Using R with ArcGIS to run factor analysis for mixed qualitative/quantitative data with the *FactorMineR* and *missMDA* packages

Before proceeding to the example, you must have the following installed on your computer:

PREREQUISITES

ArcGIS 10.3.1 or later

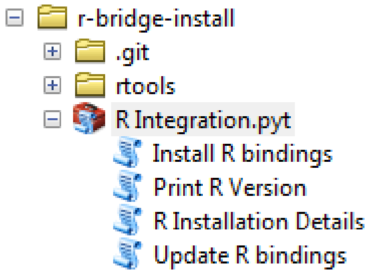
1. R Statistical Computing Software, 3.2.0 or later
    - 32-bit version required for ArcMap, 64-bit version required for ArcGIS Pro (Note: the installer installs both by default).
    - 64-bit version can be used with ArcMap by installing Background Geoprocessing and configuring scripts to run in the background.
2. R ArcGIS Bridge

SETUP INSTRUCTIONS

ArcGIS 10.3.1 or later

- In the Catalog window, navigate to the folder containing the Python Toolbox, `R Integration.pyt`. *Note*: You may have to first add a folder connection to the location that you extracted the files or downloaded via GitHub.

- Open the toolbox, which should look like this:



```
⊟ r-bridge-install
    ⊞ .git
    ⊞ rtools
    ⊟ R Integration.pyt
        Install R bindings
        Print R Version
        R Installation Details
        Update R bindings
```

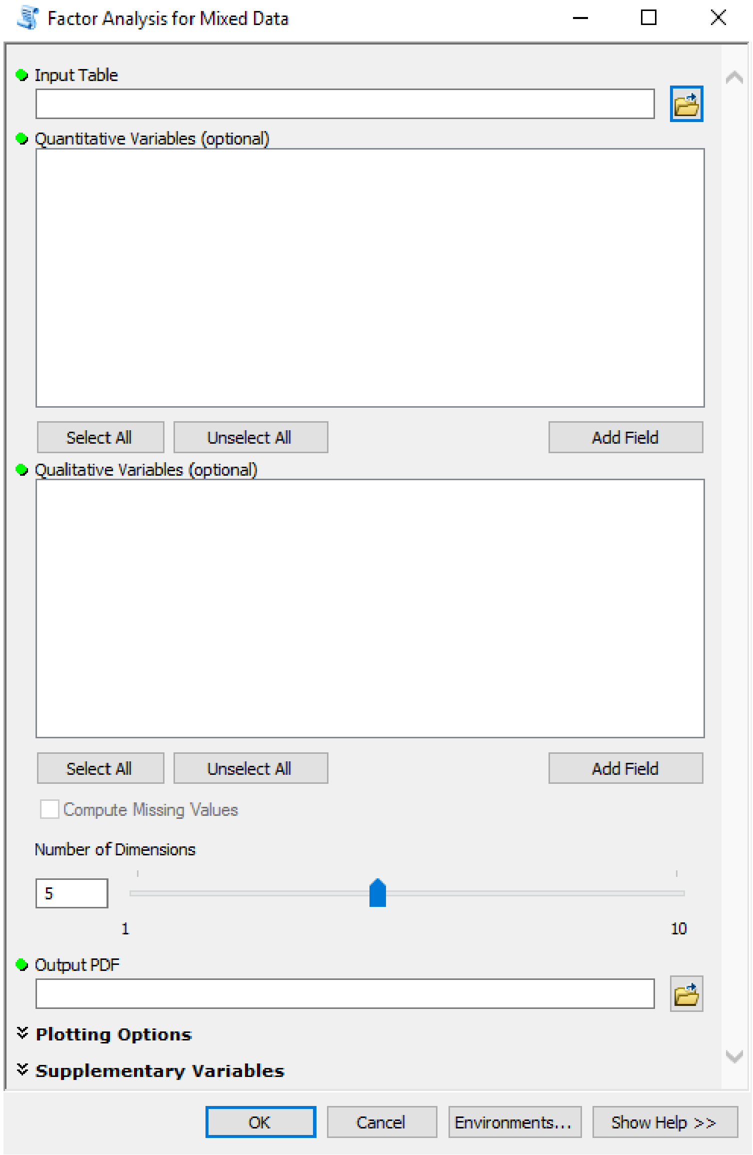
- Run the `Install R bindings` script. You can then test that the bridge is able to see your R installation by running the `Print R Version` and `R Installation Details` tools.

Background information

Tabular datasets, such as surveys, are a common data type found both within and outside the GIS sector. For example, population surveys record a large amount of socioeconomic information on individuals or households in the form of both categorical and quantitative variables. Other examples include datasets about city features like population size and density, poverty, income distribution, distance to green spaces, etc. Aside from describing socioeconomic and environmental characteristics associated with the sampled units, such datasets can be analyzed to reveal a smaller set of interpretable factors that are common (or related) to the recorded variables. At the same time, there may be clusters, groups of locations/regions whose socio-economic and environmental behavior is primarily described by a potentially small set of common underlying/latent factors. Factor analysis of mixed data (FAMD) [1], or factorial analysis of mixed data, is a useful statistical method devoted to data tables in which a group of individuals is described by a number of observed qualitative and quantitative, correlated variables. The word "*mixed*" refers to the simultaneous presence of both qualitative and quantitative data in the table. FAMD works similarly to principal component analysis (PCA) [1] for quantitative variables and as a multiple correspondence analysis (MCA) [1] for qualitative variables. FAMD can be used for investigating relationships among observed, correlated variables in terms of a potentially reduced number of unobserved variables called "factors".

The *Factor Analysis for Mixed Data* tool uses the *FactoMineR* [2][3] and *missMDA* [4] R packages to run FAMD on any combination of quantitative and qualitative variables.

Explanation of Parameters



□ Compute Missing Values

Number of Dimensions 5 (1–10)

Figure 1. *Factor Analysis for Mixed Data* tool.

Input Table: A table with attributes corresponding to qualitative and/or quantitative variables.

Quantitative Variables: Field(s) from the *Input Table* representing one or more quantitative variables.

Qualitative Variables: Field(s) from the *Input Table* representing one or more qualitative variables.

Supplementary Quantitative Variables: Field(s) from the *Input Table* representing one or more supplementary quantitative variables.

Supplementary Qualitative Variables: Field(s) from the *Input Table* representing one or more supplementary qualitative variables.

Compute Missing Values: Missing values are replaced using PCA (for quantitative only datasets), MCA (for qualitative only datasets), or FAMD (for mixed qualitative and quantitative datasets) imputation. When not checked missing values are replaced by their column average.

Number of Dimensions: Number of dimensions selected from the results of the analysis.

Color Plots By: String corresponding to the color which are used.

- Quantitative-only dataset: If “none”, no color is used for the individuals; if “ind”, a color for each individual (“ind”); if “var”, color the individuals based on a categorical variable
- Qualitative-only dataset: If “none”, one color is used for the individual, another one for the categorical variables; if “quali”, one color is used for each categorical variables; if “var”, colors are used according to the different categories of a categorical variable
- For mixed-data: If “none”, no color is used for the individuals; if “ind”, one color is used for each individual; if “var”, colors are used according to the different categories of a categorical variable

Categorical Variable: When “var” is selected, this name corresponds to the field of a categorical variable selected from the *Input Table*. The variable will be used to color final plots according to its different levels/categories.

Individuals Selection: A selection of the individuals that are drawn:

- “coord N”: select the N elements that have the highest (squared) coordinates on the 2 first dimensions drawn
- “contrib N”: select the N elements that have the highest contribution on the 2 first dimensions drawn
- “cos2 N”: select the N elements that have the highest cos2 on the 2 dimensions drawn
- “dist N”: select the N elements that have the highest distance to the center of gravity
- “cos2 N”: select the N elements that have the highest cos2 on the 2 dimensions drawn
- “dist N”: select the N elements that have the highest distance to the center of gravity

Categories Selection: A selection of the categories that are drawn:

- “coord N”: select the N elements that have the highest (squared) coordinates on the 2 first dimensions drawn
- “contrib N”: select the N elements that have the highest contribution on the 2 first dimensions drawn
- “cos2 N”: select the N elements that have the highest cos2 on the 2 dimensions drawn

- “dist N”: select the N elements that have the highest distance to the center of gravity

Add Label to Individuals: When checked, individuals drawn on the plot are labelled.

Output PDF: Creates a PDF containing graphs generated from the plot function in the *FactoMineR* package. These graphs show individual and/or variable plots using the first two dimensions extracted from the analysis. For quantitative-only datasets, a graph of the eigenvalues associated with the extracted principal components is shown in the last page. **For more information about these graphs, consult the documentation for the *FactoMineR* package**.

The tool will also display messages with the following results: (1) matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance explained by the first N dimensions; (2) matrices with all the results for the first 10 individuals (“cos2” = square cosine, “ctr” = contributions); (3) matrices with all the results for the quantitative variables (“cos2” = square cosine, “ctr” = contributions); and (4) matrices with all the results for the categorical variables (“cos2” = square cosine, “ctr” = contributions, “v.test” = criterion with a Normal distribution for over-represented categories, v.test > 0, and under-represented categories, v.test < 0);

Case study

The sample data for this tool describes spatial locations of 202 local households (Fig. 2) in the Wolong Nature Reserve, China, and a number of socio-economic variables recorded in 2005 with a targeted household survey including cropland size, household age, median income, educational level, and others [5]. **NOTE: The spatial coordinates of all household locations have been randomly fuzzed and any unique household ID code removed to maintain confidentiality of the interviewees**.

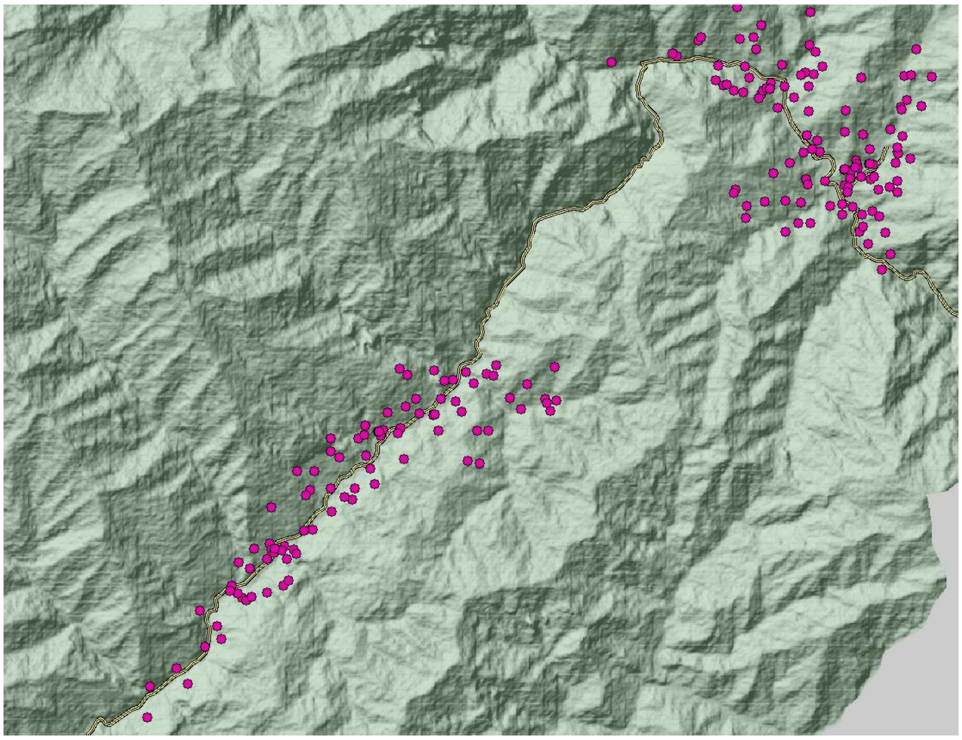
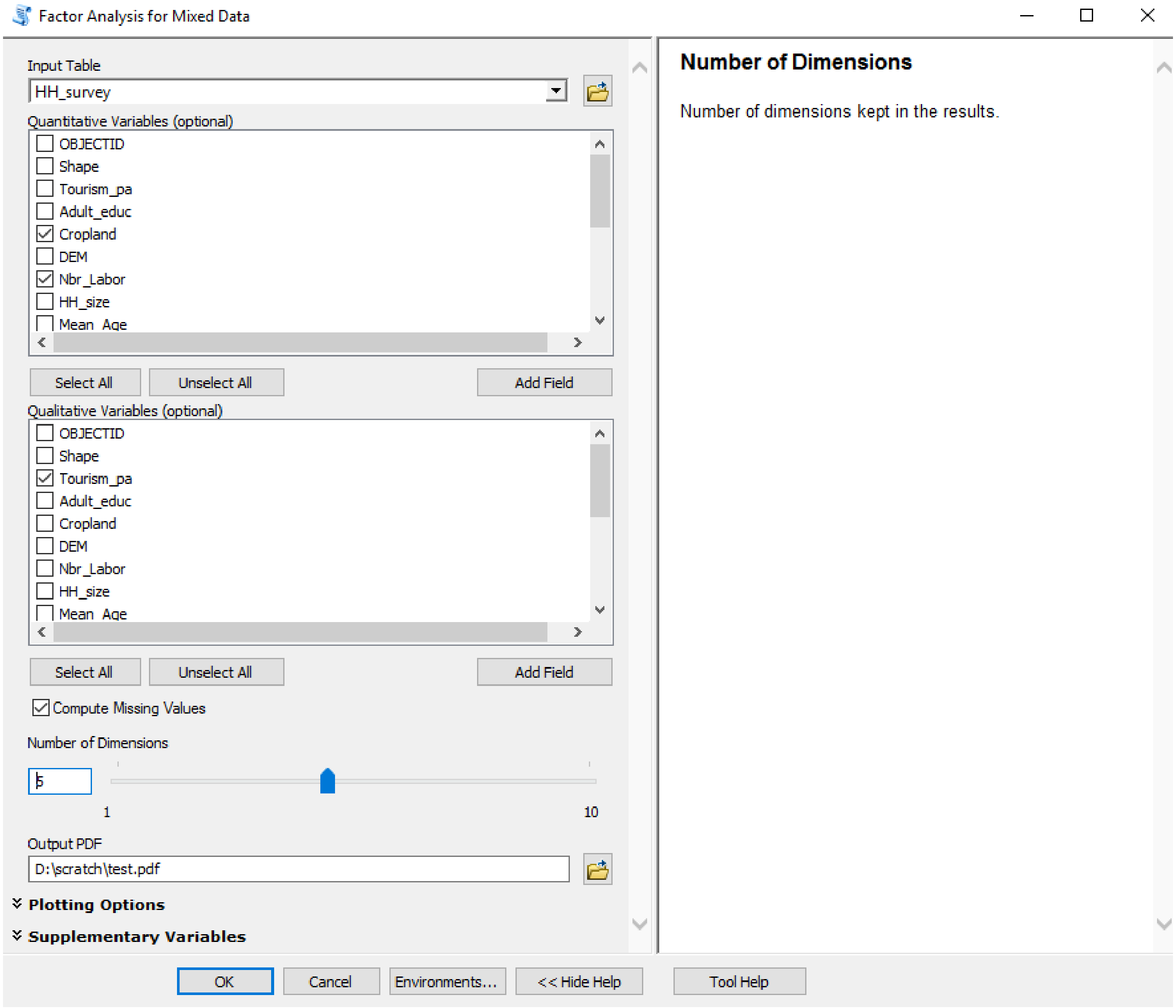


Figure 2. Locations of the 202 households surveyed in the Wolong Nature Reserve, China. Each point represents a household.

For the study, the following variables were recorded for each household:

- Direct tourism participation (0 = no; 1 yes) of any household members during the previous year. In the *HH_survey* feature class provided with this example, this field is named *Tourism_pa*.
- Maximum education level (in years) of non-student adults in the household. In the *HH_survey* feature class provided with this example, this field is named *Adult_educ*.
- The amount of cropland (in mu; 1 mu = 0.0067 ha) owned by the household for agricultural production. In the *HH_survey* feature class provided with this example, this field is named *Cropland*.
- The elevation (in meters) of the household derived from the ASTER global DEM dataset. In the *HH_survey* feature class provided with this example, this field is named *DEM*.
- The total number of laborers in the household. In the *HH_survey* feature class provided with this example, this field is named *Nbr_Labor*.
- The number of people in the household. In the *HH_survey* feature class provided with this example, this field is named *HH_size*.
- The average age of adults in the household. In the *HH_survey* feature class provided with this example, this field is named *Mean_Age*.
- Household member or immediate relative working in local government (0 = no; 1 = yes). In the *HH_survey* feature class provided with this example, this field is named *Govt_tie*.
- Household member or immediate relative being a village or group leader (0 = no; 1 = yes). In the *HH_survey* feature class provided with this example, this field is named *Lead_tie*.
- The distance (log-transformed, in meters) between the household and the main road. In the *HH_survey* feature class provided with this example, this field is named *Log_dist*.
- The total gross household income (log-transformed, in Chinese Yuan) in 1998. In the *HH_survey* feature class provided with this example, this field is named *Log_inc*.
- The township where the household is located (0 = Wolong township; 1 = Gengda township). In the *HH_survey* feature class provided with this example, this field is named *Township*.
- The presence of labor migrants in the household during 2003 (0 = no; 1 = yes). In the *HH_survey* feature class provided with this example, this field is named *Migr_labor*.
- The percentage of cropland enrolled by the household in the Grain-to-Green-Program (GTGP), subsidies paid by the Chinese government to stimulate conversion of cropland to forest land. In the *HH_survey* feature class provided with this example, this field is named *Gov_GTGP*.
- The percentage of cropland enrolled by the household in the Grain-to-Bamboo-Program (GTBP), subsidies paid by the Chinese government to convert cropland to bamboo land. In the *HH_survey* feature class provided with this example, this field is named *Gov_GTGB*.

You may be interested in investigating the relationships among a number of quantitative and qualitative variables recorded in this survey, with the goal of comprehensively describe socioeconomic characteristics associated with the interviewees. Open the tool and select the HH_survey feature class as *Input Table*. As quantitative variables, select *Gov_GTGP*, *Gov_GTGB*, *Cropland*, and *Nbr_Labor*. As qualitative variables, select *Tourism_pa*. Select a name and location on disk where to save an output pdf report file with plots of the mixed factor analysis. Leave all other tool options unaltered and click OK to run the tool (Fig. 3).



Input Table
HH_survey

Quantitative Variables (optional)
□ OBJECTID
□ Shape
□ Tourism_pa
□ Adult_educ
☑ Cropland
□ DEM
☑ Nbr_Labor
□ HH_size
□ Mean_Age

Qualitative Variables (optional)
□ OBJECTID
□ Shape
☑ Tourism_pa
□ Adult_educ
□ Cropland
□ DEM
□ Nbr_Labor
□ HH_size
□ Mean_Age

☑ Compute Missing Values

Number of Dimensions
5

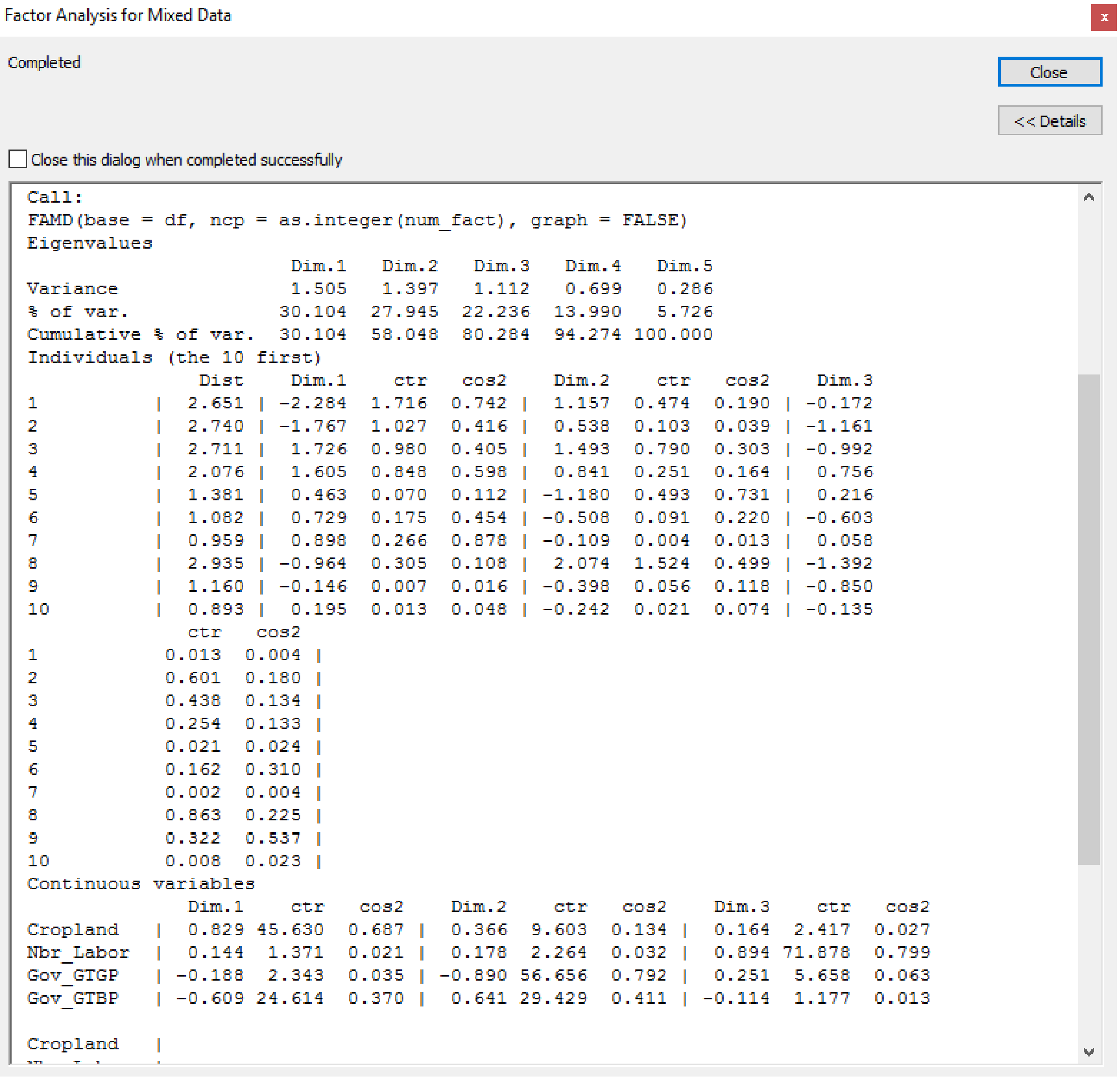
Output PDF
D:\scratch\test.pdf

Number of Dimensions

Number of dimensions kept in the results.

Figure 3. Variables and values selected in the tool interface.

If all goes smoothly, the tool identifies whether distinct groups of household units exist and if any of the variables selected are highly correlated with each other and the first two extracted dimension. If you open the result window to inspect the results (Menu > Geoprocessing > Results), under "Current Session", right-click the "Message" icon and select "View". A separate window should open showing you the entire process executed by the tool and all the estimated scores on the first N dimensions for all the selected variables, as shown in Figure 4.



Factor Analysis for Mixed Data

Completed

□ Close this dialog when completed successfully

```
Call:
FAMD(base = df, ncp = as.integer(num_fact), graph = FALSE)
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance                1.505   1.397   1.112   0.699   0.286
% of var.              30.104  27.945  22.236  13.990   5.726
Cumulative % of var.  30.104  58.048  80.284  94.274 100.000
Individuals (the 10 first)
               Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
1          |  2.651 | -2.284  1.716  0.742 |  1.157  0.474  0.190 | -0.172
2          |  2.740 | -1.767  1.027  0.416 |  0.538  0.103  0.039 | -1.161
3          |  2.711 |  1.726  0.980  0.405 |  1.493  0.790  0.303 | -0.992
4          |  2.076 |  1.605  0.848  0.598 |  0.841  0.251  0.164 |  0.756
5          |  1.381 |  0.463  0.070  0.112 | -1.180  0.493  0.731 |  0.216
6          |  1.082 |  0.729  0.175  0.454 | -0.508  0.091  0.220 | -0.603
7          |  0.959 |  0.898  0.266  0.878 | -0.109  0.004  0.013 |  0.058
8          |  2.935 | -0.964  0.305  0.108 |  2.074  1.524  0.499 | -1.392
9          |  1.160 | -0.146  0.007  0.016 | -0.398  0.056  0.118 | -0.850
10         |  0.893 |  0.195  0.013  0.048 | -0.242  0.021  0.074 | -0.135
              ctr   cos2
1            0.013  0.004 |
2            0.601  0.180 |
3            0.438  0.134 |
4            0.254  0.133 |
5            0.021  0.024 |
6            0.162  0.310 |
7            0.002  0.004 |
8            0.863  0.225 |
9            0.322  0.537 |
10           0.008  0.023 |
Continuous variables
              Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Cropland   |  0.829 45.630  0.687 |  0.366  9.603  0.134 |  0.164  2.417  0.027
Nbr_Labor  |  0.144  1.371  0.021 |  0.178  2.264  0.032 |  0.894 71.878  0.799
Gov_GTGP   | -0.188  2.343  0.035 | -0.890 56.656  0.792 |  0.251  5.658  0.063
Gov_GTBP   | -0.609 24.614  0.370 |  0.641 29.429  0.411 | -0.114  1.177  0.013

Cropland   |
```

Figure 4. Snapshot of the Messages as a result of the geoprocessing tool execution.

Within each extracted dimension, higher values indicate a higher contribution of an individual or variable to the composition of that dimension. If distinct groups of individual units (e.g. households) or variables exist, higher scores will show for different units/variables in separate dimensions. In most cases with a lot of individuals (e.g. > 200 households) it becomes hard to inspect these particular scores, thus it better to have a visual aid in the output plots from the analysis.

Individual factor map output (Fig. 5), using default options for the graphical parameters in the tool interface, shows a dispersed cloud of points. Therefore, there does not seem to be evidence of a distinct cluster of households based on the chosen socioeconomic variables. The 0-1 labelled squares in the plot show association between the categorical variable *Tourism_pa* and the first dimension (positive and negative, respectively). A small group of household units seems to be associated with the 0 value of this variable.

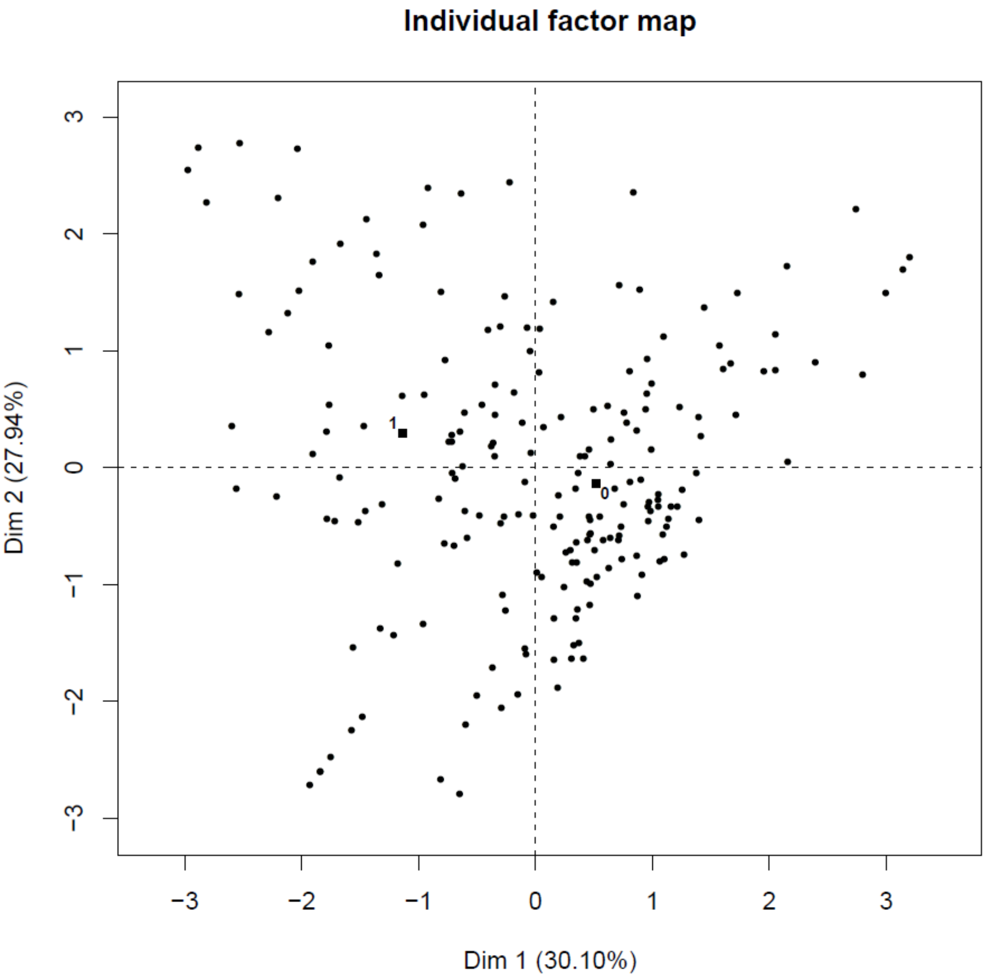


Figure 5. Individual factor map saved in the output pdf report.

The graph of the variables (Fig. 6) shows that *Gov_GTGP* is strongly associated with the second dimension, while *Cropland* and *Tourism_pa* have a medium-to-high association with the first dimension.

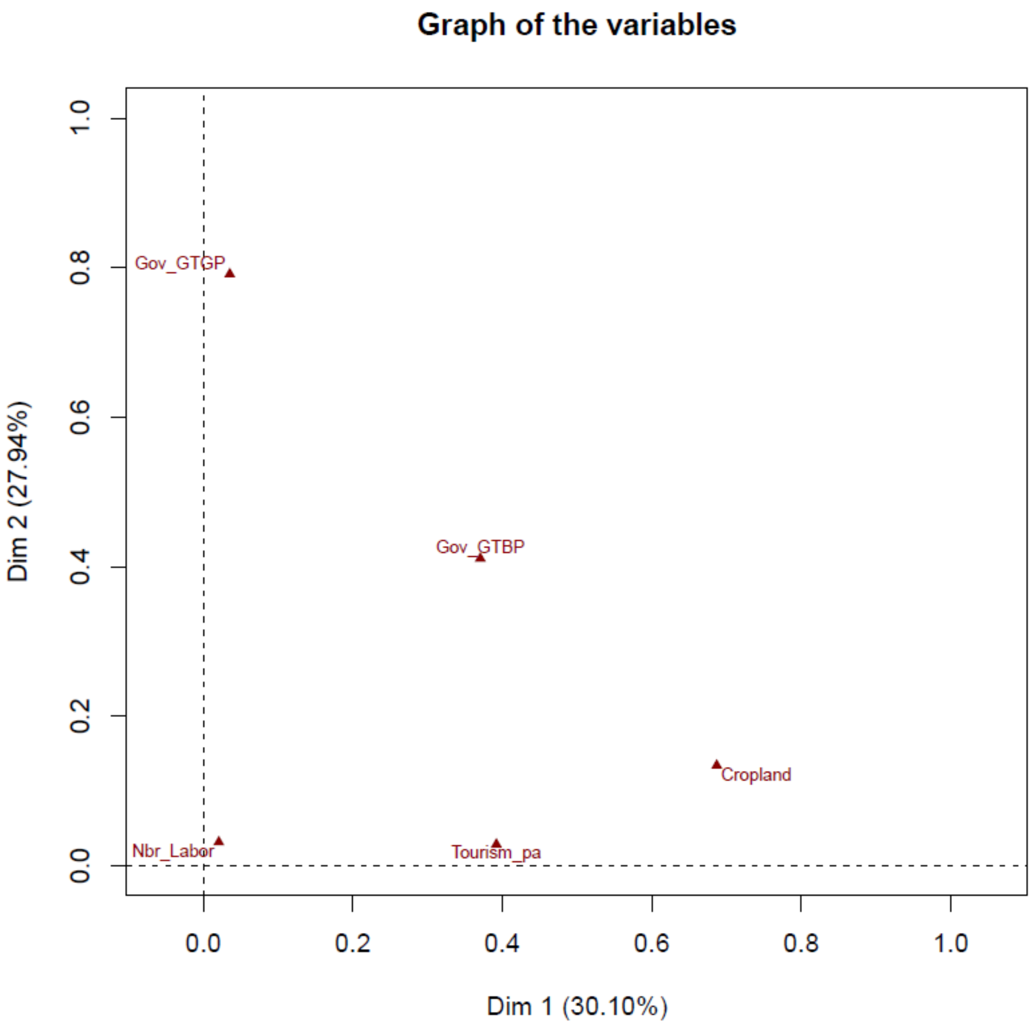


Figure 6. Graph of the variables saved in the output pdf report.

The graph of the quantitative variables on the unit circle (Fig. 7), tells which quantitative variables are mostly correlated with each other as well as with the first two dimensions. The *Gov_GTGP* has a negative correlation with the first dimension and a positive one with the second one, while the angle almost diagonal to the second quadrant on the top-left, indicates the variable is not well defined in either of those dimensions. The *Gov_GTGP* variable shows high negative association with the second dimension, while the *Cropland* shows a high correlation with the first dimension. Therefore, the first two dimensions are mostly dominated and defined by *Gov_GTGP* and *Cropland*. The plot indicates that the lower the percentage of cropland enrolled in the GTGP subsidy program (i.e. household receives less subsidies) the higher the amount of cropland owned by the household for agricultural production. Moreover, Tourism_pa helps defining the first dimension as well, with values of 1 (household members participated in tourism-related activities) negatively associated with it, i.e. higher participation in tourism corresponds to lower cropland amounts dedicated to agriculture. These conclusions are what you would expect, given that households enrolling less cropland in the GTGP program, thus receiving less subsidies, are likely to retain a higher amount of cropland for agriculture. At the same time, if the household income is partially made of tourism-related activities, it is more likely that they will have less land dedicated to agriculture and more willing to enroll it into the GTGP.

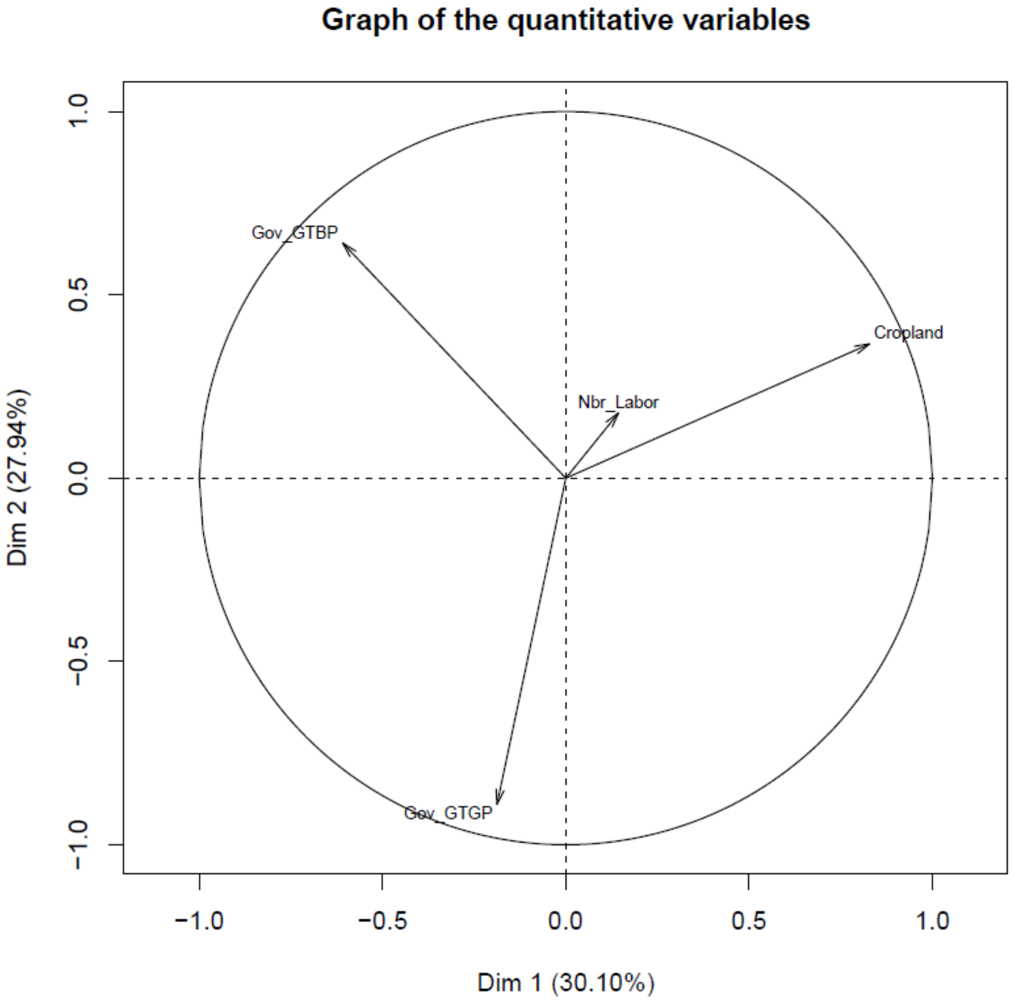


Figure 7. Graph of the quantitative variables saved in the output pdf report.

References and further reading

[1] Hair Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E. 2010. Multivariate Data Analysis, 7th Edition. Prentice Hall, 2009.

[2] Husson, F., Josse, J., and Le, S. 2016. *FactoMineR*. URL: http://factominer.free.fr/. Accessed 10/2016.

[3] Husson, F., Josse, J., Le, S., and Mazet, J. 2007. *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.04, URL: http://CRAN.R-project.org/package=FactoMineR.

[4] Husson, F., Josse, J. 2010. *missMDA*: *Handling missing values with/in multivariate data analysis (principal component methods)*. R package version 1.2, URL: http://www.agrocampus-ouest.fr/math/husson.

[5] Jianguo, L., McConnell, W., and Luo, J. 2013. *Wolong Household Study* [China]. ICPSR34365-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. http://doi.org/10.3886/ICPSR34365.v1