

# Manual for Kestimator version: 1.13

S.J. Puechmaille

Updated: 02/07/2017

Please, report any bugs or suggestions to improve the script, input or output files to: s [dot] puechmaille [at] gmail [dot] com

This script is intended to become an R package. Please, contact the author shall you be interested in contributing to it.

**Before reporting on a bug, please, make sure your input files are in the correct format.**

## TABLE OF CONTENTS

BEFORE RUNNING ANALYSES .....	1
ANALYSES .....	1
I - Sub-sampling FSTAT files .....	2
A- Input file names.....	2
B- Arguments for the 'subsampFSTAT' function.....	2
C- Output files of the 'subsampFSTAT' function.....	3
D- Example of commands to run the 'subsampFSTAT' function .....	3
II - Renaming files.....	3
III - Analysing STRUCTURE outputs .....	3
A- Input file names.....	3
B- Arguments for the 'STRUCTURE_ANALYSIS' function.....	4
C- Output files of the 'STRUCTURE_ANALYSIS' function.....	5
D- Example of commands to run the 'STRUCTURE_ANALYSIS' function.....	6
CHANGES FROM PREVIOUS VERSIONS.....	6
REFERENCES.....	7

Here we provide some information on how to estimate  $K$  (the number of clusters) from STRUCTURE V2.3.4 outputs. The estimators include:

- the Posterior Probability method (Pritchard *et al.* 2000; Pritchard & Wen 2004; Pritchard *et al.* 2010),
- the deltaK method (Evanno *et al.* 2005),
- the corrected Posterior Probability method (Puechmaille 2016),
- the corrected deltaK method (Puechmaille 2016),
- the 'MedMeaK' (median of means) method (Puechmaille 2016),
- the 'MaxMeaK' (maximum of means) method (Puechmaille 2016),
- the 'MedMedK' (median of medians) method (Puechmaille 2016),
- the 'MaxMedK' (maximum of medians) method (Puechmaille 2016).

For further details on the methods themselves and which estimates to use, please read the above mentioned papers.

## BEFORE RUNNING ANALYSES

You first need to make sure you have the necessary libraries installed. The necessary libraries are 'Rmpfr' and 'reshape'. If you use the 'subsampFSTAT' function, you will also need the 'hierfstat' library. If you do not have these libraries installed, simply type the following commands in R:

```
>install.packages("Rmpfr")
```

then select a mirror of choice and wait until the first library is installed. Then install the second library by typing:

```
>install.packages("reshape")
```

Follow by the third library if needed:

```
>install.packages("hierfstat")
```

Once you have the libraries installed, you have to load them by typing:

```
>library(Rmpfr)
```

```
>library(reshape)
```

```
>library(hierfstat)
```

Before you run any analyses, you need to make sure that you have loaded the 'Kestimator\_V1-13.R' file. In my computer, the file is located in the following folder "D:/SEB/PROJECTS\_Running/STRUCTURE/runs/" hence, to load the file, I simply run the following command in the R console:

```
>source("D:/SEB/PROJECTS/STRUCTURE/runs/Kestimator_V1-13.R ")
```

Change the path to the file (underlined above) to match your computer settings.

## ANALYSES

To run the functions, simply call them and give their arguments. A file 'Kestimator\_V1-13\_Run.R' (a simple text file) gives an example of command lines that can be typed or paste onto the R console to obtain results.

It is assumed that the lowest  $K$  value run is  $K=1$  and that  $K$  values considered are consecutive. It is also assumed that the same number of (replicate) runs has been carried out for each  $K$  value.

## I - Sub-sampling FSTAT files

### A- Input file names

The function 'subsampFSTAT' is provided for convenience in case one needs to sub-sample input files before running STRUCTURE. This function uses FSTAT input files (Goudet 1995) imported via the 'hierfstat' package (Goudet 2005) to sub-sample individuals and/or group them differently.

### B- Arguments for the 'subsampFSTAT' function

**file\_in**=Name of the FSTAT file to be sub-sampled. Note that no extension should be provided in the argument but the file is expected to have a '.dat' extension.

**path\_in**=Full path to the FSTAT file to be sub-sampled (but excluding the file name).

**path\_out**= Optional. Full path to the folder where the sub-sampled file will be written. By default, the 'path\_in' folder is used.

**ResamplingScheme**=Vector of subpopulation sizes giving for each subpopulation the number of individuals to keep. The first number in the vector will be used for the first subpopulation, the second number in the vector for the second subpopulation and so on. There must be as many numbers in the vector as there are subpopulation in the 'file\_in' input file. If a subpopulation is not to be sub-sampled, simply mention '0' as the number of individuals to sub-sample from it. As the sub-sampling is done without replacement, the number of sub-sampled individual for a subpopulation must be equal or lower than the total number of individuals in that subpopulation (if not, the function will stop and a warning message will appear).

**ResamplingSubdiv**=Optional. Mostly to be used with simulated datasets. This argument allows one to split the sub-sampled subpopulations into two or more subpopulations prior to running STRUCTURE. Such can be used to investigate the effect of prior grouping when using the LOCPRIOR option in STRUCTURE. The argument must be a vector giving, for each original subpopulation, the number of new subpopulations to create from it. If a subpopulation is not to be sub-sampled as defined via the 'ResamplingScheme' argument, simply mention '0' for that subpopulation in the vector. Note that for the sub-sampling to work, for each original subpopulation, the value in the 'ResamplingScheme' argument (i.e. number of individuals to sub-sample) must result in a whole number when divided by the value in the 'ResamplingSubdiv' (i.e. number of subpopulation in the new dataset). If this condition is not met, the function will stop and a warning message will appear. For example, the following would work:

```
ResamplingScheme<-c(10,20,30,0,50)
```

```
ResamplingSubdiv<- c(2,5,3,0,2)
```

While the following would not work:

```
ResamplingScheme<-c(10,20,30,0,50)
```

```
ResamplingSubdiv<- c(2,5,3,0,3)
```

**STRUCTURE**= Optional. Whether to also output the files in STRUCTURE format. Set to TRUE by default.

## C- Output files of the 'subsampFSTAT' function

**file\_in\_Sub.dat** = An FSTAT file (with .dat extension) with individuals sub-sampled and, if the 'ResamplingSubdiv' is provided, new subpopulations.

**file\_in\_Sub.str** = (only provided if 'STRUCTURE=TRUE'). A STRUCTURE file (with .str extension) with individuals sub-sampled and, if the 'ResamplingSubdiv' is provided, new subpopulations.

On the R console, after new files have been created, R reports on the 'Number of individuals sub-sampled' and on the 'New number of subpopulations'.

## D- Example of commands to run the 'subsampFSTAT' function

After having loaded the necessary libraries and loaded the functions from the 'Kestimator\_V1-13.R' file (cf. paragraph 'Before running analyses'), the 'subsampFSTAT' function can be run as follows:

Example 1: Sub-samples the M1\_00.dat file (with 10 subpopulations), selecting 50 individuals from all subpopulations except the 9<sup>th</sup> one for which no individuals were picked.

```
>subsampFSTAT(file_in="M1_001",path_in="D:/SEB/PROJECTS/STRUCTURE /M1/",  
  ResamplingScheme=c(rep(50,8),0,50))
```

Example 2: Does the same as example 1 above but subsequently divides each of the first 8 subpopulations in two subpopulations of equal size. The last subpopulation is not subdivided.

```
>subsampFSTAT(file_in="M1_001",path_in="D:/SEB/PROJECTS/STRUCTURE/M1/",  
  ResamplingScheme=c(rep(50,8),0,50), ResamplingSubdiv=c(rep(2,8),0,1))
```

Note that the 'rep' function in R is used to repeat numbers; for example, 'c(rep(10,4))' is the same as 'c(10,10,10,10)'

## II - Renaming files

The function 'File\_Rename' is provided for convenience in case one needs to rename many files from the same folder and more particularly, to replace, in the file name some characters with others. In an example case, if all 'f' and 'q' files begin by 'result\_job' in their file names, to make them shorter the 'result\_job' can be deleted from every file by typing in the R console:

```
>File_Rename(Folder="D:/SEB/PROJECTS/STRUCTURE/", "result_job_", "")
```

## III - Analysing STRUCTURE outputs

### A- Input file names

For each dataset, the files detailed below are needed. These files are unmodified files created by STRUCTURE v2.3.4 as long as these are requested by the user (i.e. to obtain 'q' files; in your parameters set, in the 'Advanced' tab, tick 'Print Q-hat'). The 'q' file

should contain incremental numbers (from 1 to the total number of individuals in the dataset) in the first column (make sure this is the case before analysing your data). The second column should contain the subpopulation to which each individual belongs, which may or may not correspond to the sampling origin of the individual. If 'NewPopName' is set to FALSE, the information from this second column will be used to calculate the corrected DeltaK, corrected PP and the four new estimators, MedMeaK' (median of means), 'MaxMeaK' (maximum of means), 'MedMedK' (median of medians), and 'MaxMedK' (maximum of medians). If the 'NewPopName' is set to TRUE, the new subpopulation sizes need to be provided in a vector via the "NewPopSizes" argument.

The files needed are as many 'f' AND 'q' files (out of STRUCTURE) as there are runs (argument 'Nruns') for the dataset (e.g. for a dataset tested for every  $K$  between 1 and 10 with 20 replicates per  $K$  value, 200 'f' files and 200 'q' files will be needed). These files need to be named as follows:

"DatasetName\_DatasetQualifier\_RunNumber\_q" (e.g. "M1\_run\_1\_q", "M1\_run\_2\_q", ..., "M1\_run\_3\_q").

Note that in the examples above:

*DatasetName*="M1"

*DatasetQualifier*="run"

*RunNumber*=1, 2 ... 200

'f' files are just the same but ending with 'f' instead.

## B- Arguments for the 'STRUCTURE\_ANALYSIS' function

**PercTreshGhost**=Threshold value under which the pre-defined subpopulations are considered to belong to a cluster (this parameter is used only for plotting and is set to 0.5 by default). Do not use values lower than 0.5.

**Thresh**: Threshold values under which the pre-defined subpopulations are considered to belong to a cluster (four values "0.5, 0.6, 0.7 & 0.8" are used by default). Do not use values lower than 0.5.

**InputFileBase**: Base of the name for dataset. E.g. for the 'M1' dataset, the value is "M1".

**ResFolder**: full path to the folder where the results will be added. R uses forward slashes (and not back slash).

**DatasetQualifier**: Dataset Qualifier intended to give information about the content of the dataset, especially in the situation where multiple slightly different datasets are analysed (e.g. when sub-sampling a single dataset and obtaining multiple subset datasets).

**RunDeltaK**: Whether to calculate the DelatK statistic as presented in Evanno et al. 2005 (TRUE by default).

**PlotMemb**: Whether to plot the individual membership for each run (FALSE by default). A maximum of 18 colours are set (hence a max of  $K=18$  can be plotted). Further colours can be set by changing the 'COLE' argument within the function code (i.e. in the 'Kestimator\_V1-13\_Run.R' file).

**NewPopName**: Using this option (when set to 'TRUE'), the grouping of individuals (i.e. subpopulations) can be changed after STRUCTURE was run but before the analysis of the estimate of the value of  $K$  (FALSE by default). See 'NewPopSizes' for further details.

**NewPopSizes**: When 'NewPopName' is set to TRUE, a vector of new subpopulation sizes (i.e. grouping of individuals) must be provided as a vector. E.g. 'c(10,10,10,10)' will

compute  $K$  estimates by considering 4 groups of 10 individuals each. NB: individuals assigned to the same cluster must appear consecutively in the file.

**OutPutcsv:** whether to output various results as .csv files (TRUE by default).

**NewPopNameNo:** all results will be grouped into a newly created folder named as follows: "RES\_

*DatasetQualifier\_ThreshGhost=PercTreshGhost\_ThreshMM=Thresh\_NewPopName=NewPopName-NewPopNameNo*". This allows one to specify a different number/name which will place the results in a different folder when slight changes in the run parameters are chosen (e.g. different values for the 'NewPopSizes' parameter).

## C- Output files of the 'STRUCTURE\_ANALYSIS' function

**Results\_BestK\_InputFileBase\_DatasetQualifier.csv** = A .csv file with values of DeltaK, the Posterior Probability and the likelihood of the data for the different  $K$  values considered.

**Results\_Estimators\_InputFileBase\_DatasetQualifier.csv** = A .csv file with estimates of all estimators for the different  $K$  values considered and for the 'Thresh' thresholds considered. The estimator names are followed by 'Cor' when the corrected estimate is computed (for DeltaK, LnK and PPK) and then the 'Thresh' value (for all estimates).

**Results\_Master\_InputFileBase\_DatasetQualifier\_CountInd\_K=K.csv** = (only provided if 'OutPutcsv'=TRUE); As many files as the number of considered values of  $K$ . These files contain the number of individuals belonging to the  $K$  different clusters for each run. An individual is considered as belonging to a cluster if its membership coefficient is greater than 0.5 for that cluster. 'Clus0' is a fake cluster that includes all individuals not belonging to any cluster (i.e. with a membership coefficient to each cluster  $< 0.5$ ).

**Results\_Master\_InputFileBase\_DatasetQualifier\_Spurious\_Mean.csv** = A .csv file with estimates of the number of spurious clusters (per run) for the different  $K$  values considered and for the 'Thresh' thresholds considered. Column names are formed of 'K', followed by the value of  $K$  and then the threshold value, each separated by a dot (e.g. 'K.5.0.5' is for  $K=5$  and for a threshold of 0.5). A spurious cluster is here defined as a cluster which never achieves a mean membership coefficient greater than a threshold value in any subpopulation of the dataset.

**Results\_Master\_InputFileBase\_DatasetQualifier\_Spurious\_Median.csv** = A .csv file with estimates of the number of spurious clusters (per run) for the different  $K$  values considered and for the 'Thresh' thresholds considered. Column names are formed of 'K', followed by the value of  $K$  and then the threshold value, each separated by a dot (e.g. 'K.5.0.5' is for  $K=5$  and for a threshold of 0.5). A spurious cluster is here defined as a cluster which never achieves a median membership coefficient greater than a threshold value in any subpopulation of the dataset.

**Results\_BestK\_ThreshGhost=ThreshGhost\_ThreshMM=ThreshMM\_InputFileBase\_DatasetQualifier.pdf** = A .pdf file with plots of results for the different estimators. The first page shows three graphs representing the DeltaK, likelihood of the data and Posterior Probability plots of the results as functions of the different  $K$  values. The second page shows the results of the corrected DeltaK and Posterior probability

methods for the different threshold values (given by the 'Thresh' argument). The third page shown the results of the 'MedMeaK' (median of means), the 'MaxMeaK' (maximum of means), the 'MedMedK' (median of medians) and the 'MaxMedK' (maximum of medians) methods for the different threshold values (given by the 'Thresh' argument).

**Membership\_Dataset-No.1\_InputFileBase\_DatasetQualifier.pdf** = (provided if 'PlotMemb'=TRUE). A single PDF file with plots of the membership coefficients for all  $K$  and all runs. For each run, the title summarises the results with first the  $K$  considered, followed by the corrected  $K$  (MeaMeaK), the dataset number, the run number, the probability of the data / the probability of the data for the most likely run. Below the barplot, subpopulations numbers are reported followed by the cluster number to which the subpopulation belongs to based on the Mean.Median (and a threshold of 0.5). When '0' is reported, the population does not belong to a cluster. A maximum of 18 colours are set (hence a max of  $K=18$  can be plotted). Further colours can be set by changing the 'COLE' argument within the function code (i.e. in the 'Kestimator\_V1-13\_Run.R' file).

## D- Example of commands to run the 'STRUCTURE\_ANALYSIS' function

After having loaded the necessary libraries and loaded the functions from the 'Kestimator\_V1-13.R' file (cf. paragraph 'Before running analyses), the 'STRUCTURE\_ANALYSIS' function can be run as explained below. The files used in the examples below (as well as the output files) are provided along with this document in the 'Example\_Dataset' folder.

Example 1: The number of individuals per subpopulations (10 subpopulations, 62 individuals each) are kept as originally provided

```
>STRUCTURE_ANALYSIS(Nruns=240,InputFileBase="M1_001",
  ResFolder="D:/SEB/PROJECTS/STRUCTURE/runs /M1/R-out1/",
  DatasetQualifier="Full_Pop10", PlotMemb=TRUE)
```

Results of this run will be placed in a new folder named:

```
"RES_Full_Pop10_ThreshGhost=0.5_ThreshMM=0.5-0.8_NewPopName=FALSE-1"
```

Example2: In this second example (with the same dataset as in example 1 above), the number of individuals per subpopulations (10 subpopulations, 62 individuals each) are replaced by new ones provided in the 'NewPopSizes' vector (here, 20 new subpopulations of 31 individuals each) :

```
>STRUCTURE_ANALYSIS(Nruns=240,InputFileBase="M1_001",ResFolder="D:/SEB/P
  ROJECTS/STRUCTURE/runs/M1/R-out1/", DatasetQualifier="Full_Pop10",
  PlotMemb=TRUE, NewPopName=TRUE, NewPopSizes=c(rep(c(31),20)))
```

Results of this run will be placed in a new folder named:

```
"RES_Full_Pop10_ThreshGhost=0.5_ThreshMM=0.5-0.8_NewPopName=TRUE-1"
```

## CHANGES FROM PREVIOUS VERSIONS

Changes from V1-12 to V1-13: Due to slight changes in R, the following error was appearing:

"Error in array(0, dim = c(Ndatasets, (nK + 1), NThresh), dimnames = c(NULL, : dimnames' must be a list "

This simple syntax problem (names of array dimensions) has now been fixed and Kestimator\_V1-13 now runs well with current R versions (Tested on version 3.4.0 Patched). The minor change in syntax does not affect the estimates obtained prior or after the change.

## REFERENCES

- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.
- Goudet J (1995) FSTAT (Version 1.2): a computer program to calculate F-statistics. *Journal of Heredity* **86**, 485-486.
- Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184-186.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Pritchard JK, Wen W (2004) Documentation for structure software: Version 2, p. 33. University of Chicago, Department of Human Genetics, Chicago (USA).
- Pritchard JK, Wen W, Falush D (2010) Documentation for structure software: Version 2.3, p. 38. University of Chicago, Department of Human Genetics, Chicago (USA).
- Puechmaille SJ (2016) The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources* **16**, 608-627.