on process efficiency and user-available options.

2 Installation

Software, source code, documentation and examples are available on our webpage `lasig.epfl.ch/sambada` and on GitHub `https://github.com/Sylvie/sambada/releases`.

Executable binaires

Compiled versions of Samβada are already packaged for Windows, macOS and Ubuntu.

Windows and Ubuntu

Download and expand the archive at the location of your choice.

macOS

Samβada requires the GNU Compiler Collection (GCC) version 7 to be installed on the computer. Here are the installation steps using the package manager `Homebrew`[1]:

1. Open a Terminal prompt (located at `/Applications/Utilities/Terminal.app`).
2. Browse to `Homebrew`'s home page (`https://brew.sh/`) and copy the installation command into the Terminal. At the time of writing, the command is:

[1] `https://brew.sh/`

```
/usr/bin/ruby -e "$(curl -fsSL \
https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

Press "Enter" and follow the instructions.

3. Install GCC with the Terminal command: `brew install gcc@7`.

4. Download and expand Samβada's archive at the location of your choice.

Compiling from sources

Samβada can be built from source using the GNU Build System (a.k.a Autotools) or CMake and GCC (in versions 4.4.7 to 7.x). We noticed some incompatibilities with GCC version 8 and with Clang, which are under investigation. The distribution archive `sambada-0.8.3.tar.gz` (or `sambada-0.8.3.zip`) contains all files needed for the compilation.

Building with `Make`

Provided that `make` is installed on your machine, the simplest build process consists in:

- downloading and expanding the distribution archive `sambada-0.8.3.tar.gz` (or `sambada-0.8.3.zip`) at the location of your choice;

- creating a sub-folder for separating the compiled objects from the source code, e.g. `sambada-0.8.3/build`

- running `../configure && make` in the folder `sambada-0.8.3/build` for building with your default version of GCC;

- or running `../configure CC=gcc-7 CXX=g++-7 && make` for building with GCC 7.

The executables are placed in the subfolder `binaries`.

Some further information:

- The folder containing all artefacts for the end-user can be built with `make binary-archive`.

- The script `configure` provides some specific options:

  - `sambadahostsystemname` to choose the system name used for naming the artefacts' folder;
  - `--disable-manual` to disable the compilation of the manual. A text file containing a link to the online documentation is created instead.

Building with CMake

Samβada's binaries can also be built with CMake. The process is the same except for the third step:

- running `cmake .. && make` in the folder `sambada-0.8.3/build`

Please note that this secondary build system might be subject to change.

Documentation

Software usage is explained in this manual. Theoretical background is covered by Samβada's release article Stucki et al. (2016) and Joost et al. (2007). Extended information on methods and implementation can be found in Stucki (2014) (in French).

Examples

Three sample cases are provided with the software.

DataFromManual is a tiny set of six samples with fives environmental variables and seven molecular markers. The examples from this manual are build on this dataset in order to illustrate data format, analysis workflow and distributed computing.

RandomSample is a random set of 100 georeferenced points with two environmental variables and a molecular marker. The first environmental parameter is random, while the second one is correlated to longitude in order to provide some spatial autocorrelation. The molecular marker is random.

SubsetCattleSNP contains 386 SNPs from Ugandan cattle (Vajana et al., 2018). They are already recoded for Samβada, so there is three binary markers per loci. In this case, the spatial autocorrelation takes some time to compute.

AnalysisWithPopulationStructure contains 10 SNPs from Ugandan cattle (Vajana et al., 2018). They are already recoded for Samβada, so there is three binary markers per loci. This example illustrates how to include variables representing the population structure in the analysis.

Provided you have the file tree shown on fig. 1, if you launch a shell[2] and navigate to the folder `random-data` or `subset-cattle-SNP`, examples can be run using:

```
../../binaries/sambada.exe param-random-sample.txt random-sample.txt
```

or

[2] Command Line or PowerShell on Windows, Terminal on macOS and Linux, see p. 32.

```
../../binaries/sambada.exe param-cattle.txt cattle-env.csv cattle-mark.txt
```

on Windows, and

```
../../binaries/sambada param-random-sample.txt random-sample.txt
```

or

```
../../binaries/sambada param-cattle.txt cattle-env.csv cattle-mark.txt
```

on Linux and macOS.

Note on programs' names In the main text of this manual, the programs are identified with proper nouns (`Sambada`, `Supervision` and `RecodePLINK`). However the names of the compiled programs are written in lowercase to ease typing in the Console and Terminal. This alternate spelling is also used in the command examples of this manual. Please note that Windows uses the suffix ".exe" for program names. These considerations lead to the following naming conventions:

Name in main text	Software name on	Windows	Linux & macOS
`Sambada`	⇒	`sambada.exe`	`sambada`
`Supervision`	⇒	`supervision.exe`	`supervision`
`RecodePLINK`	⇒	`recode-plink.exe`	`recode-plink`

In case of doubt, please use the names of the programs found in the directory `binaries/`. Hint: Most command line interpreters enable auto-completion of names by hitting the "TAB" key.

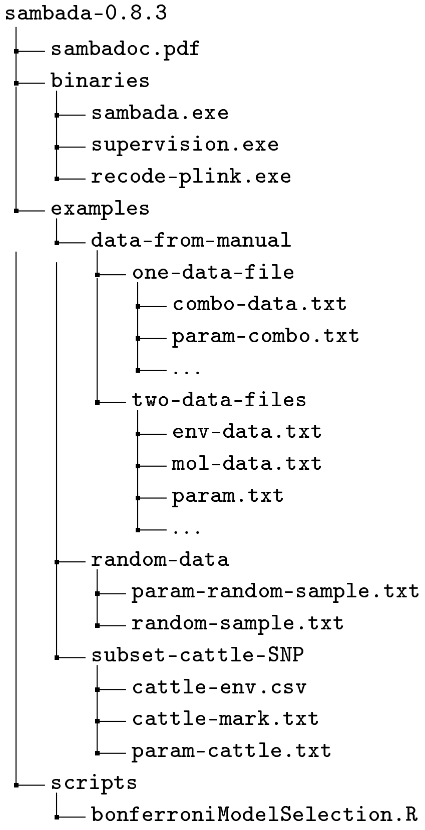


Figure 1 – Suggested file tree to run the examples.

3 Analysis overview

Three programs are available:

Samβada processes univariate and multivariate logistic models for the landscape genomics analysis and optionally measures the spatial autocorrelation in environmental and molecular datasets;

Supervision can split molecular data in blocks in order to run the analysis on several computers, and can merge the results afterwards;

RecodePLINK can translate molecular data from PLINK's to Samβada's format.

The user must provide Samβada with a parameter file to set up the analysis as well as environmental and molecular data. The workflow is summarised on fig. 2 and the data format is presented in the next section.

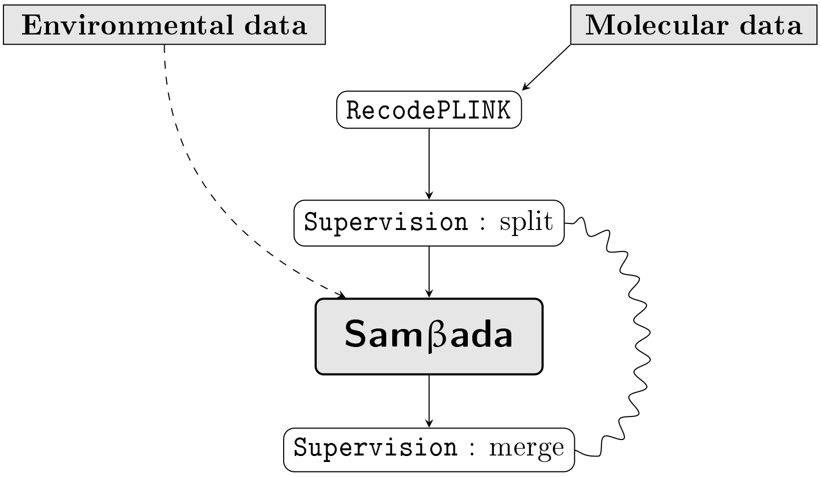


Figure 2 – Workflow of analysis. Rectangles stand for data and round-cornered figures stand for programs. Grey elements are mandatory and white ones are optional. Samβada computes correlative models and spatial autocorrelation. The two other features are optional: Supervision enables distributed computing while RecodePLINK transforms .ped/.map files to comply with Samβada's format. Arrows show processing order; Samβada input consists in environmental data (dashed line) and molecular data (solid line). The zigzag line indicates that Supervision is used before and after the main analysis.

4 Data format

Samβada's input consist of molecular and environmental data. They can be provided as a single or two separate files. Files may have any name and extension. Each line provides information for an individual, each column contains an environmental variable or a binary molecular marker. Examples are provided on fig. 3, 4 and 5. Information about data format and analysis options are specified separately in the parameter file.

Data files are organised as follows: the header line is optional and the column separator is up to the user. Sample names (identifiers) are optional, and some columns may be excluded from the analysis (for instance phenotypical information stored with the environmental data). If there is a single data file, environmental data must be provided in the first columns, and molecular data in the last ones. Sample names and coordinates are considered as environmental data. If data is split between two files, sample must be in the same order in both files. Missing data can be coded as any character string, for instance `NaN` or `?`. Fig. 3 and 4 are examples of environmental and molecular files. Fig. 5 is the combined file for the same data. The files used in this manual are distributed with Samβada. Please refer to `sambada-0.8.3/examples/data-from-manual/`.

NAME	ENV1	ENV2	ENV3	ENV4	ENV5
ID1	46	972	236	230	132
ID2	32	987	238	232	83
ID3	32	987	238	232	83
ID4	32	987	NaN	232	83
ID5	32	987	238	232	83
ID6	35	1021	235	230	87

Figure 3 – Example of environmental file (`env-data.txt`). This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/two-data-files/env-data.txt`.

5 Program use

5.1 Samβada

5.1.1 Input files

The required input for Samβada consists of environmental and molecular data, formatted as explained in sec. 4. A parameter file is needed as well to set up the analysis.

NAME	M4	M7	M8	M9	M16	M17	M18
ID1	1	1	1	0	1	1	1
ID2	0	0	0	0	0	0	NaN
ID3	0	1	0	0	0	0	1
ID4	0	0	1	1	0	1	1
ID5	0	0	1	0	0	1	0
ID6	0	1	0	0	0	1	0

Figure 4 – Example of molecular file (`mol-data.txt`). This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/two-data-files/mol-data.txt`.

NAME	ENV1	ENV2	ENV3	ENV4	ENV5	M4	M7	M8	M9	M16	M17	M18
ID1	46	972	236	230	132	1	1	1	0	1	1	1
ID2	32	987	238	232	83	0	0	0	0	0	0	NaN
ID3	32	987	238	232	83	0	1	0	0	0	0	1
ID4	32	987	NaN	232	83	0	0	1	1	0	1	1
ID5	32	987	238	232	83	0	0	1	0	0	1	0
ID6	35	1021	235	230	87	0	1	0	0	0	1	0

Figure 5 – Example of combined file for environmental and molecular data, corresponding to fig. 3 and 4 (`combo-data.txt`). Environmental data must be provided in the first columns (left part of the tabular), and molecular data in the last columns (right part). The identifier, if any, is considered as an environmental variable. This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/one-data-file/combo-data.txt`.

The parameter file contains one line per parameter, they can be specified in any order. Each line begins with the name of the current parameter, followed by the values separated by spaces.

Some parameters are mandatory, otherwise the entire line may be omitted. Any line beginning with a hash character (#) will be ignored.

Fig. 6 presents a working parameter file.

Example 1 Let's assume we want to analyse `mol-data.txt` (fig. 4) with `env-data.txt` (fig. 3). The simplest parameter file is shown on fig. 7.

Example 2 When environmental and molecular data are provided altogether, slight changes in the parameter file reflect the new data size, see fig. 8.

```
  HEADERS YES
  WORDDELIM " "
* NUMVARENV 24
* NUMMARK 120103
* NUMINDIV 804
  IDINDIV short_name ID_indiv
  SPATIAL longitude latitude SPHERICAL NEAREST 20
  AUTOCORR BOTH MARK 1000
* DIMMAX 1
* SAVETYPE END BEST 0.01
```

Figure 6 – Example of a parameter file for setting up Samβada's analysis. Each line contains an option for the computation, those marked with a sign in the margin are mandatory. The line order has no influence. In this example, the two first lines indicate that data files contain a header line and that columns are separated by spaces. The next lines state the number of environmental variables, the number of molecular markers and the number of individuals/samples. The option `IDINDIV` indicates which columns contain identifiers of individuals; here environmental and molecular data are recorded in two separated files. The next two lines address the measure of the spatial autocorrelation, with the coordinates names, which are spherical, the weighting scheme and the bandwidth; here the 20 nearest neighbours are taken into account. The analysis will include both global and local autocorrelation (`BOTH`) of molecular markers (`MARK`) and the significance will be assessed with 1,000 permutations. The next option means that the detection of selection signatures will rely on univariate models (`DIMMAX 1`). The last line indicates that results will be stored at the end of the process, that only significant models with a significant parent will be stored and that the threshold for significance is set to 1% (before Bonferroni's correction).

```
HEADERS YES
NUMVARENV 6
NUMMARK 8
NUMINDIV 6
IDINDIV NAME
DIMMAX 1
SAVETYPE END ALL
```

Figure 7 – Parameter file to analyse data from fig. 3 and 4 (`param.txt`). `NUMVARENV` and `NUMMARK` count the total number of columns in the data files. This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/two-data-files/param.txt`.

```
HEADERS YES
NUMVARENV 6
NUMMARK 7
NUMINDIV 6
IDINDIV NAME
DIMMAX 1
SAVETYPE END ALL
```

Figure 8 – Parameter file to analyse data from fig. 5 (`param-combo.txt`). Sample names are provided once, thus there is one molecular column less. This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/one-data-file/param-combo.txt`.

5.1.2 Program launch

Samβada is launched as follows if environmental and molecular data are stored in the same file:

```
sambada parameterFile dataFile
```

The command changes slightly if there are two separated input files:

```
sambada parameterFile envFile molecularFile
```

Therefore examples 1 and 2 would be launched with:

```
 sambada param.txt env-data.txt mol-data.txt
```

and

```
 sambada param-combo.txt combo-data.txt
```

respectively. Futher examples are provided on p. 5.

5.1.3 List of options

This section presents the available parameters for Samβada.

The options are presented in the following way: the first line shows the parameter name, whether it is mandatory, the list of possible values (or the expected type in parenthesis) and the default value. The paragraph is completed by a description of the option.

Data files and format

`INPUTFILE` Optional (string) -

Name(s) of the data file(s). If there are two files, indicate first the environmental file then the molecular file.
This information may also be given as an argument to the program.

`OUTPUTFILE` Optional (string) -

Base name(s) for the results file(s). If this option is omitted, the output files will be named after the molecular input file. The different ouput files are distinguished by adding suffixes ("-Out-", "-AS-", . . .), thus the input files are untouched. With this option, the results can be saved in a different folder than the data.

`HEADERS` Optional Yes / No `No`

Presence or absence of variable names. If present, they are read on the first line of the data file, otherwise the environmental variables are labelled P1, P2, P3,. . . and molecular markers are labelled M1, M2, M3,. . . .

`WORDDELIM` Optional (char) ‘ ’

Word delimiter, it must be a single character. This option applies to both molecular and environmental data, while the parameter file is assumed to be space-separated.

`LOG` Optional *1 value, see descr.* `BOTH`

Location of log information.

1	TERMINAL CONSOLE	print the log on the standard output,
	FILE	writes the log in a file with the suffix "-log",
	BOTH	uses both methods.

Data size

`NUMVARENV` Mandatory (int) -

Number of columns with environmental variables, including ignored variables and the column of identifiers (if any). If there is one input file, this counts the number of columns that do not concern molecular data[3]. If there are two input files, NUMVARENV counts the total number of columns in the environmental data file.

`NUMMARK` Mandatory (int) -

Number of columns with molecular data, including ignored data and identifiers if applicable. If there is one input file, this counts the number of columns concerning molecular markers[3]. If there are two input files, NUMMARK counts the total number of columns in the molecular data file.
For distributed analysis, this parameter must indicate the number of markers for the current block of data followed by the total number of markers[4].

`NUMINDIV` Mandatory (int) -

Number of samples included in the data file(s).

Active and inactive columns

`IDINDIV` Optional (string or int) -

Name(s) of the column(s) containing sample identifiers[5]. These optional identifiers are used to label samples in the output files for local spatial autocorrelation (otherwise the line numbers are used). If there are two data files, two names (or numbers) can be provided, the first one is for the environmental data and the second one is for molecular data. The identifier columns are automatically set as inactive. Moreover if this option is specified with two data files, the two identifiers must match on each line. (Sample must be in the same order in each file.)

`COLSUPENV` Optional (string or int) -

Name(s) of the column(s) in the environmental data to be excluded from the analysis[5]. These columns are set as inactive. (For instance, COLSUPENV can indicate columns such as the sampling date or the name of the area.)

`COLSUPMARK` Optional (string or int) -

Name(s) of the column(s) in the molecular data to be excluded from the analysis[5]. These columns are set as inactive.

`SUBSETVARENV`Optional (string or int) -

[3] In this case, NUMVARENV + NUMMARK is the total number of columns in the molecular data file. The first NUMVARENV columns contain environmental data and the following NUMMARK columns hold molecular markers.

[4] The total number of markers is used to compute the Bonferroni correction. Thus ignored data should be excluded of this total.

Name(s) of the column(s) in the environmental data to be included in the analysis while the other columns are set as inactive. The different options cumulate: the active columns are those listed here, minus those specified with COLSUPENV as well as IDINDIV.

`SUBSETMARK` Optional (string or int) -
Name(s) of the column(s) in the molecular data to be included in the analysis while the other columns are set as inactive. The different options cumulate: the active columns are those listed here, minus those specified with COLSUPMARK as well as IDINDIV.

Logistic model and results storage.

`DIMMAX` Mandatory (int) -
Maximum number of environmental variables included in the logistic models. The models with less parameters are computed as well. Use 1 for univariate models, 2 for univariate and bivariates models, . . .
Please refer to section 5.1.5 for more information on including prior knowledge in multivariate models.

`SAVETYPE` Mandatory *3 values, see descr.* -
Saving method and model selection.

1	REAL END	Storage mode: REAL saves results during processing, END writes them upon completion of computation. The second option enables sorting the models according to their Wald scores before saving.
2	ALL SIGNIF BEST	Model selection: ALL saves all models, SIGNIF saves significant models (according to the G and Wald scores) and BEST saves significant models with at least a significant parent. Notice: The option SIGNIF is deprecated and will be removed in a future release.
3	(double)	Significance threshold (p-value) for options SIGNIF and BEST. The Bonferroni correction is applied on this threshold.

Examples:

```
SAVETYPE END BEST 0.01
SAVETYPE END ALL
```

[5] The column numbers replace their names in case there is no header line.

`UNCONVERGEDMODELS` Optional Yes / No `No`

This option controls the back-up of unconverged models. If enabled, these models are saved in a separate file with the suffix "-unconvergedModels".

Population structure

`POPULATIONVAR` Optional *1 value, see descr.* `NONE`

This option indicates whether any explanatory variables represent the population structure. If present, the said population variables must be gathered in the input file, either on the left or on the right side of the group of environmental variables.

Please refer to section 5.1.5 for more information on including prior knowledge in multivariate models.

1	FIRST	population variables are on the left of the environmental variables in the input file,
	LAST	population variables are on the right of the environmental variables in the input file,
	NONE	disables the feature.

Spatial autocorrelation

`SPATIAL` Optional *5 values, see descr.* -

1	(string or int)	Column name (or number) for longitude.
2	(string or int)	Column name (or number) for latitude.
3	SPHERICAL CARTESIAN	Type of coordinates (spherical or projected).
4	DISTANCE GAUSSIAN BISQUARE NEAREST	Type of weighting scheme, see fig. 9
5	(double or int)	Bandwidth of the weighting function • Cases DISTANCE, GAUSSIAN, BISQUARE: Input type is (double). Units are in [m] for SPHERICAL coordinates; for CARTESIAN coordinates, units match those of the samples' positions. • Case NEAREST: Input type is (int).

Example: `SPATIAL X Y CARTESIAN BISQUARE 120`

`AUTOCORR` Optional *3 values, see descr.* -

This entry requires the specification of SPATIAL.

1	GLOBAL LOCAL BOTH	Type of indices to compute: Moran's I for the global spatial autocorrelation, LISA for the local one.
2	ENV MARK BOTH	Variables for the analysis.
3	(int)	Number of permutations for computing the pseudo p-values (default=99).

Example: `AUTOCORR GLOBAL BOTH 999`

`SHAPEFILE` Optional YES / NO `NO`

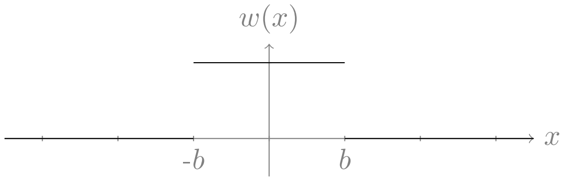
With this option, the LISA are saved as a shapefile (in addition to the usual output). This format is composed of three files: .shp, .shx and .bdf. These files can be loaded together in any GIS software to map the local autocorrelation. This entry requires the specification of SPATIAL.

5.1.4 Output

Samβada produces several output files. To illustrate the naming scheme, let us assume that the molecular data file is named `data.ext`. If the log is saved

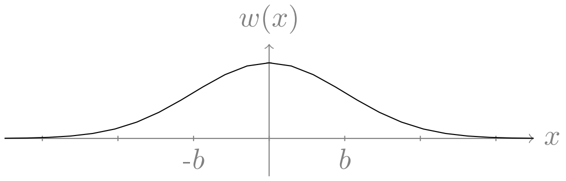
Moving window

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < b \\ 0 & \text{otherwise} \end{cases}$$



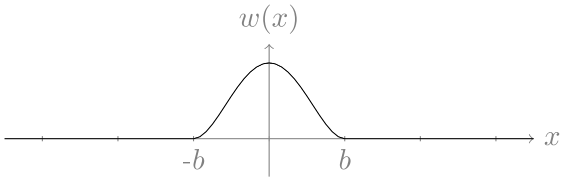
Gaussian kernel

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]$$



Bisquare kernel

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 & \text{if } d_{ij} < b \\ 0 & \text{otherwise} \end{cases}$$



N nearest neighbours

$$w_{ij} = \begin{cases} \frac{1}{N} & \text{if } j \text{ is amongst the } N \text{ nearest neighbours of } i \\ 0 & \text{otherwise} \end{cases}$$

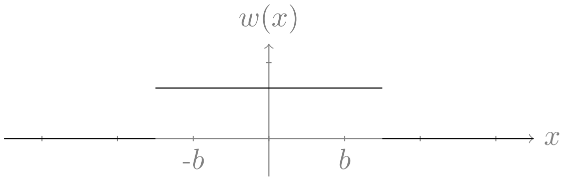


Figure 9 – Weighting schemes available for measuring spatial autocorrelation.

for future reference, the corresponding file is named `data-log.ext`. For logistic regressions, there is one file for constant models, which are not sorted (see fig. 11). There is also one file per distinct number of parameters (univariate, bivariate, trivariate models and so on, see fig. 12). In these files, models are sorted according to their Wald scores. Fig. 10 lists possible errors for logistic models. If the back-up of unconverged models is enabled, the output file is named `data-unconvergedModels.ext`. Results files are named as follows: constant models are saved in the file `data-Out-0.ext`, univariate models in the file `data-Out-1.ext`, bivariate models the file `data-Out-2.ext`, and so on. Fig. 13 provides an overview of the goodness-of-fit statistics computed by Samβada.

0	Success
1	Exponential divergence ($\boldsymbol{X\beta}$ is diverging)
2	Singular matrix (impossible to invert the information matrix)
3	Too large $\boldsymbol{\beta}$ (divergence)
4	Maximal number of iteration number reached without convergence
5	Monomorphic marker (appears in the output file for constant models)
6	Significant model with non-significant parents (multivariate analysis with option `SIGNIF`)

Figure 10 – List of possible errors for logistic models.

Marker	Loglikelihood	AverageProb	Beta_0	NumError
Hapmap43437-BTA-101873_AA	-228.2100569	0.082089552	-2.414289083	0
Hapmap43437-BTA-101873_AG	-542.450042	0.404228856	-0.387875415	0
Hapmap43437-BTA-101873_GG	-556.9893006	0.513681592	0.054740033	0
ARS-BFGL-NGS-16466_AA	-44.84132815	0.009950249	-4.600157644	0
ARS-BFGL-NGS-16466_AG	-389.8189189	0.189054726	-1.456164041	0
ARS-BFGL-NGS-16466_GG	-401.2120224	0.800995025	1.392524911	0
Hapmap34944-BES1_Contig627_1906_AA	-456.4590694	0.254975124	-1.072251619	0
Hapmap34944-BES1_Contig627_1906_AC	-555.856645	0.470149254	-0.119545151	0
Hapmap34944-BES1_Contig627_1906_CC	-472.7907257	0.274875622	-0.970024485	0

Figure 11 – Exemple of Samβada's results for constant models, there is one marker per line. The first column is the name of the molecular marker, here the locus name combined with the allele name. The following columns are the log-likelihood, the frequency of the marker, the estimate of parameter β_0 for the logistic model and the error code (0 if success). Constant models are not sorted and thus are in the same order as the markers in the input file. When considered markers are SNPs like here, there are three binary markers per locus.

Concerning the measure of spatial autocorrelation, results are stored separately for environmental data and molecular markers. In each case, there are three output files. The first one is named `Data-AS-Env.ext` (or `Data-AS-Mark.ext`) and stores Moran's I and local indicators of spatial as-

Marker	Env_1	Loglikelihood	Gscore	WaldScore	NumError	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
Hapmap41074-BTA-73520_AA	prec7	-443.11	208.53	151.72	0	0.25	0.19	0.19	0.23	0.10	890.22	912.98	-2.04	0.03
ARS-BFGL-NGS-113888_GG	prec7	-441.73	208.67	151.70	0	0.25	0.19	0.19	0.23	0.10	887.47	910.23	-2.02	0.03
Hapmap41762-BTA-117570_GG	prec7	-435.96	202.93	148.43	0	0.24	0.19	0.19	0.22	0.10	875.92	898.68	-1.86	0.03
ARS-BFGL-NGS-46098_GG	prec7	-440.04	200.82	147.60	0	0.24	0.19	0.18	0.22	0.10	884.07	906.83	-1.88	0.03
ARS-BFGL-NGS-113888_GG	latitude	-449.13	193.89	146.89	0	0.23	0.18	0.17	0.21	0.09	902.25	925.01	-0.73	0.86
Hapmap41074-BTA-73520_AA	latitude	-450.81	193.13	146.61	0	0.23	0.18	0.17	0.21	0.09	905.62	928.38	-0.75	0.85
Hapmap41762-BTA-117570_GG	latitude	-444.40	186.04	141.99	0	0.21	0.17	0.17	0.21	0.09	892.80	915.56	-0.57	0.84
ARS-BFGL-NGS-113888_GG	prec6	-455.48	181.19	138.85	0	0.21	0.17	0.16	0.20	0.09	914.95	937.71	-2.22	0.03
Hapmap41074-BTA-73520_AA	prec6	-457.38	179.99	138.13	0	0.21	0.16	0.16	0.20	0.09	918.77	941.53	-2.23	0.03
ARS-BFGL-NGS-46098_GG	latitude	-451.22	178.45	138.11	0	0.21	0.17	0.16	0.20	0.09	906.44	929.20	-0.59	0.82
Hapmap41813-BTA-27442_AA	prec7	-462.30	179.89	137.52	0	0.22	0.16	0.16	0.20	0.08	928.60	951.36	-1.92	0.03
ARS-BFGL-NGS-46098_GG	prec6	-451.51	177.87	137.27	0	0.21	0.16	0.16	0.20	0.09	907.03	929.78	-2.11	0.03
BTA-73516-no-rs_AA	prec7	-460.18	177.43	136.04	0	0.21	0.16	0.16	0.20	0.08	924.35	947.11	-1.83	0.03
Hapmap41813-BTA-27442_AA	latitude	-469.89	164.71	130.98	0	0.20	0.15	0.15	0.19	0.08	943.77	966.53	-0.76	0.76
Hapmap41762-BTA-117570_GG	prec6	-454.17	166.51	130.97	0	0.20	0.15	0.15	0.19	0.08	912.33	935.09	-1.96	0.03
ARS-BFGL-NGS-46098_GG	longitude	-458.86	163.18	130.95	0	0.18	0.15	0.15	0.18	0.08	921.72	944.48	-23.95	0.76
Hapmap41074-BTA-73520_AA	bio7	-457.07	180.61	129.73	0	0.21	0.16	0.16	0.20	0.09	918.14	940.90	-11.85	0.08
ARS-BFGL-NGS-113888_GG	bio7	-456.32	179.50	128.90	0	0.20	0.16	0.16	0.20	0.09	916.64	939.40	-11.82	0.08
BTA-73516-no-rs_AA	latitude	-468.36	161.06	128.61	0	0.19	0.15	0.14	0.18	0.08	940.72	963.48	-0.67	0.76
Hapmap28985-BTA-73836_CC	prec6	-457.78	157.45	125.68	0	0.19	0.15	0.14	0.18	0.08	919.57	942.33	1.87	-0.03
Hapmap31863-BTA-27454_GG	prec7	-474.85	155.28	123.46	0	0.19	0.14	0.14	0.18	0.07	953.70	976.43	-1.91	0.02
ARS-BFGL-NGS-46098_GG	bio7	-456.70	167.50	121.71	0	0.20	0.15	0.15	0.19	0.08	917.39	940.15	-11.35	0.08
BTA-73516-no-rs_AA	prec6	-474.90	147.99	119.50	0	0.17	0.13	0.13	0.17	0.07	953.79	976.55	-1.97	0.03
Hapmap41762-BTA-117570_GG	bio7	-460.77	153.30	113.69	0	0.18	0.14	0.14	0.17	0.07	925.54	948.30	-10.71	0.07
Hapmap28985-BTA-73836_GG	bio3	-381.27	160.94	111.21	0	0.21	0.17	0.17	0.18	0.10	766.54	789.30	19.98	-0.26
ARS-BFGL-NGS-113888_GG	bio3	-471.77	148.61	106.51	0	0.17	0.14	0.13	0.17	0.07	947.53	970.29	20.21	-0.24

Figure 12 – Exemple of Samβada's results for univariate models, there is one marker per line. The first column is the name of the molecular marker, here the locus name combined with the allele name. The second column is the name of the environmental variable. The following columns are the log-likelihood, G score, Wald score and the error code (0 if success). The five next columns are goodness-of-fit measures for the regression (pseudo-R^2). The analysis includes the AIC (*Akaike information criterion*) and BIC (*Bayesian information criterion*) as well. The two last column contain the parameters β for the regression, one constant parameter and one corresponding to the environmental variable. Results file for multivariate models contain additional columns for environmental variables (Env_2, Env_3, ...) and for regression parameters (Beta_2, Beta_3, ...).

Name	Formula	References
Akaike Informatic Criterion (AIC)	$-2l + 2p$	Akaike (1974)
Bayesian Informatic Criterion (BIC)	$-2l + p \cdot \ln n$	Schwarz (1978)
Efron pseudo-R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \pi_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Efron (1978)
McFadden pseudo-R^2	$1 - \frac{l}{l_0}$	McFadden (1974)
McFaddenAdj pseudo-R^2	$1 - \frac{l - p}{l_0}$	Smith and McKenna (2013)
CoxSnell pseudo-R^2	$1 - \left(\frac{L_0}{L}\right)^{2/n}$	Snell and Cox (1989)
Nagelkerke pseudo-R^2	$\frac{1 - \left(\frac{L_0}{L}\right)^{2/n}}{1 - L_0^{2/n}}$	Nagelkerke (1991)

Figure 13 – Goodness of fit statistics computed by Samβada for the logistic models. The first column is the name of the statistic, the second one contains the formula and the third one refers to the bibliography. The following notation is used:
n: number of samples,
y_i: value of the (binary) molecular marker for sample i,
$\bar{y}$: average value of y_i over all samples,
p: number of parameters included in the model (p = number of explanatory variables + 1),
π_i: estimated probability of occurrence of the marker for sample i $\left(\pi_i = \frac{\exp \boldsymbol{x_i}^T \boldsymbol{\beta}}{1+\exp \boldsymbol{x_i}^T \boldsymbol{\beta}} \text{ where } \boldsymbol{x_i}^T \boldsymbol{\beta} = \beta_0 + x_{i1} \cdot \beta_1 + \cdots + x_{i(p-1)} \cdot \beta_{p-1}\right)$,
$\boldsymbol{x_i}$: vector of environmental variables at location of sample i,
$\boldsymbol{\beta}$: vector of estimated regression parameters,
l: log-likelihood of the model of interest,
l_0: log-likelihood of the reference (null) model,
L: likelihood of the model of interest,
L_0: likelihood of the reference (null) model.

sociation (Anselin, 1995). If provided, the sample names appear in the first column. If both Moran's I and LISAs are computed, the first line is the global index and each subsequent line is a local index, samples are in the same order as in the data file. The second file is either named `Data-AS-Env-Sim.ext` (or `Data-AS-Mark-Sim.ext`) and stores the simulated values of the global Moran's I for each variable. This file can be used to plot their distribution and compare it to the actual value of the index. Simulation results are not stored for LISAs in order to save disk space. The third file is named `Data-AS-Env-pVal.ext` (or `Data-AS-Mark-pVal.ext`) and stores the pseudo p-values for the permutations-based significance tests. For R permutations and M events "I_{sim} is equal or more extreme than I", the p-value is $\frac{M+1}{R+1}$.

If requested, LISAs are also stored as a shapefile whose parts are named `data.shp`, `data.shx` and `data.dbf` (for data file `data.ext`). This format is read by any GIS software, for instance `QuantumGIS`[6].

5.1.5 Including prior knowledge in multivariate models

Multivariate analysis can be used to assess the effect of a combination of predictor variables on the occurrence of molecular markers. Moreover it also makes it possible to include pre-existing knowledge into the analysis, provided the data constitutes a continuous variable. In particular, if population structure was analysed beforehand and can be represented as a coefficient of membership for each individual, this information can be included in the modelling as a "population structure" variable added to the set of environmental variables. For models involving both an environmental variable and this new variable, the selection procedure will assess whether the environmental variable is associated with the genotype while taking into account the possible effect of admixture. In case there are many ancestral populations, several "population structure" variables may be included in the analysis.

In the current implementation of Samβada, the proposed approach relies on multivariate models including potentially many variables. The idea is to build models with one or several "population variables" and one environmental variable. These models will be considered as univariate when assessing their significance, since their corresponding null model will be their parent with only the "population variables". The selection procedure with the option `SAVETYPE BEST` is not convenient for taking the population structure into account, because it would retain only the models showing a significant association with all predictor variables. Thus the analysis must be run with the option `SAVETYPE ALL`. Since the computation time grows linearly with the number of markers but exponentially with the number of predictor variables for multivariate models (the power being equal to the maximum number of predictor variables included in the models, i. e. the param `DIMMAX`),

[6] `www.qgis.org`

the total computation time might be consequent. Therefore you might wish to estimate the required time by running the analysis (step 3) on a subset (for instance 1%) of the markers, before launching the computation on the whole dataset.

The procedure consists of the followings steps, assuming there are K clusters in your dataset ($K > 1$). Steps 2 to 9 are also illustrated by the example "AnalysisWithPopulationStructure".

1. Determine a set of P "population variables" representing the population structure of your dataset, for instance by taking:
   - $K - 1$ coefficients of membership, since the K^{th} coefficient is redundant with the other ones;
   - the $K - 1$ first principal axes from a PCA on the coefficients of membership, this is equivalent to the first possibility;
   - one, two or three principal axes from the PCA if they explain most of the variance in your dataset
2. Hint: Call your "population variables" by names beginning with "pop" (*e.g.* "pop1", "pop2", ...). This way Samβada will compute only relevant models in dimensions `DIMMAX-1` and `DIMMAX`, and you will gain time by speeding up the analysis and being able to skip the step 7 below.
3. Create a file of predictors variables with the E environmental variables and the P "population variables". In the input file, the columns containing the "population variables" should stick together, and form a group either on the left or on the right of the group of the environmental variables.
4. Create a parameter file with the following options:
   - `DIMMAX` $P+1$, e.g. `DIMMAX 3` if there are 2 "population variables";
   - `POPULATIONVAR FIRST` (or `LAST`) if the "population variables" are placed on the left (or on the right) of the environmental variables in the input file;
   - The option `SAVETYPE` should be set to `ALL`, e.g. `SAVETYPE END ALL`.

   The other options can be configured as usual. When selecting a subset of environmental variables with `SUBSETVARENV`, the selected "population variables" must also appear in the list. The order in which the variables will be considered during the analysis is determined by the order of the columns in the input file.
5. Run the analysis.

6. The output files for the dimensions 0 to $\texttt{DIMMAX} - 1$ are the same as usual. The output file of dimension `DIMMAX` contains two extra columns: "GscorePop" and "WaldScorePop". For the models involving all "population variables" and one environmental variable, these columns provide the G and Wald scores taking the population structure into account. The models are still sorted according to the "usual" Wald score. Model selection is done in post-processing.

7. (Skip this step if you did step 2.) Extract the models of interest:

   - optionally the "null models" of dimension P including the P "population variables";
   - the models of dimension $P + 1$ including the P "population variables" and one environmental variable;

   Although this step can be performed with `R` (or similar), it is usually faster to filter the results in the Terminal. For instance with three "population variables" called P1, P2 and P3, the command

   ```
   grep "P1 P2 P3" myfile-Out-4.txt" > filtered-results.txt
   ```

   will select the models containing the three "population variables" and one environmental variable in the output file of dimension 4.

8. Load the filtered results in your favourite statistics software.
   The last two steps are explained in more details in section 6:

9. Compute the p-values based on the G and/or Wald scores, the expected distribution is a χ^2 with one degree of freedom, since the models tested (of dimension $P+1$) include one variable more than their corresponding "null model" of dimension P.

10. Select the candidate loci while taking into account the multiple comparisons, for instance with a Bonferroni correction (potentially overconservative in this case) or with Storey and Tibshirani (2003) or Benjamini and Hochberg (1995) approaches.

5.2 Supervision

For large molecular datasets, computation workload may be distributed among several computers. To this end, `Supervision` is called prior to the analysis to split the molecular dataset in blocks (see fig. 14). The new files are named automatically on the basis of the molecular data file. The last file will contain less markers if the total number is not divisible by the block size. Each share of data is processed separately, either on the same multicore computer or on distrinct computers. Environmental data must provided to each processing node. Results are gathered afterwards so `Supervision` can merge them and produce the same output as if the whole analysis were run on a single node.

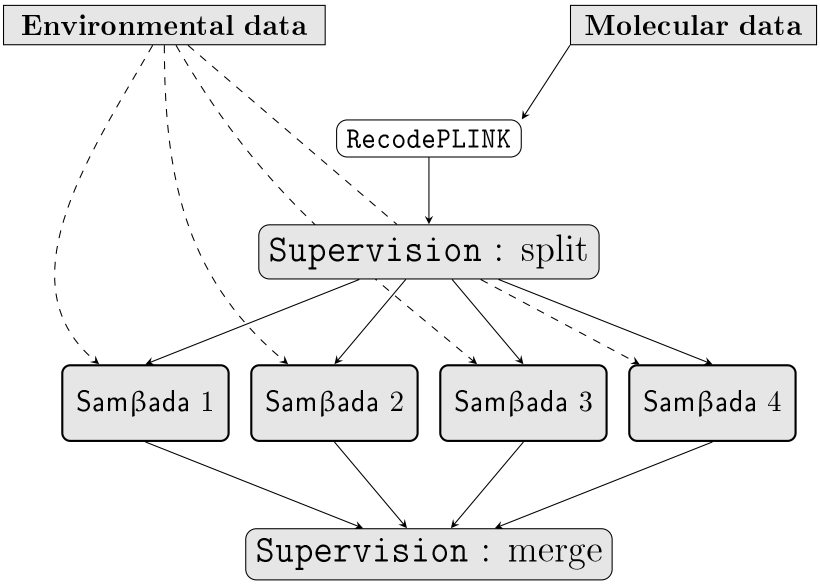


Figure 14 – Workflow of analysis for distributed computing. Rectangles stand for data and round-cornered figures stand for programs. Grey elements are mandatory and white ones are optional. Arrows show processing order for environmental data (dashed line) and for molecular data (solid line). `Supervision` is used before `Samβada` to split molecular data in blocks and afterwards to merge results.

5.2.1 Split process

Information about splitting are provided by a parameter file and the processed is launched as follow:

```
supervision parameterFile
```

The parameter file must contain the six following lines (in the same order):

`dataFile`	name of the data or molecular file (beware of trailing tabulations)
`paramFile`	name of the parameter file (not used yet)
`numEnv`	number of environmental parameters*
`numMark`	number of molecular markers
`numLines`	number of lines in the data file
`blockSize`	size of blocks of molecular data

Figure 15 – Instance of parameter file for `Supervision`.

*`numEnv` indicates the number of non-molecular columns in the file to be recoded. These variables will be copied to a separate file. If the molecular file contains no ID or environmental column, `numEnv` is to be set to 0. Please also note that the fifth parameter counts the total number of lines, including the header line if any.

New data files are named automatically. The molecular file names contains two numbers, the first one is the block number (starting from 0) and the second one is the number of the first marker in the block (starting from 0 in the first block).

Example 1 Let us assume the data of fig. 4 is stored in `mol-data.txt` and that we want blocks of four markers. The corresponding parameter file is shown on fig. 16 and the splitting is performed as follows:
`supervision param-split.txt`
The resulting data files (`mol-data-env.txt`, `mol-data-mark-0-0.txt` and `mol-data-1-4.txt`) are displayed on fig. 17.

`mol-data.txt`	
`thingy.txt`	
`1`	*numEnv*
`7`	*numMark*
`7`	*numLines*
`4`	*blockSize*

Figure 16 – Parameter file for `Supervision` in case molecular data of fig. 4 is stored in `mol-data.txt` and is to be split in blocks of four markers. This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/two-data-files/param-split.txt`.

NAME
ID1
ID2
ID3
ID4
ID5
ID6

(a) File `mol-data-env.txt`

M4	M7	M8	M9
1	1	1	0
0	0	0	0
0	1	0	0
0	0	1	1
0	0	1	0
0	1	0	0

(b) File `mol-data-mark-0-0.txt`

M16	M17	M18
1	1	1
0	0	NaN
0	0	1
0	1	1
0	1	0
0	1	0

(c) File `mol-data-mark-1-4.txt`

Figure 17 – Molecular data of fig. 4 after splitting in block of four markers.

In this case, the column NAME is considered as an environmental variable (since it is not a molecular marker). NAME will be copied to the file `mol-data-env.txt` which will not be used for Samβada's analysis. The actual environmental data (`env-data.txt`) is shown on fig. 3. The column NAME cannot be used as an identifier since it was not copied to the molecular data. Thus it must be indicated as a supplementary column to Samβada (option COLSUPENV, see p. 13).

Example 2 `Supervision` can split combined data files, containing both environmental and molecular information. Figure 18 shows the parameter file used to split data from fig. 5 (stored in `combo-data.txt`) in blocks of three markers. The splitting is performed as follows:

```
supervision param-split-combo.txt
```

The data files will be named `combo-data-env.txt`, `combo-data-mark-0-0.txt`, `combo-data-mark-1-3.txt` and `combo-data-mark-2-6.txt`, the latter will have only one column (see fig. 19). In this case, `combo-data-env.txt` will contain environmental data and can be used in Samβada's analysis. As in the previous example, the column NAME cannot be used as an identifier since it was not copied to the molecular data. Thus it must be indicated as a supplementary column to Samβada (option COLSUPENV, see p. 13).

5.2.2 Analysis with Samβada

The distributed analysis follows the same process as the single-node one. The parameter file has to be modified to include both the current and the total number of molecular markers (parameter `NUMMARK`, see p. 13). Thus if the last block has less markers than the other ones, there will be a common parameter file for the first blocks and another one for the last block. The total number of markers is used to adjust the significance threshold with the

```
combo-data.txt
thingy.txt
6                  numEnv
7                  numMark
7                  numLines
3                  blockSize
```

Figure 18 – Parameter file for `Supervision` in case molecular data of fig. 5 is stored in `combo-data.txt` and is to be split in blocks of three markers. This file is part of the distribution, see `sambada-0.8.3/examples/data-from-manual/one-data-file/param-split-combo.txt`.

Bonferroni correction (if relevant).

Example 1 Fig. 20 shows the parameter files needed to analyse molecular data from fig. 17 (b and c) with environmental data from fig. 3. For comparison, the parameter file for single node analysis is shown on fig. 7. Beside the change in the number of markers, the column NAME has to be indicated as supplementary data (COLSUPENV) since it is not available in the molecular files.

The analysis is launched with two commands:

```
sambada param-a.txt env-data.txt mol-data-mark-0-0.txt
sambada param-b.txt env-data.txt mol-data-mark-1-4.txt
```

Example 2 Fig. 21 shows the parameter files needed to analyse molecular data from fig. 19 (b, c and d) with environmental data from fig. 19a. For comparison, the parameter file for single node analysis is shown on fig. 8. Beside the change in the number of markers, the column NAME has to be indicated as supplementary data (COLSUPENV) since it is not available in the molecular files.

The analysis is launched with three commands:

```
sambada param-combo-a.txt combo-data-env.txt combo-data-mark-0-0.txt
sambada param-combo-a.txt combo-data-env.txt combo-data-mark-1-3.txt
sambada param-combo-b.txt combo-data-env.txt combo-data-mark-2-6.txt
```

5.2.3 Merge process

Once all blocks of markers have been analysed with `Samβada`, `Supervision` can merge the results. Copy every output file (whose name contains `"-Out-"`) to a single folder, then launch the program as follows:

NAME	ENV1	ENV2	ENV3	ENV4	ENV5
ID1	46	972	236	230	132
ID2	32	987	238	232	83
ID3	32	987	238	232	83
ID4	32	987	NaN	232	83
ID5	32	987	238	232	83
ID6	35	1021	235	230	87

(a) File `combo-data-env.txt`

M4	M7	M8
1	1	1
0	0	0
0	1	0
0	0	1
0	0	1
0	1	0

(b) File `combo-data-mark-0-0.txt`

M9	M16	M17
0	1	1
0	0	0
0	0	0
1	0	1
0	0	1
0	0	1

(c) File `combo-data-mark-1-3.txt`

M18
1
NaN
1
1
0
0

(d) File `combo-data-mark-2-6.txt`

Figure 19 – Molecular data of fig. 5 after splitting in blocks of three markers.

```
HEADERS YES
NUMVARENV 6
NUMMARK 4 7
NUMINDIV 6
COLSUPENV NAME
DIMMAX 1
SAVETYPE END ALL
```

(a) `param-a.txt`

```
HEADERS YES
NUMVARENV 6
NUMMARK 3 7
NUMINDIV 6
COLSUPENV NAME
DIMMAX 1
SAVETYPE END ALL
```

(b) `param-b.txt`

Figure 20 – Parameter files for distributed analysis with Samβada. File `param-a.txt` is used for data from fig. 3 and 17b. Usually there are more blocks of molecular data, and all blocks containing the same number of markers would be analysed with this set of parameters. File `param-b.txt` is used for the last block of molecular data, when it contains less markers (data from fig. 3 and 17c). These files are part of the distribution, see `param-a.txt` and `param-b.txt` in `sambada-0.8.3/examples/data-from-manual/two-data-files/`.

```
HEADERS YES
NUMVARENV 6
NUMMARK 3 7
NUMINDIV 6
COLSUPENV NAME
DIMMAX 1
SAVETYPE END ALL
```

(a) `param-combo-a.txt`

```
HEADERS YES
NUMVARENV 6
NUMMARK 1 7
NUMINDIV 6
COLSUPENV NAME
DIMMAX 1
SAVETYPE END ALL
```

(b) `param-combo-b.txt`

Figure 21 – Parameter files for distributed analysis with Samβada. Data file `combo-data.txt` has been split in fig. 19, all computations use environmental data from subfig. *a*. File `param-combo-a.txt` is used to analyse markers from subfig. *b* and *c*, while file `param-combo-b.txt` is used for the last block of molecular data (subfig. *d*). These files are part of the distribution, see `param-combo-a.txt` and `param-combo-b.txt` in `sambada-0.8.3/examples/data-from-manual/one-data-file/`.

```
supervision base-name.txt numBlock blockSize maxDimension
```

base-name.txt is the name of the molecular or combined data file which was split in blocks. numBlocks and blockSize refer to the number of data blocks (including the last one) and to the size of the "complete" ones. maxDimension is the maximum number of environmental parameters included in the models (1 for univariate analysis, 2 for bivariate analysis, ...)

Supervision merges all results, discards unconverged models (error numbers 1-5) and sort models according to their Wald score. One output file is produced for each dimension of modeling, as in the single-node Samβada's analysis. If the original data file is called base-name.txt, the results files are named base-name-res-0.txt, base-name-res-1.txt, base-name-res-2.txt, ...

Example 1 Data file mol-data.txt was split in two blocks of four markers and the analysis involved univariate models:

```
supervision mol-data.txt 2 4 1
```

Example 2 Data file combo-data.txt was split in three blocks of three markers and the analysis involved univariate models:

```
supervision combo-data.txt 3 3 1
```

Supervision also takes some optional arguments. The complete call is:

```
supervision base-name.txt numBlock blockSize maxDimension ···
   selScore scoreThreshold sortScore wordDelim
```

selScore indicates which score(s) is/are used to select significant models; possible values are G, Wald and BOTH (the default). scoreThreshold indicates the minimum score for which a model is considered as significant. This option can be used either to apply Bonferroni correction when all models were saved during the analysis or to select a subset of models, for instance for a post-processing analysis in R. selScore and scoreThreshold must be provided together. sortScore indicates which score is used to sort the models; possibles values are G, Wald (the default), AIC and BIC. wordDelim shows the current word delimiter (space is the default).

Optional arguments may be omitted from right to left. Thus the possible sets of arguments are:

```
supervision base-name.txt numBlock blockSize maxDimension
```

```
supervision base-name.txt numBlock blockSize maxDimension ···
   selScore scoreThreshold
```

```
supervision base-name.txt numBlock blockSize maxDimension ···
   selScore scoreThreshold sortScore
```

```
supervision base-name.txt numBlock blockSize maxDimension ···
   selScore scoreThreshold sortScore wordDelim
```

5.3 RecodePLINK

This tools allows the recoding of PLINK's .ped and .map files to Samβada's format (see Purcell, 2009, for further information on this format). RecodePLINK is called as follows:
`recode-plink nbSamples nbSNPs inputFile outputFile`
In case only a subset of the samples are to be used, the list of sample names may be provided in a separate file (one name per line):
`recode-plink nbSamples nbSNPs inputFile outputFile subsetFile`
Please note that RecodePLINK does not recognise comment lines in .ped/.map files at the moment. Please remove them before recoding.

6 Pre- and post-processing with R

Your analyses with Samβada may require some pre- or post-processing. This section presents a couple of examples using R[7]. These examples are distributed in the directory scripts/.

Model selection with Bonferroni correction

When performing model selection with the options SIGNIF or BEST, Samβada applies the Bonferroni correction to the provided *p*-value threshold. Models with *p*-values lower than this corrected threshold for both the G and Wald statistics are retained as significant. However you may want to use another selection method, for instance by applying several thresholds or by using only the G or the Wald score. In this case, we suggest running Samβada with a more liberal selection method (for instance with the option ALL or BEST 0.1) and then selecting the models with R. The script bonferroniModelSelection.R provides an example of model selection using the Bonferroni correction.

[7] https://www.r-project.org/

7 Frequently Asked Questions

I am lost with the Command Line!

Don't panic, there are a couple of online tutorials. For Windows, you can start with `http://www.cs.princeton.edu/courses/archive/spr05/cos126/cmd-prompt.html`. (Hint: Use `PowerShell` instead of the `Command Line`.)

MacOS and Linux `Terminals` are basically the same, see for instance: `http://www.davidbaumgold.com/tutorials/command-line/`.

My results do not make any sense!

Check whether the samples in the environmental data are in the same order as those in the molecular data. What could have happened is that one of your files got sorted by some column during data preparation.

To be continued...

References

Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control* **19**(6), 716–723, DOI: 10.1109/TAC.1974.1100705.

Anselin, Luc (1995), "Local Indicators of Spatial Association - LISA", *Geographical Analysis* **27**(2), GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, Amsterdam, Netherlands, Dec 01-05, 1993, 93–115.

Benjamini, Y and Y Hochberg (1995), "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing", *Journal of the Royal Statistical Society Series B-Statistical Methodology* **57**(1), 289–300.

Efron, Bradley (1978), "Regression and ANOVA with Zero-One Data: Measures of Residual Variation", *Journal of the American Statistical Association* **73**(361), 113–121, DOI: 10.1080/01621459.1978.10480013, eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1978.10480013.

Joost, Stéphane, Aurélie Bonin, Michael W. Bruford, et al. (2007), "A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation", *Molecular Ecology* **16**(18), 3955–3969.

McFadden, Daniel L. (1974), "Conditional logit analysis of qualitative choice behavior", in: *Frontiers in econometrics*, ed. by Zarembka Paul, New York, NY, USA: Academic Press, chap. 4, 104–142.

Nagelkerke, N. J. D. (1991), "A Note On A General Definition Of The Coefficient Of Determination", *Biometrika* **78**(3), 691–692, DOI: {10.1093/biomet/78.3.691}.

Purcell, Shaun (2009), *PLINK 1.07*, http://pngu.mgh.harvard.edu/purcell/plink/.

Schwarz, Gideon (1978), "Estimating the Dimension of a Model", *Ann. Statist.* **6**(2), 461–464, DOI: 10.1214/aos/1176344136.

Smith, Thomas J. and Cornelius M. McKenna (2013), "A Comparison of Logistic Regression Pseudo R^2 Indices", *Pseudo R 2 Indices Multiple Linear Regression Viewpoints* **39**(2), 17–26.

Snell, E.J and David R Cox (1989), *Analysis of binary data*, Second ed., vol. 32, Monographs on statistics and applied probability, London [etc.: Chapman and Hall.

Storey, J. D. and R. Tibshirani (2003), "Statistical significance for genomewide studies", *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9440–9445, DOI: {10.1073/pnas.1530509100}.

Stucki, S., P. Orozco-terWengel, B. R. Forester, et al. (2016), "High performance computation of landscape genomic models including local indicators of spatial association", *Molecular Ecology Resources* **17**(5), 1072–1089, DOI: `10.1111/1755-0998.12629`.

Stucki, Sylvie (2014), "Développement d'outils de géo-calcul haute performance pour l'identification de régions du génome potentiellement soumises à la sélection naturelle: analyse spatiale de la diversité de panels de polymorphismes nucléotidiques à haute densité (800k) chez *Bos taurus* et *B. indicus* en Ouganda", PhD thesis, Lausanne: Ecole Polytechnique Fédérale de Lausanne, DOI: `10.5075/epfl-thesis-6014`.

Vajana, Elia, Mario Barbato, Licia Colli, et al. (2018), "Combining Landscape Genomics and Ecological Modelling to Investigate Local Adaptation of Indigenous Ugandan Cattle to East Coast Fever", *Frontiers in Genetics* **9**, 385, DOI: `10.3389/fgene.2018.00385`.