

TESS version 2.3 - Reference Manual

August 2009*

Eric Durand

Chibiao Chen

Olivier François

Contents

1 Overview	2
2 What's New in Version 2.3	3
3 Method Description	3
3.1 Model Without Admixture	4
3.2 Model With Admixture (TESS > 2.0)	4
4 Dominant markers	5
5 Data Format	5
6 Using the GUI Shell	6
6.1 Main options	9
6.2 Changing the spatial network	10
6.3 Generate spatial coordinates	11
6.4 Generate geographic distances	13
6.5 Resampling individuals and loci	13
6.6 What are the dummy points?	16
6.7 Summarize project runs	16
6.8 Run existing projects	17
7 Using the Command-Line Engine	18
8 Model selection using the Deviance Information Criterion	20
9 Posterior predictive maps of admixture proportions	22
10 Post-processing TESS outputs: CLUMPP and R scripts	23
10.1 Averaging multiple runs	24
10.2 Spatial display of TESS outputs	25
11 Tutorial	25
12 FAQs	27

*contact: eric.durand@imag.fr or olivier.francois@imag.fr

1 Overview

TESS is a free computer program that implements a Bayesian clustering algorithm for spatial population genetic studies (Chen *et al.* 2007 for TESS 1.1, François *et al.* 2006). The method is based on a hierarchical mixture model where the prior distribution on cluster labels (without admixture model) or on admixture proportions (admixture model) is defined as a Hidden Markov Random Field (HMRF, no admixture model) or as a Hidden Gaussian Random Field (HGRF, admixture model) on a spatial individual network (tessellation). The program seeks population structure from individual multilocus genotypes sampled at distinct geographical locations without assuming predefined populations. It takes input data files in a format compatible with existing non-spatial clustering algorithms like STRUCTURE (Pritchard *et al.*, 2000), and returns membership probabilities and geographical cluster assignments of individuals.

TESS has been implemented using the C++ programming language. It contains a command-line engine and a Graphical User Interface (GUI) shell. The command-line engine is mainly designed for expert users who demand simplicity and flexibility and for users who need to batch-analyze a large amount of data. It accepts data files in a STRUCTURE-like format, including individual spatial coordinates in their first columns. It produces output results in both textual and graphical formats. The textual format stores the estimated membership probabilities and allele frequencies in a plain ASCII file. The text files may be post-processed using the DISTRUCT program (Rosenberg, 2004) or the CLUMPP program (Jakobsson and Rosenberg, 2007). There are four types of graphical outputs. The first graphics shows the log-likelihood history of a run, which can be used for the diagnosis of convergence of the algorithm; the second shows the estimated membership probabilities in a bar chart, as does DISTRUCT, so that users can check whether individuals share ancestry to multiple clusters or see the fraction of individual genomes assigned to each cluster (admixture model); the third displays the geographical assignment of individuals (hard clustering with geographical information); the fourth shows the convergence history of trend coefficients and can be used to diagnose their convergence. When invoked without any options, the command-line engine shows its typical usage with some explanatory notes. Mandatory options are input data file, number of individuals in the sample, ploidy, number of loci, (maximal) number of clusters, spatial interaction parameter, parameter of allele frequency model, the type of run (with or without admixture), total number of sweeps of MCMC, and burn in number of sweeps of MCMC. Other (optional) options will include the use of geographical distances between individuals, the use of dummy individuals, etc and they will be explained later.

The GUI shell can help newbies to familiarize themselves with the software, and it is generally a convenient way to use the TESS program. It provides facilities for creating and managing projects. A project is a coherent unit which groups the input data, the algorithmic parameter settings, and the output results altogether. Projects are saved in chosen folders automatically. By interacting with the GUI shell, users can check their data, specify the parameter settings, run the MCMC algorithm, and visualize the results without mastering the usage of the command-line engine.

TESS also provides a couple of additional features. It is able to start from a pre-specified clustering pattern obtained from the neighbor-joining algorithm, continue from a previous run and it allows its user to modify the automatically generated neighborhood system (accessible only via the GUI shell). It contains a routine that can help users to generate individual spatial coordinates from population sample coordinates or spatial

ranges and a routine that can compute geographic distances between individuals. It can perform multiple runs with distinct numbers of clusters, it can display a summary of all runs and sort them by their values of the Deviance Information Criterion (DIC), a statistical measure of the model prediction capabilities.

2 What’s New in Version 2.3

Version 2.3 improves version 2.2 and implements important changes to version 1.x and 2.0 including

- an improved model of admixture based on trend surfaces and on a conditional autoregressive model (a spatial Markov model similar to the Potts model used in the version without admixture),
- a neighbor-joining tree as the starting configuration for the initial clustering (optional),
- a display of run summaries, including the computation of the DIC for each run and the possibility to export output results to the CLUMPP format.
- the ability to compute prediction maps for the admixture model, predicting admixture values for geographic sites where no individual has been sampled.
- support of new data input format (“one row per individual”).
- support of dominant data for diploid individuals.

3 Method Description

Here we give a brief description of the methods used in the software. Denoting by (s_i) , $i = 1, \dots, N$, the set of observed sites, each s_i is surrounded by points which are closer to s_i than to any other sites. This set of points is known as the Dirichlet cell. We say that two sites are neighbors if their corresponding cells share a common edge. The TESS program is based on a hierarchical MRF model whose neighborhood system is obtained from the Voronoi tessellation (François *et al.*, 2006; Chen *et al.*, 2007). The program offers the possibility to modify the neighborhood system by connecting additional sites or by breaking links between sites. This option may be useful in order to include known geographical barriers, and it generally allows users to specify their particular individual network. The default weights on the network are set to one. For irregular sampling designs, it is possible and useful to incorporate weights that depend on geographic distance between sampling sites. This can be done via the “Compute Geographic Distance” option before creating projects. In the case where geographic distances are available, the weight between individuals i and j is set to $w_{ij} = \exp(-d_{ij}/\theta)$ where d_{ij} is the distance between individuals i and j and θ is a scale parameter (default value is the average distance between individuals).

The data $z = (z_{i\ell})$ consist of N multilocus genotypes obtained from individuals located at the sampled sites. An individual genotype z_i records paired alleles at L loci, where the number of possible alleles at locus ℓ is equal to J_ℓ . Although we are assuming diploid individuals, the program is also able to handle haploid individuals, by setting the ploidy parameter to $A = 1$ (default $A = 2$).

3.1 Model Without Admixture

We denote by c_i the cluster from which the individual i originates, and we assume the existence of at most K_{\max} clusters ($c_i \in \{1, \dots, K_{\max}\}$). In practice the constant K_{\max} could be considered larger than the true (or presumed true) number of clusters, K . As in the STRUCTURE program, the TESS program performs the statistical inference of the multidimensional parameter (c, f) with $c = \{c_i\}_{i=1, \dots, N}$ the cluster labels, $f = (f_{k\ell j})$, $k = 1, \dots, K_{\max}$, $\ell = 1, \dots, L$, and $j = 1, \dots, J_\ell$ the allele frequencies. The priors on allele frequencies are Dirichlet distributions $\mathcal{D}(\lambda, \dots, \lambda)$. The prior distribution on the set of cluster configurations is defined as a Gibbs distribution

$$\pi(c) = \exp[\psi U(c)]/Z, \quad c \in \{1, \dots, K_{\max}\}^N, \quad (1)$$

where ψ is a nonnegative parameter called the *interaction parameter*, $U(c)$ is the number of neighboring pairs that share the same labels in c , and Z is a normalizing constant called the *partition function*. With ψ equal to 0, this HMRF model assumes a non-informative spatial prior, and then corresponds to the (no admixture, uncorrelated allele frequencies) clustering model of Pritchard *et al.* (2000) which can be considered as a special case of our model. Typical values of the interaction parameter could be taken in the range $\psi \in (0.5, 1)$ for $K = 2 - 10$. Inferences on (c, f) are carried out by simulating the posterior distribution $\pi(c, f|z)$ through a Markov Chain Monte Carlo (MCMC) sampling algorithm.

3.2 Model With Admixture (TESS > 2.0)

In the model with admixture (Durand *et al.*, 2009), we assume that the individual genomes arise from the admixture of at most K_{\max} (potentially) unobserved parental populations. We estimate the fraction of individual i 's genome, q_{ik} , that originates in cluster k . The admixture model assumes that the q_{ik} are spatially autocorrelated, so that neighboring individuals are more similar than distant ones. The model used is

$$q_{i.} \sim \mathcal{D}(\alpha_{i1}, \dots, \alpha_{iK}),$$

where

$$\log(\alpha_{.k}) \sim \mathcal{N}(\mu_{.k}, \sigma_k^2 (Id - \psi W)^{-1}),$$

$\mu_{.k}$ is a mean trend effect, W is a weight matrix defined on the Voronoi tessellation, Id is the identity matrix, and ψ is a spatial interaction parameter. For all individuals $i \in \{1, \dots, N\}$, and all cluster labels $k \in \{1, \dots, K\}$, we denote $y_{ik} = \log(\alpha_{ik})$ and we have

$$y_{ik}|y_{jk}, j \neq i \sim \mathcal{N}\left(\mu_{ik} + \psi \sum_{j=1}^N w_{ij} (y_{jk} - \mu_{jk}), \sigma_k^2\right),$$

where μ_{ik} is the i th coordinate of the mean $\mu_{.k}$ and w_{ij} represents the weight of the interaction between individuals i and j . We have $\mu_{ik} = \tilde{X}_i \beta_k$, where \tilde{X}_i is a vector containing monomials of latitude and longitude (see section Main options to choose the degree), and β_k is a vector of regression coefficients which is estimated by the algorithm. If the degree of the trend surface is ≥ 1 , $\mu_{.k}$ models geographic trends. The term $\sum_{j=1}^N w_{ij} (y_{jk} - \mu_{jk})$ represents the part of randomness that can be captured by spatial autocorrelation, and that is likely to result from isolation-by-distance. In this new model, the spatial interaction parameter ψ represents the strength of the spatial autocorrelation. Note that this *interaction parameter* has not the same meaning as in the HMRF model, and the MCMC

algorithm is able to estimate ψ . It may be beneficial to run the model without admixture before trying to estimate admixture, because it could provide an upper bound on the number of clusters in the data. In the presence of clines and if the samples are irregularly spaced, the model without admixture might create patches along the clines. Performing runs of the admixture model may then be useful to further determine the presence of clines. The DIC could then be helpful to decide which patterns best explain the genetic data.

The above described model is called the CAR model, and TESS contains an alternative modelling of admixture, called the BYM model, which is described in (Durand *et al.*, 2009).

4 Dominant markers

For dominant markers, such as AFLPs, it is not possible to distinguish between the heterozygote genotypes and the dominant homozygote genotype. Following (Falush *et al.*, 2007) and starting from version 2.3, we implement a model to deal with dominant markers. For each loci, we assume that there may be a single allele that is recessive to all other alleles, while all other markers are codominant. Full details are given in (Falush *et al.*, 2007). In order to perform these computations the algorithm must be told which allele (if any) is recessive at each locus. This is done by checking the box “Row of recessive alleles” when creating a new project (see section Using the GUI Shell), and including a single row of L integers in the input `le`, between the (optional) extra lines and the genotypes. This row indicates the recessive allele at each of the L loci in the data set (see section Data Format). If at a given locus all the markers are codominant then the recessive value at that locus must be set to the missing data value (which must be a negative integer). If the phenotype is unambiguous, then it is coded in the input `le` as it is. If it is ambiguous then it is coded as homozygous for the dominant allele(s). For example, assume that allele A is recessive versus alleles B and C, and that B and C are codominant. Then, phenotype A is coded AA, B is coded BB, BC is coded BC. The genotypes AB, AC, etc are illegal in the input `le` when A is recessive.

5 Data Format

TESS allows the user to enter his/her data in two distinct formats. The first one (the same as in TESS 1.x) uses the same data format as STRUCTURE. This format uses two rows to represent the spatial coordinates and the multilocus genotype of one diploid individual. Besides the “pure data”, ie (XY)spatial coordinates + genotypic data, there can be additional rows or columns present in the data file for informational or other purposes. XY coordinates could be input either as longitude-latitude degrees or as standard coordinates in the metric system. However, when attempting to predict on an ascii-raster map, it is crucial that the first coordinate is longitude, and the second one latitude. TESS accepts population sample coordinates instead of individual coordinates. However, TESS will automatically add a small perturbation to population coordinates, so that all individuals have distinct coordinates. See also the “Generate Spatial Coordinates” paragraph to generate individual coordinates from population coordinates. All missing data should be represented by a single negative integer. The content of a data file should look like this:

```
No Info  X      Y      L1  L2  L3  L4  L5  L6  L7  L8  L9  L10
```

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10			
01	Info1	103.4	35.1	120	128	-9	129	156	234	148	124	182	98
01	Info1	103.4	35.1	120	128	-9	129	142	228	142	118	182	98
02	Info2	96.8	53.6	128	128	124	137	156	234	142	124	182	98
02	Info2	96.8	53.6	120	128	124	129	142	234	142	112	182	98
03	Info3	79.4	47.4	128	128	146	135	142	228	144	124	182	98
03	Info3	79.4	47.4	120	128	124	129	142	228	134	124	182	98
04	Info4	99.2	40.8	120	128	146	135	142	234	142	124	182	98
04	Info4	99.2	40.8	120	124	124	129	142	228	134	112	182	98
05	Info5	79.8	67.3	120	128	146	129	156	228	144	124	182	98
05	Info5	79.8	67.3	120	128	124	129	142	228	142	116	182	98
06	Info6	92.5	79.8	120	124	146	129	142	228	146	124	182	98
06	Info6	92.5	79.8	120	124	124	129	142	228	142	112	182	98
07	Info7	100.2	61.9	120	128	146	129	142	234	140	124	186	98
07	Info7	100.2	61.9	120	124	124	129	142	228	140	112	182	98
08	Info8	89.4	52.1	126	124	146	129	142	228	146	124	186	98
08	Info8	89.4	52.1	120	124	144	129	142	228	140	112	180	98
09	Info9	93.0	55.8	120	128	-9	129	156	234	146	116	-9	98
09	Info9	93.0	55.8	120	128	-9	129	142	228	142	112	-9	98
10	Info0	89.3	55.7	128	128	146	135	156	228	142	124	182	98
10	Info0	89.3	55.7	128	126	124	135	142	228	134	112	182	98

This example contains genotypic data from 10 diploid individuals at 10 loci and missing data are represented by “-9”. Note that TESS only makes use of the last columns in the data. Therefore, for this example data file, the “Number of Extra Rows” should be set to 1 and the “Number of Extra Columns” should be set to 2. The line containing $r1, \dots, r10$ is optional. Each $ri, i = 1 \dots 10$ denotes the recessive allele for locus i . It must be an integer. If no recessive allele is found at locus i , ri should be set to the missing data value (which must be a negative integer).

TESS also tolerates data stored in a format with one single row for each individual. For diploid individuals, twice the number of loci (plus two columns) are then required for encoding the whole data set, which should look like this (only the two first individuals and the 3 first loci are reported):

No	Info	X	Y	L1,1	L1,2	L2,1	L2,2	L3,1	L3,2
r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
01	Info1	103.4	35.1	120	120	128	128	-9	-9
02	Info2	96.8	53.6	128	120	128	128	124	124

Note that the recessive alleles (optional) are reported in the same way for the two data formats.

6 Using the GUI Shell

Screenshots presented in this section were taken under Linux. The actual GUI may differ visually a little bit under Microsoft Windows.

The GUI shell provides a convenient way to use TESS. It also helps newbies to familiarize themselves with the software. With help of the GUI shell, there is no need for users to understand and remember the command-line options. The GUI shell will call

the command-line engine internally and present the analysis results to users visually. The GUI shell can be launched by double-click “tessgui.exe” in the TESS home directory.

When using the GUI shell, users work with projects. A project is a coherent unit which groups the input data, the algorithmic parameter settings, and the output results altogether. To create a new project, access the menu “File⇒New Project...” or the corresponding button on the tool bar (see Figure 1). The GUI shell will show the “New Project” dialog box asking the user to key in the required project information (see Figure 2).

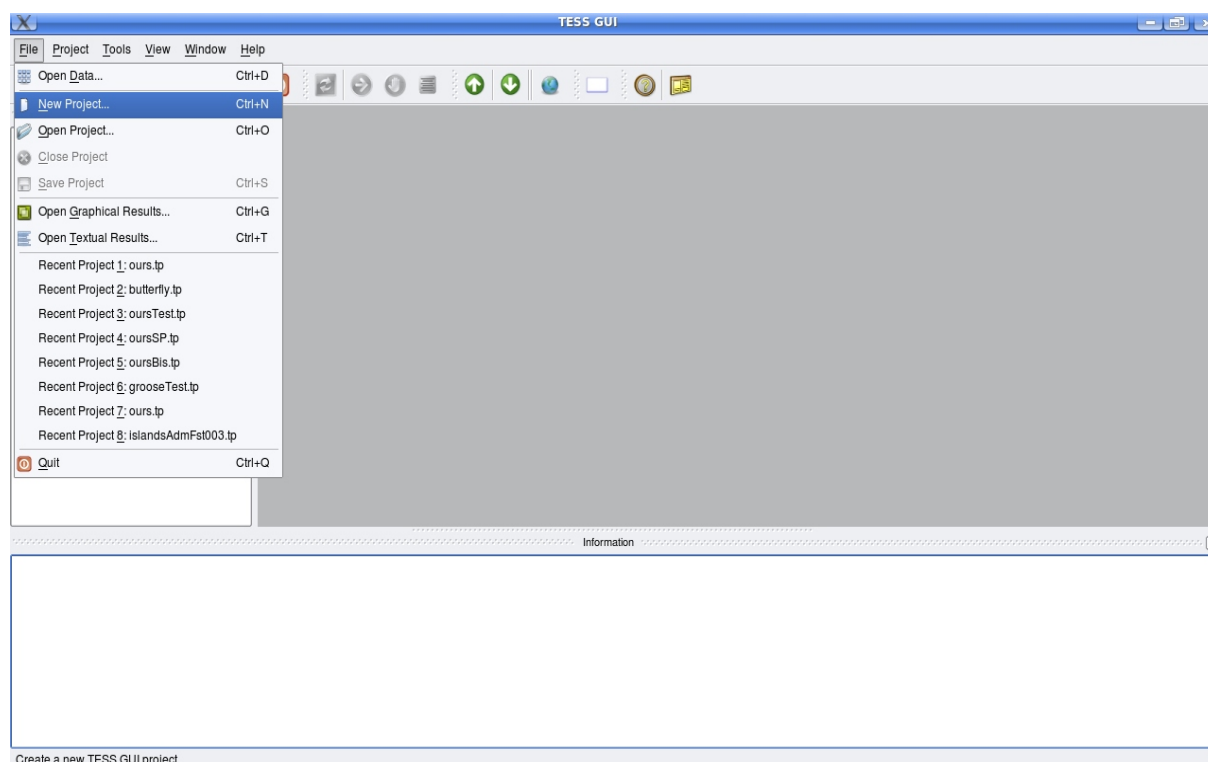
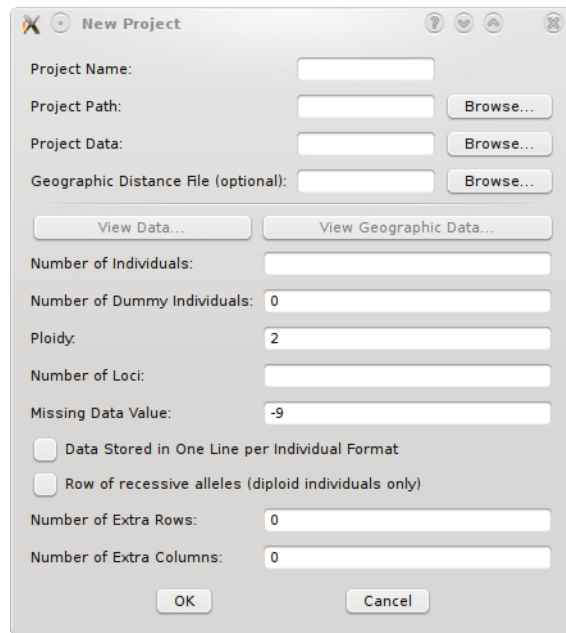


Figure 1: Using the GUI Shell - First Step: Creating Project

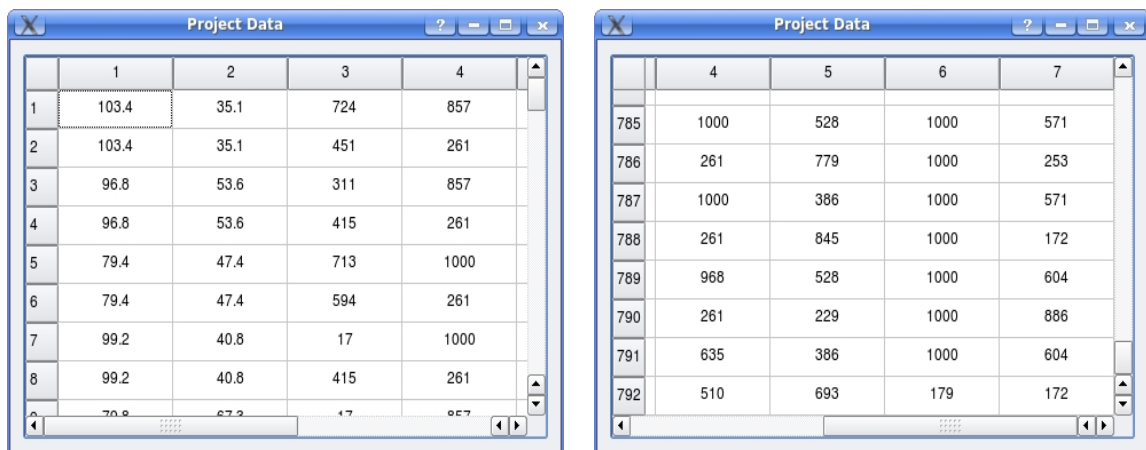
The user should name his/her project and choose the project path and data (tip: using the “Browse...” is a convenient way to input this information). Note that TESS requests its user to put his/her data in the same directory of the project file itself. We also recommend users to store different projects using separate directories for clear organization of information, although TESS does not request this. The user also needs to input the information and format of the project data. These include number of individuals, ploidy, number of loci, missing data value, number of extra rows, and number of extra columns. If the user is not clear of this information, he/she can always click the “View Data...” button to check the data first (see Figure 3). From the left picture in Figure 3, we can see there is no extra row or extra column and the first two columns store the spatial coordinates of individuals; from the right picture in Figure 3, we know there are 792 rows, therefore the number of individuals should be 396 and there are 7 columns, hence the number of loci should be 5. After viewing the data, user can close the “Project Data” window and continue to input information. Users should input information and format of the data with care. Although the software can catch most common errors, it is impossible to prevent users from making mistakes and wrong inputs may result in strange analysis results. The user can also enter a geographic distance file, which should contain pairwise



A dialog box titled "New Project" with the following fields and buttons:

- Project Name: [text box]
- Project Path: [text box] **Browse...**
- Project Data: [text box] **Browse...**
- Geographic Distance File (optional): [text box] **Browse...**
- View Data...** **View Geographic Data...**
- Number of Individuals: [text box]
- Number of Dummy Individuals: **0**
- Ploidy: **2**
- Number of Loci: [text box]
- Missing Data Value: **-9**
- ☐ Data Stored in One Line per Individual Format
- ☐ Row of recessive alleles (diploid individuals only)
- Number of Extra Rows: **0**
- Number of Extra Columns: **0**
- OK** **Cancel**

Figure 2: Input information for a new project.



Two windows titled "Project Data" showing data tables. The left window shows columns 1-4 and rows 1-8. The right window shows columns 4-7 and rows 785-792.

	1	2	3	4
1	103.4	35.1	724	857
2	103.4	35.1	451	261
3	96.8	53.6	311	857
4	96.8	53.6	415	261
5	79.4	47.4	713	1000
6	79.4	47.4	594	261
7	99.2	40.8	17	1000
8	99.2	40.8	415	261

	4	5	6	7
785	1000	528	1000	571
786	261	779	1000	253
787	1000	386	1000	571
788	261	845	1000	172
789	968	528	1000	604
790	261	229	1000	886
791	635	386	1000	604
792	510	693	179	172

Figure 3: View data to check the number of individuals, the ploidy, the number of loci, the missing data value, the number of extra rows, and the number of extra columns.

Figure 4: Input the level of spatial influence (with or without admixture), the number of runs, the maximal number of clusters, the total number of sweeps, the burn in number of sweeps, and the other parameters.

geographic distances between individuals. This file is optional. If the user does not enter a geographic distance file, uniform weights will be used on the neighborhood diagram. Otherwise, the neighborhood will incorporate weights that depend on the geographic distances between individuals (see section Method description). When finishing inputs, click the “OK” button to confirm the creation of the new project.

When the project is created, the GUI shell will automatically load it and show its data to the user. The user can then start to run the project by accessing the menu “Project⇒Run...” or the corresponding button on the tool bar. The GUI shell will show the “New Run(s)” dialog box asking the user to key in the required run information (see Figure 4).

6.1 Main options

The user should input the statistical method for estimating membership coefficients and allele frequencies (admixture or no-admixture model), the number of runs, the maximal number of clusters, the spatial interaction parameter, the parameter of allele frequency model, the total number of sweeps, and the burnin number of sweeps. We recommend starting with the no-admixture model. He/She can also choose whether to continue from the run with the lowest DIC (this option is disabled for the first run, since the configuration file is not available yet). He/She can start from the clustering pattern obtained by a neighbor-joining algorithm. Except the number of runs, the user can just accept the suggested values during the first trials.

The menu allows users to switch to the admixture model (default is a no-admixture model). In the admixture model, the user can choose the degree of trend surface (from 0 to 3), and the initial CAR variance. He/She can also choose whether to infer the CAR variance from data or not. Also, the user can modify the scale parameter θ (see sections Method description and Generate geographic distances) by clicking on the “Advanced

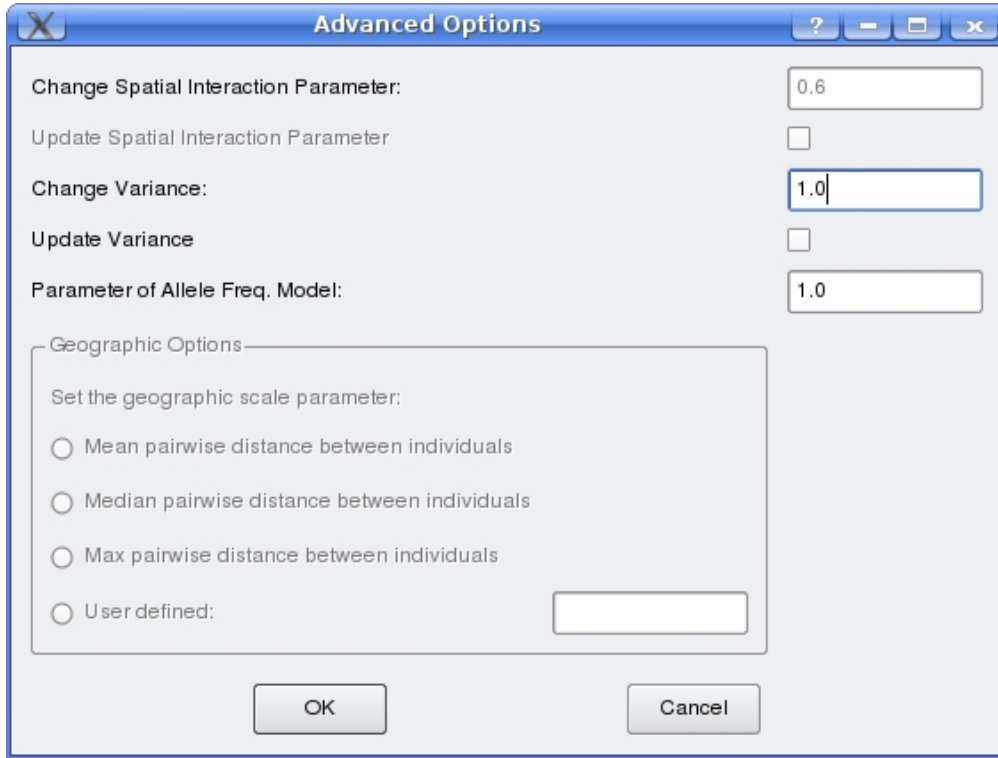


Figure 5: Set the advanced options: change the spatial interaction parameter, the model variance, the parameter of allele frequency model, and the geographic options.

Options” button (Figure 5).

The GUI shell can batch-run a project with same parameters and a range of values for K_{\max} . To batch-run with different parameters, the user still needs to use the command-line engine. After filling this information, the user can click the “OK” button to confirm run of the project with the specified parameters. The GUI shell will call the command-line engine to run the project. It will also create a sub-directory in the project directory for each run and saved the results of the run in that sub-directory. When a project is running, the GUI shell will display relevant information in the “Information” window. After a run is finished, user can double-click items under the run name in the “Project” window to load and view the results (see Figure 6).

Finally, the user can input a map in the ASCII-RASTER format where he/she wants to predict admixture coefficients (this option is available in the admixture model only). If used, the program will output one file per cluster (K files) in the output folder. Those files are named *"datafile"*Prediction_cluster_k.txt (where *"datafile"* is the name of the file containing data, e.g. example.dat), and they are in the ASCII-RASTER format. However, note that the prediction might be inaccurate if the run was too short, because the regression coefficients of the trend might have not converged yet.

6.2 Changing the spatial network

Modifying the TESS network may be useful for including realistic features within the analysis, like the existence of geographical barriers. The user can modify the generated neighborhood system by accessing the menu “Project⇒Modify Neighbor System...” or the corresponding button on the tool bar. The GUI shell will load the neighborhood system

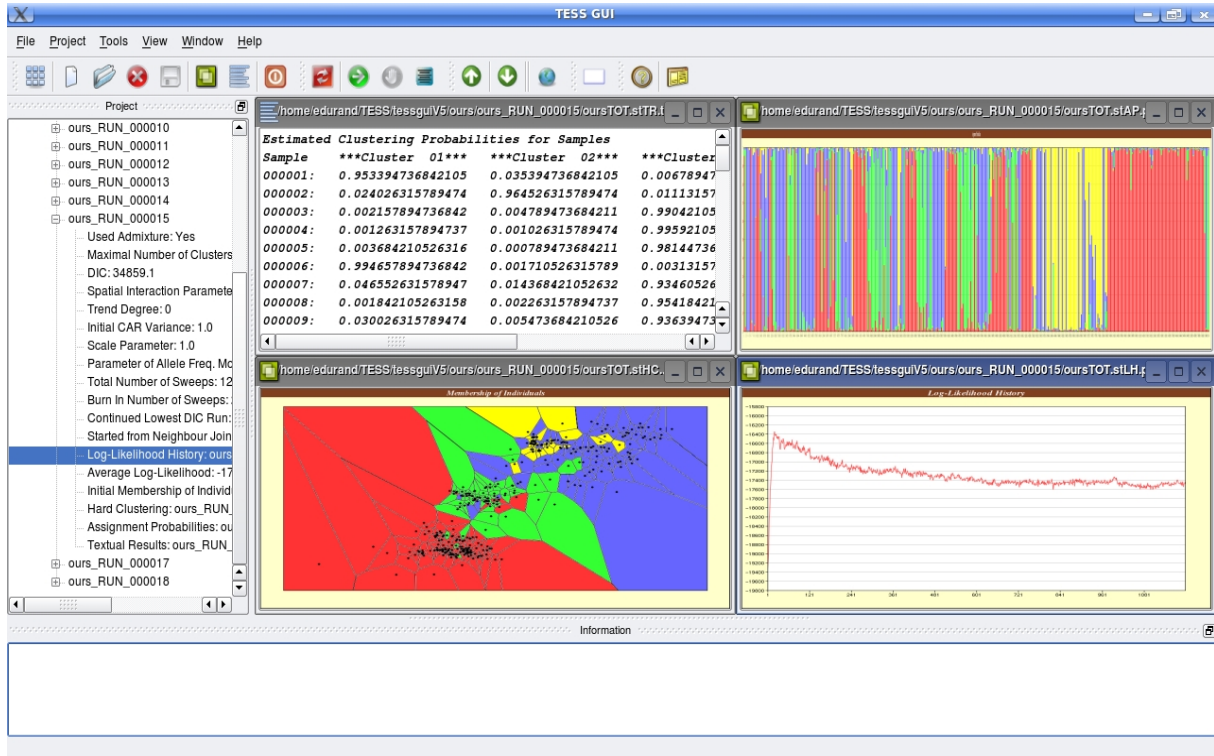


Figure 6: Check the results by double-click items under the run name in the “Project” window.

diagram and show the “Modify Neighborhood System” dialog box (see Figure 7). This can be done before or after runs of project. Modified neighborhood system will be used in subsequent runs. To modify the neighborhood system, first choose an individual; its current neighbors will be displayed in the “Neighbors” window. The user can remove neighbors for the chosen individual by select and move neighbors from the “Neighbors” window to the “Removal List” window; he/she can also add neighbors by select and move candidates from the “Candidates” window to the “Addition List”. After modification for one individual, the user can click “Apply” to accept the modification and continue to modify neighborhood for other individuals. At any time, the user can click “OK” to accept or click “Cancel” to reject the overall modifications. Alternatively, introducing dummy points may be a more direct way of modifying the spatial prior network of TESS.

6.3 Generate spatial coordinates

Many users have spatial coordinates available for population samples only, whereas TESS requests individual coordinates. Ideally users should use individual birth places as standard input to TESS in addition to the genotypic data. Nevertheless, the program can work (and can be useful) with partial information. It can manage population sample coordinates replicated for each individual in the population sample. In this case, individuals from the same population will be treated as unrelated. Be aware that the hard-clustering output will be misleading when using population sample coordinates, because only the value of the first individual labelled in the sample will be used to color the Voronoi cell. The membership coefficient bar chart will nevertheless be correct. The program will always work better with individual coordinates, and we then recommend to slightly modify the sample geographic coordinates, **before starting a project**.

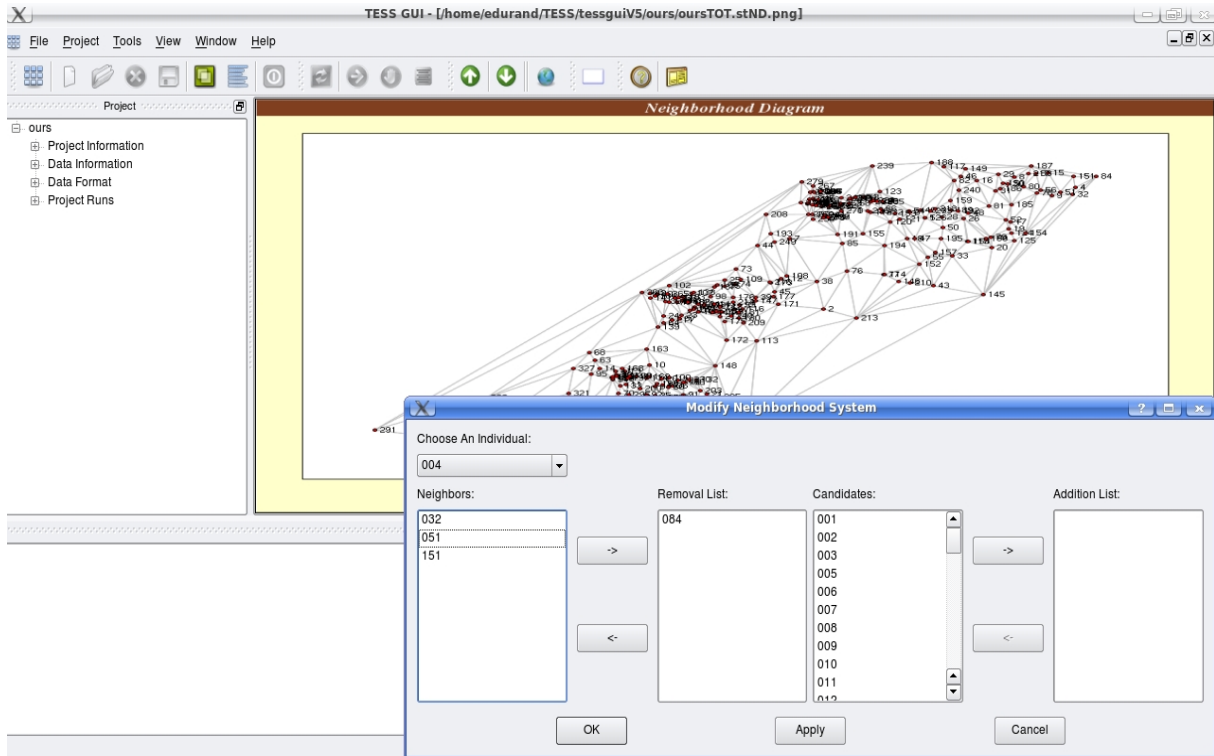


Figure 7: Visually modify the neighborhood system.

The “Generate Spatial Coordinates” option (Figure 8) can be used to generate random spatial coordinates from population sample coordinates or from prescribed ranges of coordinates. This option should be used prior to any run. Simulations from prescribed ranges of spatial coordinates can be a convenient way to use TESS when the individual birth places are not available to the study. The user will be asked to input a 7 column data file containing information on the number of individuals in each sample, the min X coordinate, the max X coordinate and a SD value, the min Y coordinate, the max Y coordinate and a SD value for each sample. The following example of a spatial coordinate file has 2 population samples with 4 and 5 individuals. The X range in pop 1 is (7.4, 9.2) and the Y range in pop 2 is (3.1, 6.5). The program will draw random coordinates within the prescribed range

```
4 7.4 9.2 1.0 3.1 6.5 1.0
5 1.4 6.1 1.0 7.2 9.5 1.0
```

If $X_{\min} = X_{\max}$, the program will use the SD parameter to sample from a normal distribution.

```
4 7.4 7.4 1.0 3.1 6.5 1.0
5 1.4 6.1 1.0 7.2 9.5 1.0
```

The program also asks the user to input a genotype matrix (see the Data Format section) and then combines the simulated spatial coordinates and the multilocus genotypes into a file with the correct TESS format.

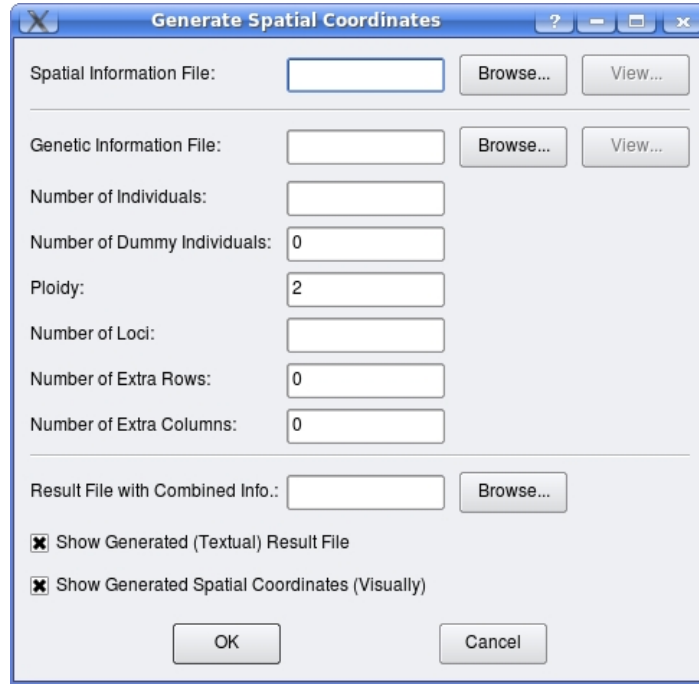


Figure 8: The “Generate Spatial Coordinates” option.

6.4 Generate geographic distances

TESS enables users to compute pairwise geographic distances between individuals from their spatial coordinates. The utility is accessible via the Tool menu (see Figure 9). It allows users to compute great circle distances or Euclidian distances between individuals. The user can input a file containing the whole data set ((extra rows and columns) + coordinates + genotypes) or a file with spatial coordinates only. Both input file formats (one line per individual or 2 lines per diploid individuals) are supported. If two individuals are too close to each other (which may happen when using perturbed coordinates), the program may not be able to compute the great circle distance because of numerical issues. In this case, the program computes the Euclidian distance instead (which should be roughly the same at short ranges). The geographic distances should be computed prior to creating a TESS project. Users cannot add the geographic distances to an existing project. If the geographic distance file is available, the algorithm will weight the Voronoi neighborhood with $w_{ij} = \exp(-d_{ij}/\theta)$ (see section Method description), where d_{ij} is the geographic distance and θ is a scale parameter (see section Main options). Otherwise, the algorithm sets $w_{ij} = 1$ for all pair of vertices i, j in the TESS network.

6.5 Resampling individuals and loci

This option of the program enables the user to create subsets of data with either specified or randomly chosen sets of individuals and loci (see Figure 10)

Generate Geographic Distances

☒ Compute Great Circle Distances ☐ Compute Euclidian Distances

File Containing Spatial Information (TESS format): **Browse...** **View...**

☐ Data Stored in One Line per Individual Format

Number of Individuals:

Ploidy:

Number of Extra Rows:

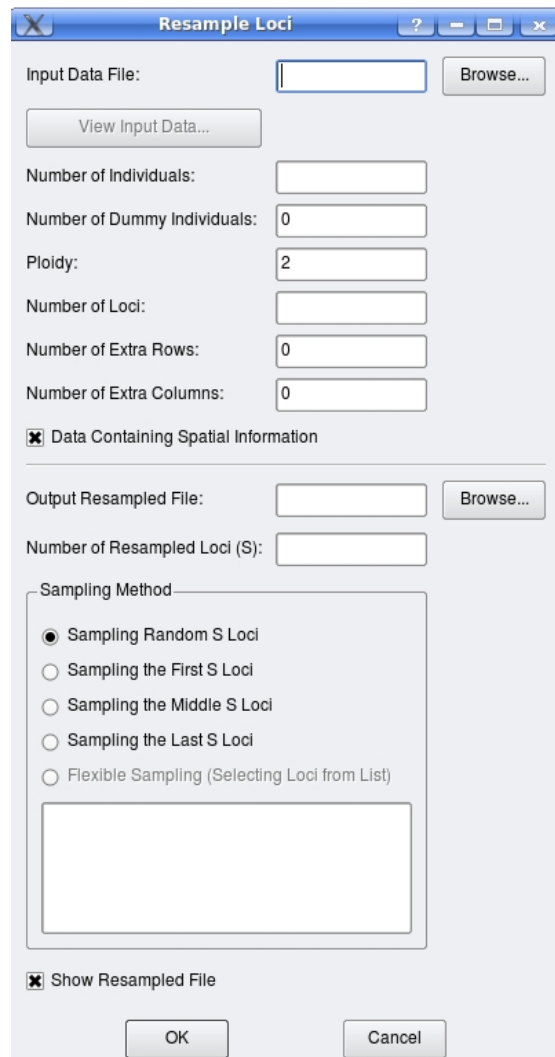
Number of Extra Columns:

Result File with Geographic Distances: **Browse...**

☒ Show Generated (Textual) Result File

OK **Cancel**

Figure 9: The “Generate Geographic Distances” option.



The image shows a software dialog box titled "Resample Loci". It contains several input fields and checkboxes for configuring a resampling process. The fields include "Input Data File", "Number of Individuals", "Number of Dummy Individuals", "Ploidy", "Number of Loci", "Number of Extra Rows", and "Number of Extra Columns". There are also checkboxes for "Data Containing Spatial Information" and "Show Resampled File". A "Sampling Method" section contains five radio button options: "Sampling Random S Loci" (selected), "Sampling the First S Loci", "Sampling the Middle S Loci", "Sampling the Last S Loci", and "Flexible Sampling (Selecting Loci from List)". A text area is located below the sampling methods. At the bottom are "OK" and "Cancel" buttons.

Resample Loci

Input Data File: Browse...

View Input Data...

Number of Individuals:

Number of Dummy Individuals:

Ploidy:

Number of Loci:

Number of Extra Rows:

Number of Extra Columns:

☒ Data Containing Spatial Information

Output Resampled File: Browse...

Number of Resampled Loci (S):

Sampling Method

☒ Sampling Random S Loci

☐ Sampling the First S Loci

☐ Sampling the Middle S Loci

☐ Sampling the Last S Loci

☐ Flexible Sampling (Selecting Loci from List)

☒ Show Resampled File

OK Cancel

Figure 10: The resampling option.

6.6 What are the dummy points?

Using dummy points is another way to modify the TESS network topology. Dummy individuals should indeed represent points where no population can be sampled (i.e. major water-masses, high mountain ranges, desert, etc). Dummy points could also be placed between very distant samples, and may cut the links between individuals at the opposite sides of the dummy cell. Dummy points will actually create additional white cells which may annihilate the local correlations between individuals.

Dummy points must be placed at the end of the data set, and they must be input in the last rows. They can be coded as standard individuals. Because their genotypic data will be ignored, a convenient way to input dummy points is by using the missing value symbol at all loci (eg -9). Here is an example data set with 10 diploid individuals and 2 additional dummy individuals, placed at the end of the data set. In this example, empty cells will be located at points with spatial coordinates (84.5, 45.7) and (92.3, 52.7).

No	Info	X	Y	L1	L2	L3	L4	L5	L6	L7	L8	L9	LA
01	Info1	103.4	35.1	120	128	-9	129	156	234	148	124	182	98
01	Info1	103.4	35.1	120	128	-9	129	142	228	142	118	182	98
02	Info1	96.8	53.6	128	128	124	137	156	234	142	124	182	98
02	Info1	96.8	53.6	120	128	124	129	142	234	142	112	182	98
03	Info1	79.4	47.4	128	128	146	135	142	228	144	124	182	98
03	Info1	79.4	47.4	120	128	124	129	142	228	134	124	182	98
04	Info1	99.2	40.8	120	128	146	135	142	234	142	124	182	98
04	Info1	99.2	40.8	120	124	124	129	142	228	134	112	182	98
05	Info1	79.8	67.3	120	128	146	129	156	228	144	124	182	98
05	Info1	79.8	67.3	120	128	124	129	142	228	142	116	182	98
06	Info1	92.5	79.8	120	124	146	129	142	228	146	124	182	98
06	Info1	92.5	79.8	120	124	124	129	142	228	142	112	182	98
07	Info1	100.2	61.9	120	128	146	129	142	234	140	124	186	98
07	Info1	100.2	61.9	120	124	124	129	142	228	140	112	182	98
08	Info1	89.4	52.1	126	124	146	129	142	228	146	124	186	98
08	Info1	89.4	52.1	120	124	144	129	142	228	140	112	180	98
09	Info1	93.0	55.8	120	128	-9	129	156	234	146	116	-9	98
09	Info1	93.0	55.8	120	128	-9	129	142	228	142	112	-9	98
10	Info1	89.3	55.7	128	128	146	135	156	228	142	124	182	98
10	Info1	89.3	55.7	128	126	124	135	142	228	134	112	182	98
D1	Dummy	84.5	45.7	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
D1	Dummy	84.5	45.7	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
D2	Dummy	92.3	52.7	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
D2	Dummy	92.3	52.7	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9

6.7 Summarize project runs

The summary table allows the user to quickly access the principal information of all runs, including their value of K_{\max} and of the DIC. It is also linked to the main window of the GUI. For example, clicking on the path of the log-likelihood history in the summary table displays it in the main window. The table also allows the user to sort runs by their values of the DIC or K_{\max} , or simply according to their labels by clicking on the corresponding column label. The "Lower Range" and "Upper Range" drop-down menus allow the user to

select which runs to display in the summary table. By default, all runs are displayed. The average DIC is computed over the selected lower and upper runs. The summary table is dynamically updated. One can still use TESS while the summary window is open, and the summary table will be automatically populated with any new run. Likewise, deleting a run automatically removes it from the table. Figure 11 illustrates the summary window.

This window also provides a tool to easily export runs to the CLUMPP format. CLUMPP can only average estimates of models with the same value of K_{\max} . A drop-down menu allows the user to select a particular value of K_{\max} . A list of candidate runs is then automatically filled with runs between the "lower" and "upper" ranges that correspond to the selected value of K_{\max} . The user can then further filter the candidate runs by choosing to keep only a percentage of them using the "Select Lowest DIC runs" drop-down menu. For example, selecting 10% in this menu will automatically remove all runs from the candidate list that are not within the 10% runs with the lowest DIC values. Then, the user selects runs from the candidate list (multi-selection is enabled) and moves them to the "Runs to Export" list by clicking on the right-arrow button. It is still possible to remove runs from the rightmost list by selecting them and clicking of the left-arrow button. Finally, the user chooses a file to store the exported runs in the CLUMPP format (the popfile). The file will be created if it does not exist. If the "Write CLUMPP paramfile" button is ticked, an additional file called "paramfile" is created in the same directory as the selected popfile. This file contains all the parameters needed to run CLUMPP. The user simply needs to copy the popfile and the paramfile to the CLUMPP directory and run the CLUMPP software.

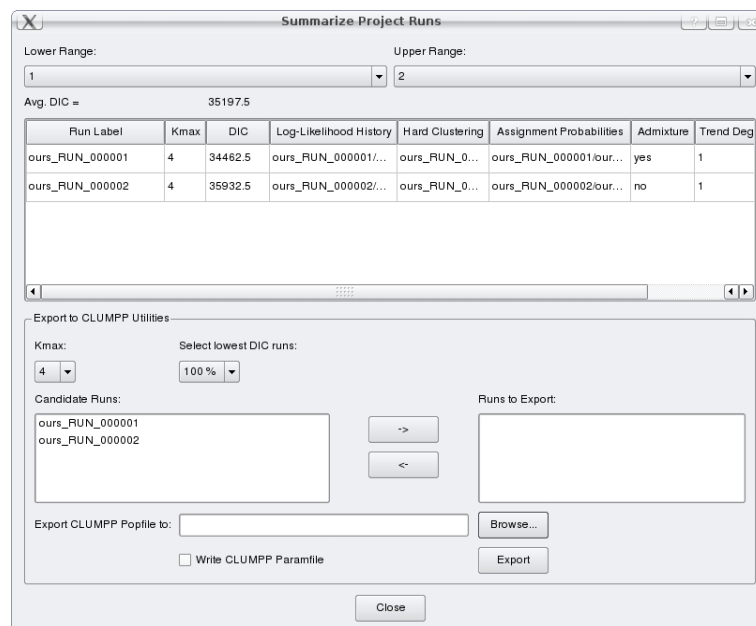


Figure 11: Summary of project runs

6.8 Run existing projects

By accessing the "File" menu, user can also load and run existing (or recent) projects, view data, view graphical results, and view textual results (see Figure 12).

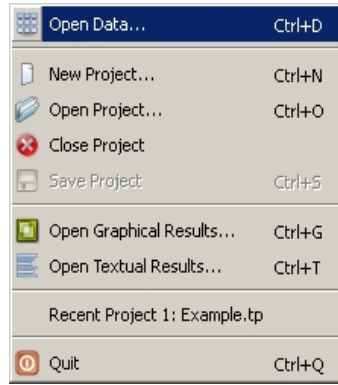


Figure 12: File Menu

7 Using the Command-Line Engine

In fact, there are three command-line engines: one for the model without admixture, called ‘tessm.exe’, and the other two for the admixture models, named ‘tessmAdmCAR.exe’ (CAR model) and ‘tessmAdmBYM.exe’. First, we detail the command-line engine for the model without admixture. It can be invoked in the following way: on the Windows platform, first launch the ‘Command Prompt’, then change directory to the TESS home and invoke the ‘tessm.exe’ by typing ‘tessm’ followed by its options. When there are no options given to tessm.exe, it will show its typical usage (see Figure 13) and exit. Let us see an example: say that there is a data file ‘Data’ in the same directory of tessm.exe, which contains 500 diploid individuals genotyped at 10 loci. Assuming there are at most 8 subpopulations, we set the interaction parameter of HMRP to 0.6 and the parameter of Dirichlet allele frequency model to 1.0. If we want to run the MCMC algorithm (no admixture model) for a total 12,000 sweeps with the first 2,000 sweeps discarded as burn-in, we can use the following command:

```
tessm -FData -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
```

The program will start to run and report its progress and timing. Four result files will be produced in the same directory: ‘DataLH.png’, ‘DataAP.png’, ‘DataHC.png’, and ‘DataTR.txt’. DataLH.png shows the log-likelihood history of the run; DataAP.png contains a bar chart that shows the estimated assignment probabilities of individuals to different clusters; DataHC.png shows the final geographical assignment of individuals; DataTR.txt contains the estimated assignment probabilities and allele frequencies in text format. User can also check ‘DataVD.png’ for the Voronoi tessellation of the data, ‘DataND.png’ for the generated neighborhood system based on the Voronoi tessellation, and ‘DataIM.png’ for initial geographical assignment of individuals. These files are for information only. Use a picture viewer to view the ‘.png’ files and a text editor to view the ‘.txt’ files, or use the GUI shell to manipulate the command-line engine (then users can analyze their data and visualize the results without using any external utilities).

A configuration file called ‘DataCF’ is also generated each time a run obtains a lower DIC value (always generated for the first run). The file can help the user to continue a run by utilizing the results generated from a previous run. To continue from the run with the lowest DIC value, issue this command:

```
tessm -FData -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000 -pDataCF
```

```

C:\Documents and Settings\Olivier\Bureau\tess>tessm
Typical Usage: tessm -FF -NN -AA -LL -KK -PP -DD -SS -BB
Options can be specified in any order.

Required Options:
-FF: F = File Name of Input Data File
      For file name contains spaces, input it as "F".
-NN: N = Number of Individuals
-aa: A = Ploidy (1 = Haploid, 2 = Diploid, ...)
-LL: L = Number of Loci
-KK: K = Number of Clusters
-PP: P = Interaction Parameter of HMMRF
-DD: D = Parameter of Dirichlet Allele Frequency Model (no F-model)
-SS: S = Total Number of Sweeps of MCMC
-BB: B = Burn In Number of Sweeps of MCMC

Optional Options:
-qq: q = with or without (Default) admixture
      q = y, Yes
      q = n, No (Default)
-ff: f = with or without (Default) F-model
      f = y, Yes
      f = n, No (Default)
-xx: x = Parameter alpha for admixture, Default 1.0
-ll: l = Parameter Lambda for F-model, Default 1.0
-rr: r = Number of Extra Rows in Data File, Default: 0
-cc: c = Number of Extra Columns in Data File, Default: 0
-dd: d = Number of Dummy Individuals, Default: 0
-ii: i = Folder Name of Input Data File, Default: Current Folder
      For folder name contains spaces, input it as "i".
-oo: o = Folder Name of Output Result Files, Default: Current Folder
      For folder name contains spaces, input it as "o".
-pp: p = Configuration File for Continue from Previous Run, Default: NULL
      For file name contains spaces, input it as "p".

C:\Documents and Settings\Olivier\Bureau\tess>

```

Figure 13: TESS Command-Line Usage (without admixture)

By continuing from a previous run, a user can gradually improve the analysis result until his/her satisfaction. This can save a lot of time when he/she is exploring a large, complex data set.

To read data file from a sub-directory “D Dir” and put result files into a sub-directory “R Dir”, use the following command:

```
tessm -FData -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000 -i"D Dir" -o"R Dir"
```

For file (directory) name contains spaces, enclose it into quotes, as did in the above example. The data directory and result directory must already exist (TESS won’t create these directories).

Should a user have a large number of data files to analyze, he/she is recommended to use the command-line engine and he/she should put the commands into a batch file (“.bat”). In this way, the data can be automatically analyzed in an unattended manner. The content of an example batch file can be:

```

tessm -FData0 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData1 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData2 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData3 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData4 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData5 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData6 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData7 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData8 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000
tessm -FData9 -N500 -A2 -L10 -K8 -P0.6 -D1.0 -S12000 -B2000

```

Now, we shortly describe the command-line engine for the model with admixture. It can be invoked in the following way: on the Windows platform, first launch the “Command Prompt”, then change directory to the TESS home and invoke “tessmAdmCAR.exe”

(or "tessmAdmBYM.exe") followed by its options. When no options are given to tessmAdmCAR.exe, it will show its typical usage and exit (see Figure 14).

As an example, we want to analyze the same data set as illustrated for the program without admixture: say that there is a data file "Data" in the same directory of tessm.exe, which contains 500 diploid individuals genotyped at 10 loci. Assuming there are at most 8 subpopulations, we set the initial spatial interaction parameter to 0.6, the parameter of Dirichlet allele frequency model to 1.0, and the trend degree to 2. In addition, we want the algorithm to estimate the spatial interaction parameter. If we want to run the MCMC algorithm (admixture model) for a total 12,000 sweeps with the first 2,000 sweeps discarded as burn-in, we can use the following command:

```
tessmAdmCAR -FData -N500 -A2 -L10 -K8 -P0.6 -T2 -D1.0 -S12000 -B2000 -upy
```

The "-upy" argument tells TESS to update the spatial interaction parameter. The program will produce all the result files described for the program without admixture. In addition, it will produce a file named "DataBH.png" which shows the convergence history of the regression coefficients. It will also create a text file named "DataBetaHat.txt" which contains the estimated regression coefficients.

8 Model selection using the Deviance Information Criterion

In order to choose between alternative models, TESS calculates the Deviance Information Criterion (DIC) for each run. The DIC is a statistical measure of the model prediction capabilities. It is computed as the model deviance penalized by an estimate of the effective number of parameter, which is a measure of model complexity (Spiegelhalter *et al.*, 2002). The DIC is well-adapted to Markov chain Monte Carlo simulations, and its principle is that models with smaller DIC should be preferred to models with larger DIC. The DIC may be used either to select the number of clusters, or, for example, to decide whether spatial autocorrelation is necessary to explain the data or not. Assume that θ is a vector containing all the algorithm parameters. Let us denote $D(\theta)$ the model deviance for the parameter set θ , computed as -2 times the log-likelihood, and p_D the effective number of parameters in the model. We have

$$\text{DIC} = \bar{D} + p_D ,$$

where \bar{D} denotes the posterior mean of the deviance. There are different ways to compute p_D . We follow (Spiegelhalter *et al.*, 2002; Gelman *et al.*, 2003) by setting

$$p_D = \bar{D} - D(\bar{\theta}) ,$$

i.e., the difference between the posterior mean of the deviance and the deviance taken at the posterior mean of the parameters. Roughly speaking, p_D represents the gain in fit expected when estimating the model parameters.

The DIC can be used to select the maximal number of cluster (model without admixture) or the maximal number of parental populations (admixture model), K_{\max} , that best suits the data, in a way similar to the ΔK criterion traditionally used for STRUCTURE (Evanno *et al.*, 2005). Alternatively, the DIC can be used to compare different admixture models (or no-admixture models). For example, one can use the DIC to compare models

```

C:\WINDOWS\system32\cmd.exe
C:\Documents and Settings>cd ..
C:\>tessmAdmCAR.exe
Typical Usage: tessm -FF -NN -AA -LL -KK -PP -DD -SS -BB
Options can be specified in any order.

Required Options:
-FF: F = File Name of Input Data File
      For file name contains spaces, input it as "F".
-NN: N = Number of Individuals
-AA: A = Ploidy (1 = Haploid, 2 = Diploid, ...)
-LL: L = Number of Loci
-KK: K = Number of Clusters
-TT: T = Degree of Trend
-PP: P = sPatial interaction parameter (default: 1.0)
-DD: D = Parameter of Dirichlet Allele Frequency Model (no F-model)
-SS: S = Total Number of Sweeps of MCMC
-BB: B = Burn In Number of Sweeps of MCMC

Optional Options:
-rr: r = Number of Extra Rows in Data File, Default: 0
-cc: c = Number of Extra Columns in Data File, Default: 0
-ii: i = Folder Name of Input Data File, Default: Current Folder
      For folder name contains spaces, input it as "i".
-oo: o = Folder Name of Output Result Files, Default: Current Folder
      For folder name contains spaces, input it as "o".
-pp: p = Configuration File for Continue from Highest Likelihood Run, Default: N
ULL
      For file name contains spaces, input it as "p".
-dd: d = Number of dummy individuals, Default: 0
-uu: u = Update sigma?
      u = y, Yes (Default)
      u = n, No
-upup: up = Update Psi?
      up = y, Yes
      up = n, No (Default)
-xx: x = admixture parameter, Default: 30.0
-mm: m = Name of file containing map (ascii-raster format) to predict admixture
      coefficients on (optional)
      For file name contains spaces, input it as "m".
-nn: n = number of coordinates to use in regression. Default: 2.
      If n = 1, the second coordinate is used. n=0 dis the same as T=0
-ss: s = Shuffle update order for MCMC?
      s = y, Yes
      s = n, No (Default)
-jj: j = Init the run with Neighbor Joining tree
      j = y, Yes
      j = n, No (Default)
-gg: g = Name of file containing Geographic distances between individuals (optio
nal)
      For file name contains spaces, input it as "g".
-scsc: sc = scale parameter for geographic distances. User predefined value or e
nter one. Ignored if no geographic distance file is given.
      sc = m, Mean distance between individuals (default)
      sc = d, median distance between individuals
      sc = x, maX distance between individuals
      sc = value of the scale parameter (e.g. 10.4)
-spsp: sp = SPecial data format: one individual = one row
      sp = y, Yes
      sp = n, No (Default)
-bsbs: bs = Burn-In Structure, Default: 0
C:\>_

```

Figure 14: TESS Command-Line Usage (with admixture)

with different trend degrees. However, we do not recommend to use the DIC to compare between the admixture and the no-admixture models.

An important intrinsic feature of imposing spatially structured priors is the possibility for the MCMC algorithm to eliminate a number of spurious clusters automatically. When we input a maximum of K_{\max} clusters to the model, the effective number of cluster in the data may be a smaller value, K . In this case, the DIC sometimes selects models in which K_{\max} is greater than K . Section 11 illustrates the use of the DIC to perform a full TESS analysis.

9 Posterior predictive maps of admixture proportions

The admixture models implemented in TESS allow us to predict expected admixture proportions on every point of a map. To do so, the user needs to provide a map in the ASCII-raster format, which is a standard way of encoding geographical information into a file. The ASCII-raster format encodes the map into a matrix, which represents a row by row series of space-delimited ASCII depth values. Each element of the matrix represents the depth value of a particular location on the map. Such a map can be imported from a standard Geographic Information System (GIS). The basic structure of an ASCII-raster map has the header information at the beginning of the file followed by the cell value data:

```
NCOLS xxx

NROWS xxx

XLLCENTER xxx | XLLCORNER xxx

YLLCENTER xxx | YLLCORNER xxx

CELLSIZE xxx

NODATA_VALUE xxx

row 1

row 2

...

row n
```

Here is an example of ASCII-raster map (only the first line is displayed here) :

```
#NCOLS    188
#NROWS    132
#XLLCENTER 0
#YLLCENTER 0
#CELLSIZE 1
```

```

#NODATA_VALUE -99
-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
-99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
-99 -99 -99 -99 -99 -99 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-99 -99 1 -99 -99 -99 1 1 1 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99

```

The first six lines of the ASCII-raster file describes the map parameters. Their meanings are listed in table 1 below.

Parameter	Description	Requirements
NCOLS	Number of cell columns	Integer greater than 0
NROWS	Number of cell rows	Integer greater than 0
XLLCENTER or XLLCORNER	X coordinate of the origin (by center or lower left corner of the cell)	Match with Y coordinate type
YLLCENTER or YLLCORNER	Y coordinate of the origin (by center or lower left corner of the cell)	Match with X coordinate type
CELLSIZE	Cell size	Greater than 0
NODATA_VALUE	Location with no data (where no prediction will be computed)	Optional Default is -9999

Table 1: ASCII-raster parameters explained.

Users input the map as a run parameter in TESS (see section 6.1). It is crucial that the individual coordinates in TESS input data are ordered as longitude followed by latitude for the prediction to work properly. The program will then generate K_{\max} new files in the output directory. Each file contains the predicted value on every cell of the map that is not equal to NODATA_VALUE. Typically, NODATA_VALUE can code for ocean if the studied species is terrestrial. The user can then use a GIS to display the posterior predictive map. Alternatively, one can use the R software (for instance using the `image()` function and replacing the NODATA_VALUES by NA values). Figure 15 shows an example of a posterior predictive map displayed using the R software. It corresponds to a simulation analysis with $K = 3$ clusters. The 3 clusters are displayed on the same map.

10 Post-processing TESS outputs: CLUMPP and R scripts

In this section, we describe how to post-process TESS outputs. In particular, we explain briefly how to average multiple runs using CLUMPP (Jakobsson and Rosenberg, 2007) to correct for label switching. In addition, we show how to display TESS (or CLUMPP) output spatially.

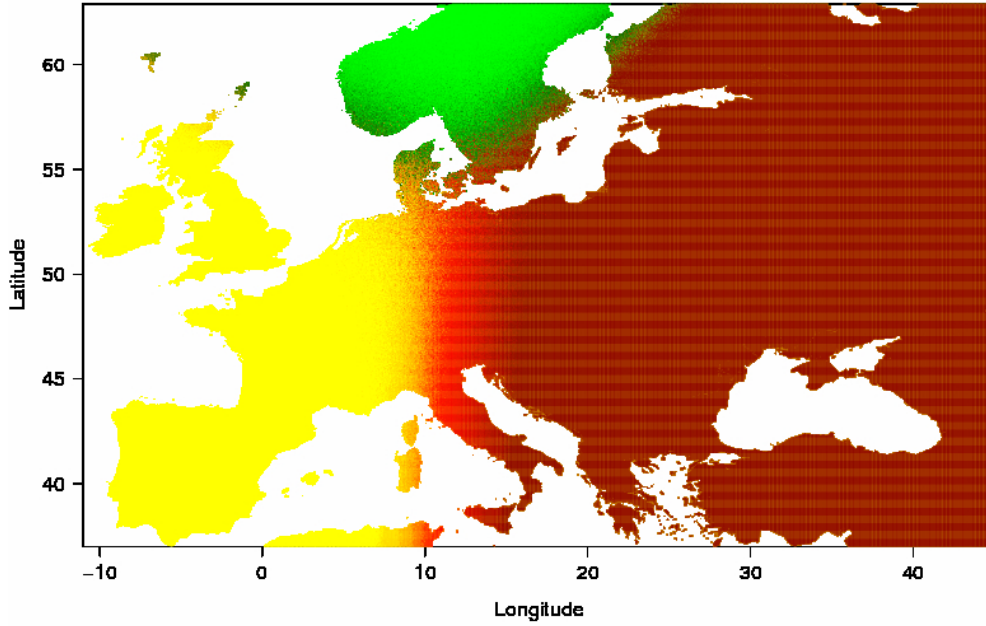


Figure 15: Posterior predictive map for a data set simulating a demographic expansion in Europe with $K = 3$ clusters. The map was drawn with the R software.

10.1 Averaging multiple runs

A general shortcoming with MCMC algorithms is that it is difficult to assess convergence. Indeed, the state space of clustering algorithms is typically huge and the algorithms will most likely fail to explore it entirely within a reasonable computing time. To overcome this difficulty, a good strategy is to run the algorithm several times (independently) with the same parameter values (typically, the same value for the maximal number of clusters, K_{\max}) and to average the results over the different runs. However, this cannot be done in a straightforward way because of label switching in independent runs. Because the cluster labels are arbitrarily chosen, a particular label will not always be associated to the same clustered individuals in distinct runs. For example, assume we are analyzing a data set containing 100 individuals. We run TESS with $K_{\max} = 2$ and we detect that the 50 first individuals are assigned to cluster 1, the other ones being grouped into cluster 2. Then, we run TESS again with $K_{\max} = 2$, detecting the same structure. However, this time the first 50 individuals are assigned to cluster 2 and the other ones to cluster 1. Obviously, we cannot directly average the results from the two runs (for instance, the average membership coefficients would be .5 in both clusters for all individuals), and one would need to permute the individuals first. To do so in an automatic fashion, Jakobsson and Rosenberg (2007) developed the software CLUMPP. It takes an input file containing the estimated membership coefficients for multiple independent runs, and averages them after correcting for label switching. TESS graphical user interface (GUI) provides a tool to easily export multiple runs into a CLUMPP input file. See section 6.7 for more details.

10.2 Spatial display of TESS outputs

In order to display TESS estimations of admixture proportions spatially, we provide an R script that interpolates expected admixture proportions on every point on a grid. The script can be found in the file `R Script/krigAdmixProportions.R` within the TESS directory. The interpolation technique is known as universal kriging. We do not provide technical details on universal kriging here. The interested reader can find further details and references in (Ripley, 1981). Figure 16 shows an example of such an interpolation. In order to use the R script, one first needs to store the estimated admixture proportions in a text file in the CLUMPP output format. There are two ways to do so. The first (and easiest) one is to directly copy-paste the estimates from TESS textual results into a new text file. The other option is to use the "Export to CLUMPP" feature implemented in TESS (see section 6.7). Let us assume we have created a file named *estimates.txt* with one of the two methods, which contains the estimated individual admixture proportions in the CLUMPP format. Also, let us denote *data.txt* the file containing the TESS data. Furthermore, let us assume that the two files are placed in the `R Script` directory (which is inside the TESS directory). Here is a typical instruction sequence to run the script:

- Launch R and change its working directory to `R Script/`
- Load the files in R and store them into internal variables by typing: `data <- read.table("data.txt"); estimates <- read.table("estimates.txt")`
- Load the script by typing: `source("krigAdmixProportions.R")`
- Run the script by typing: `krigAdmixProportions(data,estimates)` (for the default options)
- The script will automatically draw one map per cluster

The `krigAdmixProportions` function has several options, which we detail here. The full list of arguments with their default values is

```
krigAdmixProportions(data, estimates, twoRows=TRUE, extraCol=0, onePlot=TRUE, drawAxes=FALSE)
```

Table 2 sums up the different arguments of the kriging function.

11 Tutorial

In this section, we provide a short TESS tutorial. In particular, we highlight how to choose K_{\max} for two particular examples.

The first example consist of a simulated data set which has already been studied in (Chen *et al.*, 2007). It consists of a group of individuals structured into five subpopulations with a pairwise F_{ST} equal to 0.04. Each subpopulation contains 100 individuals genotyped at 10 independent microsatellite loci. The genotypes are drawn from subpopulation-specific allele frequencies. The geographic coordinates where simulated from Gaussian distributions, so that the five subpopulations where organized in a star shape on a ring. The regions of the five subpopulations overlap geographically. We ran TESS without admixture for K_{\max} ranging from 2 to 9 for 10,000 sweeps, discarding the first 5,000. The

Parameter	Description	Requirements
data	R data frame (or matrix) containing TESS data. Alternatively, can contain coordinates only.	Must be correctly formatted
estimates	R data frame (or matrix) containing estimated admixture proportions.	Must be correctly formatted
twoRows	Is "data" stored in two row per individual? If false, data contains one row per individual.	Boolean (default: TRUE). Set it to FALSE for haploid individuals
extraCol	Number of extra columns in the file "data" (columns before the coordinates)	Positive integer (default: 0)
onePlot	If TRUE, all maps are drawn in the same window. Otherwise, one plot per cluster	Boolean (default: TRUE)
drawAxes	If TRUE, x and y axis are drawn.	Boolean (default: FALSE)

Table 2: Parameters of the KrigAdmixProportions function explained.

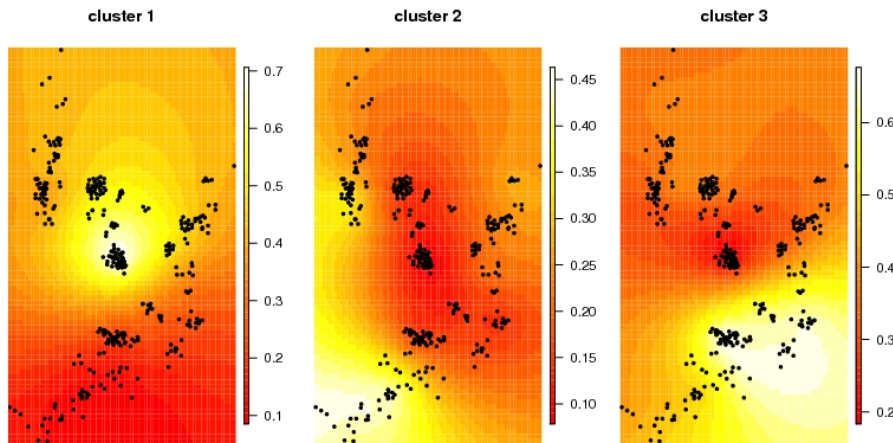


Figure 16: Spatial interpolation of admixture proportions for an example data set analyzed with TESS.

other parameters were set to their default values. For each value of K_{\max} , we computed the DIC, and we plotted its value against K_{\max} . The membership coefficients displayed in figure 17 (left) provide unambiguous evidence of 5 main clusters in the data. The regularization implied by the spatial prior is visible on figure 17(C). Indeed, no clear additional cluster is detected for $K_{\max} > 5$. The DIC curve decreases sharply and then exhibits a plateau at $K_{\max} = 5$ (figure 17(A)). To emphasize the connection with the ΔK method traditionally used to select K_{\max} with STRUCTURE, figure 17 (right) displays the same analysis performed with STRUCTURE. It also correctly concludes that $K_{\max} = 5$.

The second example consists of a real data set. It contains 76 individuals genotyped at 822 loci, from the *Arabidopsis thaliana* species. Keeping the same analysis scheme as

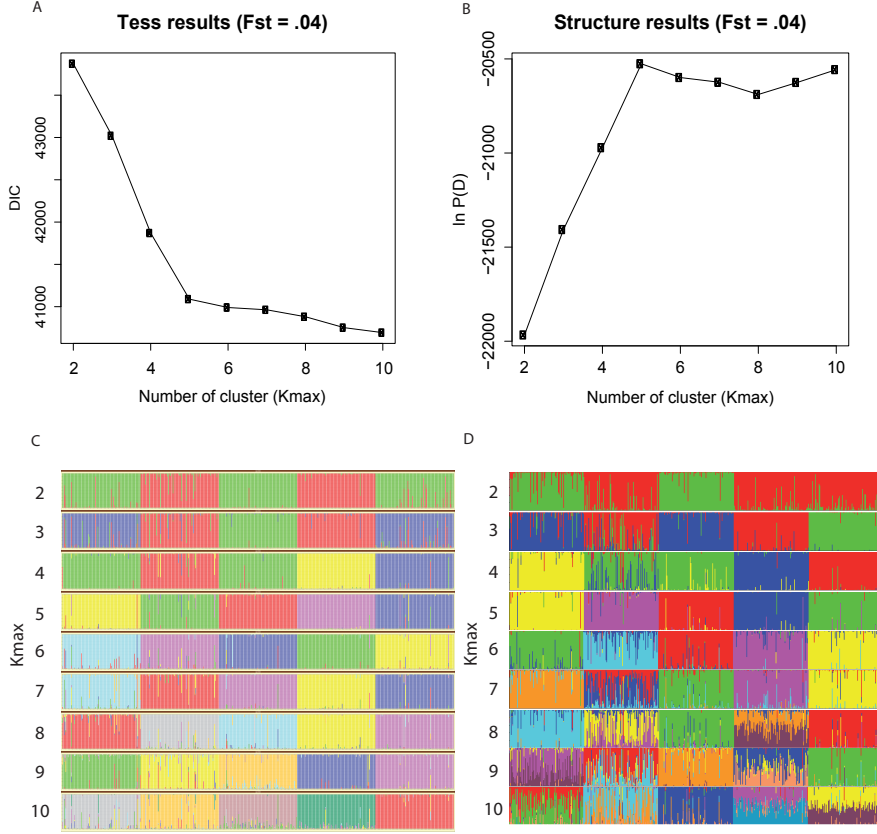


Figure 17: Analysis of an example data set (5-islands, $F_{st} = .04$) A) DIC for 9 TESS runs with K_{max} ranging from 2 to 10. The plateau starting at $K_{max} = 5$ is an indication that the number of clusters is $K = 5$. B) Logarithm of evidence, $\log P(D)$, for 9 STRUCTURE runs with K ranging from 2 to 10, Evanno et al criterion indicates that there are 5 clusters in the data set. C) Posterior estimates of cluster membership for TESS. For $K_{max} \geq 5$, 5 main clusters are visible. D) Posterior estimates of cluster membership for STRUCTURE.

in the first example, we varied K_{max} from 2 to 9. As this is a real data set, for which the "true structure" is not known, we ran the admixture model of TESS 100 independent times for each value of K_{max} . Each run consisted of 50,000 sweeps with a burn-in period of 30,000. For each value of K_{max} , we computed the DIC. We averaged the estimated admixture coefficients over the 10% runs with the lowest values of the DIC using the software CLUMPP. Figure 18 shows that the DIC selects $K_{max} = 4$. Remark that the admixture estimates provide evidence for four almost identical populations for all K_{max} in 4-9. The use of CLUMPP is greatly facilitated by using the GUI shell of TESS, which allows to export directly runs to the CLUMPP format. This functionality is described in section 6.7.

12 FAQs

1. How long should I run TESS on my data set?

Answer: The default values suggest using very short runs, but we generally rec-

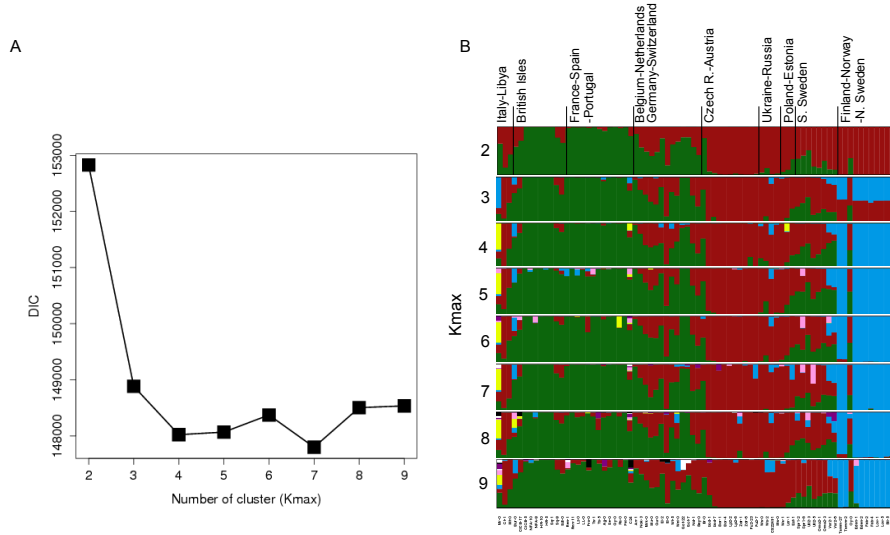


Figure 18: Analysis of the *Arabidopsis thaliana* data set (76 individuals genotyped at 822 loci). A) DIC as a function of K_{\max} . Each dot corresponds to the average DIC value computed over the 10 lowest DIC values. B) Posterior estimates of individual admixture coefficients for distinct values of K_{\max} .

ommend to use longer runs (eg, burnin = 10,000 sweeps and runing period = 50,000 sweeps). Preliminary (short) runs may be useful to calibrate length. The no-admixture program usually converges fast.

2. What is the interaction parameter and which values should I use?

Answer: In the model without admixture, the interaction parameter ψ is requested to implement the TESS prior distribution on cluster configurations. For large ψ (eg, 10) individuals are likely to be put into a single cluster, whatever the genetic data, which may then have weak influence on the posterior distribution. For $\psi = 0$, spatial data have no influence on the posterior. Good values for ψ are related to phase coexistence in the Potts Model on a random graph. We suggest using ψ in the range (0.5, 1.0) but larger values may sometimes lead to better results for very difficult data sets. Try 3 different values and be patient. In the admixture model, ψ is the intensity of the effect of spatial autocorrelation, that is large ψ would favor patchy configurations. The MCMC algorithm is able to update the spatial interaction parameter in the admixture model.

3. How do I determine the number of clusters?

Answer: For the model without admixture, start with $K_{\max} = 2$, then increase the number of cluster until the barplot stabilizes. The DIC should also stabilize – or slowly vary – close to the correct value. The general strategy is to gradually increment K_{\max} and then plot the DIC value against K_{\max} . In TESS, the actual number of cluster, K , may be less than K_{\max} . Analyze the results on the basis of the DIC and their barplots. Be aware that this may not be the perfect solution.

4. How many runs should I perform?

Answer: For moderate size data sets, we recommend running hundreds of runs, and then interpret the results from the runs having small DICs.

5. How can I perform Bayesian estimation and summarize the results?

Answer: Keep the 20% lowest DIC runs, and use CLUMPP to perform averages over these runs. The results may differ from the “Hard Clustering” colored tessellation of TESS. These averages may generally be closer to a Bayesian estimate than estimates from a single run. You may use DISTRUCT to display the estimated membership coefficients. You may also use the kriging functions provided by the program R to interpolate them spatially, and overlay geographical maps to the results (R package ‘maps’ and ‘spatial’ or ‘fields’ recommended). R scripts are available inside the TESS directory. They may also produce nice outputs.

6. Should I modify the neighborhood?

Answer: It may be a good idea to try including realistic features and to remove the very long edges in the TESS network. Weighting by geographic distance or using dummy points may be efficient alternative ways to do this.

7. Does removing a link impose a geographical barrier?

Answer: No. Removing a link between two individuals means that these individuals are not related a priori. Genetic evidence may be strong enough so that these two individuals may be assigned to a same cluster. If all the links were removed, the program would behave like STRUCTURE.

8. Does TESS include isolation by distance?

Answer: TESS includes decay of correlation of membership coefficients with distance within clusters, and may indicate the presence of genetic discontinuities between clusters. They are two possible distinct cases. (1) We are in presence of small but real spatial genetic discontinuities, then TESS may detect clusters with fewer loci or smaller differentiation than necessary without spatial information; (2) We are in the presence of clines and sampling is regular, then TESS may detect less clusters than STRUCTURE (even with more loci). When the number of loci is very large, both programs will tend to produce the same output.

9. Does TESS correct for uneven sampling?

Answer: To a moderate extent, yes. But if sampling is highly irregular, no hope.

References

- Chen, C., Durand, E., Forbes, F., and François, O. (2007). Bayesian Clustering Algorithms Ascertaining Spatial Population Structure: A New Computer Program and a Comparison Study. *Molecular Ecology Notes*, **7**, 747–756.
- Durand, E., Jay, F., Gaggiotti, O. E., and François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, **26**(9), 1963–1973.
- Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**(8), 2611–2620.
- Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**(4), 574.
- François, O., Ancelet, S., and Guillot, G. (2006). Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population Genetics. *Genetics*, **174**, 805–816.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.
- Jakobsson, M. and Rosenberg, N. A. (2007). CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics*, **23**, 1801–1806.
- Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959.
- Ripley, B. D. (1981). *Spatial statistics*. Wiley New York.
- Rosenberg, N. A. (2004). DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.
- Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.