Operating principles of the FactoQGIS tool

Florent Demoraes, December 2018
UMR ESO CNRS 6590, Université Rennes 2, France
https://perso.univ-rennes2.fr/florent.demoraes

License, languages, software versions and script content

License, languages, and software versions

FactoQGIS was developed under GNU General Public License v2.0 and works with QGIS 2.18 and R 3.5.1 version or newer. Developments are underway to be able to keep on using R scripts in more recent releases of QGIS (release 3.0 and newer) thanks to the Processing R Provider add-on [1]. To run R scripts in QGIS, R must of course be installed on the computer. FactoQGIS is mainly based on the FactoMineR package developed by François Husson et al. (2009). It also makes secondary use of factoextra, stringr, openxlsx, R2HTML and corrplot packages. These packages must be previously installed in the R software (or via R Studio). To execute R scripts, QGIS uses the Processing module (Graser & Olaya, 2015), which is itself based on the Python subprocess module. The FactoQGIS tool consists of two files which are available on GitHub[2]. The first file "Typological_Analysis_PCA_PCA_and_HAC.rsx" contains the script. The second file "Typological_Analysis_PCA_PCA_and_HAC.rsx.help" contains the help. These files must be stored in the folder: C:\Users\...\...\ qgis2\processing\rscripts

A script in 7 steps

The script header contains the python parameters associated with the arguments to be filled by the user in the dialog box. Below the header, the R script begins, which is itself broken down into 7 parts as shown below.

1 - Loads the packages necessary to execute the script.
2 - Retrieves in R objects the parameters entered by the user in the dialog box and converts them into argument values for R functions.
3 - Imports the dataset (the attribute table of the layer) and creates a dataset corresponding only to the active quantitative variables.

[1] This adds-on is developed by North Road: https://github.com/north-road/qgis-processing-r (accessed on May 27, 2019). However, this add-on till now does not provide the multiple field selection option which is required for PCA. Furthermore, contextual help is no longer available.
[2] https://github.com/ESO-Rennes/FactoQGIS (accessed on May 27, 2019).

4 – Launches the PCA, calculates the results (tables, plots) in different formats.

5 – Launches the HAC, calculates the results (tables, plots) in different formats.

6 – Appends the results in an html file that pops up automatically at the end of the process.

7 – Creates a layer which contains in its attributes a column with the cluster the spatial units belong to, resulting from the typology.

The FactoQGIS dialog box

FactoQGIS is accessible from the QGIS toolbox (Figure 1).

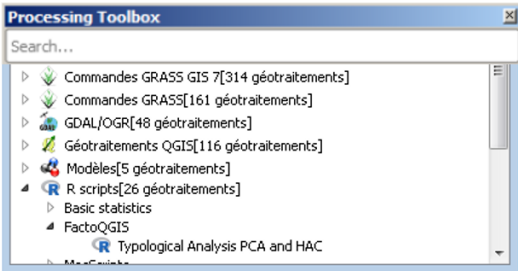

Figure 1. FactoQGIS in the QGIS toolbox

In the dialog box (Figure 2), the user must enter 14 parameters. The first 11 ones are the input parameters and are mandatory. Some have default values. The last three ones are the output parameters and are optional. If the user does not specify any name for the output files, the latter will be saved in a temporary folder.

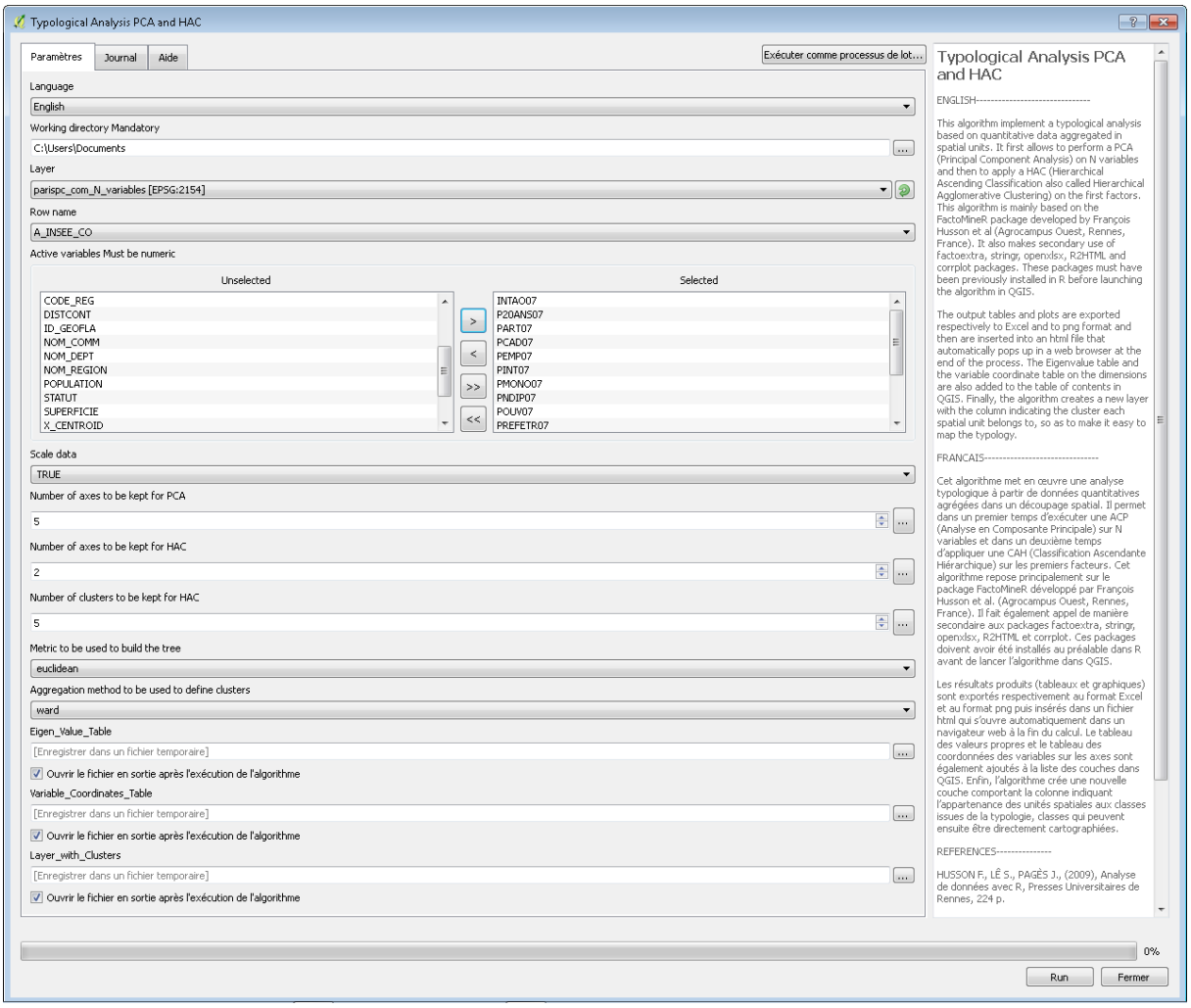

Figure 2. The FactoQGIS tool dialog box

Input parameters

1 - Language

French or English. This parameter will define the language to be applied to the captions of the tables and plots in the output html file.

2 - Working directory

This field is mandatory. The path to the working directory must be short and must not contain any special characters or spaces. All the output tables and plots will be stored in it.

3 - Layer

Layer on which to apply the PCA and the HAC. The attribute table of this layer must contain quantitative variables. This layer must be loaded in QGIS.

4 - Row name

Field that contains the identifier of the spatial units. This ID will then appear on the factor maps and is also required for merging data in the end of the algorithm.

5 - Active variables

Actives variables on which the PCA will be performed. Must be numeric. The active variables which appear in the figure 2 are detailed in table 2.

6 - Scale data

Option to scale and center the data. Should be applied in the vast majority of the cases, especially when the unit variance is very different between the variables.

7 - Number of axes to be kept for PCA

Number of axes to be kept for PCA. 5 is the default value. Generally, we keep the N first factors which explain at least 95% of the inertia. It is recommended to first let the default value and to check the Eigen values table and the scree plot. If needed you can change the default value and perform a second time the PCA.

8 - Number of axes to be kept for HAC

Number of axes to be kept for HAC. 2 is the default value. Generally, we keep the N first factors which explain at least 80% of the inertia so as to get a more stable clustering. It is recommended to first let the default value and to check the Eigen values table and the scree plot. If needed you can change the default value and perform a second time the HAC.

9 - Number of clusters to be kept for HAC

Number of clusters to be kept for HAC. 5 is the default value. It is recommended to first let the default value and to check the hierarchical tree. If needed you can change the default value and perform a second time the HAC.

10 - Metric to be used to build the tree

Metric to be used for calculating dissimilarities between individuals. The currently available options are "euclidean" and "manhattan". Euclidean distances are root sum-of-squares of differences, and manhattan distances are the sum of absolute differences. Default value is "euclidean".

11 - Aggregation method to be used to define clusters

Clustering method. The four methods implemented are "average" (unweighted pair-group arithmetic average method), "single" (single linkage), "complete" (complete linkage), and "ward" (Ward's method). Ward's method is the most commonly used and is the default value.

Outputs

12 – Eigen Value Table

Eigen values table which gives for each variable its part to the global inertia. This table is automatically added to the table of contents in QGIS and is also exported to an Excel table sheet.

13 – Variable Coordinates Table

Table which gives the coordinates of each variable on the axes. This table is automatically added to the table of content in QGIS and is also exported to an Excel table sheet.

14 – Layer with Clusters

Output vector layer with the column indicating the cluster each spatial unit belongs to. This layer is automatically added to the table of contents in QGIS so as to make it easy to map the typology.

Outputs format and output files summary

Table 1 shows for each of the results produced by the FactoQGIS tool its format, whether a file is created in the workspace, whether it is added to the html file and whether it is added to the table of contents in QGIS. Most of the results (tables and plots) are inserted in an html file that automatically pops up in a web browser at the end of the process.

Table 1. Summary of the outputs created by FactoQGIS

Output	Format	Output file(s) stored in the working directory	Appended to the html file	Added to the table of contents in QGIS
Table of the Eigen values	xlsx, csv	x	x	x
Scree plot (Gain of inertia)	png	x	x	
First factorial map showing the variables (axes 1 and 2)	png, pdf	x	x	
Variable Coordinates Table	xlsx, csv	x		x
Quality of the representation of the variables (Cos2)	png	x	x	
First factorial map showing the coordinates of the individuals (dimensions 1 and 2)	png, pdf	x	x	
Hierarchical cluster tree	png	x	x	
Hierarchical cluster tree on the first factor map	png	x	x	
Bar plots showing the variables which best describe the clusters*	png	x	x	
Tables giving the description of the clusters by the variables			x	
Layer with Clusters	shp	x (only if a name was given by the user)		x

** Only the variables with a v-test $\geq |1.96|$ are plotted.*