**Algorithm description**

Works with:  
Processing R Provider: version 2.x and newer   
QGIS: version 3.4 and newer  
R: version 3.5.3 and newer  
  
This algorithm makes it possible to implement a typological analysis based on quantitative data aggregated in spatial units. It first allows to perform a PCA (Principal Component Analysis) on N variables and then to apply a HAC (Hierarchical Ascending Classification) on the first factors. This algorithm is mainly based on the FactoMineR package developed by François Husson et al (Agrocampus Ouest, Rennes, France).

The output tables and plots are exported respectively to Excel and to png format and then are inserted into an html file that automatically pops up in a web browser at the end of the process. The eigenvalue table and the variable coordinate table on the axes are also added to the table of content in QGIS. Finally, the algorithm creates a new layer which contains an attribute field indicating the cluster each spatial unit belongs to, so as to make it easy to map the typology. For this purpose, a style file in QML format is created and can be used as default rendering (the colors of the clusters are the same as those on the 3D hierarchical tree).

The algorithm offers several setting options with default values.  
  
REFERENCES---------------  
  
BENZECRI, J.P., (1973) L'Analyse des données, Dunod, 619 p. ISBN 2-04-007225-X  
CORNILLON P.A., GUYADER A., HUSSON F., JEGOU N., JOSSE J., KLUTCHNIKOFF N., LE PENNEC E., MATZNER-LØBER E., ROUVIERE L. & THIEURMEL B. (2018). R pour la statistique et la science des données. Presses Universitaires de Rennes, 1 ed. 415 p., ISBN : 978-2-7535-7573-8  
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)  
GREENACRE M. J.; BLASIUS J. (2006), Multiple Correspondence Analysis and Related Methods. CRC press. ISBN 978-1-58488-628-0.  
HUSSON F., LÊ S., PAGÈS J., (2009), Analyse de données avec R, Presses Universitaires de Rennes, 224 p. ISBN 978-2753509382  
LE ROUX B., ROUANET H. (2005), Geometric Data Analysis - From Correspondence Analysis to Structured Data Analysis, Springer Netherlands, 475 p. ISBN 978-1-4020-2236-4  
LEBART L., PIRON M., MORINEAU A., (2006), Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données, Dunod, 464 p.  
  
See also:  
https://hal.archives-ouvertes.fr/hal-02181611  
https://www.rdocumentation.org/packages/FactoMineR/versions/1.41   
https://github.com/ESO-Rennes/FactoQGIS

**Input parameters**

**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, plots and the QML style file will be stored in it.

**Input 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.

**Spatial Units ID**

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.

**Active Quantitative Variables**

Actives variables on which the PCA will be performed. Must be numeric.

**Additional Quantitative Variables**

Additional quantitative variable(s) used to interpret the results. This/these variable(s) does (do) not participate in the construction of the axes.

**Additional Qualitative Variables**

Additional qualitative variable(s) used to interpret the results. This/these variable(s) does (do) not participate in the construction of the axes.

**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.

**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.

**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 classification. 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.

**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 2D hierarchical tree. If needed you can change the default value and perform a second time the HAC.

**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".

**Aggregation method to be used to define clusters**

Clustering method. The four methods implemented are "average" ([unweighted pair-]group [arithMetic] average method, aka ‘UPGMA’), "single" (single linkage), "complete" (complete linkage), and "ward" (Ward's method). Ward's method is the default value.

**Outputs**

**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 content in QGIS and is also exported to an Excel table sheet.

**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.

**Layer with Clusters**

Output vector layer with the attribute field indicating the cluster each spatial unit belongs to so as to easily mapping the typology. The style file in QML format created by the algorithm can be used as default rendering (the colors of the clusters are the same as those on the 3D hierarchical tree).

Algorithm author: Florent DEMORAES (ESO-Rennes - UMR CNRS 6590, Université Rennes 2) et étudiants du M2 SIGAT (Master 2 Système d'Information Géographique et Analyse des Territoires, Université Rennes 2, Promotions 2018-2019 et 2019-2020)

Help author: Florent DEMORAES (ESO-Rennes - UMR CNRS 6590, Université Rennes 2)

Algorithm version: V2 - Novembre 2019