# FACTOQGIS: A GUI TOOL BASED ON AN R SCRIPT TO PERFORM GEOMETRIC DATA ANALYSIS IN A FREE AND OPEN SOURCE GIS

**Florent Demoraes**, Univ Rennes, CNRS, ESO - UMR 6590, F-35000 Rennes, France
florent.demoraes@univ-rennes2.fr

**Marc Souris**, UMR Unité des Virus Emergents (UVE : Aix-Marseille Univ – IRD 190 – Inserm 1207 – IHU Méditerranée Infection), Marseille, France
marc.souris@ird.fr

*FactoQGIS is an algorithm that allows the implementation of a geometric analysis of multidimensional data in QGIS. More specifically, this tool was designed to easily perform a typological analysis on quantitative data aggregated in spatial units. This method is broadly used in geography but it was up to now executed out of GIS environments, in specific statistical software. FactoQGIS is a tool which precisely fills this gap among GIS functionalities. It first performs a PCA (Principal Component Analysis) and second a HAC (Hierarchical Ascending Classification) on the first factors. FactoQGIS is based on an R script that mainly uses the FactoMineR package developed by François Husson et al. (Agrocampus Ouest, Rennes, France). The results (tables and plots) are exported respectively in Excel and png format and then inserted into an html file that automatically pops up in a web browser at the end of the process. The algorithm also creates a new layer with a column indicating the cluster each spatial unit belongs to, so as to make it easy to map the typology. FactoQGIS is accessible from a graphical user interface directly in the QGIS environment. It will be of particular interest to geographers and to any users who wish to simply build and map a multidimensional typology without knowing the R language. To illustrate how FactoQGIS works, we performed as an example, a typological analysis on socio-demographic data that are aggregated by "arrondissements" and "communes" in Paris.*

***Keywords: Geometric Data Analysis; Typological Analysis; Aggregated data in spatial units; R script; Free and Open Source GIS***

## INTRODUCTION

Today there are increasingly rich and varied data sources, a large part of which are aggregated in geographical divisions (administrative units, census tracts, watersheds, etc.). This mass of information requires specific analytical methods to generate knowledge to support public policies, to define environmental management orientations, to steer the development of projects and also to guide geomarketing strategies of companies. The need of analytical methods therefore concerns a wide range of public stakeholders (local authorities, state administrations), semi-public organizations (agricultural offices, urban planning agencies) and private actors (design office, data visualization and communication consulting companies working for polling institutes, newspaper and magazine publishing houses). Among analytical methods, the definition of synthetic profiles or major types among spatial units is often sought to highlight, at a given time, daily mobility patterns (Demoraes et al., 2013), rural structures (Walsh, 2000), urban districts composition (Metzger, 2001), electoral behavior (Rivière, 2012), or to differentiate social-ecological

units (Hanspach et al., 2016), etc. It is also useful to track the trajectory of spatial units over time, so as to monitor the urban transformations (Piron, et al., 2004), or the dynamics of household flows (Robson, et. al., 2009). In this perspective, typological analysis, which is part of the field of Geometric Data Analysis (Benzécri J.-P., 1973; Le Roux B. & Rouanet H., 2004; Greenacre M. & Blasius J., 2006), is of primary interest. First it allows trends to be identified in data sets. Second it allows clusters to be created that bring together similar spatial units. More precisely, the typological analysis combines a Factor Analysis (such as Principal Component Analysis, Multiple Correspondence Analysis, Factor Analysis of Mixed Data) and a Hierarchical Ascending Classification (also called Hierarchical Agglomerative Clustering) based on the first factorial dimensions (see Lebart L. et al., 2006; Husson et al., 2009).

Another observation concerns more specifically the software offer, which can be schematically divided into two main groups:

- software solutions dedicated to the analysis of statistical data (R, SPAD, SPSS, Stata, etc.) that allow the implementation of geometric data analysis techniques,
- GIS software (QGIS, OpenJump, gvSIG, SavGIS, ArcGIS, Mapinfo, etc.) that allow to handle data with a spatial component, such as aggregated data in spatial units, and to produce cartographic representations, including typology maps.

These two major toolsets are generally not very interoperable and require format conversions and a whole series of imports and exports to move from one to the other. However, this dichotomy needs to be qualified. Software companies or developer communities have developed relatively integrated software solutions. As for statistical data analysis software, the R software offers the possibility of implementing a workflow integrating spatial data management (rgdal/sp/rgeos and sf packages), typological analysis (FactoMineR, ade4 and cluster packages), and mapping tools (ggplot2, rCarto, maptools, ggmap and cartography packages). However, this workflow requires a good knowledge of the R language.

As for GIS software, among the proprietary solutions, we can mention the implementation of the PCA and multivariate clustering in ArcGIS. However, the use of these two methods is limited to the analysis of multi-band satellite imagery. In SavGIS[1], PCA can be applied both on vector and raster layers, but HAC is not available. In QGIS, there are equivalent tools also dedicated to image analysis through the GRASS functionalities. QGIS also makes it possible to execute R scripts[2], some of which are dedicated to factor analysis (PCA, MCA, FAMD, based on the ade4 and FactoMineR packages) and clustering (HAC with the cluster package). These scripts, like all the other features available in the QGIS toolbox, can be appended to a workflow that can be designed in the graphical modeler. However, the analysis is extremely broken up: it is necessary to execute a first script to get the correlation circle, a second script to get the factorial map with the variables, a third script to get the factorial map with the individuals and a fourth one to get the contribution plot. In addition, with these scripts, the PCA can only be applied to 4 variables, which greatly limits their interest. Concerning the HAC script, it must be applied to 5 variables and cannot therefore be used to create a typology on the first two or three factors obtained from a previous PCA.

[1] http://www.savgis.org/SavGIS/accueil.html (accessed on May 28, 2019)

[2] The whole list is available at https://github.com/qgis/QGIS-Processing/tree/master/rscripts (accessed on May 13, 2019)

These observations led us to develop a tool called FactoQGIS that meets the following criteria:

- A tool that can be easily executed from a graphical interface, without scripting,
- A tool integrated into a free and open source GIS software that is widely used,
- A tool that produces outputs that are easy to interpret and that can be directly used in QGIS,
- A tool with detailed contextual help and default settings, the most commonly used

FactoQGIS will hence be of particular interest for an academic purpose especially for geographers, urban planners and GIS analyst students.

## OPERATING PRINCIPLES OF FACTOQGIS

### License, languages, software versions and script content

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 [3]. 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[4]. 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

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.
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.

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

[3] This add-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.
[4] https://github.com/ESO-Rennes/FactoQGIS (accessed on May 27, 2019)

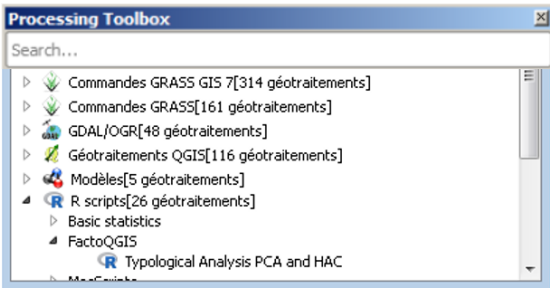



**Fig. 1. FactoQGIS in the QGIS toolbox**

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. The parameters are detailed in the appendices at the end of the article.

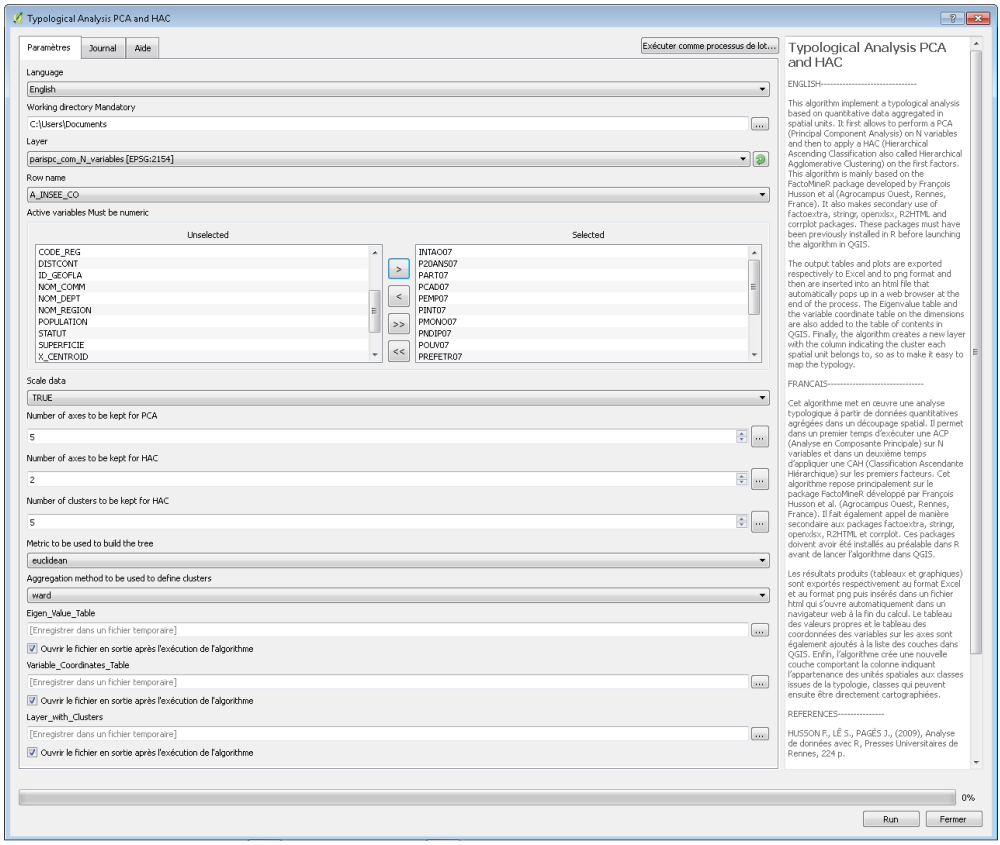



**Fig. 2. The FactoQGIS tool dialog box**

## Outputs format and output files

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 an attribute indicating the cluster each spatial unit belongs to | shp | x (only if a name was given by the user) | | x |

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

## APPLICATION OF FACTOQGIS TO ANALYZE THE SOCIO-DEMOGRAPHIC SPATIAL PATTERN OF PARIS

To illustrate how FactoQGIS works, we performed as an example, a typological analysis on data which are provided along with the book written by Commenges et al. (2014)[5] and which correspond to a sample of the French 2007 census data. These data are aggregated by "arrondissements" and "communes" in Paris and its first outer ring. The dataset includes 143 spatial units scattered over 4 administrative "départements"[6]. With regard to a series of 14 socio-demographic indicators (Table 2), the authors aimed at understanding how Paris and its first outer ring are organized and at pointing out the similarities and dissimilarities between spatial units.

[5] http://framabook.org/docs/Respace/RetEspace_Donnees.zip
[6] See map of Paris and its first outer ring: https://fr.wikipedia.org/wiki/Fichier:Petite_couronne.svg

**Table 2. List of the 14 socio-demographic indicators used in the typological analysis**

| Label | Description |
|---|---|
| **INTAO07** | **Proportion of temporary workers (employed labor force)** |
| **P20ANS07** | **Proportion of people under 20 years of age (total population)** |
| **PART07** | **Proportion of craftsmen (employed labor force)** |
| **PCAD07** | **Proportion of executives (employed labor force)** |
| **PEMP07** | **Proportion of employees (employed labor force)** |
| **PINT07** | **Proportion of intermediate occupations (employed labor force)** |
| **PMONO07** | **Proportion of single-parent families (families)** |
| **PNDIP07** | **Proportion of non-graduates (population > 15 years old)** |
| **POUV07** | **Proportion of workers (employed labor force)** |
| **PREFETR07** | **Proportion of households with a head being a foreign person (households)** |
| **PRET07** | **Proportion of retirees (employed labor force)** |
| **REFEROUI07** | **Percentage of votes in favor for the YES in the European referendum (2006)** |
| **RFUCQ07** | **Median income (current euro)** |
| **TXCHOM07** | **Proportion of unemployed (labor force)** |

*Source: INSEE (2007), presented in Commenges et al. (2014)*

In this article, the objective is no to present in detail the analysis of the Parisian socio-demographic spatial pattern which was already done by Commenges et al., but just to illustrate how FactoQGIS works and how it was applied to these data. In that respect, the following screenshots show the results as they appear in the output html file and every result is associated with a brief comment.

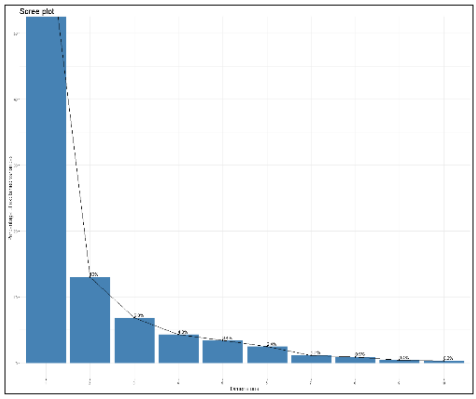
The PCA allows to summarize observations which are defined in a space of p variables to a space of p principal components. The advantage of this method is to reduce the dimensions and eliminate collinearity between variables. The components (also called factors, dimensions or axes) correspond to linear combinations of all the indicators analyzed. The principal components are synthetic variables that identify the main differentiation factors within the initial table. Several metrics can be used to characterize these components, starting with the eigenvalues. The latter reflect the inertia of the point cloud explained by each factor. The sum of these eigenvalues gives the total variance (also called inertia). To get an idea of how the variables are structured, it is necessary to examine the relative part of variance for each component as well as their cumulative part. In our example, these indications are contained in the table 3 and are also represented in graphical form in a scree plot (Figure 3).

It appears that 67% of the information contained in the 14 initial attributes is summarized by the first factor (see variance percentage in table 3). The discriminating power of the following axes is relatively low (Figure 3). The two first dimensions summarize 80 % of the total inertia and the first fifth dimensions almost 95% (see cumulative variance percentage in table 3). This information confirms that the default values for the $7^{th}$ and $8^{th}$ input parameters in the dialog box (Figure 2) are the good ones and that they do not need to be changed in this case.

**Table 3. Eigen values for each dimension**

| | eigenvalue | variance.percent | cumulative.variance.percent |
|---|---|---|---|
| Dim.1 | 9.4e+00 | 6.7e+01 | 6.7e+01 |
| Dim.2 | 1.8e+00 | 1.3e+01 | 8.0e+01 |
| Dim.3 | 9.5e-01 | 6.8e+00 | 8.7e+01 |
| Dim.4 | 6.0e-01 | 4.3e+00 | 9.1e+01 |
| Dim.5 | 4.7e-01 | 3.4e+00 | 9.4e+01 |
| Dim.6 | 3.4e-01 | 2.4e+00 | 9.7e+01 |
| Dim.7 | 1.6e-01 | 1.1e+00 | 9.8e+01 |
| Dim.8 | 1.2e-01 | 8.8e-01 | 9.9e+01 |
| Dim.9 | 5.5e-02 | 3.9e-01 | 9.9e+01 |
| Dim.10 | 4.0e-02 | 2.8e-01 | 1.0e+02 |
| Dim.11 | 2.9e-02 | 2.1e-01 | 1.0e+02 |
| Dim.12 | 2.2e-02 | 1.6e-01 | 1.0e+02 |
| Dim.13 | 1.4e-02 | 1.0e-01 | 1.0e+02 |
| Dim.14 | 1.3e-03 | 9.6e-03 | 1.0e+02 |

**Figure 3. Scree plot (gain of inertia)**




The first factorial map is defined by the first two dimensions (Figure 4). The first axis (dim. 1) clearly differentiates on the left part of the plot, spatial units with a high proportion of executives (PCAD07) and high incomes (RFUCQ07), and on the right part of the plot, spatial units characterized by variables indicating a greater socio-economic disadvantage (such as a high unemployment rate, TXCHOM07).

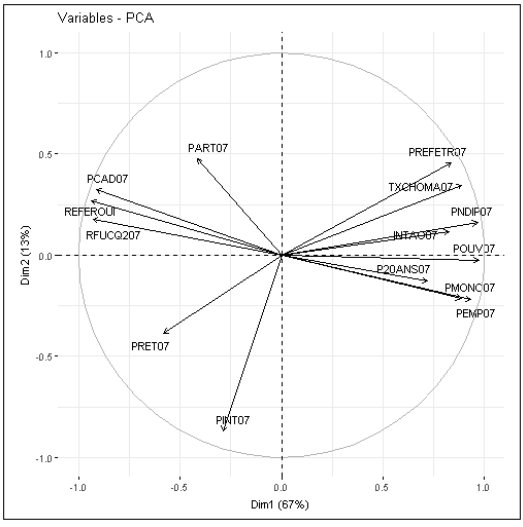



**Figure 4. First factorial map showing the variables (dimensions 1 and 2)**

The quality of the representation of the variable on the dimensions is another important metrics to characterize the components. The quality is measured by the square cosine ($cos^2$) of the angle between a variable and its projection on an axis. This square cosine is calculated for every dimension. Figure 5 illustrates this quality. The bigger and darker the circle, the better the quality. Strictly speaking, only well projected elements should be interpreted.

For example, the variable POUV07 has a high quality on the first dimension. This means that the $cos^2$ is very close to 1 and the angle is very close to 0 (POUV07 vector is almost aligned with the dimension 1 in figure 4). On its part, the variable PINT07 has a low quality on the first dimension. This variable is much less well projected on the dimension 1 (see its high angle with dimension 1 in figure 4).

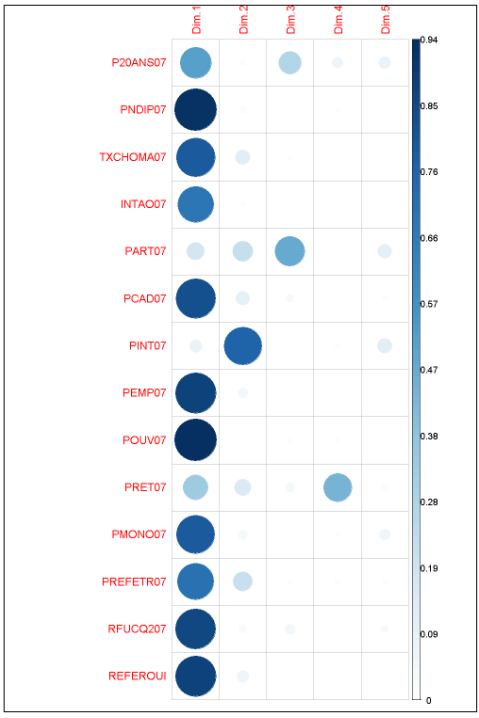



**Figure 5. Quality of the representation of the variables on the dimensions (Cos²)**

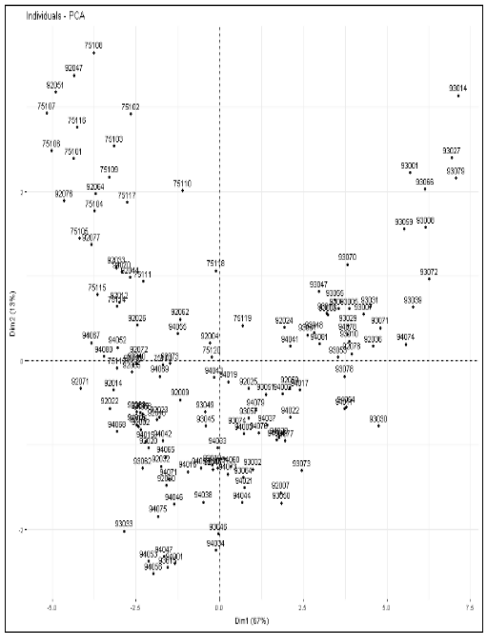



**Figure 6. First factorial map showing the coordinates of the individuals (dimensions 1 and 2)**

Figure 6 is the first factorial map showing the coordinates of the individuals and can be overlapped with the first factorial map showing the variables (figure 4). Profiles seem to be emerging: on the first axis, we can distinguish on the left, spatial units with ID starting with 75 (inner Paris) which have a high proportion of executives and high incomes (see figure 4) and on the right, spatial units with ID starting with 93 (Seine Saint-Denis) characterized by variables indicating a greater socio-economic disadvantage (see figure 4).

Hierarchical cluster trees allow to get much accurate profiles. Clustering algorithms aim at defining groups of individuals (spatial units, in our case) that are homogeneous in terms of their statistical attributes. Groups are homogeneous if the statistical individuals in each group are as similar as possible within each group. The algorithm in the example returns 5 optimal clusters (Figure 7).

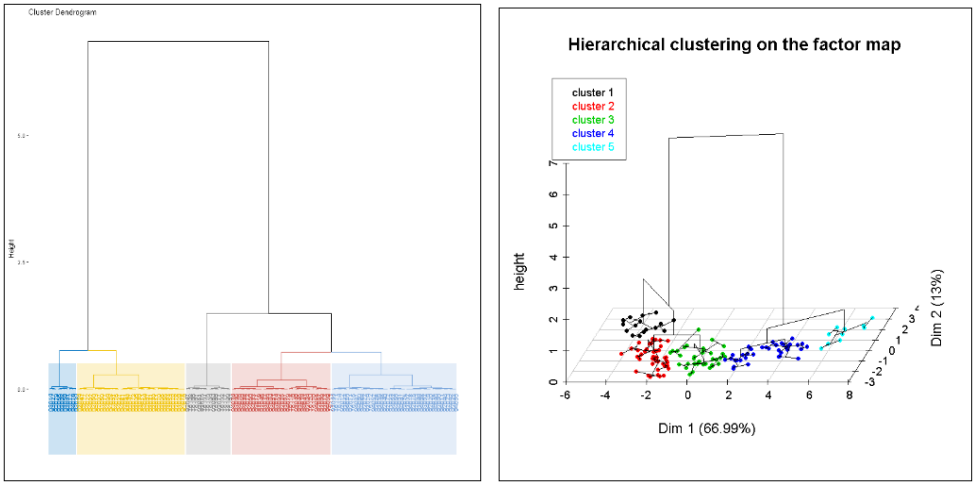
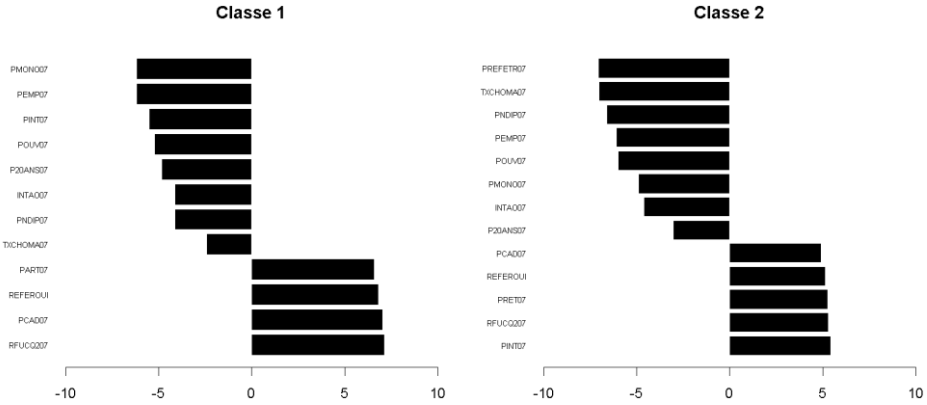



**Figure 7. Hierarchical cluster trees**

The following bar plots (figure 8) show for each cluster the values taken by the variables in comparison to the overall mean (table 4). These values are useful to qualify each cluster. Only the variables associated with a v-test >|1.96| are significant and are therefore plotted.



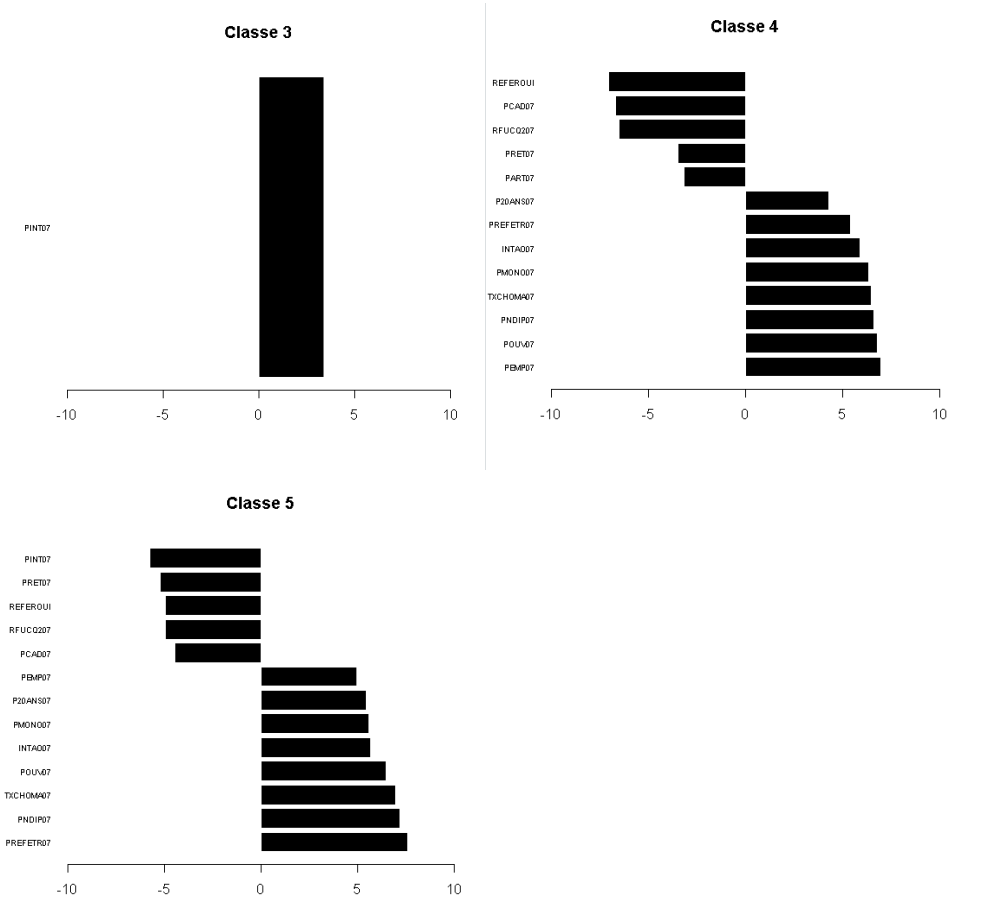



**Figure 8. Bar plots showing the variables which best describe the clusters (only the variables with a v-test >|1.96| are plotted)**

- Cluster 1 corresponds to wealthy municipalities (Median income - RFUCQ high; Proportion of executives - PCAD high; Percentage of votes for YES in the 2006 European referendum - REFEROUI high; Percentage of craftsmen - High proportion).
- Cluster 2 corresponds to upper middle-class municipalities with a high proportion of intermediate professions (PINT07), rather high median incomes (RFUCQ07), an over-representation of retirees (PRET07), a percentage of votes for YES in the 2006 European referendum - REFEROUI rather high. On the other hand, the proportion of households with a head being a foreign person (PREFETR), the proportion of non-graduates (PNDIP) and the proportion of unemployed people (TXCHOM) is low.
- Cluster 3, in the center of the graph (figure7), has only an over-representation of intermediate professions as a particular characteristic.

- Cluster 4 corresponds to lower middle-class municipalities with a high proportion of employees (PEMP07) and workers (POUV07), as well as a high proportion of unqualified (PNDIP07) and unemployed people (TXCHOM).
- Class 5 includes particularly disadvantaged municipalities.

1

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| RFUCQ207 | 7.136 | 33451.31 | 21686.64 | 5872.162 | 6973.155 | 0 |
| PCAD07 | 7.0243 | 49.5 | 25.951 | 2.2638 | 14.1798 | 0 |
| REFEROUI | 6.827 | 76.5062 | 52.3035 | 4.2268 | 14.9946 | 0 |
| PART07 | 6.5859 | 7.25 | 4.8811 | 1.5207 | 1.5214 | 0 |
| TXCHOMA07 | -2.4288 | 9.125 | 11.4615 | 1.3636 | 4.0689 | 0.0151 |
| PNDIP07 | -4.1155 | 10 | 18.8182 | 1.7678 | 9.0628 | 0 |
| INTAO07 | -4.1354 | 0.6875 | 1.3427 | 0.4635 | 0.6701 | 0 |
| DISTCONT | -4.6416 | 4.7828 | 10.4849 | 4.4925 | 5.196 | 0 |
| P20ANS07 | -4.8515 | 21.0625 | 25.6434 | 4.6296 | 3.9937 | 0 |
| POUV07 | -5.2392 | 5.125 | 15.1329 | 1.1659 | 8.0794 | 0 |
| PINT07 | -5.505 | 20.1875 | 25.4196 | 2.0377 | 4.0199 | 0 |
| PEMP07 | -6.179 | 17.8125 | 28.6224 | 1.8445 | 7.3995 | 0 |
| PMONO07 | -6.1847 | 6.3125 | 10.4545 | 0.9164 | 2.8327 | 0 |

2

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| PINT07 | 5.4139 | 28.1591 | 25.4196 | 3.4241 | 4.0199 | 0 |
| RFUCQ207 | 5.3079 | 26345.68 | 21686.64 | 2413.373 | 6973.155 | 0 |
| PRET07 | 5.2822 | 16.8409 | 15.1538 | 2.3543 | 2.5373 | 0 |
| REFEROUI | 5.1529 | 62.0295 | 52.3035 | 6.8878 | 14.9946 | 0 |
| PCAD07 | 4.9169 | 34.7273 | 25.951 | 8.529 | 14.1798 | 0 |
| P20ANS07 | -3.0429 | 24.1136 | 25.6434 | 2.9559 | 3.9937 | 0.0023 |
| INTAO07 | -4.6013 | 0.9545 | 1.3427 | 0.2083 | 0.6701 | 0 |
| PMONO07 | -4.9078 | 8.7045 | 10.4545 | 1.1981 | 2.8327 | 0 |
| POUV07 | -5.9856 | 9.0455 | 15.1329 | 2.763 | 8.0794 | 0 |
| PEMP07 | -6.0851 | 22.9545 | 28.6224 | 3.1691 | 7.3995 | 0 |
| PNDIP07 | -6.6142 | 11.2727 | 18.8182 | 2.1676 | 9.0628 | 0 |
| TXCHOMA07 | -7.0246 | 7.8636 | 11.4615 | 1.3072 | 4.0689 | 0 |
| PREFETR07 | -7.0488 | 9.0909 | 16.3497 | 2.4292 | 8.1809 | 0 |

5

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| PREFETR07 | 7.569 | 35.3 | 16.3497 | 5.2924 | 8.1809 | 0 |
| PNDIP07 | 7.1682 | 38.7 | 18.8182 | 4.2673 | 9.0628 | 0 |
| TXCHOMA07 | 6.937 | 20.1 | 11.4615 | 1.578 | 4.0689 | 0 |
| POUV07 | 6.4576 | 31.1 | 15.1329 | 2.6627 | 8.0794 | 0 |
| INTAO07 | 5.6435 | 2.5 | 1.3427 | 0.5 | 0.6701 | 0 |
| PMONO07 | 5.5892 | 15.3 | 10.4545 | 1.6155 | 2.8327 | 0 |
| P20ANS07 | 5.4462 | 32.3 | 25.6434 | 2.3685 | 3.9937 | 0 |
| PEMP07 | 4.9359 | 39.8 | 28.6224 | 2.0396 | 7.3995 | 0 |
| PCAD07 | -4.4822 | 6.5 | 25.951 | 1.9105 | 14.1798 | 0 |
| RFUCQ207 | -4.9436 | 11136.7 | 21686.64 | 1094.014 | 6973.155 | 0 |
| REFEROUI | -4.9626 | 29.53 | 52.3035 | 3.4954 | 14.9946 | 0 |
| PRET07 | -5.2205 | 11.1 | 15.1538 | 1.6401 | 2.5373 | 0 |
| PINT07 | -5.787 | 18.3 | 25.4196 | 1.4177 | 4.0199 | 0 |

4

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| PEMP07 | 6.9691 | 35.8158 | 28.6224 | 2.6343 | 7.3995 | 0 |
| POUV07 | 6.7936 | 22.7895 | 15.1329 | 3.7216 | 8.0794 | 0 |
| PNDIP07 | 6.5967 | 27.1579 | 18.8182 | 4.1646 | 9.0628 | 0 |
| TXCHOMA07 | 6.4659 | 15.1316 | 11.4615 | 2.2144 | 4.0689 | 0 |
| PMONO07 | 6.3086 | 12.9474 | 10.4545 | 1.5381 | 2.8327 | 0 |
| INTAO07 | 5.9062 | 1.8947 | 1.3427 | 0.552 | 0.6701 | 0 |
| PREFETR07 | 5.4125 | 22.5263 | 16.3497 | 4.5116 | 8.1809 | 0 |
| P20ANS07 | 4.2774 | 28.0263 | 25.6434 | 2.2418 | 3.9937 | 0 |
| PART07 | -3.1599 | 4.2105 | 4.8811 | 0.8322 | 1.5214 | 0.0016 |
| PRET07 | -3.4831 | 13.9211 | 15.1538 | 1.5455 | 2.5373 | 5e-04 |
| RFUCQ207 | -6.5239 | 15340.74 | 21686.64 | 1642.102 | 6973.155 | 0 |
| PCAD07 | -6.7071 | 12.6842 | 25.951 | 4.1174 | 14.1798 | 0 |
| REFEROUI | -7.0472 | 37.5632 | 52.3035 | 5.1588 | 14.9946 | 0 |

**Table 4. Tables giving the description of the clusters by the variables (only the variables with a v-test >|1.96| are listed)**

The joint analysis of the map (figure 9) and the previous plots and tables shows a set of well-off municipalities in the West of Paris and the Hauts-de-Seine (cluster 1 in black and cluster 2 in red). On the other hand, the northern part of the study area, corresponding to most of the communes of Seine-Saint-Denis concentrates more social disadvantage (cluster 4 in dark blue and cluster 5 in light blue). On its side the Val-de-Marne has a very contrasting situation with 4 clusters out of 5.

## DISCUSSION AND CONCLUSION

FactoQGIS is an algorithm that allows a seamless execution of geometric data analysis in the QGIS environment. More specifically, this tool was designed to perform a typological analysis on quantitative data aggregated in spatial units. This tool relies on R packages and can easily be executed from a GUI. It is of particular interest to geographers and to any users who wish to simply build and map in a free and open source GIS software, a multidimensional typology without knowing the R language. The tool presented in this article is a first release. Several avenues for improvement are being considered. For

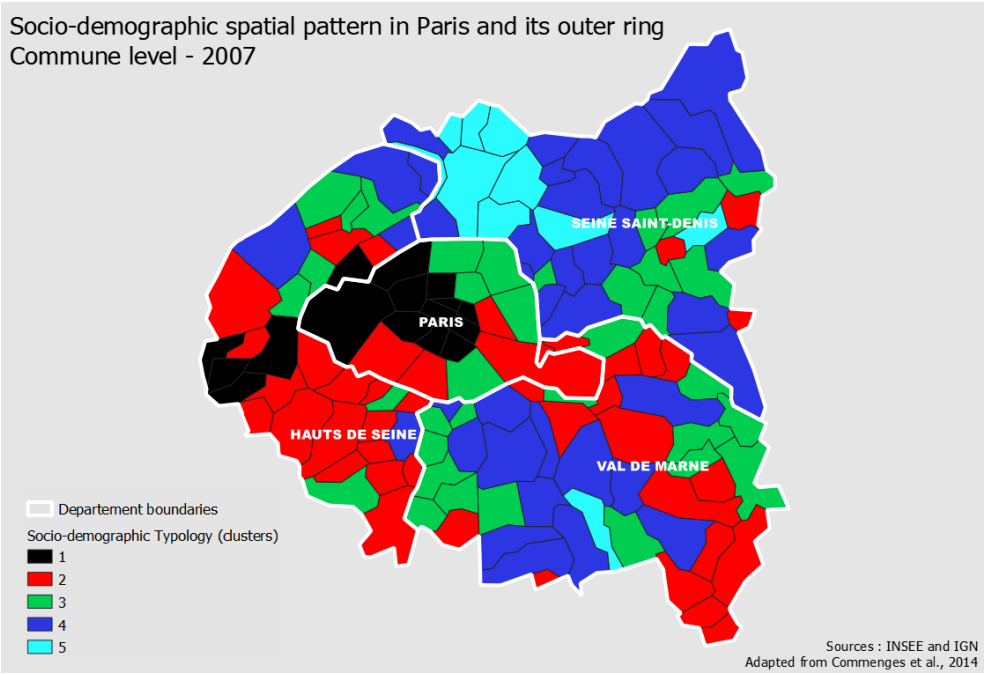



**Figure 9. Output map portraying the typology in QGIS**

example, we plan to add the possibility of choosing supplementary quantitative and qualitative variables. The option of applying a partitioning method based on the k-means before performing the HAC is also planned. This will allow FactoQGIS to be used on large datasets (several thousand spatial units). Finally, additional developments will make it possible to integrate other types of factor analysis, in particular MCA and FAMD, in order to easily produce typology maps whatever the kind of input data. The scripts of all these forthcoming developments will be provided to the community under GNU GPL license in GitHub.

ACKNOWLEDGEMENTS

Students of the second year of the SIGAT Master Degree 2018-2019 (Université Rennes 2, France) and Mégane Bouquet (UMR ESO 6590 CNRS, Rennes, France)

**References**

1. Benzécri, J.P., (1973) *L'Analyse des données*, Dunod, 619 p. ISBN 2-04-007225-X
2. Commenges, H. (dir.), (2014) *R et espace - Traitement de l'information géographique*. Groupe ElementR- Framabook. Available online: https://framabook.org/r-et-espace/ (accessed on May 13, 2019)

3. Demoraes, F.; Gouëset, V.; Piron, M.; Figueroa, O.; Zioni, S. (2013) Desigualdades socioterritoriais e mobilidades cotidianas nas metrópoles de América Latina: uma comparação entre Bogotá, Santiago de Chile e São Paulo. *Revista dos Transportes Públicos - ANTP*, Planejamento urbano, 35 (134), 9-30. Available online: https://halshs.archives-ouvertes.fr/halshs-01110019 (accessed on May 13, 2019)
4. Graser, A.; Olaya, V. (2015) Processing: A Python Framework for the Seamless Integration of Geoprocessing Tools in QGIS. Vol. 4, *ISPRS Int. J. Geo-Information*, 2219-2245. Available online: https://doi.org/10.3390/ijgi4042219 (accessed on May 13, 2019)
5. Greenacre M. J.; Blasius J. (2006). Multiple Correspondence Analysis and Related Methods. CRC press. ISBN 978-1-58488-628-0.
6. Hanspach, J.; Loos, J.; Dorresteijn, I.; Abson, D.; Fischer, J. (2016) Characterizing social–ecological units to inform biodiversity conservation in cultural landscapes, *Diversity and Distributions*, 1-12, Available online: https://doi.org/10.1111/ddi.12449 (accessed on May 13, 2019)
7. Husson, F.; Lê, S.; Pagès, J., (2009) *Analyse de données avec R*, Presses Universitaires de Rennes. 224 p. ISBN 978-2753509382
8. Le Roux, B.; Rouanet, H. (2004) *Geometric Data Analysis - From Correspondence Analysis to Structured Data Analysis*, Springer Netherlands, 475 p. ISBN 978-1-4020-2236-4
9. Lebart, L.; Piron, M.; Morineau, A. (2006) *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données*, Dunod. 464 p. ISBN 978-2100496167
10. Metzger, P. (2001) *Perfiles ambientales de Quito*. Quito: MDMQ; IRD, 117 p. ISBN 9978-41-682-X, Available online: http://www.documentation.ird.fr/hor/fdi:010026340 (accessed on May 13, 2019)
11. Piron, M.; Dureau, F.; Mullon, C. (2004) Dynamique du parc de logements à Bogota : Analyse par typologies multi-dates. *Cybergeo*, 1-23, Available online: https://journals.openedition.org/cybergeo/2925 (accessed on May 13, 2019)
12. Rivière, J. (2012) Mapping votes and social inequalities: the case of Paris and its inner suburbs, *Metropolitics*, 1-8, Available online: https://www.metropolitiques.eu/Mapping-votes-and-social.html (accessed on May 13, 2019)
13. Robson, B.; Lymperopoulou, K.; Rae, (2009) A. A typology of the functional roles of deprived neighbourhoods, Centre for Urban Policy Studies, Manchester University, Department for Communities and Local Government, 63 p. ISBN: 978-1-4098-1017-9, Available online: https://webarchive.nationalarchives.gov.uk/20120919132719/http:/www.communities.gov.uk/documents/communities/pdf/1152966.pdf (accessed on May 13, 2019)
14. Walsh, J. (2000) *Irish Rural Structure and Gaeltacht Areas*. National Spatial Strategy Report. Maynooth and Brady Shipman Martin. Available online: http://www.irishspatialstrategy.ie/docs/report10.pdf (accessed on May 13, 2019)

ONLINE RESOURCES

1. Blog on how to execute R scripts in QGIS 3.0 and later. Available online: https://github.com/north-road/qgis-processing-r/releases/tag/v0.0.2 (accessed on May 13, 2019)
2. List of the R scripts that can be executed from the QGIS Toolbox. Available online: https://github.com/qgis/QGIS-Processing/tree/master/rscripts (accessed on May 13, 2019)
3. Documentation of the FactoMineR package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/FactoMineR/versions/1.41 (accessed on May 13, 2019)
4. Documentation of the factoextra package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/factoextra/versions/1.0.5 (accessed on May 13, 2019)
5. Documentation of the stringr package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/stringr/versions/1.3.1 (accessed on May 13, 2019)

6. Documentation of the openxlsx package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/openxlsx/versions/4.1.0 (accessed on May 13, 2019)
7. Documentation of the R2HTML package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/R2HTML/versions/2.3.2 (accessed on May 13, 2019)
8. Documentation of the corrplot package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/corrplot/versions/0.84 (accessed on May 13, 2019)
9. Data set used in this article as an example to illustrate how FactoQGIS works. Available online: http://framabook.org/docs/Respace/RetEspace_Donnees.zip (accessed on May 13, 2019)

APPENDICES

*FactoQGIS 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 factorial 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.

*FactoQGIS 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.

*Settings and activation of R scripts in QGIS*
The procedure is available online:
https://docs.qgis.org/2.18/en/docs/training_manual/processing/r_intro.html (accessed on May 13, 2019)

*Writing new processing algorithms as python scripts*
The procedure is available online:
https://docs.qgis.org/2.8/en/docs/user_manual/processing/scripts.html (accessed on May 13, 2019)

## Authors

**Florent Demoraes,** Lecturer at the University of Rennes 2 (areas: Latin American metropolises, teaching activities, teaching topics, GIS theory, basics of spatial analysis, etc.), Director of ESO Laboratory-Rennes - UMR CNRS 6590 Spaces and Societies. He is a member of the Scientific Council of the American Institute in Rennes. Head of ESO collection in HAL with Stéphane Lorette (ESO-Nantes). Head of IT Management Committee - ESO-Rennes Laboratory. Scientific area of research includes geomatics, spatial analysis, spatial statistics, mobility in space, residential segregation and social inequalities (spatial dimension).

**Dr. Marc Souris,** is senior research director at the Institut de Recherche pour le Développement (IRD), in the UMR 190 "Unité des virus émergents". He holds a PhD in Computer Science and a Habilitation to Lead Research. Mathematician and computer scientist by training, his work at IRD mainly concerns information sciences applied to geography and epidemiology. Since 1983, he has been developing research, innovation, software, and teaching in geomatics: geographic information systems, spatial analysis, statistics, modelling. He is the main author of the SavGIS GIS software package (www.savgis.org). The thematic applications that it has carried out during its scientific activities have focused mainly on the issue of risk, whether "natural" or "health": epidemiology, health and environment, risk analysis and prevention for natural disasters. He has also worked in the fields of urban planning, land use planning, geology and archaeology. He has taught the foundations of geomatics and spatial analysis for epidemiology and geography of health in Master II and PhD programs at the University of Paris Ouest-Nanterre La Défense and at the Asian Institute of Technology (Thailand).