POPS short tutorial

Flora Jay

September 8, 2011

flora.jay@imag.fr

Contents

1 Introduction

This tutorial aims to help users to run POPS graphical user interface and POPS command-line engines on a simulated data set. It highlights several POPS features but is not exhaustive. Several features are shared with the software TESS and are explained in the TESS manual [1].
http://membres-timc.imag.fr/Flora.Jay/ or
http://membres-timc.imag.fr/Olivier.Francois/
Statistical models implemented in POPS are described in [4].

2 Simulated data

The data file "example.txt" is provided along with the software, in the directory "POPS/Example/". It contains environmental and genetic data. The genetic data were simulated by resampling markers from the genetic data of a plant species from the Intrabiodiv database [2] and adding a small perturbation to those markers. The data file contains 269 rows (1 row for header and 1 row for each haploid individual) and the following 90 columns:

1) individual number,

2) temperature at sampled site,

3) longitude,

4) latitude,

5+) 86 genetic markers.

3 Create a new project and load input file

To create a new project, access the menu **"File⇒ New Project..."** or click on the corresponding icon on the tool bar. The GUI shell will show the "New Project" dialog box asking the user to key in the required project information (Figure 1). The user should **name his/her project and choose the project path and data**. The user also needs to **input the information and format of the project data**. These include number of individuals, ploidy, number of loci, number of qualitative covariates, number of quantitative covariates, presence of spatial coordinates (spatial coordinates are mandatory to run the admixture version), 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 (Figure 2). From the left picture in Figure 2, we can see there is 1 extra row and 1 extra column, the second column stores temperature, which is a quantitative covariates, then two columns store the spatial coordinates of individuals, which are required for running the admixture models of POPS. From the right picture in Figure 2, we know there are 269 rows, therefore the number of individuals should be 268 (since they are haploid and the first row is a header) and there are 90

columns, hence the number of loci should be 86. After viewing the data, users can close the "Project Data" window and continue to input information.

When this task is complete, click the "OK" button to confirm the creation of the new project.

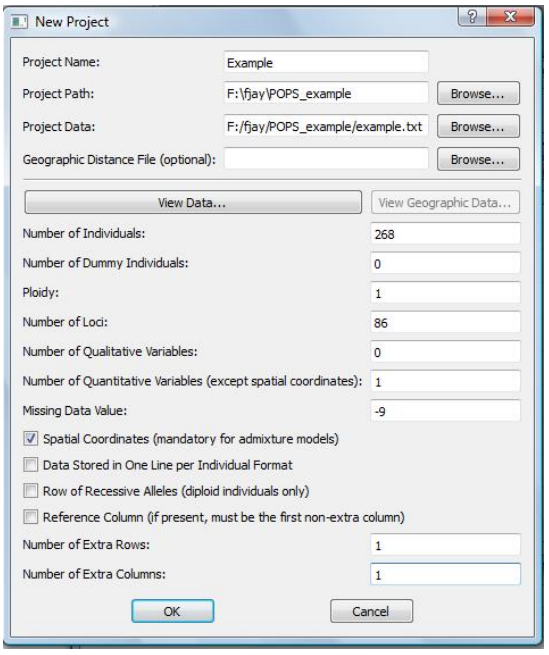

Figure 1: Input information for a new project.

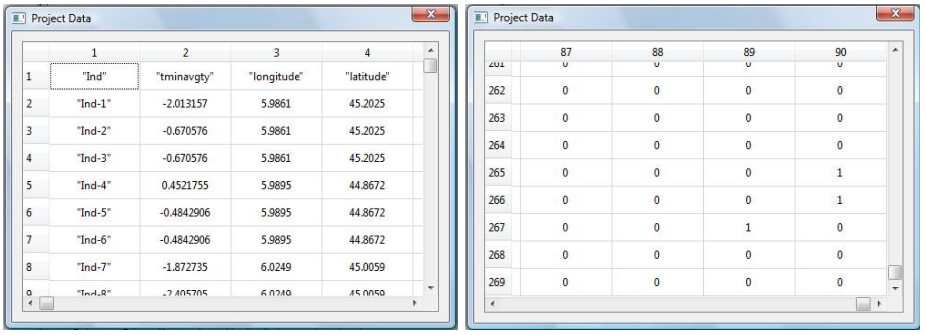

Figure 2: View input file to check data information (number of individuals, number of loci, extra rows ...).

4 Check the new project

When the project is created, the GUI shell automatically loads it and shows the input data to the user. The user can **check that the file information is correct by looking at the color code** : extra information should be highlighted in gray, qualitative covariates in yellow, quantitative covariates in green, spatial coordinates in blue and genetic data in white (Figure 3). If one column or row is wrongly highlighted, a mistake has been

made when giving data information and the user must create a new project with correct settings. In addition, project and data information appears in the left tree widget, entitled Project (left panel in Figure 3).

If spatial coordinates are available, the user can display diagrams build from input data by double-clicking on "Voronoi" and "Neighborhood" items in the section "Project Information" of the tree widget. Diagrams are displayed in the right panel (Figure 3).

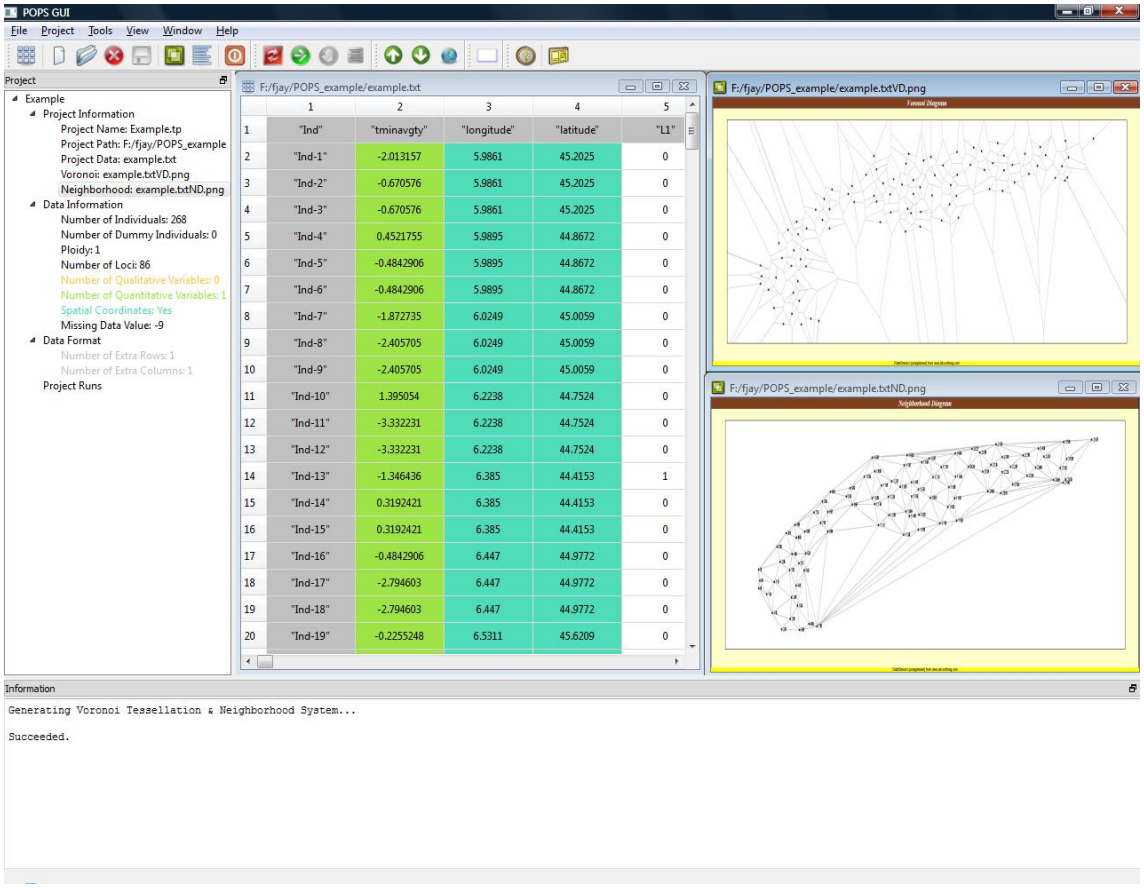

Figure 3: To obtain this screen view after loading a project, double-click on the items "Voronoi" and "Neighborhood" in the tree widget (left panel). Then, click on "Tile" in the dropdown menu "Window". Last, right-click on the images and choose the "Fit to Window" option.

5 Set parameters and run POPS

The user can start to run the project by accessing the menu **"Project⇒ Run..."** or click on the corresponding icon on the tool bar. Figure 4 displays settings to launch POPS using models with admixture. These settings will launch 2 runs for each value of the maximal number of cluster K ranging between 2 and 3. The MCMC runtime is set to 1,200 sweeps and the burn-in period to 200 sweeps. The degree of the spatial trend surface is set to 1. A click on "Advanced Options" button opens a new window where extra choices can be made. Here we keep the default advanced options.

Click the "OK" button to start the analysis. Progression of runs can be followed on the "Information" window (at the bottom of the main window).

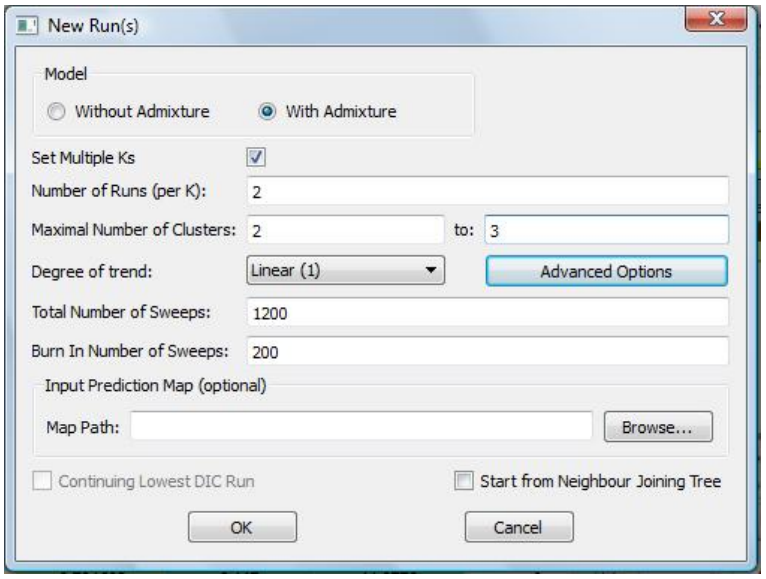

Figure 4: Input settings for new run(s).

6 View results

POPS creates a sub-directory in the project directory for each run and saves the results of the run in that sub-directory. After a run is completed, the user can read input parameters and output scores (averaged log-likelihood, DIC, correlation between estimated and predicted coefficients) under the run name in the section "Project Runs" of the tree widget "Project". **He can also load the textual results file and display graphical results by double-clicking items in the tree widget** (Figure 5). From **"Project $\Rightarrow$ Summarize Project Runs"** (or from the corresponding icon on the tool bar), the user can open the "Summarize Runs" window (Figure 6). In this window, a summary table allows users to quickly access the summaries of all runs (K, trend degree, model used, DIC, correlation score) and to order runs according to these values. The summary table 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 the graphic in the main window. Users can use the "Plot DIC" button to display a plot of DIC values as a function of K for all runs in the summary table.

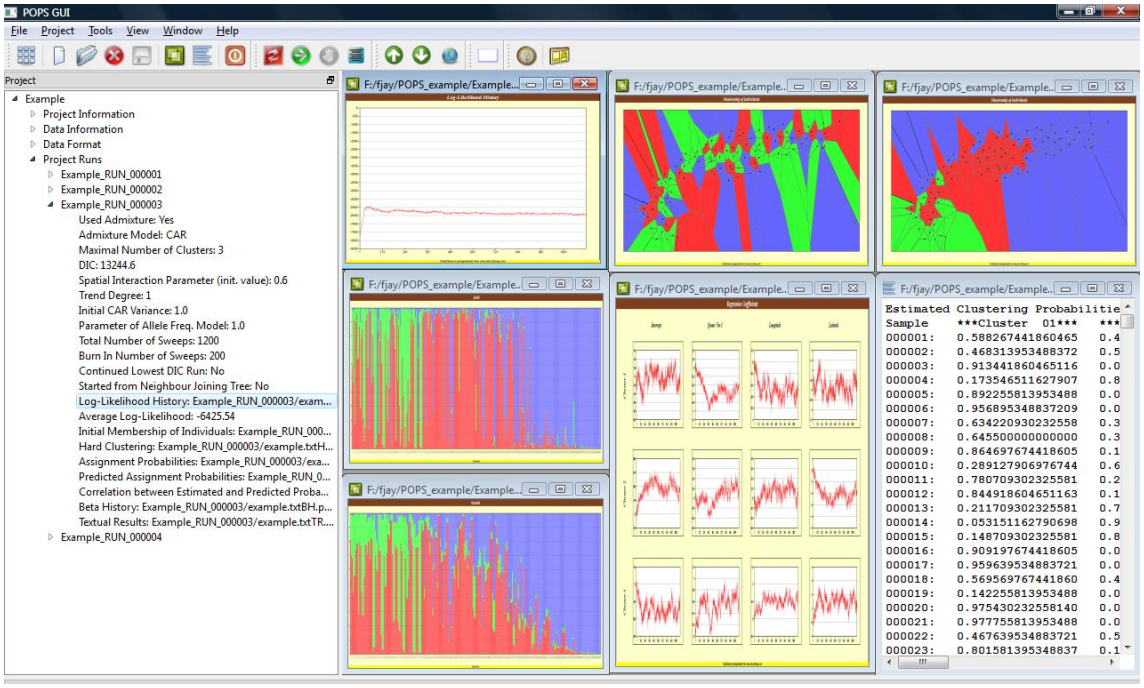

Figure 5: All graphical results of a run can be displayed in the main window using the "Project" tree widget. Here we clicked on the items labeled "Log-likelihood History", "Initial Membership of Individuals", "Hard Clustering", "Assignment Probabilities", "Predicted Assignment Probabilities", "Beta History", and "Textual Results", under the third run name "Example_RUN_000003".

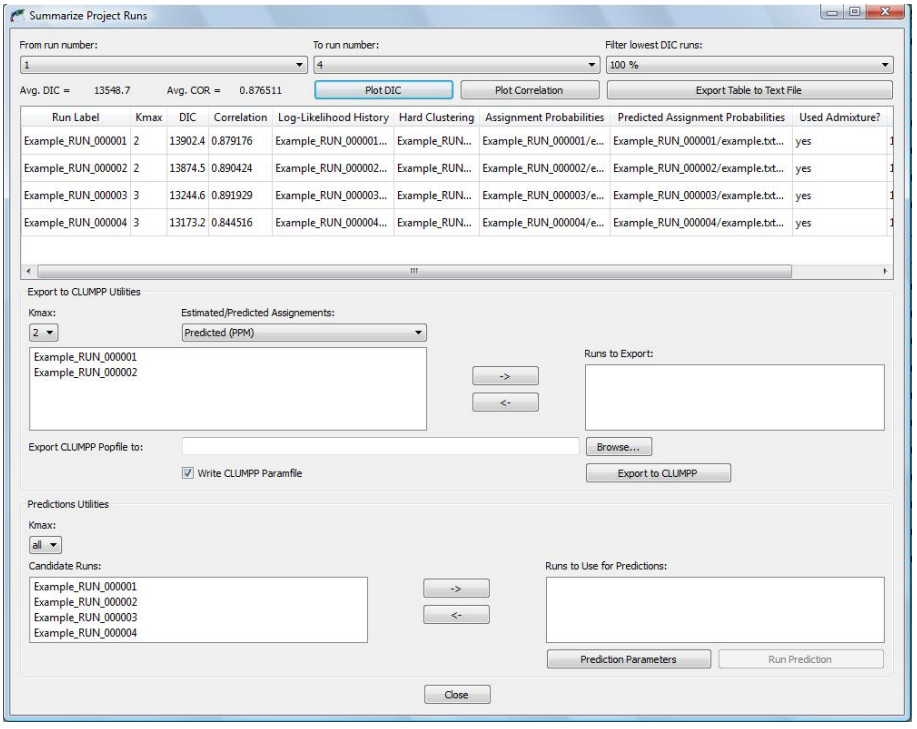

Figure 6: Summarize window can be opened from "Project ⇒ Summarize Project Runs". Information of runs previously launched is display in a table. Runs to display can be chosen using the lower run number, upper run number and DIC filter combo boxes. We ordered the runs according to their DIC, by clicking on the DIC column header.

7 Prediction utilities

Assume now that the quantitative covariate (temperature) contained in `"example.txt"` is increased by 3 units (3°C) for each individual. The new set of covariates is stored in `"futureEnv.txt"`. This file contains 4 columns (sample id, increased temperature, longitude, latitude). To forecast population genetic structure for these new covariates, the user first need to choose one or several runs from which he wants to use the estimated parameters. This can be done in the **"Prediction Utilities" section of the "Summarize" window** (see Figure 7). Then the file containing the new set of environmental covariates should be loaded. For that, the user has to click on the **"Prediction Parameters" button**, fill in the window as detailed on Figure 8 and click on "Ok". Then, **he/she can launch the prediction computation by clicking on the "Run Predicion" button in the "Summarize Project Runs" Window**. Predictions are computed for each selected run, and are accessible from the "Project" tree widget (Figure 9).

Figure 7: Section "Prediction Utilities" of the Summarize window. Here we chose to use parameters estimated in the runs 000003 and 000004 for the predictions. For that, we selected them in the "Candidate Runs" list and moved them to the "Runs to Use for Predictions" list by clicking the right-arrow button.

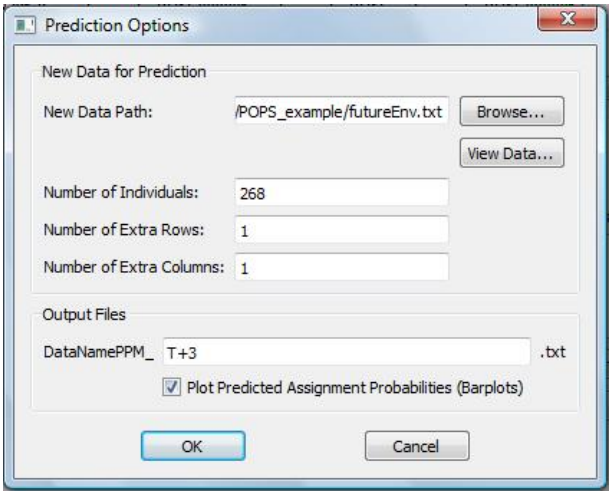

Figure 8: Information for a new set of covariates. Input a file name and corresponding format details. Choose a name for the outputs (textual files and optionally graphical files). Here, the outputs will have the suffix "PPM_T+3.txt" and "PPM_T+3.png". If the box is checked, graphical outputs will be generated (barplots of the predicted assignments probabilities).

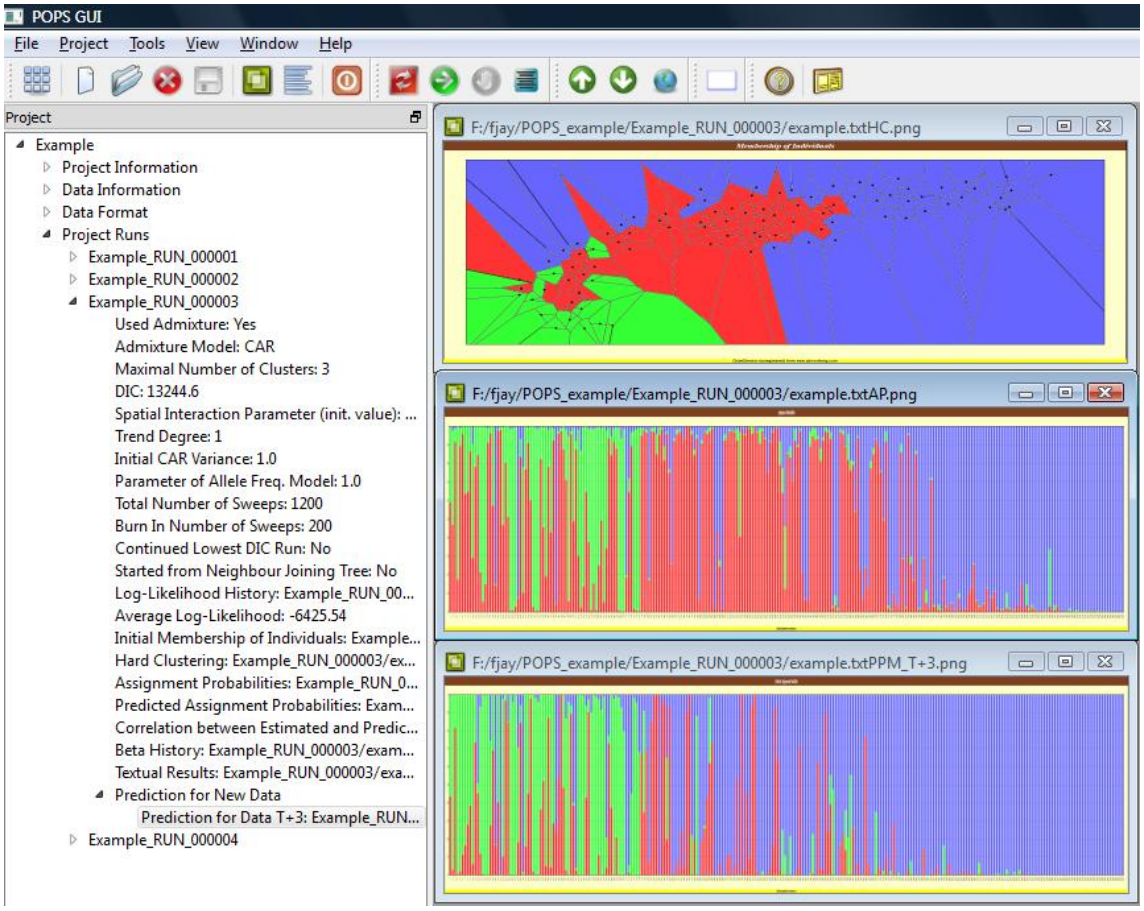

Figure 9: Run predictions can be displayed by clicking on the label "Prediction for Data T+3" in the "Project" tree widget ("Project Runs" section → run name → "Prediction for New Data"). Here, 3 graphical outputs are displayed for the run "Example_RUN_000003": the hard clustering map (top), the assignments probabilities (middle), the predicted probabilities under a 3°C increase (bottom).

8 Post-processing

8.1 Average multiple runs

To average multiple runs with the same number of clusters, Jakobsson and Rosenberg developed the software `CLUMPP` [3]. Assignment probabilities or run predictions can be exported to the `CLUMPP` format using the "Clumpp Utilities" section in the "Summarize Window". More details are provided in `TESS` manual [1].

8.2 Using `R` to plot maps or calculate correlation between barplots

We provide `R` functions to display `POPS` outputs on maps using kriging techniques and to calculate correlation between barplots.

To load `POPS` functions for post-processing using `R` software, the user has to launch `R`, and execute

```
# Install required packages
install.packages("fields")
install.packages("RColorBrewer")
# Change the current work directory to the one containg POPSutilities
setwd("pathonyourcomputer/POPS/R/")
source("POPSutilities.r")
```

Then, he/she can execute the following script that we used to

- calculate the correlation between current population genetic structure and structure forecasted for a 3°C temperature increase
- display estimated coefficients and predicted coefficients spatially (see Figure 10)

The script is also available in the file "POPS/R/scriptExample.r"

```
# You should copy the files you want to analyse in the work directory
# Or replace the lines below by correct file names
# e.g. "MyProject/Example_RUN_000003/example.txtTR.txt"
estimationFile="example.txtTR.txt"
predictionFile="example.txtPPM_T+3.txt"

# Calulate rate of turnover between current population genetic structure
# and structure forecasted for temperature increase
correlationFromPops(file1=estimationFile,file2=predictionFile,nind=268)

# Load spatial coordinates
data=read.table("example.txt",header=T)
coord=data[,c("longitude","latitude")]

# Create a grid from these coordinates
# with 100 pixels for longitude, 100 for latitude
grid=createGrid(min(coord[,"longitude"]),max(coord[,"longitude"]),
         min(coord[,"latitude"]),max(coord[,"latitude"]),100,100)
constraints=NULL

# OR  Create a grid from an ascii raster file
# ascii raster files can for example be downloaded from NOAA website
# http://www.ngdc.noaa.gov/mgg/geodas/geodas.html
# or Global Map website
# http://www.globalmap.org/english/index.html
asciiFile="alps.asc"
grid=createGridFromAsciiRaster(asciiFile)
# To display only altitudes above 1000:
constraints=getConstraintsFromAsciiRaster(asciiFile,cell_value_min=1000)

# Plot estimated coefficients and predicted coefficients maps
par(mfrow=c(2,1))
mapsFromPops(file=estimationFile,nind=268,coord=coord,grid=grid,
    constraints=constraints,method="max",main="Present")
```

```
mapsFromPops(file=predictionFile,nind=268,coord=coord,grid=grid,
    constraints=constraints,method="max",main="Temperature increase: +3")

#Display legend
x11()
displayLegend(K=3)
```

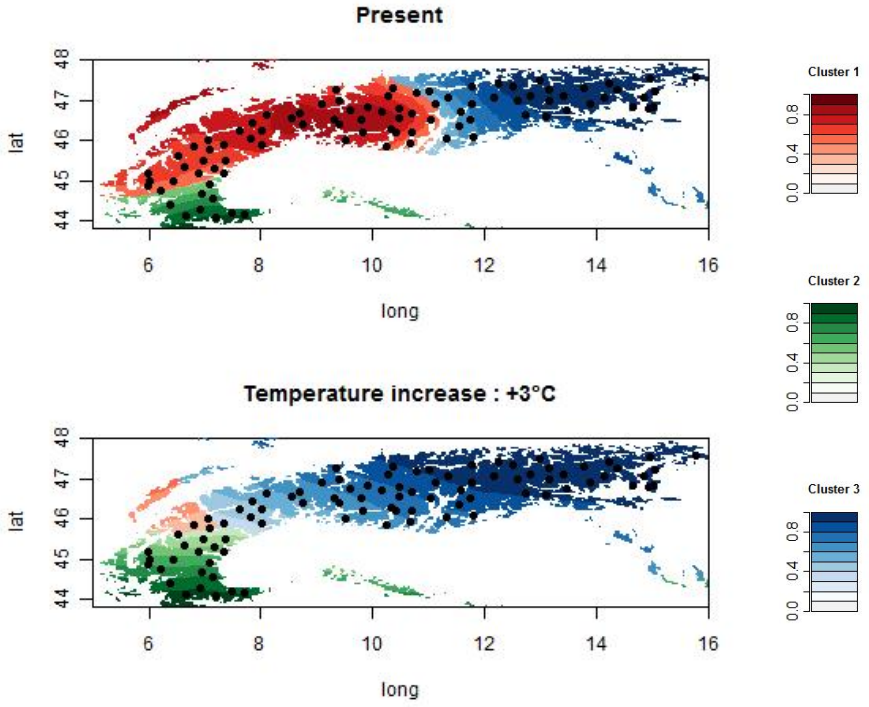

Figure 10: Maps of admixture coefficients and their prediction for a 3°C increase in temperature are displayed spatially.

9 POPS command-line options

POPS is based on two command-line engines: one for the models without admixture, and the other for the models with admixture. We describe the main commands for both programs. When there are no options given to POPS, it will show its typical usage and exit. Here are the main options, that can be specified in any order.

9.1 Run POPS via the command-line engine

Required Parameters

-F	File Name of Input Data File
-N	Number of Individuals
-A	Ploidy (1 = Haploid, 2 = Diploid, ...)
-L	Number of Loci
-K	Maximal Number of Clusters
-XL	Number of Qualitative Variables
-X	Number of Quantitative Covariates (other than spatial coordinates)
-T	Degree of Trend (-1: No spatial coordinates in datafile, 0: Spatial coordinates present but not used, 1,2,3: Degree 1, 2, or 3)
-D	Parameter of Dirichlet Allele Frequency Model
-S	Total Number of Sweeps of MCMC
-B	Burn In Number of Sweeps of MCMC
-P	Spatial Interaction Parameter (for admixture models only)

Optional Parameters

-r	Number of Extra Rows in Data File
-c	Number of Extra Columns in Data File
-i	Folder Name of Input Data File (default: Current Folder)
-o	Folder Name of Output Result Files (default: Current Folder)
-orun	Suffix to Append to Output Result Files Names e.g. a Run Number (-orun1 or -orun0001) or a Specific Run Name (-orunAdm002)
-sp	Special Data Format: 1 individual = 1 row (-spy: yes, -spn: no, default)
...	Run pops|more and popsAdm|more or see POPS manual for extra options

Note that if a file name contains spaces, it should be input as "file name".
The command pops (or pops.exe on Windows system) runs models without admixture, whereas popsAdm (or popsAdm.exe) runs admixture models.

We give command-lines to run POPS analyses of the data file "example.txt" that is provided along with the software, in the directory "POPS/Example/".

The data contain 268 genotyped haploid individuals (-N268 -A1), and 1 extra row -r1. The columns consist in (in order): 1 an extra column (identifier for samples, -c1), 1 quantitative covariate (temperature -X1), 2 columns for longitude and latitude, 86 columns for genetic data (-L86). Assuming there are at most 3 clusters (-K3), we set the parameter of the Dirichlet allele frequency model to 1.0 (-D1.0). To run the MCMC algorithm for a total of 1,000 sweeps (-S1000) with the first 200 sweeps discarded as burn-in period (-B200), and setting the degree of the spatial trend surface to 1 (-T1), we use the following command

```
pops -Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000
-B200 -r1 -c1 -iExample -oExample -orun001
```

Output results will be stored in the directory "Example" (`-oExample`) and the string `"_RUN001_"` will be appended to the output names (`-orun001`).

To run `POPS` with admixture using the same data and the same parameters, we additionnally specify the spatial interaction parameter (e.g. `-P0.6`)

popsAdm `-Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000 -B200` **`-P0.6`** `-r1 -c1 -iExample -oExample` **`-orunAdm001`**

The string `"_RUNAdm001_"` will be appended to the output names (`-orunAdm001`).

9.2 Use `POPS` prediction utilities

Prediction utilities are available from the command-line `pops -predict`, whether you are using admixture models or not. Running `pops -predict` will show typical usage and exit.

Here, we give command lines to forecast genetic structure under a 3°C increase (see section 7 for more details). New values of covariates for 268 individuals are stored in "Example/futureEnv.txt" (`-newFExample/futureEnv.txt -newN268`) . This file contains also 1 extra row (`-newR1`) and 1 extra column (`newC1`).

We first forecast the structure using regression coefficients estimated from the run "example.txt_RUN001". This run uses a model without admixture (`-admno`) and with a linear spatial trend (`-T1`). The regression coefficients are stored in the file with suffix "BetaHat.txt" (`-BETAExample/example.txt_RUN001_betaHat.txt`).

```
pops -predict -BETAExample/example.txt_RUN001_betaHat.txt
-newFExample/futureEnv.txt -newN268 -newR1 -newC1 -XL0 -X1 -T1
-K3 -admno -oExample/example.txt_RUN001_PPM_T+3
```

Values of predicted membership coefficients will be stored in "Example/example.txt_RUN001_PPM_T+3.txt" and displayed on a barplot in "Example/example.txt_RUN001_PPM_T+3.png".

Now, we forecast the structure using the run with admixture that we launched before ("Example/example.txt_RUNAdm001"). The new command-line is

```
pops -predict -BETAExample/example.txt_RUNAdm001_betaHat.txt
-newFExample/futureEnv.txt -newN268 -newR1 -newC1 -XL0 -X1 -T1
-K3 -admyes -oExample/example.txt_RUNAdm001_PPM_T+3
```

Maps. If you want to display predicted membership/admixture coefficients on a map, you should report to the section 8.

References

[1] E. Durand, C. Chen, and O. François. Tess version 2.3 - reference manual. `http://membres-timc.imag.fr/Olivier.Francois/manual.pdf`, 2009.

[2] F. Gugerli, T. Englisch, H. Niklfeld, A. Tribsch, Z. Mirek, M. Ronikier, N.E. Zimmermann, R. Holderegger, and P. Taberlet. Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation - a project synopsis. *Perspectives in Plant Ecology, Evolution and Systematics*, 10(4):259 – 281, 2008.

[3] M. Jakobsson and N.A. Rosenberg. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806, 2007.

[4] F. Jay, E.Y. Durand, O. François, and M.G.B. Blum. POPS: A software for prediction of population genetic structure using latent regression models. *Manuscript*, 2011.