FactoQGIS help

Florent Demoraes, December 2018

This is the content of the file named "**Typological_Analysis_PCA_PCA_and_HAC.rsx.help**" that must be stored in the following folder: C:\Users\...\...\ qgis2\processing\rscripts

The help is available both in English and French.

Algorithm author: Florent Demoraes
Help author: Florent Demoraes
Algorithm version: V1 - December 2018

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)

Algorithm description

ENGLISH----------
This algorithm implements a typological analysis based on quantitative data aggregated in spatial units. First, it performs a PCA (Principal Component Analysis) on N variables and second, a HAC (Hierarchical Ascending Classification also called Hierarchical Agglomerative Clustering) on the first factors. This algorithm is mainly based on the FactoMineR package developed by François Husson et al (Agrocampus Ouest, Rennes, France). It also makes secondary use of factoextra, stringr, openxlsx, R2HTML and corrplot packages. These packages must have been previously installed in R before launching the algorithm in QGIS. 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 dimensions are also added to the table of contents in QGIS. Finally, the algorithm creates a new layer with the column indicating the cluster each spatial unit belongs to, so as to make it easy to map the typology.

FRANCAIS----------
Cet algorithme met en œuvre une analyse typologique à partir de données quantitatives agrégées dans un découpage spatial. Il permet dans un premier temps d'exécuter une ACP (Analyse en Composante Principale) sur N variables et dans un deuxième temps d'appliquer une CAH (Classification Ascendante Hiérarchique) sur les premiers facteurs. Cet algorithme repose principalement sur le package FactoMineR développé par François Husson et al. (Agrocampus Ouest, Rennes, France). Il fait également appel de manière secondaire aux packages factoextra, stringr, openxlsx, R2HTML et corrplot. Ces packages doivent avoir été installés au préalable dans R avant de lancer l'algorithme dans QGIS. Les résultats produits (tableaux et graphiques) sont exportés respectivement au format Excel et au format png puis insérés dans un fichier html qui s'ouvre automatiquement dans un navigateur web à la fin du calcul. Le tableau des valeurs propres et le tableau des coordonnées des variables sur les axes sont également ajoutés à la liste des couches dans QGIS. Enfin, l'algorithme crée une nouvelle couche comportant la colonne indiquant l'appartenance des unités spatiales aux classes issues de la typologie, classes qui peuvent ensuite être directement cartographiées.

REFERENCES----------
1. BENZECRI, J.P., (1973) L'Analyse des données, Dunod, 619 p. ISBN 2-04-007225-X
2. 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)
3. HUSSON F., LÊ S., PAGÈS J., (2009), Analyse de données avec R, Presses Universitaires de Rennes, 224 p. ISBN 978-2753509382
4. 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
5. LEBART L., PIRON M., MORINEAU A., (2006), Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données, Dunod, 464 p.

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)

Input parameters

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.

Français ou anglais. Cela définira la langue à appliquer aux titres des tableaux et graphiques dans le fichier html en sortie.

Working directory Mandatory

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.

Ce champ est obligatoire. Le chemin vers l'espace de travail doit être court et ne pas

comporter de caractères spéciaux ou d'espaces.
Tous les tableaux et graphiques en sortie seront enregistrés dans ce dossier.

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.
Couche sur laquelle appliquer l'ACP et la CAH. La table attributaire de cette couche doit contenir des variables quantitatives. Cette couche doit être ouverte dans QGIS.

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.
Champ contenant l'identifiant des unités spatiales. Cet ID apparaîtra sur les plans factoriels et est nécessaire aussi pour joindre les données en fin d'algorithme.

Active variables Must be numeric
Actives variables on which the PCA will be performed. Must be numeric.
Variables actives sur laquelle l'ACP sera calculée. Doivent être au format numérique.

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.
Option pour centrer-réduire les données. Doit être appliquer dans la plupart des cas et surtout lorsque les ordres de grandeur des variables sont très différents.

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.
Nombre d'axes à garder pour l'ACP. 5 est la valeur par défaut. En général on garde les N premiers facteurs qui expliquent au moins 95% de l'inertie.
Il est recommandé d'exécuter une première fois l'algorithme avec cette valeur par défaut, de vérifier le tableau des valeurs propres et le graphe des gains d'inertie.
Si cela se justifie, vous pouvez changer la valeur par défaut et relancer l'ACP.

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.
Nombre d'axes à garder pour la CAH. 2 est la valeur par défaut. En général on garde les N premiers facteurs qui expliquent au moins 80% de l'inertie afin d'obtenir une classification plus stable.
Il est recommandé d'exécuter une première fois l'algorithme avec cette valeur par défaut et de vérifier le tableau des valeurs propres et le graphe des gains d'inertie. Si cela se justifie, vous pouvez changer la valeur par défaut et relancer la CAH.

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.
Nombre de classes pour la CAH. 5 est la valeur par défaut. Il est recommandé d'exécuter une première fois l'algorithme avec cette valeur par défaut et de vérifier le dendrogramme.
Si cela se justifie, vous pouvez changer la valeur par défaut et relancer la CAH.

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".
Métrique à utiliser pour calculer les dissemblances entre les individus. Les options actuellement disponibles sont " euclidean " et "manhattan".
Les distances euclidiennes sont la somme des racines-carrés des différences, et les distances de Manhattan sont la somme des différences absolues. La valeur par défaut est " 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.
Méthode d'agrégation des individus pour la classification. Les quatre méthodes mises en œuvre sont "average", "single" (liaison simple), "complete" (liaison complète), et "ward" (méthode de Ward). La méthode de Ward est la valeur par défaut.

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 contents in QGIS and is also exported to an Excel tablesheet.
Tableau des valeurs propres qui donne pour chaque variable sa part dans l'inertie total. Cette table est automatiquement ajoutée à la table des matières dans QGIS et est également exportée vers une feuille de calcul Excel.

Variable_Coordinates_Table
Table which gives the coordinates of each variable on the axes. This table is automatically added to the table of contents in QGIS and is also exported to an Excel tablesheet.
Tableau qui donne les coordonnées de chaque variable sur les axes. Cette table est automatiquement ajoutée à la table des matières dans QGIS et est également exportée vers une feuille de calcul Excel.

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 to make it easy to map the typology.
Couche vectorielle en sortie avec la colonne indiquant la classe à laquelle appartient chaque unité spatiale afin de cartographier facilement la typologie.