SesIndexCreatoR: An R Package for Socioeconomic Indices Computation and Visualization

Benoît Lalloué
EHESP & Lorraine University

Séverine Deguen
EHESP

Jean-Marie Monnez
Lorraine University

Cindy Padilla
EHESP

Wahida Kihal
EHESP

Denis Zmirou-Navier
EHESP & Lorraine University

Nolwenn Le Meur
EHESP

Abstract

This vignette corresponds to a submission to Journal of Statistical Software.

In order to study social inequalities, indices can be used to summarize the multiple dimensions of the socioeconomic status. As a part of the Equit'Area Project, a public health program focused on social and environmental health inequalities, a statistical procedure to create (neighborhood) socioeconomic indices was developed. This procedure uses successive principal components analyses to select variables and create the index. In order to simplify the application of the procedure for non specialists, the R package **SesIndexCreatoR** was created. It allows the creation of the index with all the possible options of the procedure, the classification of the resulting index in categories using several classical methods, the visualization of the results, and the generation of automatic reports.

Keywords: socioeconomic status, multidimensional index, principal component analysis, hierarchical classification, R.

1. Introduction

When studying social inequalities, it is generally interesting to take into account the socioeconomic status (SES) of an individual, a neighborhood or a region rather than consider only one socioeconomic variable such as educational level or income. However, socioeconomic status is a complex and multidimensional concept which encompasses many aspects such as employment, income, education, housing and social bonds. All of these aspects can themselves be represented by various variables. To synthesize and consider these different aspects, one solution is to create a SES index.

There are already many existing SES indices, especially at the neighborhood level (Jarman 1983; Morris and Carstairs 1991; Carstairs 1995; Salmond *et al.* 1998; Eibner and Sturm 2006; Messer *et al.* 2006; Bell *et al.* 2007; Fukuda *et al.* 2007; Pampalon *et al.* 2009). However, most of them use a small number of variables, combine variables with simple methods (such as Z-score) and/or select variables only from the litterature, which seems inappropriate for the purpose of the Equit'Area Project, a public health program focused on social and environmental health inequalities (`http://www.equitarea.org`), as detailed elsewhere (Lalloué *et al.* 2013). Thus, a new statistical procedure to create neighborhood socioeconomic indices was developed. Basically, this procedure does not create an index from a set of determined and precise variables, but aims to select, from a large data set, variables which will compose the SES index. It is based on several successive principal component analyses and the whole procedure is detailed in the aforementioned article. It has already been successfully used in several analyses aiming to study health or environmental inequalities (Padilla *et al.* 2013a,b).

Compared to other existing approaches to compute indices, our procedure is a slightly more complex to understand and apply, especially for non statisticians. Therefore we have implemented our model in a R (R Core Team 2013) package, named **SesIndexCreatoR**. The package is freely available on the website of the Equit'Area project and on CRAN. The purpose of this package is to give tools as simple as possible to perform the procedure while keeping the various possibilities it offered, like using different data mining methods, adding illustrative units, or performing only one step of the procedure. Moreover, once the index is created, users can display all the results of the different analyses both in text and graphical output, and generate a report summary.

In this paper we present and illustrate the use of the **SesIndexCreatoR** package for Lille agglomeration (a large French metropolitan area). For further examples we recommend reading the works by Padilla et al as mentioned above.

2. Material and methods

2.1. Data

The example data provided in the **SesIndexCreatoR** package concerns one large city in France, Lille (Nord Pas de Calais region, northern France), and some adjacent municipalities. The statistical unit is the sub-municipal French census block groups (called IRIS) defined by the National Institute of Statistics and Economic Studies (INSEE). These units have an average of 2,000 inhabitants and are constructed to be as homogeneous as possible in terms of socio-demographic characteristics and land use. Census block groups (BGs) are divided into three distinct categories: housing, economical activity and miscellaneous. Housing BGs are the most common, economical activity BGs include at least 1,000 employees and at least twice as many employees as residents, and miscellaneous BGs are specific wide areas sparsely populated (leisure parks, port areas, forest, etc.). As activity and miscellaneous BGs have some particular profiles due to the way they are defined, they are treated in the example as illustrative units (meaning that they are not part of the procedure but will have an index

value). For confidentiality and distribution reasons, the real BGs idenficators are replaced in the example data set with a simple number from 1 to 234 (which is the number of BGs of the area).

Socioeconomic data are taken from the 1999 national census (source: INSEE) and provide counts of population, households and residences at BG scale covering all the social, economic and demographic aspects. Median income at the BG scale is taken from a second database: the "Revenus fiscaux des ménages" database (source: INSEE-DGI). Using this raw data, 37 variables are defined at the BG scale based on the INSEE definitions. These variables are chosen to be representative of the theoretical concept of SES and in line with the variables most often used in the literature, or that could be considered as linked with the SES concept. All variables are related to family structure, household type, immigration status, employment, income, education and housing (more details are available in Table 1 and Table 2). Some of the variables are intentionally redundant and represent the same notion, in view to determine which best represents this notion (using the algorithm implemented in the proposed package). In our example, there are two such groups: 7 variables of unemployment and 3 variables of labor force. We also note there are an unexpectedly high number of missing values for median income but, willing to keep this variable in the analysis, we filled missing values with the average value of the adjacent BGs.

2.2. SES index creation

The SES index creation procedure is detailed in Lalloué *et al.* (2013). Basically, it follows three successive steps :

1. *Study of the redundant variables.* As already mentioned, several variables represent the same notion and we want to determine which best represented this notion. Therefore, one variable is selected for each group by applying principal component analysis (PCA) to each of the groups of redundant variables. The selected variable for each group is the one with the largest correlation with the first component of the PCA on the group.

2. *Selection of the variables.* A PCA or a multiple factor analysis (MFA) on the remaining variables (i.e., non redundant variables and variables selected in step 1) is used to select the variables with a contribution to the first component larger than the average one, i.e., variables that were best correlated with the first component. The choice of PCA or MFA depends on the willingness to give the same weight in the analysis to each domain (MFA) or not (PCA).

3. *Construction of the index.* A final PCA is carried out including the variables selected in step 2. Provided that the first component of this PCA could be interpreted as a "SES component", it is used to calculate the socioeconomic index as the reduced first component.

3. The SesIndexCreatoR package

The **SesIndexCreatoR** package depends on the **FactoMineR** (Husson *et al.* 2013; Lê *et al.* 2008) and **class** (Venables and Ripley 2002) packages. In particular, most of data analysis and visualization functions, such as principal component analysis or hierarchical clustering, used in this package come from **FactoMineR**. We thus refer the user to the **FactoMineR** package

Domain	*Variable name*	*Description*
BG type	`Type`	Census block group type (H: housing ; A: activity ; D: miscellaneous ; Z: one BG municipality) [c]
Family and Household	`UnderAge25`	People under the age of 25 in the total population
	`OverAge65`	People over the age of 65 in the total population
	`SingleParentFamilies`	Single-parent families in the total population
	`HouseholderAlone`	Householders living alone in the total population
Immigration	`ForeignPop`	Foreign people in the total population
Employment and income	`LabourForce`	People in the labor force in the total population [a]
	`MenLabourForce`	Men in the labor force in the total male population [a]
	`WomenLabourForce`	Women in the labor force in the total female population [a]
	`UnemployedTotal`	Unemployed people in the labor force [b]
	`UnemployedForeigners`	Unemployed foreigners in the labor force [b]
	`UnemployedAge1524`	Unemployed people in the 15-24 years old labor force [b]
	`UnemployedOverAge50`	Over 50 years old unemployed people in the labor force [b]
	`UnemployedMen`	Unemployed people in the male labor force [b]
	`UnemployedWomen`	Unemployed people in the female labor force [b]
	`UnemployedMore1Year`	People unemployed for more than 1 year in the labor force [b]
	`SelfEmployed`	Self-employed (independent workers, employers, etc.) in the labor force
	`InsecureJobs`	People with unstable jobs in the labor force (apprentices, trainees, temporary jobs, etc.)
	`SteadyJobs`	People with steady jobs in the labor force
	`MedianIncome`	Median Income per consumption unit (in euros per year) [c]

Table 1: Description of 37 socioeconomic variables available for the Lille agglomeration at the census block group scale, by domain. (Unless stated otherwise, variables are proportions expressed in % ; [a] Redundant group "labor force" ; [b] Redundant group "unemployment" ; [c] Not a proportion)

Domain	*Variable name*	*Description*
Education	AttendingSchool	People 6-15 years old attending school in the 6-15 years old population
	NoDiplomas	People with no diploma (and not studying) in the 15 years old and more population
	BasicGeneralQualifications	People with basic or intermediate general or vocation qualifications (and not studying) in the 15 years old and more population
	GeneralCertificates	People with general or vocational maturity certificates (and not studying) in the 15 years old and more population
	LowerTertiaryEducation	People with at least a lower tertiary education (and not studying) in the 15 years old and more population
	HigherEducationalDegree	People with a higher educational degree (and not studying) in the 15 years old and more population
	Students	Students in the 15 years old and more population
Housing	IndividualHouse	Individual houses in the main residences
	MultipleDwellingUnits	Multiple dwelling units in the main residences
	NonOwner	Non-owner-occupied in the main residences
	SubsidizedHousing	Subsidized housing in the main residences
	BuiltBefore1968	Main residences built before 1968
	Builtafter1990	Main residences built after 1990
	Less40m2	Main residences less than 40m^2
	Larger150m2	Main residences larger than 150m^2
	ParkingSpace	Main residences with a parking space (garage or other)
	WithoutCar	Households without a car
	TwoOrMoreCars	Households with 2 or more cars

Table 2: Description of the 37 socioeconomic variables available for the Lille agglomeration at the census block group scale, by domain (continued). (Unless stated otherwise, variables are proportions expressed in %).

and its manual for details on PCA and HC functions outputs. The sources and binaries of the package **SesIndexCreatoR** are available on the Equit'Area website or on CRAN and the installation is standard.

Because the package is also aimed to be used by R novice users, the example data are not included as R dataset but as a text file, in order to show in the package's manual how to import a file.

Function	Description
`ClassifHC`	Internal function: Classification with Hierarchical Clustering (HC)
`ClassifInt`	Internal function: Classification by intervals
`ClassifQuant`	Internal function: Classification by quantiles
`plot.SesClassif`	Plot the results of the classification of a socioeconomic index
`plot.SesIndex`	Plot the results of the construction of a socioeconomic index
`print.SesClassif`	Print the classification of a socioeconomic index results
`print.SesIndex`	Print the creation of a socioeconomic index results
`SelectVar`	Internal function: Selection of variables
`SesClassif`	Create categories from a socioeconomic index
`SesIndex`	Creation of a Socio-Economic Index
`SesReport`	Creation of a report for SesIndex and SesClassif functions
`SesStep1`	Internal function: perform the first step of the creation of the socioeconomic index

Table 3: Functions available in **SesIndexCreatoR** 1.0-1.

SesIndexCreatoR is composed of three main functions and several visualizing and internal functions (see Table 3):

- The `SesIndex` function creates a socioeconomic index such as defined in the Equit'Area project. It is possible to choose the starting set of variables, the potential redundant groups of variables, the potential supplementary units, the method of selection (PCA or MFA) and the step of the procedure to perform. Results include the final index and all the results of the intermediate steps.

- The `SesClassif` function creates socioeconomic categories, based on a socioeconomic index created by `SesIndex` function, with different technics such as hierarchical clustering, quantiles or equals subdivisions. Results include both a table with the original data set with class of each unit and the results of the classification technic (cut points, classes particularities, ...).

- The `SesReport` function creates a .html file with a report summarizing the results of the different steps of the creation of a socioeconomic index with the `SesIndex` function and, if any, the classification of the index using the `SesClassif` function. This function also allows to create a .csv file containing the original data set and the index and, if any, the classification.

4. Example

First, the socioeconomic data from the text file are imported in a data frame:

```
R> library("SesIndexCreatoR")
R> SesData <- read.table(
+    system.file("extdata","SesData.txt", package = "SesIndexCreatoR"),
+    header=TRUE,sep="\t", row.names=1)
```

The SesData.txt contains 37 socioeconomic variables and 1 type variable (giving the type of BG) for each BG of the Lille municipality and adjacent municipalities, as describe in Section 2.1. Then, the SesData dataframe has 234 rows representing the BGs and 38 columns representing the variables.

As the `SesIndex` function needs vectors or lists of variables' names as arguments, we then extract the different vectors and lists needed to call the function (with redundant groups). The first line of the following code chunk allows to extract the names of the variables to analyse as a vector. The remaining lines extract the names of the variables in the two groups of redundant variables (see Table 1) and create a list containing the two vectors of names for the groups of redundant variables.

```
R> varnames <- colnames(SesData)[2:ncol(SesData)]
R> group1 <- grep("+Unemployed", colnames(SesData), value=TRUE)
R> group2 <- grep("+LabourForce", colnames(SesData), value=TRUE)
R> groupvarnames <- list(group1, group2)
```

In order to consider activity and miscellaneous BGs as illustrative units, we extract the names of the corresponding rows (in our example, A is for "Activity" and D for "Miscellaneous" types of BGs) :

```
R> illus <- rownames(SesData[SesData[,"Type"] %in% c("A", "D"),])
```

It is "now" possible to create a socioeconomic index described in Materiel and methods using `SesIndex`. Here, we will create a socioeconomic index using all the 3 steps. Two groups of redundant variables are defined in *groupvarnames* and several BGs are set illustrative. By default, all the 3 steps are performed and step 2 uses a PCA.

```
R> index <- SesIndex(SesData, varnames=varnames, groupvarnames=groupvarnames,
+                    sup=illus)
```

```
R> plot(index, choice="ind", label="none")
```

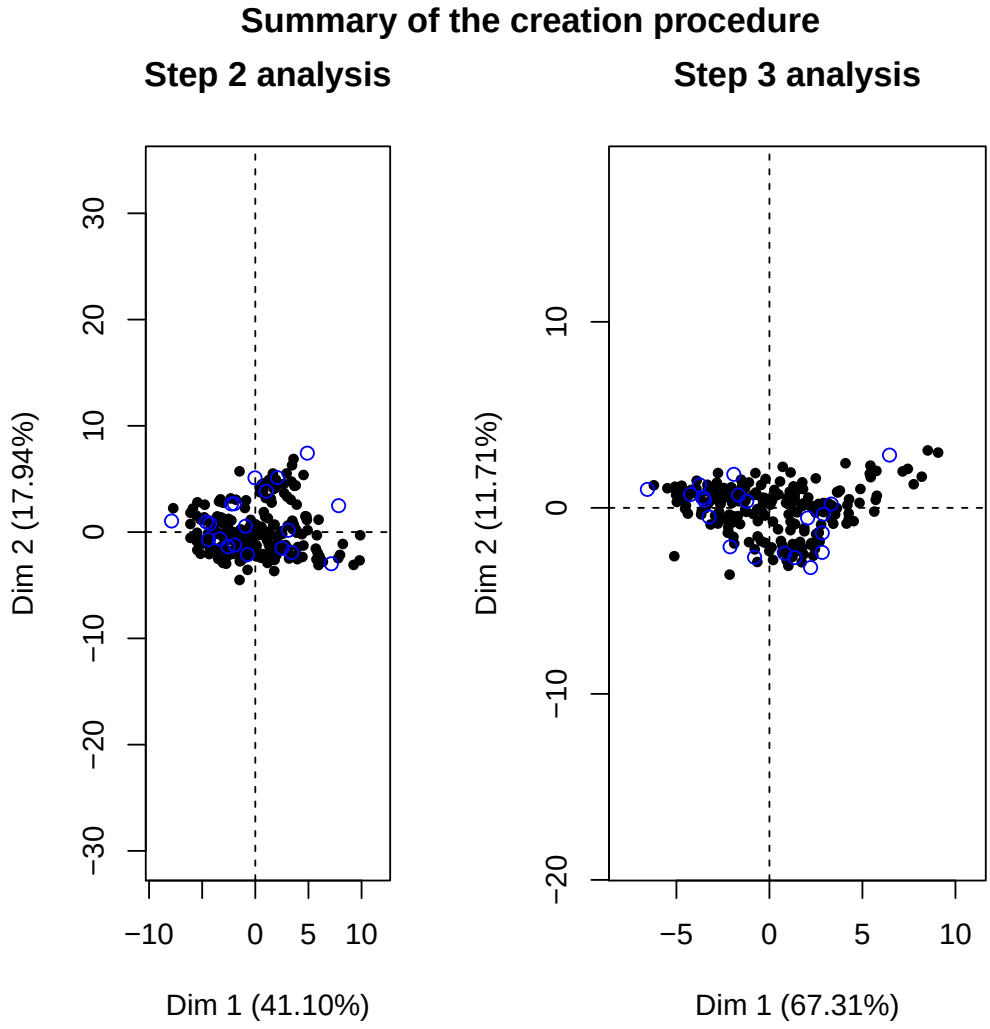


Figure 1: Synthetic view of the graphical outputs for individuals.

Once the index is created, we want to explore the results of the procedure. For instance, among the groups of redundant variables listed in Table 1 (Unemployment and Labor Force), the variables representing the best these groups and selected by our procedure are:

```
R> index$step1$selection
```

```
[1] "UnemployedTotal" "LabourForce"
```

Or, among the list of variables in Tables 1 and 2 (except the redundant variables dropped at step 1), the variables selected to compose the SES index for Lille agglomeration are:

```
R> index$step2$selection
```

```
 [1] "ForeignPop"            "UnemployedTotal"
 [3] "InsecureJobs"          "SteadyJobs"
 [5] "SingleParentFamilies"  "NoDiplomas"
 [7] "IndividualHouse"       "MultipleDwellingUnits"
 [9] "ParkingSpace"          "NonOwner"
[11] "WithoutCar"            "TwoOrMoreCars"
[13] "SubsidizedHousing"     "MedianIncome"
```

```
R> plot(index, choice="var", step=2)
```

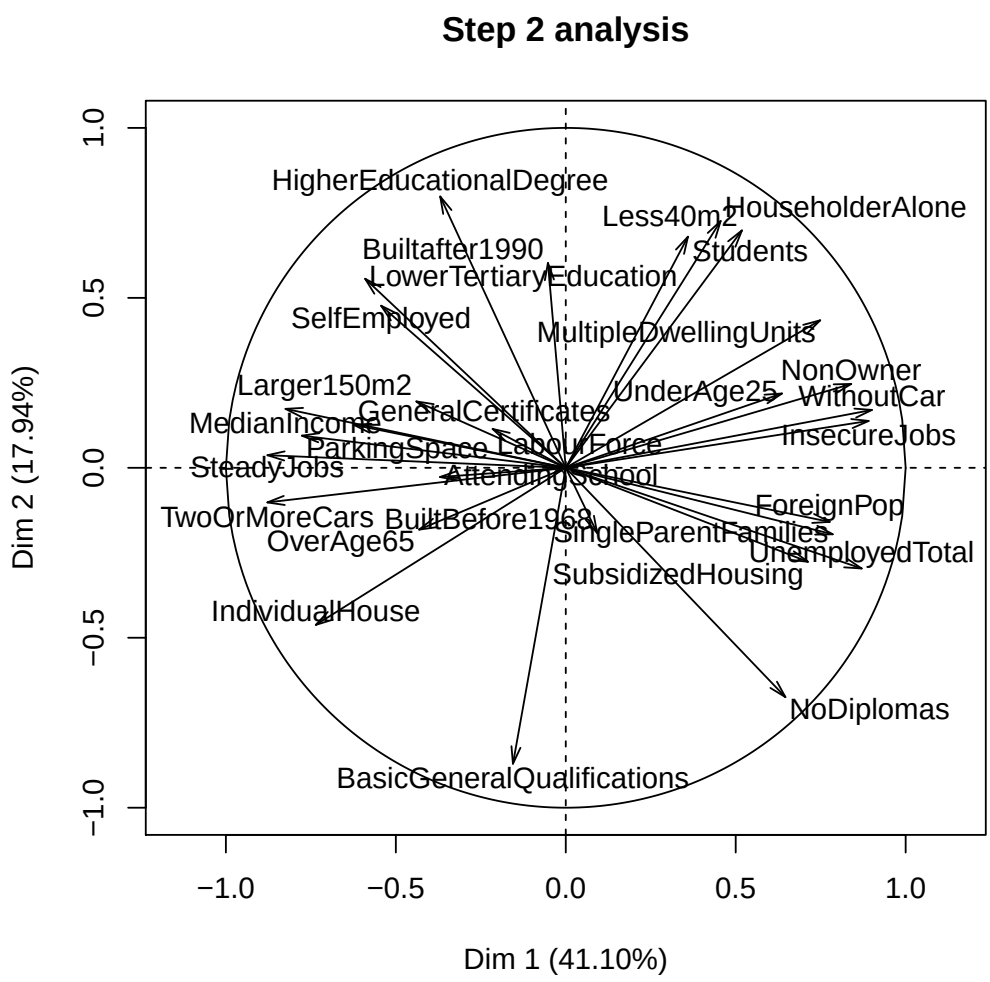


Figure 2: Correlation circle for the second step.

It is also possible to obtain detailed results of the data mining technics, like the correlation coefficients of the variables with the two first components of the second step analysis:

```
R> index$step2$analysis$var$coord[,c(1,2)]
```

```
                                  Dim.1       Dim.2
UnderAge25                   0.63630110  0.21862821
OverAge65                   -0.43163887 -0.18186596
ForeignPop                   0.77678724 -0.15910609
LabourForce                 -0.21536615  0.11435159
UnemployedTotal              0.87073008 -0.29624249
SelfEmployed                -0.54334383  0.47689134
InsecureJobs                 0.89130598  0.13786739
SteadyJobs                  -0.87777733  0.03703831
SingleParentFamilies         0.78556166 -0.19555464
NoDiplomas                   0.64635948 -0.67461860
HouseholderAlone             0.45573846  0.72622501
AttendingSchool             -0.37025379 -0.02709806
BasicGeneralQualifications -0.15542464 -0.87073704
GeneralCertificates         -0.62380039  0.12642215
LowerTertiaryEducation      -0.59037219  0.55626000
HigherEducationalDegree     -0.36942886  0.79823028
Students                     0.35951016  0.68000017
IndividualHouse             -0.73553777 -0.46219873
MultipleDwellingUnits        0.74861629  0.43400686
BuiltBefore1968              0.09240263 -0.19191144
Builtafter1990              -0.05250108  0.60279988
ParkingSpace                -0.77646906  0.09453152
NonOwner                     0.83998699  0.24679532
Less40m2                     0.51827452  0.69853014
Larger150m2                 -0.43976155  0.19532641
WithoutCar                   0.90096827  0.17029037
TwoOrMoreCars               -0.87800029 -0.10183450
SubsidizedHousing            0.71268195 -0.27645939
MedianIncome                -0.82471346  0.17342006
```

Or the proportion of variance explained by the four first components of the final step:

```
R> index$step3$analysis$eig[1:4,]
```

```
       eigenvalue percentage of variance
comp 1  9.4233115              67.309368
comp 2  1.6390996              11.707854
comp 3  1.0678887               7.627776
comp 4  0.5014364               3.581689
       cumulative percentage of variance
comp 1                           67.30937
comp 2                           79.01722
comp 3                           86.64500
comp 4                           90.22669
```

The above outputs are especially interesting to understand the procedure of variable selection. We can see in these results that the variables of total unemployment and total labor force were respectively selected from the groups of redundant unemployment variables and labor force variables. Then, for these two groups only these two variables were kept in the next steps. We can see in the selection from the step 2 that only variables with the highest correlations with the first component were selected. Here, 14 variables out of 29 were kept for the final step and the construction of the SES index. Eventually, the first component of the final step PCA performed on these 14 variables explained more than 67% of the total variance.

```
R> plot(index, choice="var", step=3)
```

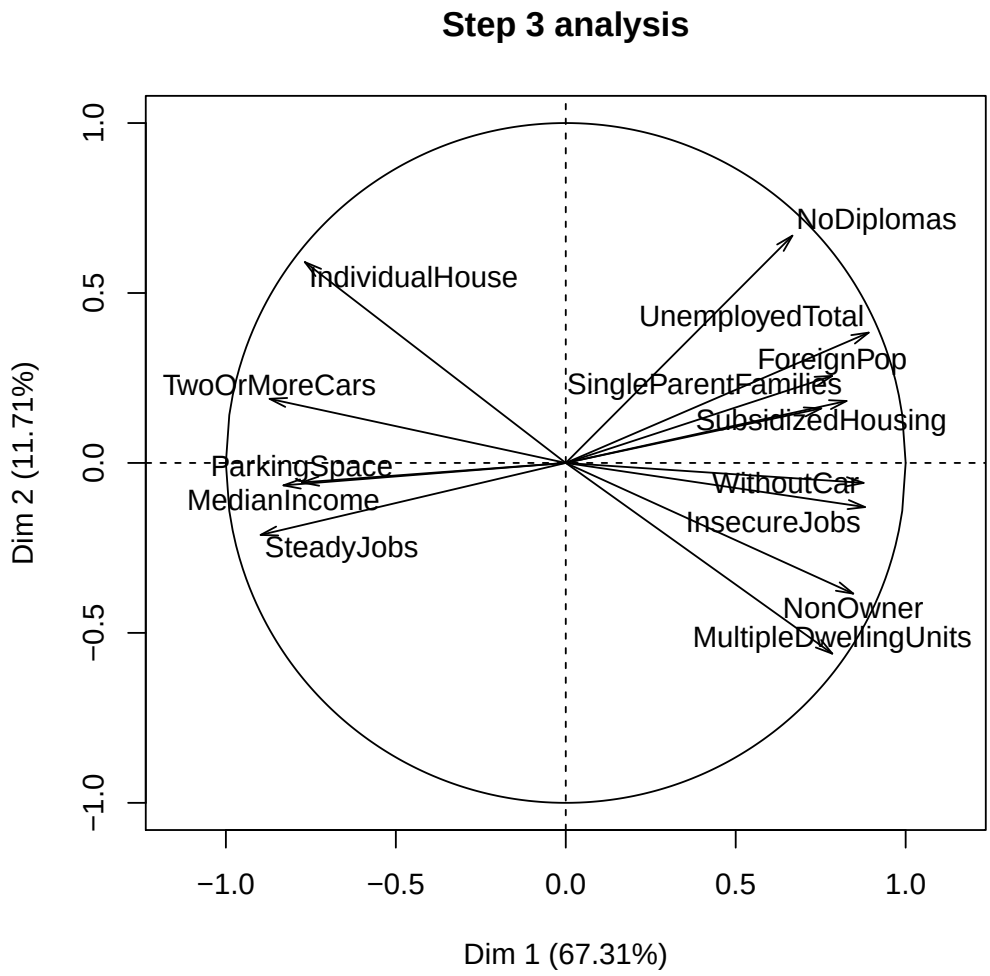


Figure 3: Correlation circle for the final step.

Some graphical outputs can be seen in Figure 1 to Figure 3. Figure 1 is a synthetic view of the projection of the BGs on the first principal components of the PCA performed in step 2

and step 3. Black dots represent active units whereas blue circles represent illustrative units (i.e., BG of the economical activity or miscellaneous types). Due to the number of units, BG labels are not displayed here but are activated by default. The step 3 part of the figure allows to see that BG are mainly along the first component and have not an extremely important variability along the second component.

Figure 2 gives the circle of correlations of the PCA performed in step 2. Most of the variables seem to have a good correlation with the first component, both positively and negatively, whereas as the correlations with the second component are mainly positive (except for two variables). A few variables (5) are not well represented on this plane and may have higher correlations with the third or fourth component. On this figure, a first opposition between "variables of deprivation", at the right, and "variables of favor", at the left, can be seen.

Finally, Figure 3 shows the circle of correlations of the PCA performed in step 3. The opposition between the "deprivation" and the "favor" variables is clear, with a high positive correlation between the first component and proportions of non-owner, unemployment, insecure jobs, person without diploma, subsidized housing, ... and a high negative correlation between the first component and proportion of steady jobs, individual houses, ... The first component of this PCA can then be interpreted as a SES component and be used as a SES index.

We now want to create categories from the socioeconomic index. We use a hierarchical clustering followed by a k-nearest neighbor (k-nn) algorithm. We decide to have an automatic number of classes (i.e., to cut the hierarchical clustering tree where the relative loss of inertia is the highest) :

```
R> categories <- SesClassif(index)
```

Others possibilities currently in the `SesClassif` function are to create classes with hierarchical clustering without k-nn consolidation, with quantiles or with equal range of values.

We can summarize some characteristics of the different categories using simple functions. For instance, it is possible to compare variables average values in eahc category and the overall mean :

```
R> for (i in 1:3) {
+    print(paste("Category",i))
+    print(round(categories$analysis$desc.var$quanti[[i]][,c(2,3,6)],2))
+  }
```

```
[1] "Category 1"
                    Mean in category Overall mean p.value
TwoOrMoreCars                   0.33         0.21       0
IndividualHouse                 0.71         0.45       0
SteadyJobs                      0.73         0.65       0
ParkingSpace                    0.60         0.43       0
MedianIncome                27529.06     21986.21       0
NoDiplomas                      0.11         0.16       0
ForeignPop                      0.02         0.05       0
```

```
SubsidizedHousing                0.08         0.26       0
UnemployedTotal                  0.10         0.16       0
SingleParentFamilies             0.11         0.17       0
MultipleDwellingUnits            0.25         0.52       0
WithoutCar                       0.16         0.29       0
InsecureJobs                     0.09         0.13       0
NonOwner                         0.35         0.58       0
[1] "Category 2"
                     Mean in category Overall mean p.value
MultipleDwellingUnits             0.66         0.52    0.00
NonOwner                          0.69         0.58    0.00
WithoutCar                        0.35         0.29    0.00
InsecureJobs                      0.14         0.13    0.00
SingleParentFamilies              0.18         0.17    0.02
SteadyJobs                        0.63         0.65    0.01
MedianIncome                  19693.52     21986.21    0.00
ParkingSpace                      0.35         0.43    0.00
IndividualHouse                   0.30         0.45    0.00
TwoOrMoreCars                     0.14         0.21    0.00
[1] "Category 3"
                     Mean in category Overall mean p.value
UnemployedTotal                   0.33         0.16       0
ForeignPop                        0.14         0.05       0
SingleParentFamilies              0.28         0.17       0
SubsidizedHousing                 0.74         0.26       0
NoDiplomas                        0.30         0.16       0
InsecureJobs                      0.18         0.13       0
WithoutCar                        0.47         0.29       0
NonOwner                          0.90         0.58       0
MultipleDwellingUnits             0.85         0.52       0
IndividualHouse                   0.13         0.45       0
TwoOrMoreCars                     0.08         0.21       0
ParkingSpace                      0.19         0.43       0
MedianIncome                  12624.39     21986.21       0
SteadyJobs                        0.46         0.65       0
NULL
```

```
R> plot(categories$analysis, choice="map", label="none", draw.tree=F)
```

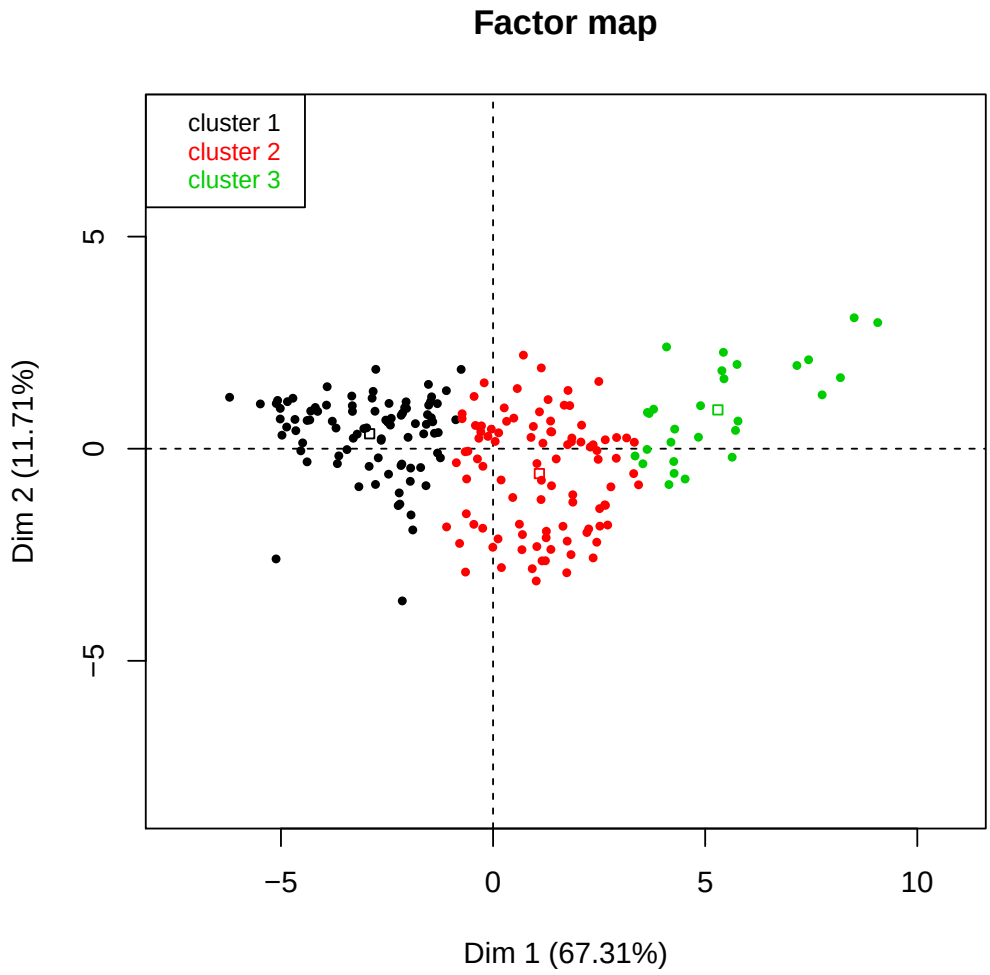


Figure 4: Plot of the individuals by categories.

We can see that the optimal number of categories (according to the inertia criterion) was 3. The description of these categories showed that they are organised by decreasing socioeconomic status. Indeed, category 1 has higher average values of variables like median income or proportion of steady jobs, and lower average values of proportion of unemployed people or proportion of subsidized housing; whereas category 3 has lower values of median income and higher values of unemployment. Figure 4 shows the projection of these categories on the two first axes of the final PCA (note that it is also possible to use directly `plot(categories)` to have both the dendrogram and the projection of the units).

Eventually, we want to export the detailed results of all the three steps of creation of the SES index and of the classification. We also want to export a data file containing the index

and the categories. To do so, the `SesReport` function is used (since this function create several files, an exemple of the resulting report is available in appendix). By default, files are created in the current working directory with basename "SesReport" (which can be change as arguments of the `SesReport` function).

```
R> SesReport(categories)
```

5. Conclusion

In this article we presented the **SesIndexCreatoR** package, designed to easily create socio-economic indices with a reproductible statistical procedure. One originality of this procedure compared to other existing indices lays in selecting the final variables for the index by usage of data mining techniques rather than only information gleaned from a literature review, allowing to discard part of the subjectivity that may influence the choice of the variables. This data driven approach allows the data "speak by themself".

The **SesIndexCreatoR** package allows to apply this procedure in a versatile way, by specifying which steps of the procedure should be runned (for instance only step 2 if the aim is to compare selection of variables between metropolitan areas without create indices, or only step 3 if one wants all the introduced variables to be in the index), adding illustrative units or selecting the method used. Once the index created, several tools are available to visualize, synthetize, explore and export the results in a convenient way for further utilization.

We project to extend the package in the future and among other improvements we foresee to implement others methods of classification, to add more tools to help the interpretation of the results, or to allow other ways of visualization (such as mapping). However, these improvement will be made according to users' returns and needs.

References

Bell N, Schuurman N, Hayes MV (2007). "Using GIS-Based Methods of Multicriteria Analysis to Construct Socio-Economic Deprivation Indices." *International Journal of Health Geographics*, **6**, 17.

Carstairs V (1995). "Deprivation Indices: Their Interpretation and Use in Relation to Health." *Journal of Epidemiology and Community Health*, **49 Suppl 2**, S3–8.

Eibner C, Sturm R (2006). "US-Based Indices of Area-Level Deprivation: Results from HealthCare for Communities." *Social Science & Medicine (1982)*, **62**(2), 348–359.

Fukuda Y, Nakamura K, Takano T (2007). "Higher Mortality in Areas of Lower Socioeconomic Position Measured by a Single Index of Deprivation in Japan." *Public Health*, **121**(3), 163–173.

Husson F, Josse J, Le S, Mazet J (2013). ***FactoMineR**: Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.25, URL http://CRAN.R-project.org/package=FactoMineR.

Jarman B (1983). "Identification of Underprivileged Areas." *British Medical Journal (Clinical research ed.)*, **286**(6379), 1705–1709.

Lalloué B, Monnez JM, Padilla C, Kihal W, Le Meur N, Zmirou-Navier D, Deguen S (2013). "A Statistical Procedure to Create a Neighborhood Socioeconomic Index for Health Inequalities Analysis." *International Journal for Equity in Health*, **12**(1), 21.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R Package for Multivariate Analysis." *Journal of statistical software*, **25**(1), 1–18.

Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, Elo I, Burke JG, O'Campo P (2006). "The Development of a Standardized Neighborhood Deprivation Index." *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **83**(6), 1041–1062.

Morris R, Carstairs V (1991). "Which Deprivation? A Comparison of Selected Deprivation Indexes." *Journal of Public Health Medicine*, **13**(4), 318–326.

Padilla C, Deguen S, Lalloué B, Zmirou-Navier D, Viera V (2013a). "Cluster Analysis of Social and Environment Inequalities of Infant Mortality. A Spatial Study in Small Areas Revealed by Local Disease Mapping in France." *Science of the Total Environment*, **454-455**, 433–441.

Padilla C, Lalloué B, Pies C, Lucas E, Zmirou-Navier D, Deguen S (2013b). "An Ecological Study to Identify Census Blocks Supporting a Higher Burden of Disease: Infant Mortality in the Lille Metropolitan Area, France." *Maternal and Child Health Journal*, pp. 1–9.

Pampalon R, Hamel D, Gamache P, Raymond G (2009). "A Deprivation Index for Health Planning in Canada." *Chronic Diseases in Canada*, **29**(4), 178–191.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Salmond C, Crampton P, Sutton F (1998). "NZDep91: A New Zealand Index of Deprivation." *Australian and New Zealand Journal of Public Health*, **22**(7), 835–837.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

A. Appendix: Automatic report generated with SesReport

Creation and classification of a socioeconomic index

This report was automatically generated by the SesReport function of the R package SesIndexCreatoR .

Index

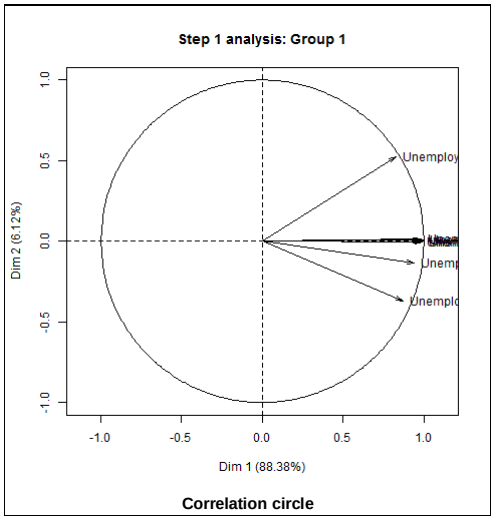
Step 1: Reduction of the redundant groups

A Principal Component Analysis is performed for each redundant group and only the variable with the highest correlation with the first component is kept.
The results of the analysis of each group are :

- **Group 1**

The following results are obtained for the first component of the PCA :

Variable	Coord	Cos2	Contrib
UnemployedForeigners	0.87	0.75	12.2
UnemployedTotal	0.99	0.98	15.92
UnemployedAge1524	0.83	0.68	11.02
UnemployedOverAge50	0.94	0.88	14.21
UnemployedMen	0.98	0.96	15.47
UnemployedWomen	0.98	0.95	15.43
UnemployedMore1Year	0.99	0.97	15.75



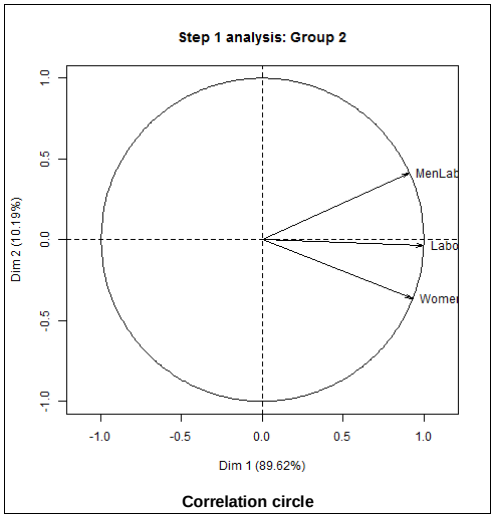
The variable selected for group 1 is "*UnemployedTotal*"

- **Group 2**

The following results are obtained for the first component of the PCA :

Variable	Coord	Cos2	Contrib
LabourForce	1	0.99	37.01

Variable	Coord	Cos2	Contrib
MenLabourForce	0.91	0.83	30.79
WomenLabourForce	0.93	0.87	32.21



The variable selected for group 2 is "*LabourForce*"

Step 2: Selection of the variables

A Principal Component Analysis is performed on the variables kept after Step 1. Only variables with a correlation higher than the average correlation will be kept after Step 2.
The results of the analysis are :

10 first eigenvalues

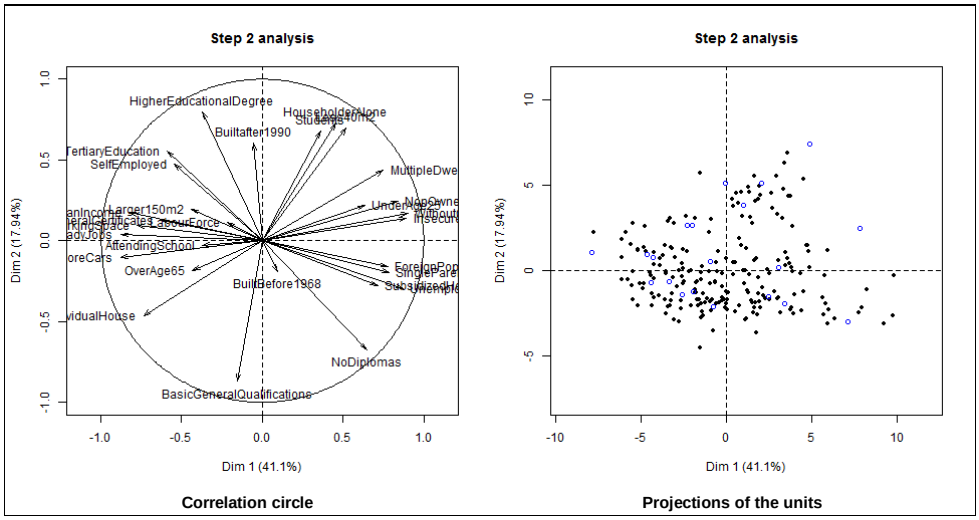
	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 1	11.92	41.1	41.1
comp 2	5.2	17.94	59.05
comp 3	2.67	9.2	68.25
comp 4	2.24	7.73	75.97
comp 5	1.34	4.63	80.6
comp 6	1.02	3.53	84.13
comp 7	0.97	3.36	87.49
comp 8	0.64	2.19	89.68
comp 9	0.61	2.11	91.79
comp 10	0.46	1.6	93.39

Summary of the two first components

Variable	Coord 1	Cos2 1	Contrib 1	Coord 2	Cos2 2	Contrib 2
UnderAge25	0.64	0.4	3.4	0.22	0.05	0.92
OverAge65	-0.43	0.19	1.56	-0.18	0.03	0.64
ForeignPop	0.78	0.6	5.06	-0.16	0.03	0.49
LabourForce	-0.22	0.05	0.39	0.11	0.01	0.25
UnemployedTotal	0.87	0.76	6.36	-0.3	0.09	1.69
SelfEmployed	-0.54	0.3	2.48	0.48	0.23	4.37
InsecureJobs	0.89	0.79	6.66	0.14	0.02	0.37
SteadyJobs	-0.88	0.77	6.46	0.04	0	0.03
SingleParentFamilies	0.79	0.62	5.18	-0.2	0.04	0.73
NoDiplomas	0.65	0.42	3.5	-0.67	0.46	8.75

Variable	Coord 1	Cos2 1	Contrib 1	Coord 2	Cos2 2	Contrib 2
HouseholderAlone	0.46	0.21	1.74	0.73	0.53	10.14
AttendingSchool	-0.37	0.14	1.15	-0.03	0	0.01
BasicGeneralQualifications	-0.16	0.02	0.2	-0.87	0.76	14.57
GeneralCertificates	-0.62	0.39	3.26	0.13	0.02	0.31
LowerTertiaryEducation	-0.59	0.35	2.92	0.56	0.31	5.95
HigherEducationalDegree	-0.37	0.14	1.14	0.8	0.64	12.25
Students	0.36	0.13	1.08	0.68	0.46	8.89
IndividualHouse	-0.74	0.54	4.54	-0.46	0.21	4.11
MultipleDwellingUnits	0.75	0.56	4.7	0.43	0.19	3.62
BuiltBefore1968	0.09	0.01	0.07	-0.19	0.04	0.71
Builtafter1990	-0.05	0	0.02	0.6	0.36	6.98
ParkingSpace	-0.78	0.6	5.06	0.09	0.01	0.17
NonOwner	0.84	0.71	5.92	0.25	0.06	1.17
Less40m2	0.52	0.27	2.25	0.7	0.49	9.38
Larger150m2	-0.44	0.19	1.62	0.2	0.04	0.73
WithoutCar	0.9	0.81	6.81	0.17	0.03	0.56
TwoOrMoreCars	-0.88	0.77	6.47	-0.1	0.01	0.2
SubsidizedHousing	0.71	0.51	4.26	-0.28	0.08	1.47
MedianIncome	-0.82	0.68	5.71	0.17	0.03	0.58

In red: variables selected in the second step.



The variables selected in the step 2 are :
ForeignPop, UnemployedTotal, InsecureJobs, SteadyJobs, SingleParentFamilies, NoDiplomas, IndividualHouse, MultipleDwellingUnits, ParkingSpace, NonOwner, WithoutCar, TwoOrMoreCars, SubsidizedHousing, MedianIncome

---

Step 3: Creation of the index

The final step of the creation perform a Principal Component Analysis on the variables selected in Step 2. The socio-economic index is therefore the first component of the final PCA, *if this component can be interpreted as a socio-economic component.*
The results of the final analysis are :

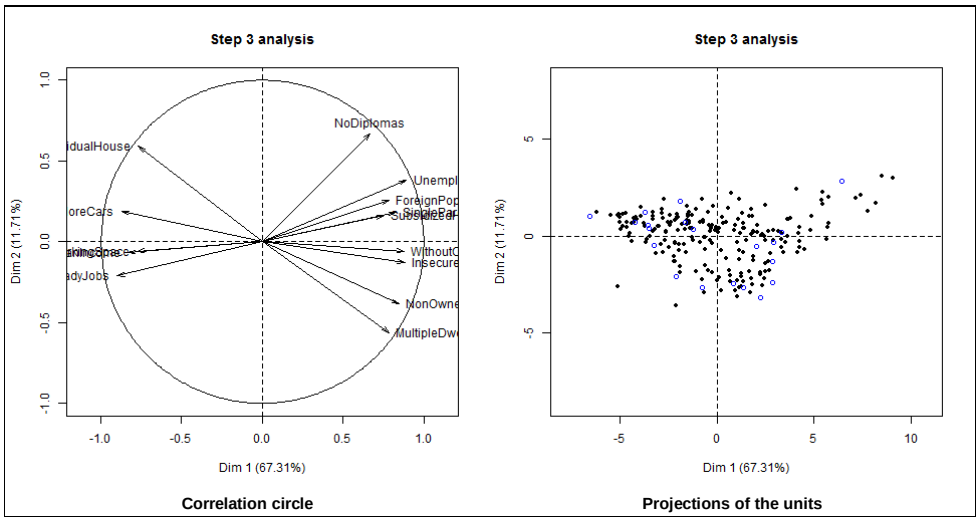
10 first eigenvalues

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 1	9.42	67.31	67.31
comp 2	1.64	11.71	79.02
comp 3	1.07	7.63	86.64

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 4	0.5	3.58	90.23
comp 5	0.37	2.63	92.86
comp 6	0.26	1.89	94.74
comp 7	0.18	1.3	96.04
comp 8	0.16	1.12	97.16
comp 9	0.13	0.9	98.05
comp 10	0.1	0.74	98.8

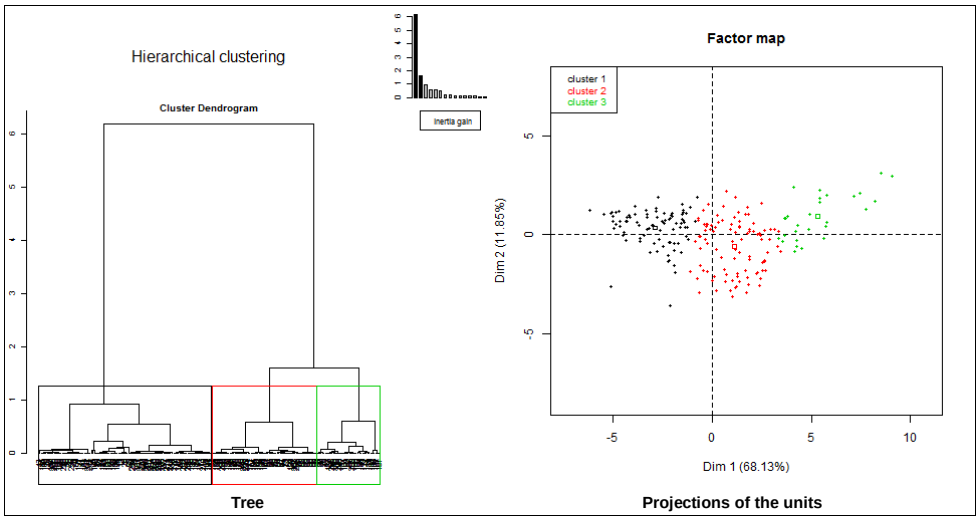
Summary of the two first components

Variable	Coord 1	Cos2 1	Contrib 1	Coord 2	Cos2 2	Contrib 2
ForeignPop	0.78	0.61	6.52	0.26	0.07	4
UnemployedTotal	0.89	0.79	8.43	0.38	0.15	8.97
InsecureJobs	0.88	0.78	8.23	-0.13	0.02	1.03
SteadyJobs	-0.9	0.81	8.55	-0.21	0.04	2.74
SingleParentFamilies	0.83	0.68	7.24	0.18	0.03	2.03
NoDiplomas	0.67	0.44	4.72	0.67	0.45	27.28
IndividualHouse	-0.77	0.59	6.26	0.59	0.35	21.31
MultipleDwellingUnits	0.78	0.62	6.53	-0.56	0.31	19.21
ParkingSpace	-0.78	0.6	6.4	-0.06	0	0.2
NonOwner	0.85	0.72	7.59	-0.38	0.15	9.02
WithoutCar	0.88	0.77	8.15	-0.06	0	0.21
TwoOrMoreCars	-0.87	0.76	8.06	0.19	0.04	2.17
SubsidizedHousing	0.75	0.57	6	0.16	0.03	1.57
MedianIncome	-0.83	0.69	7.32	-0.07	0	0.27



Classification

After the creation of the socioeconomic index, a classification is performed using HC+knn. 3 classes are created using this technic.



The following tables contain descriptions of each classe by the variables selected in the socioeconomic index (only variables with a significant difference between the class and the whole sample are shown):

Description by the variables of the class 1

Variable	Mean in category	Overall mean	Sd in category	Overall Sd
TwoOrMoreCars	0.34	0.21	0.11	0.13
IndividualHouse	0.72	0.45	0.22	0.31
SteadyJobs	0.73	0.65	0.05	0.11
ParkingSpace	0.6	0.43	0.19	0.22
MedianIncome	27590.41	22080.58	6806.45	7238.27
NoDiplomas	0.11	0.15	0.05	0.09
ForeignPop	0.02	0.05	0.02	0.05
SubsidizedHousing	0.07	0.25	0.1	0.3
UnemployedTotal	0.1	0.16	0.03	0.09
SingleParentFamilies	0.11	0.16	0.05	0.07
InsecureJobs	0.09	0.13	0.03	0.05
MultipleDwellingUnits	0.24	0.52	0.21	0.32
WithoutCar	0.16	0.29	0.07	0.15
NonOwner	0.35	0.58	0.18	0.26

Description by the variables of the class 2

Variable	Mean in category	Overall mean	Sd in category	Overall Sd
MultipleDwellingUnits	0.67	0.52	0.22	0.32
NonOwner	0.7	0.58	0.14	0.26
WithoutCar	0.36	0.29	0.1	0.15
InsecureJobs	0.15	0.13	0.04	0.05
SingleParentFamilies	0.18	0.16	0.05	0.07
SteadyJobs	0.62	0.65	0.06	0.11
MedianIncome	19785.34	22080.58	3824.47	7238.27
ParkingSpace	0.34	0.43	0.15	0.22
IndividualHouse	0.29	0.45	0.21	0.31
TwoOrMoreCars	0.14	0.21	0.05	0.13

Description by the variables of the class 3

Variable	Mean in category	Overall mean	Sd in category	Overall Sd
UnemployedTotal	0.33	0.16	0.09	0.09
ForeignPop	0.15	0.05	0.07	0.05
SubsidizedHousing	0.76	0.25	0.29	0.3

Variable	Mean in category	Overall mean	Sd in category	Overall Sd
SingleParentFamilies	0.28	0.16	0.04	0.07
NoDiplomas	0.3	0.15	0.08	0.09
NonOwner	0.9	0.58	0.09	0.26
WithoutCar	0.48	0.29	0.11	0.15
InsecureJobs	0.18	0.13	0.02	0.05
MultipleDwellingUnits	0.85	0.52	0.13	0.32
IndividualHouse	0.14	0.45	0.12	0.31
TwoOrMoreCars	0.08	0.21	0.03	0.13
ParkingSpace	0.19	0.43	0.1	0.22
MedianIncome	12566.52	22080.58	2041.17	7238.27
SteadyJobs	0.46	0.65	0.09	0.11

Description by the variables of the class NA

Variable	Mean in category	Overall mean	Sd in category	Overall Sd
WithoutCar	0.87	0.29	NA	0.15
MultipleDwellingUnits	1	0.52	NA	0.32
NonOwner	0.91	0.58	NA	0.26
NoDiplomas	0.04	0.15	NA	0.09
SteadyJobs	0.5	0.65	NA	0.11
IndividualHouse	0	0.45	NA	0.31
TwoOrMoreCars	0	0.21	NA	0.13
UnemployedTotal	0	0.16	NA	0.09
ParkingSpace	0	0.43	NA	0.22

Affiliation:

Benoît Lalloué
EHESP Rennes, Sorbonne Paris Cité, France & Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, France & IECL, Institut Elie Cartan Nancy, CNRS : UMR 7502, Lorraine University, INRIA BIGS, France
EHESP - Département Epidémiologie, Biostatistiques and Sciences de l'Information
Avenue du Professeur Léon Bernard
35043 Rennes cedex, France
E-mail: benoit.lalloue@ehesp.fr

Séverine Deguen
EHESP Rennes, Sorbonne Paris Cité, France & Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, France
E-mail: severine.deguen@ehesp.fr

Jean-Marie Monnez
IECL, Institut Elie Cartan Nancy, CNRS : UMR 7502, Lorraine University, INRIA BIGS, France
E-mail: jean-marie.monnez@univ-lorraine.fr

Cindy Padilla
EHESP Rennes, Sorbonne Paris Cité, France & Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, France
E-mail: cindy.padilla@ehesp.fr

Wahida Kihal
EHESP Rennes, Sorbonne Paris Cité, France
E-mail: wahida.kihal@ehesp.fr

Denis Zmirou-Navier
EHESP Rennes, Sorbonne Paris Cité, France & Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, France & Lorraine University, Medical School, France
E-mail: denis.zmirou@ehesp.fr

Nolwenn Le Meur
EHESP Rennes, Sorbonne Paris Cité, France & UMR936 INSERM, Université de Rennes 1, France
E-mail: nolwenn.lemeur-rouillard@ehesp.fr