Potec Data Set Analysis

MVA project

Diego Garcia-Olano, Elsa Mullor

27/06/2014

Content :

I – Problem Dataset and Description 3
II- Pre-processing..3
III- MCA 5
IV- Clustering 12
V – Prediction 15
VI – Conclusion 19
VII – Appendix 20
1. Catdes results 20
2. Prediction errors with different processing of outliers 23
2. R code 24

I – Problem Dataset and Description

For our problem we decided on the POTEC dataset, which is based on the Adult dataset that can be found at https://archive.ics.uci.edu/ml/datasets/Adult in the UCI Machine Learning Repository. Also known as the "Census Income" data set, the data set contains 32561 individuals information along 15 variables taken from the 1994 US Census data.

The task at hand then is to predict whether an individual's income exceeds $50,000 dollars per year. The binary target variable "target" contains values of either "<=50K" to denote the individual makes less than or equal to 50,000 dollars a year or ">50K" denoting they make more than that amount. The target is fairly imbalanced as only 24% of the population makes more than 50 thousand dollars a year.

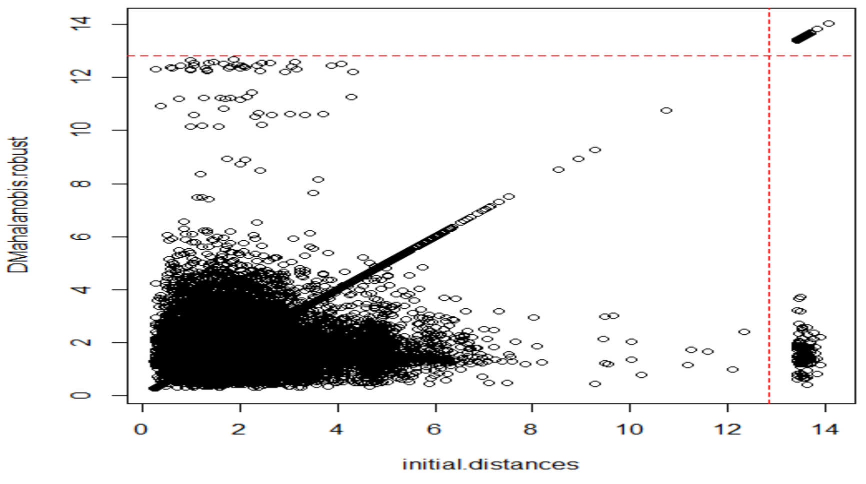
The dataset variables consists of the following variables and values for each:

> workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
> fnlwgt: continuous (a weight originally set by initial data handlers)
> education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc- acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
> education-num: continuous (number of years of schooling)
> marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, etc.
> occupation: Tech-support, Craft-repair, Other-service, Sales, Exec- managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm- clerical, Farming-fish, Transport-moving, Priv-house-serv, Protective-serv, Armed Forces.
> relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
> race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
> sex: Male, Female
> capital-gain: continuous (per year)
> capital-loss: continuous (per year)
> hours-per-week: continuous (per week)
> native-country: United-States, Camb odia, England, Puerto-R ico, Canada , Germany, Outlying-US(Guam-USV I- etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand- Netherlands.

II- Pre-processing

We take an initial look at our mostly categorical data and notice that three of our variables ("workclass", "occupation" & "native.country") contain some values of "?", so we set a new level of "level_NA" for each and map those values to that.

After that we run an algorithm for outlier detection using initial mahalanobis distances between individuals as compared with a robust derived mahalanobis distance calculation between points, and obtain the following plot. This outlier detection is made using only the continuous variables.



We determine that there are thirty eight outliers, located in the upper right portion of the prior plot. As the Mahalanobis distance should follow a Chi-Square distribution (with 5 degrees of freedom according to the number of continuous variables), we set as outliers the points for which the robust mahalanobis distance is higher than the 97.5%-quantile of this distribution. The red lines show this value.

These outlier individuals have in common that they all make over 50k a year, but more importantly that they all answered that there “capital gains” per year was 99999 which is irregular for that variable. From here on out, we will assign these individuals with a very low weight so that their captial gains responses don’t unduly influence the analysis.

We next discretize ("fnlwgt", "education.num", "capital.gain", "capital.loss" "hours.per.week") into quartiles. We also discretize the “age” variable into 5 groups (under 25, 26 to 34, 35 to 49, 50 to 62, 63 and up).

In the end we are left we with a dataset as follows:

```
> summary(potec)
            age                      workclass                         fnlwgt              education                   education.num
 age:25 and under: 6411   Private         :22696   fnlwgt [1.23e+04,1.18e+05]:8141   HS-grad     :10501   education.num [1,9]   :14754
 age:26 to 35    : 8514   Self-emp-not-inc: 2541   fnlwgt (1.18e+05,1.78e+05]:8140   Some-college: 7291   education.num (9,10]  : 7291
 age:36 to 49    :10574   Local-gov       : 2093   fnlwgt (1.78e+05,2.37e+05]:8140   Bachelors   : 5355   education.num (10,12]: 2449
 age:50 to 64    : 5726   level_NA        : 1836   fnlwgt (2.37e+05,1.48e+06]:8140   Masters     : 1723   education.num (12,16]: 8067
 age:65 and up   : 1336   State-gov       : 1298                                     Assoc-voc   : 1382
                          Self-emp-inc    : 1116                                     11th        : 1175
                          (Other)         :  981                                     (Other)     : 5134
              marital.status            occupation          relationship                    race             sex
  Divorced             : 4443   Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311   Female:10771
  Married-AF-spouse    :   23   Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039   Male  :21790
  Married-civ-spouse   :14976   Exec-managerial:4066   Other-relative:  981   Black             : 3124
  Married-spouse-absent:  418   Adm-clerical   :3770   Own-child     : 5068   Other             :  271
  Never-married        :10683   Sales          :3650   Unmarried     : 3446   White             :27816
  Separated            : 1025   Other-service  :3295   Wife          : 1568
  Widowed              :  993   (Other)        :9541
            capital.gain                  capital.loss                 hours.per.week            native.country       target
 capital.gain 0    :29849   capital.loss 0        :31042   hours.per.week [1,40)  : 7763   United-States:29170   <=50K:24720
 capital.gain (0,1]: 2712   capital.loss (0,4.36]: 1519   hours.per.week 40      :15217   Mexico       :  643   >50K : 7841
                                                          hours.per.week (40,45]: 2442   level_NA     :  583
                                                          hours.per.week (45,99]: 7139   Philippines  :  198
                                                                                          Germany      :  137
                                                                                          Canada       :  121
                                                                                          (Other)      : 1709
```

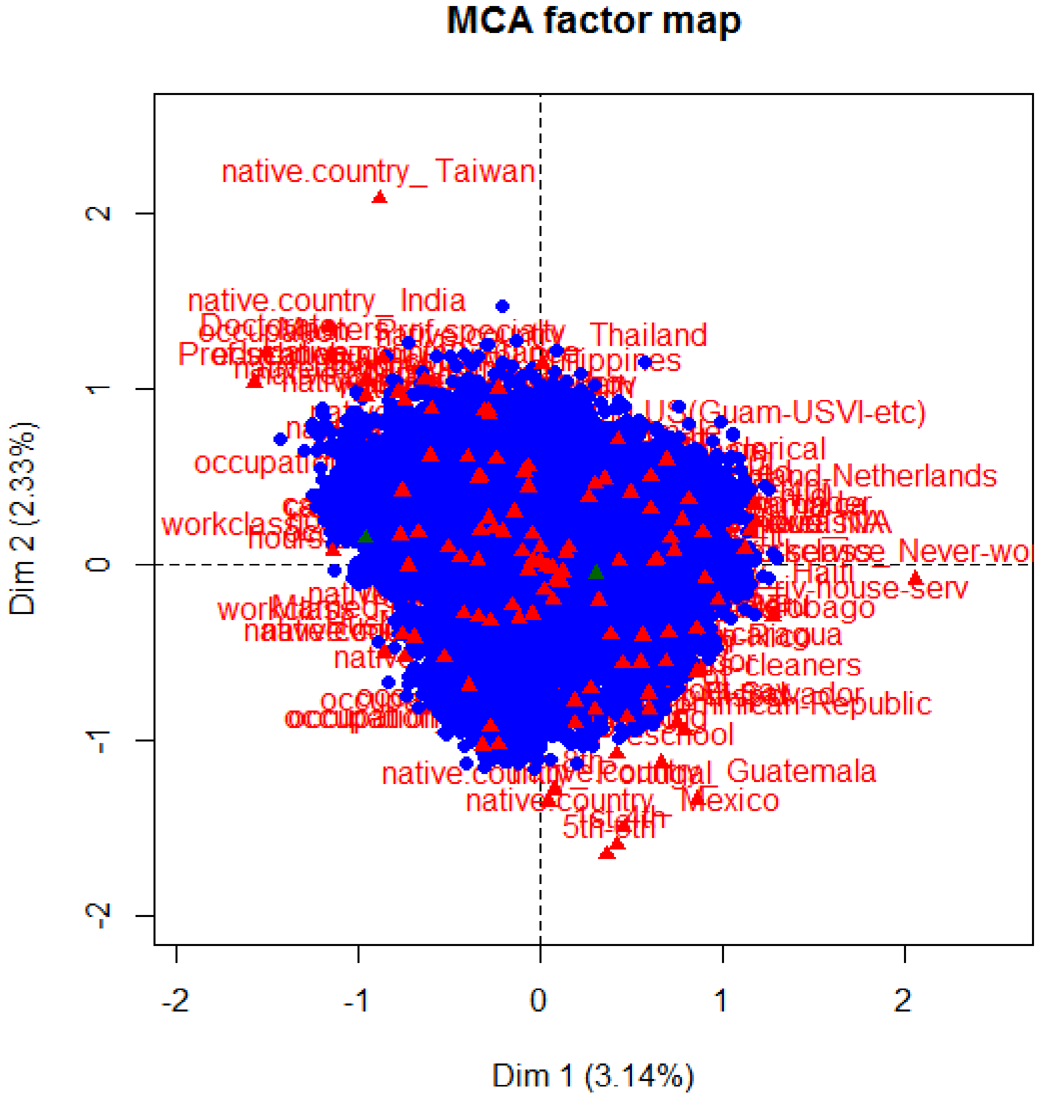
From our data summary, we can see the most occurring individuals are between 36 and 49, work for private employers, have at least a high school education, are husbands, white, work 40 hours a week, consider the US their native country and make less than 50 thousand dollars a year.

III- MCA

Now all of our variables are factors, and so is the target. It is a natural idea to perform an MCA on this dataset. Besides, this will allow us to detect non-linear relationship between variable. As we have a lot of individuals and quite many modalities too, it would be a more adapted method than applying a PCA.

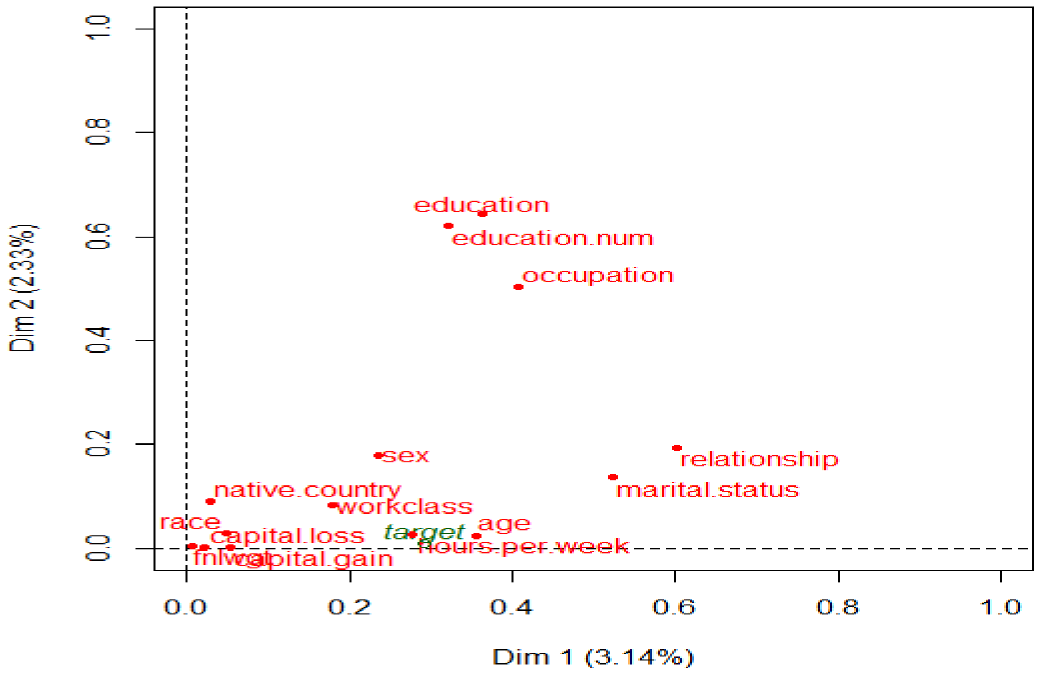
As we only have 14 variables and we assume that they all play a role to predict the target, we will only set the target variable as illustrative. So we have 14 active variables. We will also set the weights of the MCA as 1 for all individuals except the outliers whom will be assigned a weight of 0.00001.

The followings are some plots of the MCA results.

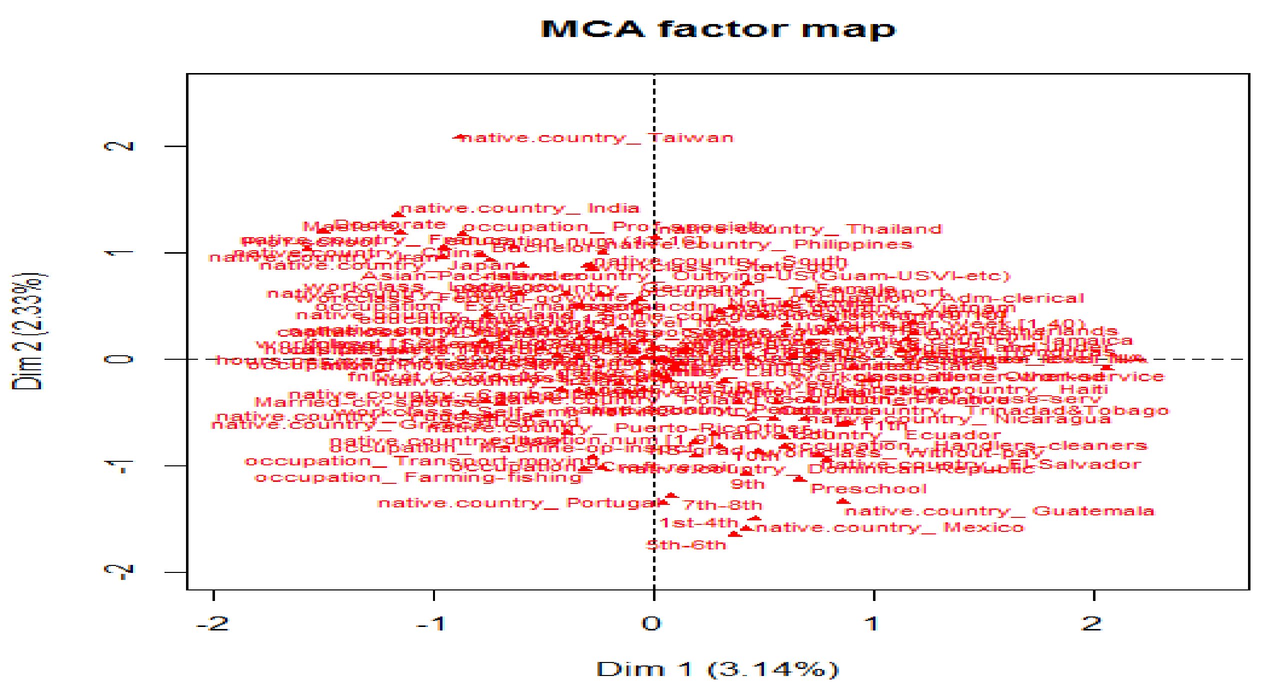


The first plot represents all individuals and all variables (active and illustrative), it is very loaded and we cannot really extract information from it except from the global distribution and some outlier modalities.

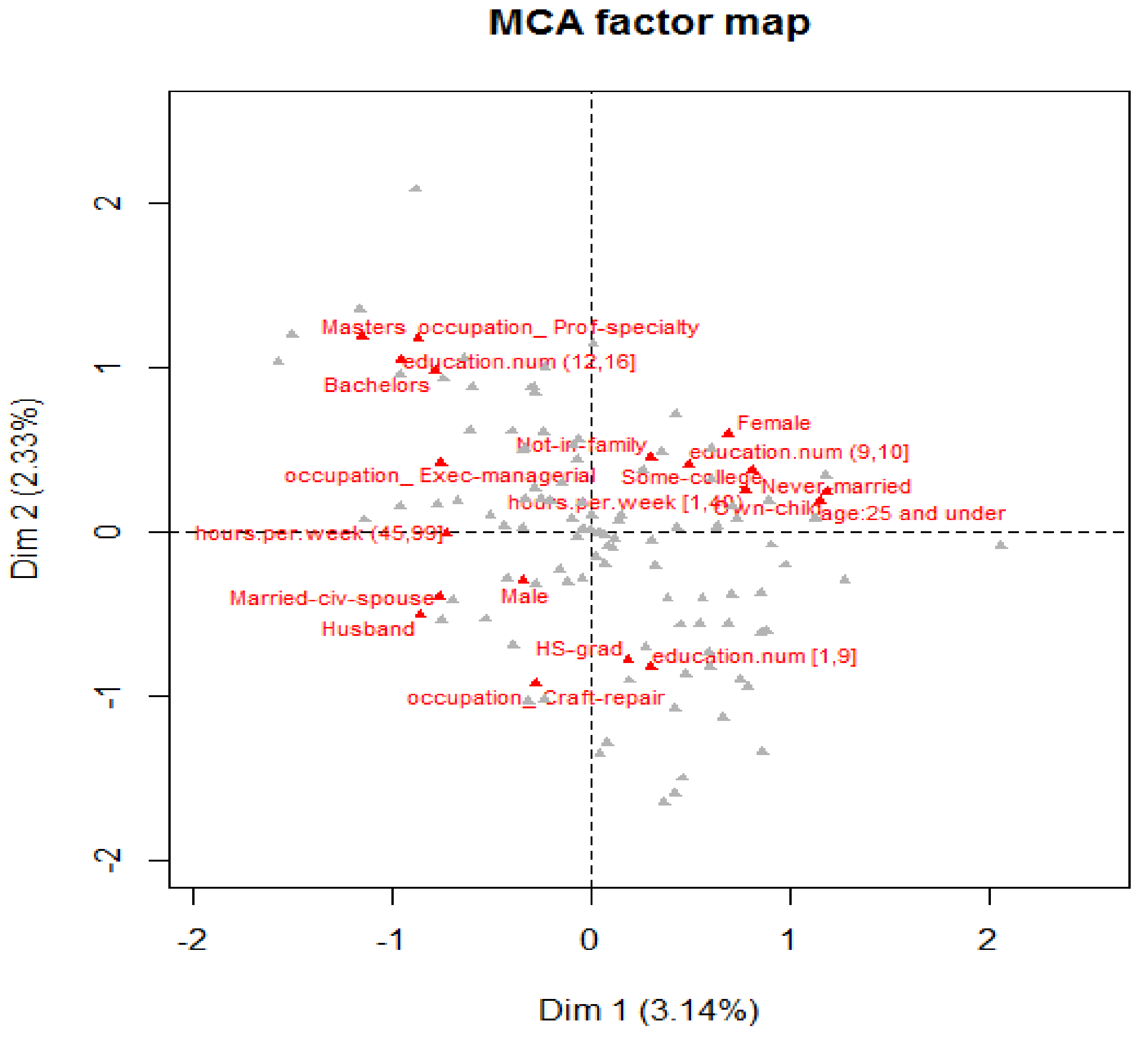
The second plot below is easier to read. It represents the variables according to their correlations (cos square) to the dimensions. We can see that the most relevant modalities are “education”, “education.num” and “occupation” (further from the center). Those three variables are also the most correlated with the second dimension. On the other hand “relationship” and “marital.status” are highly correlated with the first dimension. The “target” variable is also very close to the axis of the first dimension. Anyway to interpret the distribution of individuals, the latent concepts and the clustering, we are more interested in the distribution of modalities.



The next plot representing all active variables is quite difficult to read. We can easily detect outlier modalities but the more central ones are overlapping. To have a better understanding we will plot only the variables that contributed most to the dimensions.

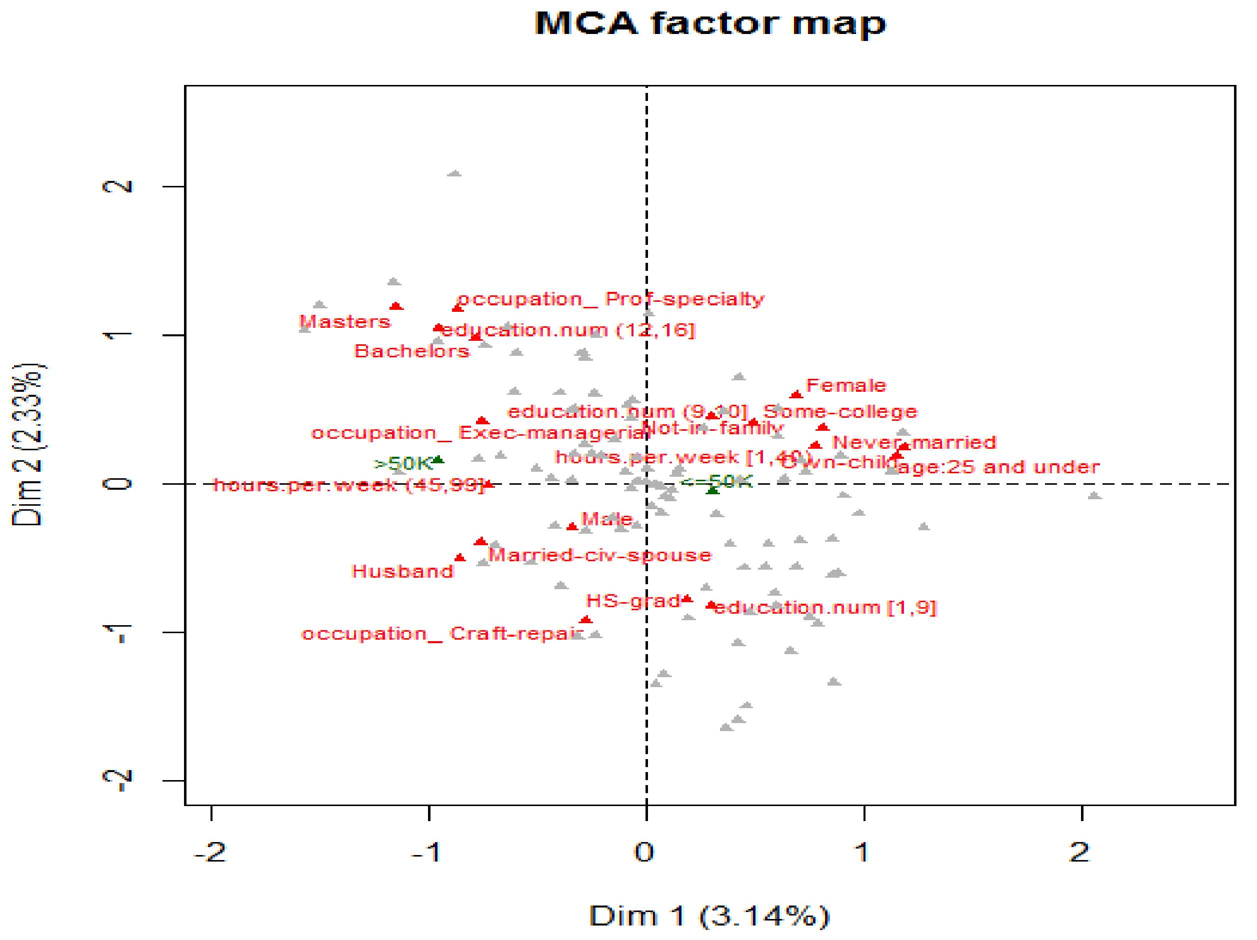


The following plot shows the 20 variables that contributed most to the dimensions. We can already notice some pattern in the distribution of modalities. Like the opposition of "female" and "male" or the curve described by the number of studying years. Also the worked hours per week seem to follow a straight line along the first dimension. Very young people would be in the middle right of the plot, whereas highly educated ones will be at the top left.

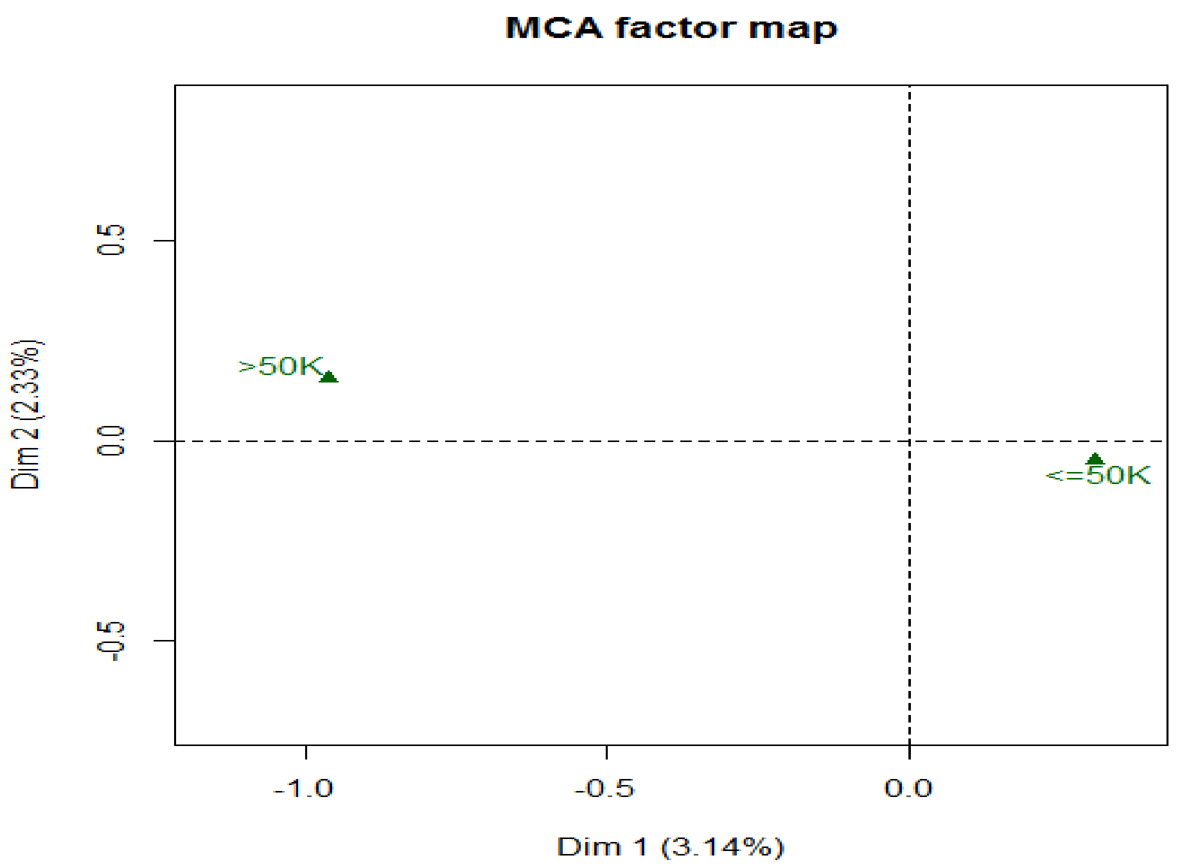


We can also plot the 20 variables that are most correlated to the dimensions. The two modalities of the target variable appear on this plot. They are distributed over the first dimension axis and we can see that the "left" part of the plot would be the one containing people earning more than 50K a year while the "right" part would be people earning less than 50K a year. The first dimension could also be the dimension of wealth.

Apart from the target the other variables are approximately the same as the ones which contributed most to the dimensions. We can keep reading information from this plot. For instance, we can see that the female modality is quite high on the second dimension but in the right part of the plot, while the male modality is on the contrary quite low on the second dimension but in the left part of the plot. So comparing those positions with the "education.num" evolution on the plot, and the target distribution, we can conclude that women are globally more educated but will make less money, while men are less educated but will globally make more money. We can also notice that people under 25 are very unlikely to make more than 50K a year.



The last plot only shows illustrative variables (here the target). We can notice that the modality "<=50K" is quite central (even if a little bit on the right) so it will be a common modality.



To obtain a more formal description of the dimensions we use the 'dimdesc' function. Here are the results we obtained (keeping only the p-values equal to zero). The firsts tables show the '$quali' result, that is the relevant variables, whereas the second ones will show the '$category' that is the most relevant modalities. We are more interested in the second one, although the variables can also give interesting information.

The first dimension is very discriminative (even if the percentage of variation explained is only 3.14%). First this dimension separates the target modalities. So the first dimension is the dimension of "wealth". We can point out that a positive capital gain will also be negatively correlated with the first dimension (on the left) and that makes sense as we can assume that only wealthy people will make any kind of 'capital gain'. The first dimension also separates the working hours modalities: on the right there will be people working less (between 1 and 40 hours a week) certainly including people who do not have a job. On the left there will be people working a lot (between 45 and 99 hours a week). Furthermore we can see that education is also distributed on this dimension, with two lower modalities on the right and higher education in the left part of the plot. This kind of information is directly related to the professional situation and what one is earning a year. But the first dimension also bears some 'social' information. First the age categories: in the right there will be younger people, while in the left there are middle age ones (36 to 64; so not retired people). This is coherent with the professional and financial discrimination (as young people often do not work, and we can expect that the salary of someone will reach a maximum when he is between 36 and 50). Finally we can notice that people who did not give their profession (occupation_level_NA) are in the right part of the plot too, as is the 'female' modality whereas high professions (managerial…) will be on the left part (negative correlation).

Dim1 Qualitative

	R2	p.value
age	0.35	0.00
workclass	0.18	0.00
education	0.37	0.00
education.num	0.32	0.00
marital.status	0.52	0.00
occupation	0.41	0.00
relationship	0.60	0.00
race	0.05	0.00
sex	0.24	0.00
capital.gain	0.06	0.00
hours.per.week	0.28	0.00
target	0.29	0.00

Dim1 Categorical

	Estimate	p.value
<=50K	0.31	0.00
hours.per.week [1,40)	0.44	0.00
capital.gain 0	0.21	0.00
Female	0.25	0.00
Own-child	0.42	0.00
occupation_level_NA	0.49	0.00
education.num (9,10]	0.28	0.00
education.num [1,9]	0.19	0.00
age:25 and under	0.54	0.00
>50K	-0.31	0.00
hours.per.week (45,99]	-0.31	0.00
capital.gain (0,1]	-0.21	0.00
Male	-0.25	0.00
Husband	-0.59	0.00
occupation_ Prof-specialty	-0.50	0.00
occupation_ Exec-managerial	-0.44	0.00
Married-civ-spouse	-0.53	0.00
education.num (12,16]	-0.43	0.00
Prof-school	-0.77	0.00
Masters	-0.56	0.00
Doctorate	-0.73	0.00
Bachelors	-0.38	0.00
age:50 to 64	-0.24	0.00
age:36 to 49	-0.25	0.00

The second dimension is clearly the dimension of education, with highest education on top and lower ones at the bottom. It also differentiates women and men. Some professions also appears as relevant, mostly because they are professions requiring high/low (if there is a positive/negative correlation) level of education.

Dim2 qualitative

	R2	p.value
workclass	0.08	0.00
education	0.64	0.00
education.num	0.62	0.00
marital.status	0.14	0.00
occupation	0.50	0.00
relationship	0.19	0.00
sex	0.18	0.00
native.country	0.09	0.00

Dim2 category

	Estimate	p.value
Female	0.19	0.00
occupation_ Prof-specialty	0.56	0.00
education.num (12,16]	0.34	0.00
Some-college	0.28	0.00
Prof-school	0.55	0.00
Masters	0.61	0.00
Doctorate	0.62	0.00
Bachelors	0.53	0.00
Male	-0.19	0.00
Husband	-0.26	0.00
occupation_ Craft-repair	-0.34	0.00
education.num [1,9]	-0.45	0.00
HS-grad	-0.22	0.00
7th-8th	-0.44	0.00
5th-6th	-0.59	0.00

The third dimension will give a finer separation of age modalities. It opposes retired people (over 65 years old) to more middle age ones (26 to 49). It also opposes some education levels but in a different way as before. The 'extremes' level of education will be positively correlated while the 'middle' one (10 to 12 years of education) will be negatively correlated.

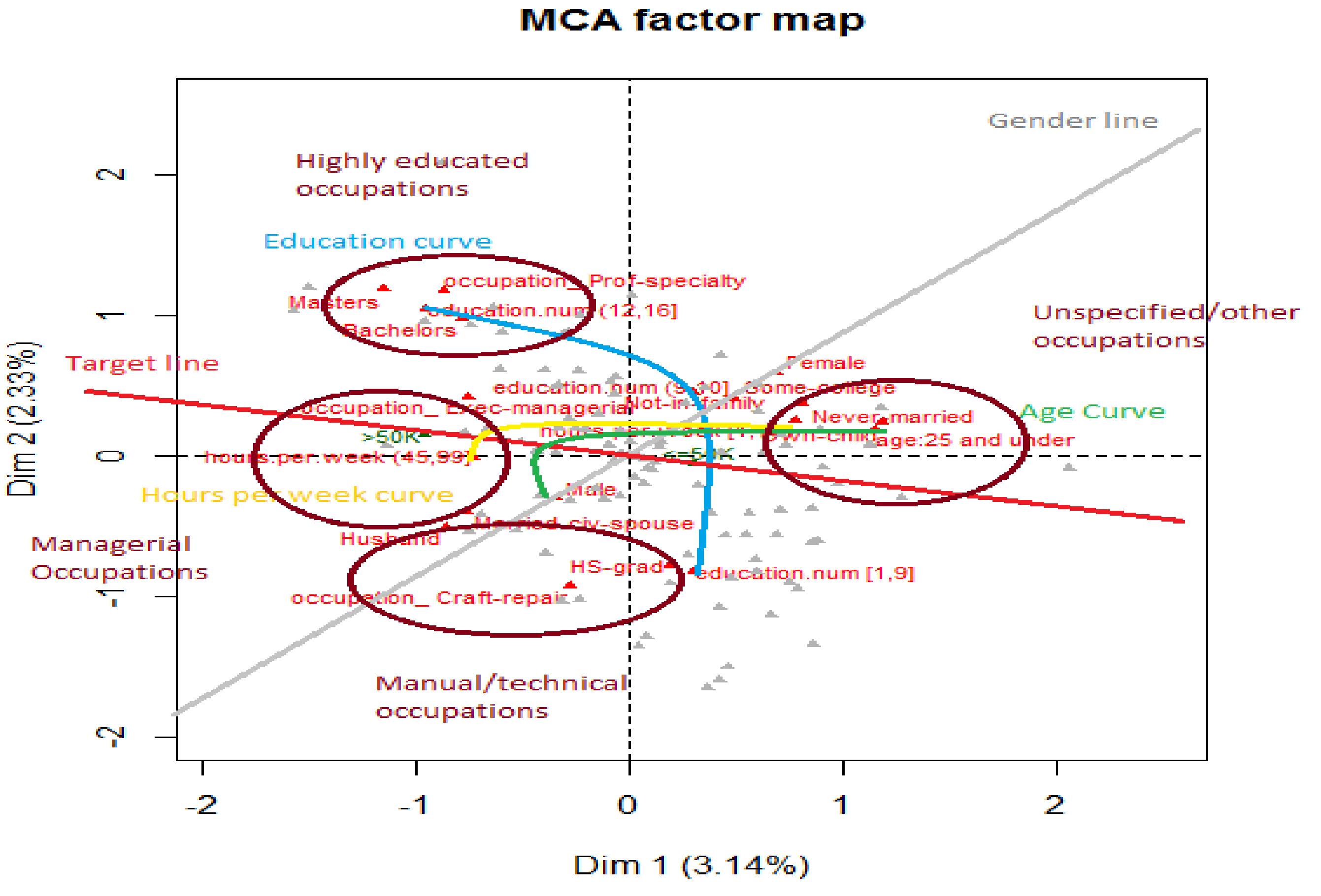
Dim3 qualitative

	R2	p.value
age	0.16	0.00
workclass	0.44	0.00
education	0.31	0.00
education.num	0.28	0.00
marital.status	0.16	0.00
occupation	0.47	0.00
relationship	0.20	0.00
sex	0.06	0.00
hours.per.week	0.07	0.00

Dim3 category

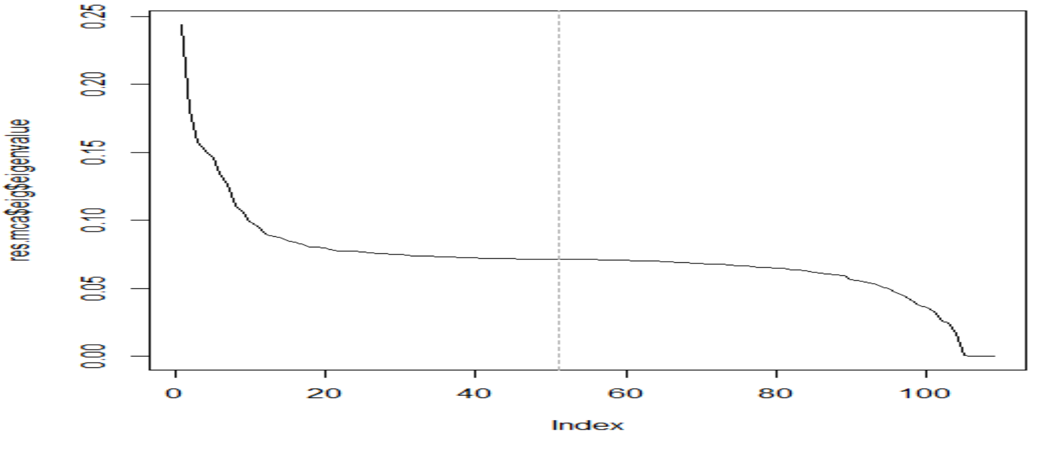
	Estimate	p.value
Male	0.10	0.00
Own-child	0.24	0.00
Husband	0.15	0.00
occupation_level_NA	1.02	0.00
education.num (12,16]	0.25	0.00
education.num (9,10]	0.24	0.00
workclass_level_NA	0.76	0.00
age:65 and up	0.41	0.00
Female	-0.10	0.00
Unmarried	-0.35	0.00
education.num (10,12]	-0.58	0.00
Assoc-voc	-0.70	0.00
Assoc-acdm	-0.72	0.00
age:36 to 49	-0.22	0.00
age:26 to 35	-0.20	0.00

We can make a plot to summarize these ideas and to give a 'visual' of latent concepts, but the plot is a little bit loaded… We can point out that in a society where there will be sex-equality, the red line and the grey line would be perfectly perpendicular, (they should be mediatrices of each other) so the angle between them is an inequality measure (or its sine)! We can also point out that even if the 'male' modality is low on the second dimension while the ‘female’ one is higher, it is not so clear that men are less educated than women because of the curved line of education levels.



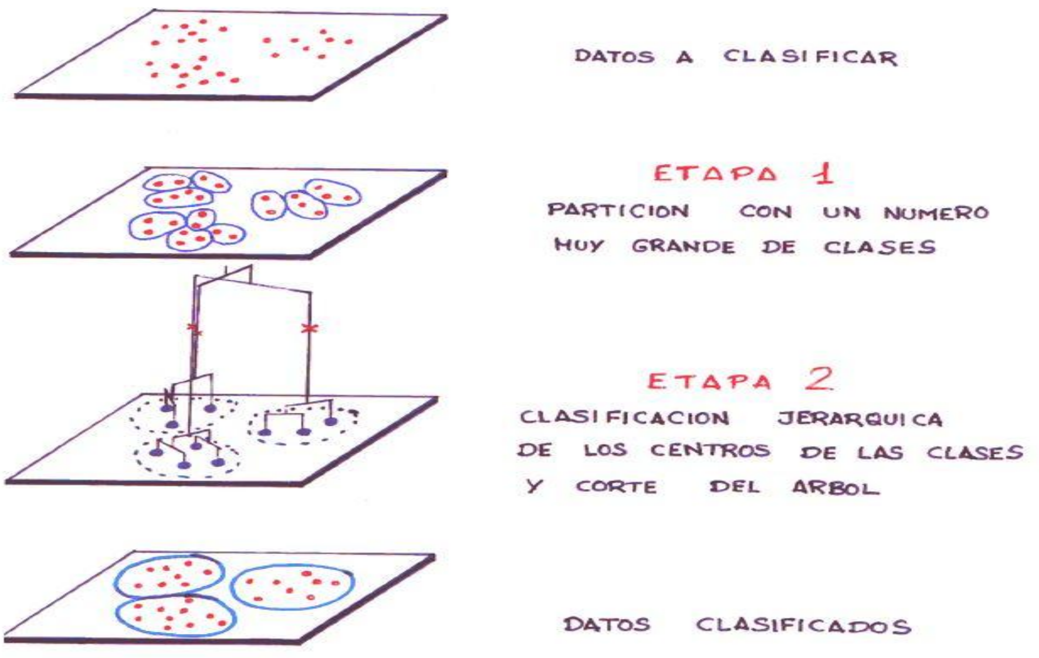
Now we want to apply clustering on our data to create groups of people. For that we need to decide how many dimensions we want to keep from the MCA analysis. First we can plot the eigenvalues according to the dimensions. It total we have 109 dimensions. There are several rules we could use to select the dimensions. We can keep the dimensions for which the eigenvalue is higher than the mean. We chose to keep the dimensions for which the eigenvalue is higher than one over the number of active variables (in this case 14). With this rule we keep 51 dimensions.

In the following plot, the cut is indicated by the vertical gray line.

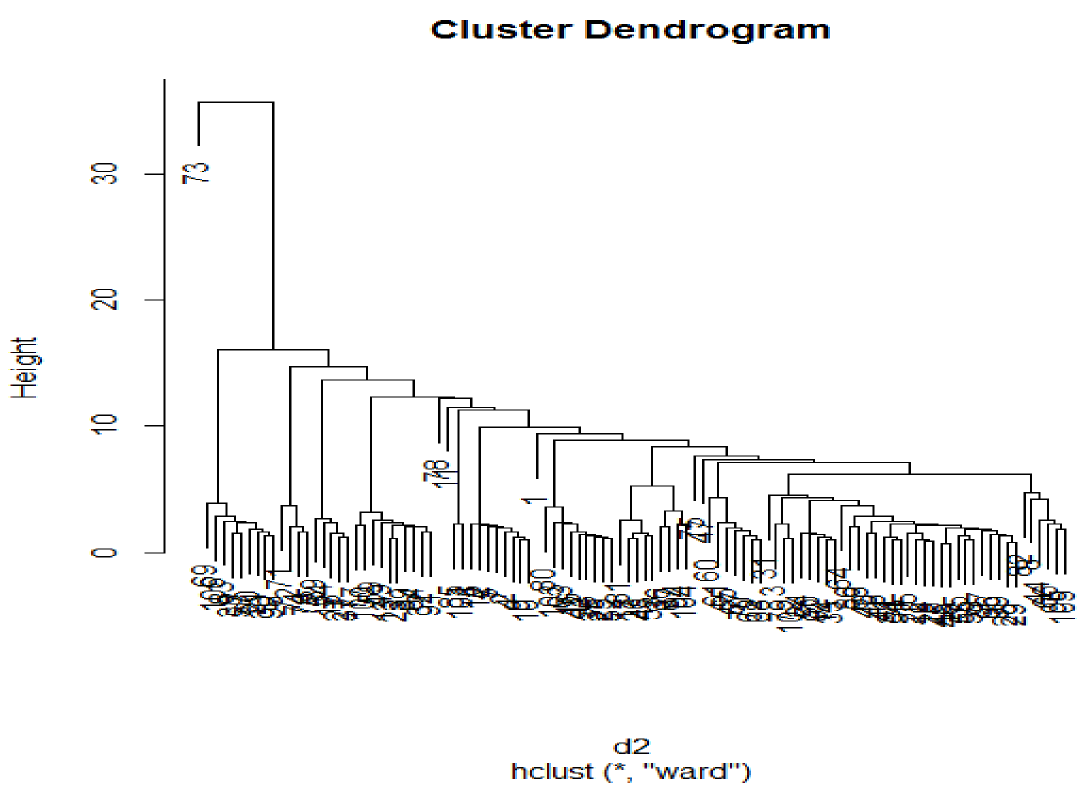


IV- Clustering

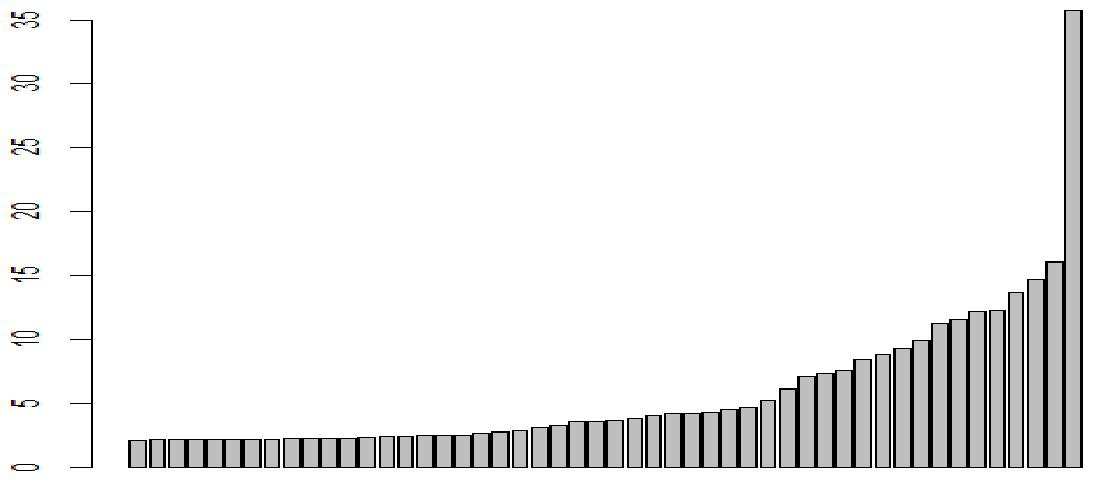
After deciding to keep the first 51 dimensions, we then re-run MCA (fixing the number of dimension kept to 51) and store only the significant dimensions into Psi. Because our dataset is relatively large, we progress with the following strategy to handle clustering:



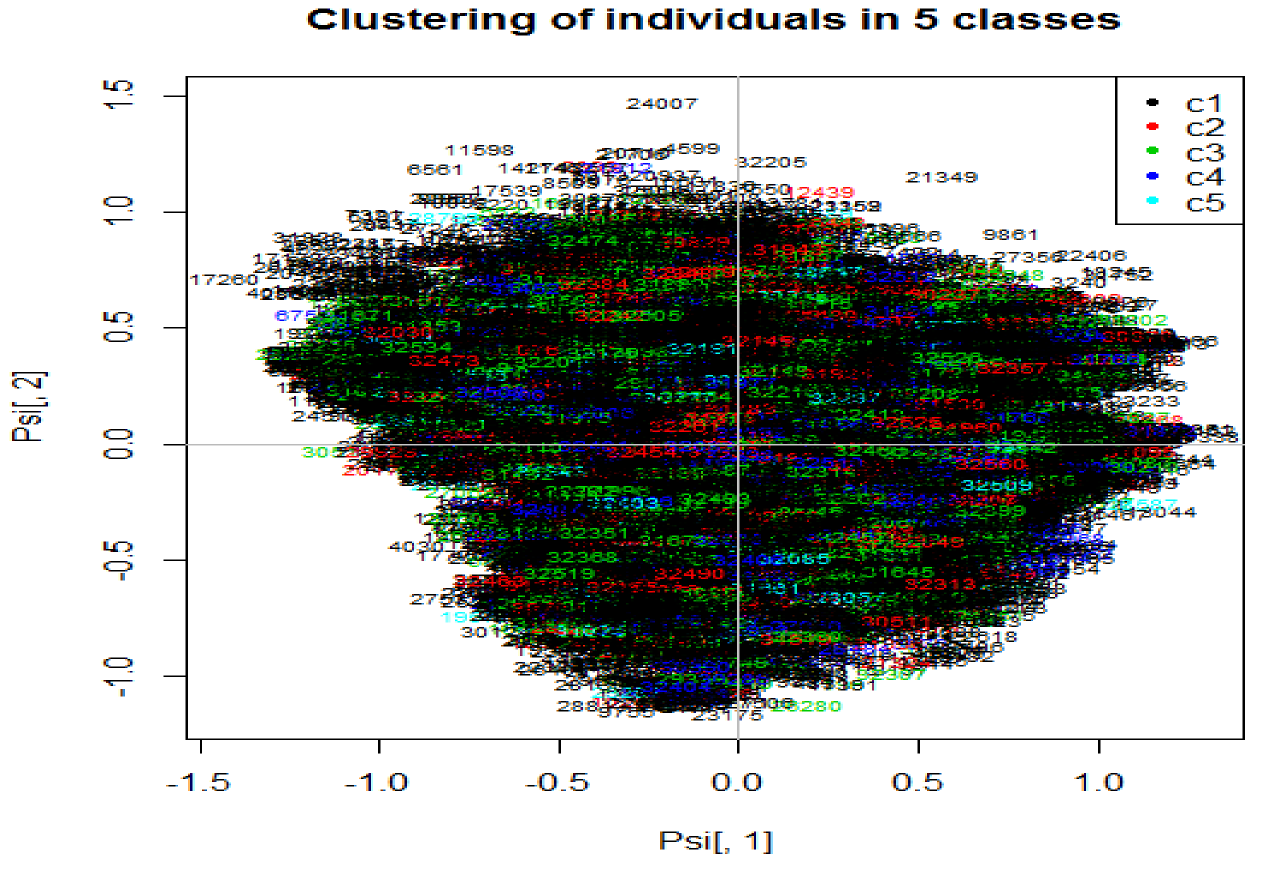
We decide to first perform two separate runs of the k-means algorithm on Psi giving each the same arbitrarily large number of clusters (12 in our instance) to look for. We then do a hierarchical clustering upon the centroids of crossing these 2 kmeans partitions using the ward distance criterion.



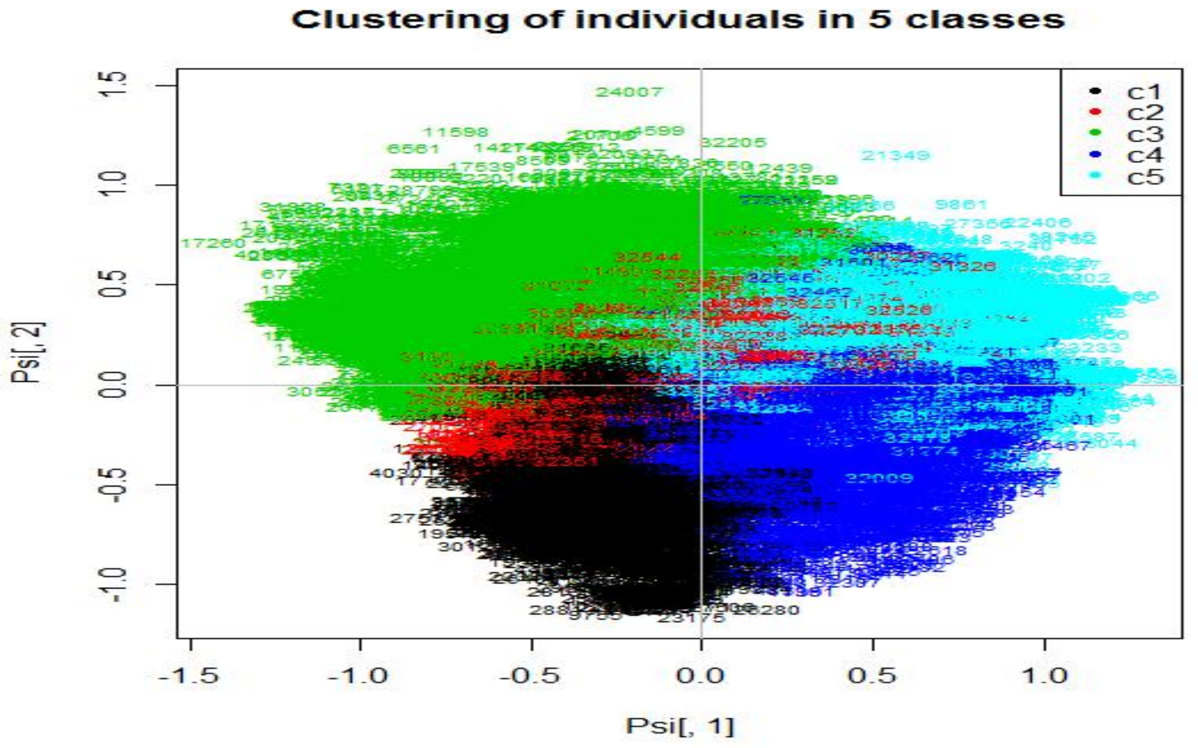
Upon looking at the results, we do a barplot of the heights of the jumps between different clusterings, and decide that taking 5 clusters seems reasonable.



We then prune our results at a depth of 5, recalculate the centers of gravity given 5 clusters, and plot our findings which shows the clustering of individuals into 5 clusters.



This plot is not very well seperated so we decided to run kmeans again (consolidation of the clustering) on our data this time giving it an input of 5 clusters to find beginning from the centroids obtained in the prior step



This time the clusters are much better separated. We then run the function catdes on our clusters to get a description of what characterizes them (see **Appendix** for full listing).

cluster 1: **White Working Class Men (25%)**

Male , White, Married, self employed, age 36 to 50
Blue collar occupations: craft-repair, transport-moving, farming-fishing, etc
High school, Some-college
Works 45 to 99 hours a week

cluster 2: **Vocational School Tech-support** (7%)

education = Assoc-voc, assoc-admin

cluster 3: **The Haves** (24.3%)

Occupation: Prof-specility, Exec-managerial
education: Masters, Bachelors, Doctorate
Works 45 to 99 hours a week
Husbands, age 36 to 49
workclass: Local gov, State gov, self employed
capital gains and capital losses

cluster 4: **The Diverse Workers** (25%)

Female, under 25, Black, Mexican,
High School
working class occupation,
works under 40 hours a week

cluster 5: **The Unemployed** (17.7 %)

Female, occupation: NA, workingclass: NA
under 25, unmarried / divorced, Some college

V - Prediction

We chose to use prediction trees. Our first choice was to use C4.5 (which is a multi-way tree using theory of information: the splits are chosen according to entropy). Mostly because the dataset information included error rates obtained using standard algorithms and C4.5 had pretty good results. Additionally, we studied prediction trees in class. As a comparison of this model we also implemented a CART tree (Classification and Regression Tree). The two libraries used in R are RWeka and rpart respectively.

The validation protocol is the following:

First we will split our data in two parts (a training part that will contain 2/3 of the data and a testing part with 1/3). The training data will be used to optimize the parameters and to build the final model. The testing data are only used to compute the final validation error.

Then, for each model (C4.5 and rpart) we will optimize the parameters on the training data. To do so, we use 10 fold cross validation. Setting a range for our parameters, we evaluate the 10 fold CV error (the training and model are built with 9/10 of the data and the testing error is evaluated with 1/10) for every one of those and we will select the parameter that gives the lowest error.

The train function in R helps doing this loop.

The parameters we want to optimize are:

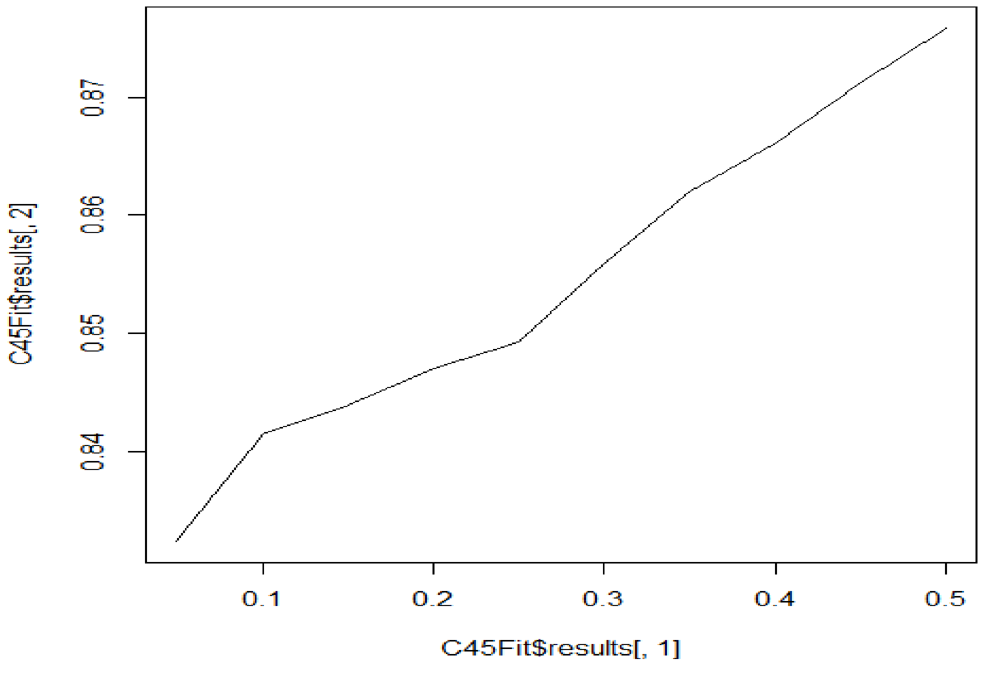
- For C4.5: C the Confidence threshold for pruning the tree
- For CART: cp the complexity parameter

We will optimize ‘C’ from 0 to 0.5 by 0.1, and ‘cp’ from 0.001to 0.01 by 0.0005.

The following screenshot of R results describe the model and the accuracy obtained for the different parameters. In the C4.5 tree we are testing 10 parameters.

```
> C45Fit  # accuracy keep increasing with C, so final model kept: 0.5
C4.5-like Trees

21707 samples
   14 predictors
    2 classes: ' <=50K', ' >50K'

No pre-processing
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 19536, 19536, 19536, 19537, 19536, 19536, ...

Resampling results across tuning parameters:

  C     Accuracy  Kappa  Accuracy SD  Kappa SD
  0.05  0.832     0.502  0.0126       0.0406
  0.1   0.841     0.543  0.00953      0.0214
  0.15  0.844     0.55   0.00965      0.0227
  0.2   0.847     0.56   0.00995      0.0218
  0.25  0.849     0.565  0.00987      0.0234
  0.3   0.856     0.586  0.00972      0.0202
  0.35  0.862     0.606  0.00892      0.0181
  0.4   0.866     0.618  0.00755      0.0143
  0.45  0.871     0.632  0.00777      0.0165
  0.5   0.876     0.646  0.00711      0.0141

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was C = 0.5.
```

This is the accuracy curve for C4.5 model, when optimizing the parameter. We can see that the accuracy keeps increasing. The best parameter will be C=0.5.




```
> C45Fit$bestTune
     C
10 0.5
```

Once we have selected the best parameter, we can build the model with the whole training data (no more 10 fold CV), and the best parameter.

From this model we are then able to compute the training and testing error.

model.tree<-J48(target~., data=potec, subset=learn, control=Weka_control(C=0.5), na.action=NULL)

Train

```
> pred.learn<-predict(model.tree, data=potec[learn])
> tab<-table(pred.learn,potec$target[learn])
> (error.learn<-100*(1-sum(diag(tab))/nlearn))
[1] 12.0422
> tab

pred.learn  <=50K  >50K
     <=50K  15473  1659
     >50K     955  3620
```

The Training error is 12%

Test

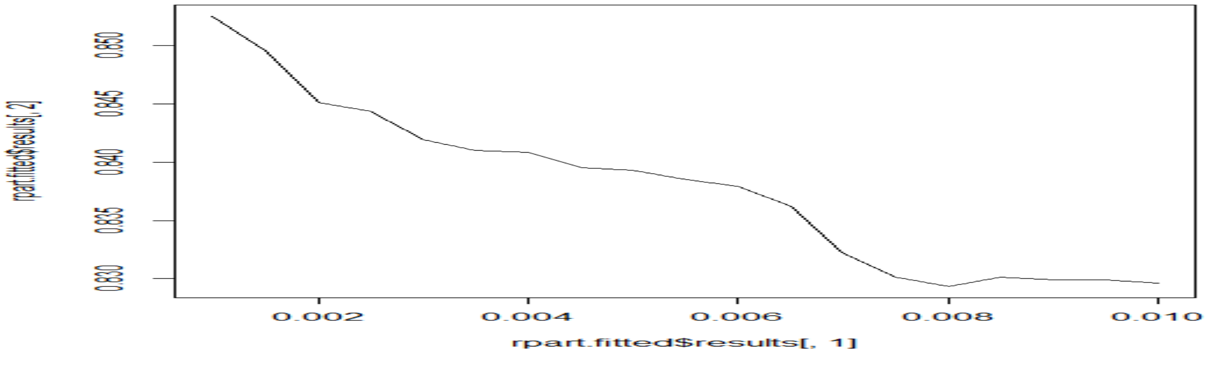
```
> pred.test<-predict(model.tree, newdata=potec[-learn,])#subset=-learn
> tab<-table(pred.test,potec$target[-learn])
> (error.test<-100*(1-sum(diag(tab))/ntest))
[1] 17.03519
> tab

pred.test  <=50K  >50K
    <=50K   7513  1070
    >50K     779  1492
```

The Test error is 17%. So there is an over-fitting of the model, the training error is a little bit optimistic.

Then we do the same process with CART tree. Here we are testing 19 parameters.

```
> rpart.fitted
CART

21707 samples
   14 predictors
    2 classes: ' <=50K', ' >50K'

No pre-processing
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 19536, 19536, 19536, 19536, 19536, 19536, ...

Resampling results across tuning parameters:

  cp       Accuracy  Kappa  Accuracy SD  Kappa SD
  0.001    0.853     0.576  0.0104       0.0286
  0.0015   0.85      0.566  0.011        0.0264
  0.002    0.845     0.56   0.0121       0.0261
  0.0025   0.844     0.558  0.0127       0.027
  0.003    0.842     0.552  0.011        0.0218
  0.0035   0.841     0.551  0.0115       0.0256
  0.004    0.841     0.551  0.0115       0.0256
  0.0045   0.84      0.545  0.00995      0.0218
  0.005    0.839     0.542  0.0101       0.0238
  0.0055   0.838     0.54   0.00917      0.0219
  0.006    0.838     0.538  0.00887      0.0207
  0.0065   0.836     0.535  0.0102       0.0205
  0.007    0.832     0.522  0.0108       0.0214
  0.0075   0.83      0.521  0.00965      0.0221
  0.008    0.829     0.52   0.00972      0.0218
  0.0085   0.83      0.521  0.00817      0.0199
  0.009    0.83      0.519  0.00797      0.0181
  0.0095   0.83      0.519  0.00797      0.0181
  0.01     0.83      0.517  0.00785      0.018

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cp = 0.001.
```

As the parameters plays an opposite role (the confidence threshold as opposed to the complexity parameter), in this case the accuracy curve has a decreasing behavior. The best parameter is 0.001.




```
> rpart.fitted$bestTune
1 0.001
```

With this parameter and the whole training set, we build the final model, and compute error rates.

```
> p1 <- rpart(target ~ ., data=learndata,control=rpart.control(cp=0.001), weights=w[learn])
```

<u>Train</u>

```
> pred.learn<-predict(p1, data=learndata, type="class")
> tab<-table(pred.learn,learndata$target)
> (error.learn<-100*(1-sum(diag(tab))/nlearn))
[1] 14.65426
```

The training error is 14.7%, which is slightly worst than with the C4.5 model.

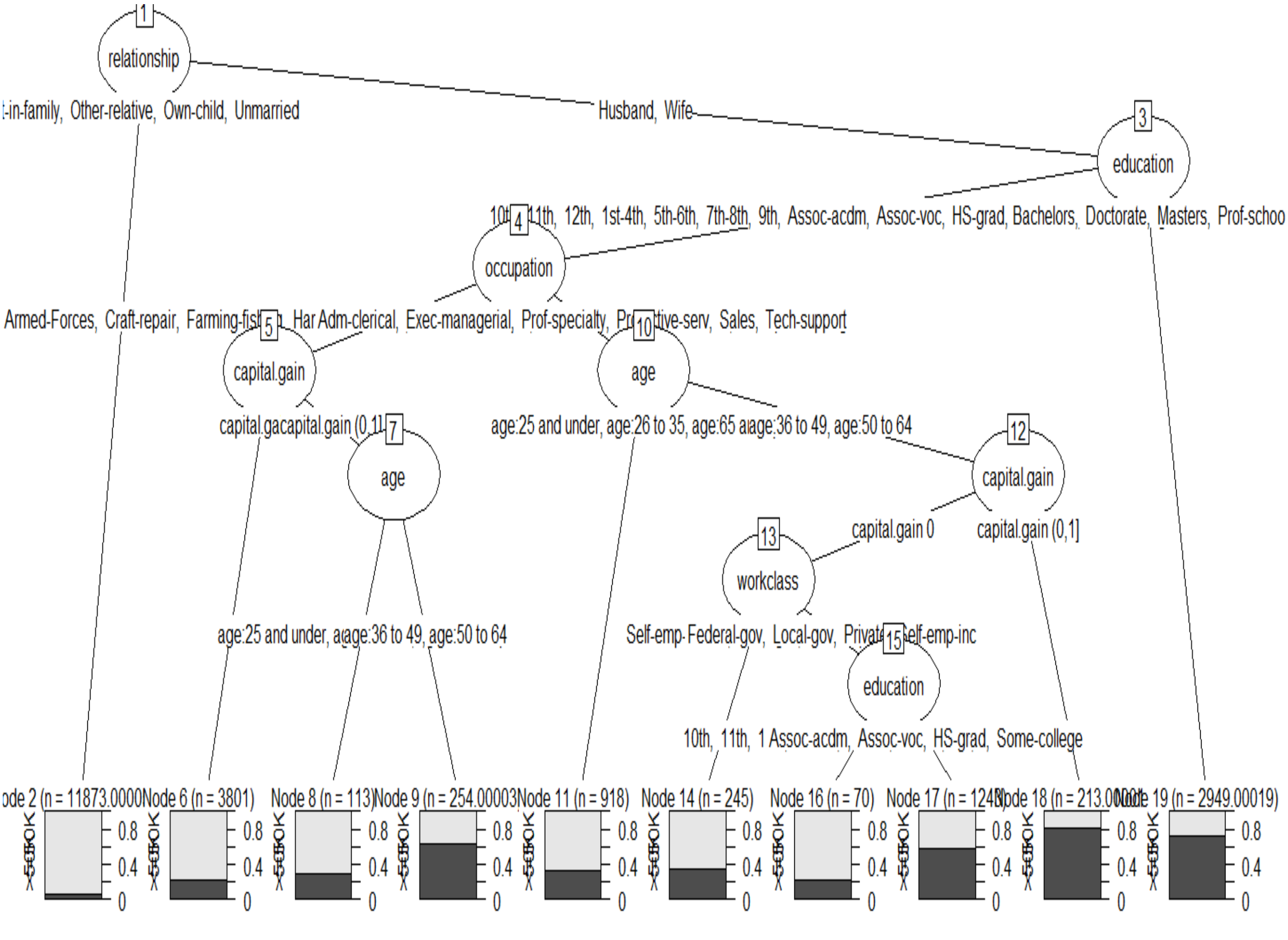
<u>Test</u>

```
> pred.test<-predict(p1, newdata=potec[-learn,], type="class")#subset=-learn
> tab.test<-table(pred.test,potec$target[-learn])
> (error.test<-100*(1-sum(diag(tab.test))/ntest))
[1] 15.93882
```

The testing error is 16%. This time, this is less than the previous model. We computed the validation error on both model because as we are using only two models this isn't costly. Nonetheless we should select our best final model according to the training error obtained. Therefore we would have chosen the C4.5 tree (but in this case we know that because of over-fitting this model is actually worst).

Here follow a plot of the tree obtained with CART algorithm (as we had issues plotting the C4.5 tree from RWeka library) and a complexity parameter non-optimal (cp=0.05) because otherwise the tree was too loaded. This plot is just here to ease the interpretation and give a visual.

One of the main advantages of prediction tree is their interpretability. Indeed here we can understand very quickly what is happening to make a decision. The cuts are done on 'relationship', 'occupation', 'age', 'capital.gain' and 'workclass'. The dark grey represents the modality '>50K' while the light grey represents '<=50K'. For instance, following branches we can see that someone who is married and highly educated will have a high probability to make more than 50K a year (leave 19, bottom right). On the other hand, someone unmarried or not in family will be very likely to make less than 50K a year (leave 2, bottom left).



Finally, we also have to mention that we had issues to modify the weights with C4.5 method, while it was possible with the CART algorithm. In our studies of the impact the outliers can have (they represent only a 0.12% of the individuals which is quite negligible); we tried to apply the algorithm building the model without the outliers (just suppressing the corresponding rows) and to test them whether on the testing data also suppressing outliers or on the whole remaining data (test data and outliers). In both cases we got a slightly better training error (which makes sense as we get rid of difficult cases) and a slightly worst validation error. What was strange was that the validation error obtained with the outliers was smaller than the one without outliers. As the differences were quite small (less than 0.3%) and it impacted neither the best parameter optimization nor the comparison of the best model, we chose to keep our results as they were. Another solution would have been to set the corresponding value of 'Gain' (99999) to NA and to impute it using Mice method, transforming therefore outliers into more 'normal' individuals. The different results obtained are in the appendix.

VI - Conclusion

Going over all the task of this work we can say that they are complementary and necessary. First the pre-processing of the data gives us a first overview and understanding of the topic. Besides before going into further processing we need to clean the data and put it in the desired format. Then MCA will give us a deeper understanding of the data and most importantly, of how variables are related to each other, and to individuals. This step is very important and also very relevant to the clustering. Indeed the clustering allows us to group individuals but we have to link those groups to the modalities to understand who are the individuals in each group.
Then the prediction is a different task, once we have understood the data, we can try to build a model to predict the target. Though we can see that some of the global behaviors we noticed observing the data (with MCA or with clustering) are retrieved in the prediction trees. The final predicting models are not so bad, as we have validation errors around 17% for the selected model (and 14% for the CART tree). If we consider that we are predicting the year income of individuals with only few information (14 categorical variables), we couldn't expect very low errors. Indeed, explaining income with those few variables has led us to have very stereotyped results. We can remark as a final conclusion that our results, on the interpretation point of view, have to be considered taking some step back and keeping in mind that they only describe tendencies.

VII – Appendix

1. Catdes results

CatDes<-catdes(pot,num.var=15)

```
$'1'
                                              Cla/Mod     Mod/Cla      Global       p.value     v.test
sex= Male                                 37.46672786 98.84973968 66.92054912  0.000000e+00        Inf
relationship= Husband                     60.72917456 97.00932316 40.51779736  0.000000e+00        Inf
occupation= Craft-repair                  53.52525006 26.56495944 12.58867971  0.000000e+00        Inf
marital.status= Married-civ-spouse        54.30689103 98.47439157 45.99367341  0.000000e+00        Inf
education.num=education.num [1,9]         40.45682527 72.27267224 45.31187617  0.000000e+00        Inf
education= HS-grad                        42.28168746 53.75953505 32.25023801  0.000000e+00        Inf
occupation= Transport-moving              57.92110207 11.19990314  4.90464052 7.863172e-177  28.352200

age=age:50 to 64                          39.15473280 27.14614360 17.58545499 3.053490e-143  25.482859
workclass= Self-emp-not-inc               44.35261708 13.64571982  7.80381438 2.172460e-104  21.697368
hours.per.week=hours.per.week (45,99]     33.43605547 28.90180409 21.92500230  1.473502e-67  17.366745
race= White                               26.96649410 90.82213343 85.42735174  4.477581e-63  16.763963
occupation= Farming-fishing               48.99396378  5.89659765  3.05273180  1.913304e-59  16.259488
education= 7th-8th                        55.10835913  4.31044921  1.98396855  2.940733e-59  16.233130
occupation= Machine-op-inspct             40.05994006  9.71061872  6.14845981  3.545308e-50  14.895123
education.num=education.num (9,10]        31.31257715 27.64257174 22.39181843  7.427174e-39  13.038114
education= Some-college                   31.31257715 27.64257174 22.39181843  7.427174e-39  13.038114
age=age:36 to 49                          29.80896539 38.16442669 32.47443260  8.587570e-37  12.670762
hours.per.week=hours.per.week 40          28.38930144 52.30657465 46.73382267  8.019758e-32  11.739247

sex= Female                                0.88199796  1.15026032 33.07945088  0.000000e+00       -Inf
relationship= Unmarried                    0.31921068  0.13318804 10.58321305  0.000000e+00       -Inf
relationship= Own-child                    0.55248619  0.33902409 15.56463254  0.000000e+00       -Inf
relationship= Not-in-family                1.15593016  1.16236833 25.50597340  0.000000e+00       -Inf
marital.status= Never-married              0.46803332  0.60540017 32.80918891  0.000000e+00       -Inf
marital.status= Divorced                   1.23790232  0.66594019 13.64515832  0.000000e+00       -Inf
education.num=education.num (12,16]        0.08677327  0.08475602 24.77503762  0.000000e+00       -Inf
education.num=education.num (10,12]        0.00000000  0.00000000  7.52126777  0.000000e+00       -Inf
education= Bachelors                       0.09337068  0.06054002 16.44605510  0.000000e+00       -Inf
age=age:25 and under                       4.85103728  3.76558905 19.68919873  0.000000e+00       -Inf


$'2'
                                            Cla/Mod     Mod/Cla     Global       p.value     v.test
education.num=education.num (10,12]     100.000000 100.00000000  7.5212678  0.000000e+00        Inf
education= Assoc-voc                    100.000000  56.43119641  4.2443414  0.000000e+00        Inf
education= Assoc-acdm                   100.000000  43.56880359  3.2769264  0.000000e+00        Inf
occupation= Tech-support                 21.443966   8.12576562  2.8500353  1.601999e-42  13.666841

education= 10th                           0.000000   0.00000000  2.8653911  6.852434e-33 -11.945515
education= 11th                           0.000000   0.00000000  3.6086115  2.146532e-41 -13.476643
education= Masters                        0.000000   0.00000000  5.2916065  6.595852e-61 -16.464543
education= Bachelors                      0.000000   0.00000000 16.4460551 3.675568e-200 -30.185462
education.num=education.num (9,10]        0.000000   0.00000000 22.3918184 1.380558e-282 -35.922099
education= Some-college                   0.000000   0.00000000 22.3918184 1.380558e-282 -35.922099
education.num=education.num (12,16]       0.000000   0.00000000 24.7750376 1.656772e-317 -38.093256
education.num=education.num [1,9]         0.000000   0.00000000 45.3118762  0.000000e+00       -Inf
education= HS-grad                        0.000000   0.00000000 32.2502380  0.000000e+00       -Inf
```

```
$'3'
                                          Cla/Mod     Mod/Cla      Global       p.value    v.test
occupation= Prof-specialty             75.217391304 39.33308071 12.71459722  0.000000e+00       Inf
education.num=education.num (12,16]    98.103384158 99.96210686 24.77503762  0.000000e+00       Inf
education= Prof-school                 99.479166667  7.23759000  1.76898744  0.000000e+00       Inf
education= Masters                     99.071387115 21.56119742  5.29160652  0.000000e+00       Inf
education= Bachelors                   97.535014006 65.97195908 16.44605510  0.000000e+00       Inf
occupation= Exec-managerial            48.647319233 24.98421119 12.48733147 2.477382e-288 36.288270
education= Doctorate                   99.515738499  5.19136036  1.26838856 5.181326e-252 33.907406
hours.per.week=hours.per.week (45,99]  36.923938927 33.29544019 21.92500230 4.348762e-163 27.214723
marital.status= Married-civ-spouse     29.787660256 56.34710117 45.99367341 5.678733e-100 21.224450
```

```
capital.gain=capital.gain (0,1]         41.556047198 14.23519010  8.32898253  2.810151e-95  20.710076
relationship= Husband                   30.038656863 50.05683971 40.51779736  7.254105e-87  19.755080
age=age:36 to 49                        30.745224135 41.06353417 32.47443260  1.354550e-76  18.522706
workclass= Local-gov                    41.662685141 11.01427308  6.42793526  3.997424e-73  18.087510
workclass= State-gov                    43.066255778  7.06075534  3.98636406  1.178661e-51  15.120930
capital.loss=capital.loss (0,4.36]      39.828834760  7.64178350  4.66509014  1.070824e-42  13.696129
workclass= Self-emp-inc                 42.114695341  5.93659214  3.42741316  2.477974e-40  13.294902
race= Asian-Pac-Islander                42.444658325  5.57029178  3.19093394  7.517808e-39  13.037189
hours.per.week=hours.per.week (40,45]   34.889434889 10.76165214  7.49976966  3.052831e-34  12.201465
sex= Male                               26.135842129 71.93381331 66.92054912  3.698872e-28  11.002936

sex= Female                             20.629468016 28.06618669 33.07945088  3.698872e-28 -11.002936
relationship= Other-relative             9.785932722  1.21258052  3.01280673  4.257818e-32 -11.792678
occupation= Farming-fishing              9.255533199  1.16205633  3.05273180  3.913086e-35 -12.367607
native.country= Mexico                   5.132192846  0.41682455  1.97475508  7.694412e-40 -13.209883
education= 5th-6th                       0.000000000  0.00000000  1.02269586  2.969871e-41 -13.452661
hours.per.week=hours.per.week 40        20.884537031 40.14146773 46.73382267  7.430188e-42 -13.554713
capital.loss=capital.loss 0             23.555183300 92.35821650 95.33490986  1.070824e-42 -13.696129
relationship= Unmarried                 14.683691236  6.39130984 10.58321305  1.805462e-48 -14.630091
occupation= Adm-clerical                14.960212202  7.12391057 11.57826848  2.755830e-50 -14.911951
education= 12th                          0.000000000  0.00000000  1.32981174  1.614022e-53 -15.400902
race= Black                             13.348271447  5.26714665  9.59429993  4.279436e-57 -15.924570
workclass=level_NA                       9.858387800  2.28621953  5.63864746  2.740295e-59 -16.237462
occupation=level_NA                      9.820944113  2.28621953  5.66014557  7.400905e-60 -16.317582
education= 9th                           0.000000000  0.00000000  1.57857560  1.741855e-63 -16.819991
education= 7th-8th                       0.000000000  0.00000000  1.98396855  8.631347e-80 -18.914698
workclass= Private                      21.197567853 60.76796766 69.70301895  2.205843e-85 -19.581912
hours.per.week=hours.per.week [1,40)    16.114904032 15.80143994 23.84140536  2.320297e-88 -19.928139
capital.gain=capital.gain 0             22.747830748 85.76480990 91.67101747  2.810151e-95 -20.710076
occupation= Handlers-cleaners            3.649635036  0.63155236  4.20748749 4.418948e-102 -21.451528
occupation= Transport-moving             4.445835942  0.89680435  4.90464052 8.188346e-108 -22.056972
education= 10th                          0.107181136  0.01263105  2.86539111 5.134651e-113 -22.592510
education= Assoc-acdm                    0.000000000  0.00000000  3.27692638 2.552338e-132 -24.477704
education= 11th                          0.000000000  0.00000000  3.60861153 6.313491e-146 -25.723919
occupation= Machine-op-inspct            3.696303696  0.93469749  6.14845981 1.233483e-149 -26.053327
relationship= Own-child                 10.793212313  6.90918277 15.56463254 1.172802e-151 -26.231156
education= Assoc-voc                     0.000000000  0.00000000  4.24434139 3.700556e-172 -27.970625
occupation= Other-service                5.250379363  2.18517115 10.11946808 3.967336e-207 -30.711741
age=age:25 and under                    10.310403993  8.34912214 19.68919873 3.589808e-215 -31.308370
occupation= Craft-repair                 6.172237131  3.19565492 12.58867971 8.374722e-233 -32.578113
education.num=education.num (10,12]      0.000000000  0.00000000  7.52126777 1.146796e-310 -37.677814
education.num=education.num (9,10]       0.013715540  0.01263105 22.39181843  0.000000e+00       -Inf
education.num=education.num [1,9]        0.013555646  0.02526209 45.31187617  0.000000e+00       -Inf
education= Some-college                  0.013715540  0.01263105 22.39181843  0.000000e+00       -Inf
education= HS-grad                       0.009522903  0.01263105 32.25023801  0.000000e+00       -Inf


$‘4‘
                                            Cla/Mod     Mod/Cla      Global      p.value     v.test
sex= Female                             39.60635039 52.12609971 33.07945088 0.000000e+00        Inf
occupation= Other-service               55.38694992 22.29960899 10.11946808 0.000000e+00        Inf
marital.status= Never-married           40.84994852 53.32355816 32.80918891 0.000000e+00        Inf
```

```
education.num=education.num [1,9]      54.96136641 99.08357771 45.31187617  0.000000e+00        Inf
education= HS-grad                     54.36625083 69.75806452 32.25023801  0.000000e+00        Inf
workclass= Private                     30.97902714 85.91153470 69.70301895  0.000000e+00        Inf
relationship= Own-child                44.45540647 27.52932551 15.56463254 5.872174e-237  32.870143
relationship= Unmarried                49.53569356 20.85777126 10.58321305 3.684415e-236  32.814277
age=age:25 and under                   41.33520512 32.38025415 19.68919873 5.704108e-226  32.091983
education= 11th                        65.44680851  9.39638319  3.60861153 8.048895e-194  29.698416
relationship= Not-in-family            37.31487056 37.86656891 25.50597340 2.770522e-183  28.870892
marital.status= Divorced               43.01147873 23.35043988 13.64515832 1.147375e-174  28.176118
relationship= Other-relative           60.55045872  7.25806452  3.01280673 1.037125e-125  23.849000
hours.per.week=hours.per.week [1,40)   35.60479196 33.77321603 23.84140536 7.647292e-125  23.765227
occupation= Handlers-cleaners          50.94890511  8.52883675  4.20748749  1.430151e-97  20.962927
education= 10th                        56.27009646  6.41495601  2.86539111  1.152393e-93  20.530393
marital.status= Separated              53.85365854  6.74486804  3.14793772  1.483758e-88  19.950507
race= Black                            40.04481434 15.28592375  9.59429993  4.069816e-83  19.314330
education= 12th                        68.36027714  3.61681329  1.32981174  1.287134e-80  19.014765
marital.status= Widowed                52.16515609  6.32942326  3.04966064  1.462592e-76  18.518574
capital.gain=capital.gain 0            26.35599183 96.12658847 91.67101747  1.735854e-73  18.133430
occupation= Machine-op-inspct          42.60739261 10.42277615  6.14845981  1.269772e-69  17.637490
occupation= Priv-house-serv            83.22147651  1.51515152  0.45760265  3.127251e-50  14.903507
occupation= Adm-clerical               35.19893899 16.21456500 11.57826848  9.638997e-49  14.672728
education= 9th                         50.97276265  3.20136852  1.57857560  1.183171e-36  12.645601
native.country= Mexico                 46.34525661  3.64125122  1.97475508  7.908572e-32  11.740427
capital.loss=capital.loss 0            25.69422073 97.45845552 95.33490986  4.824292e-29  11.185088

capital.loss=capital.loss (0,4.36]     13.69321922  2.54154448  4.66509014  4.824292e-29 -11.185088
hours.per.week=hours.per.week (40,45]  15.47911548  4.61876833  7.49976966  3.893110e-33 -11.992431
workclass= Self-emp-inc                 8.33333333  1.13636364  3.42741316  1.680599e-48 -14.634966
education= Doctorate                    0.00000000  0.00000000  1.26838856  4.942120e-53 -15.328373
workclass= Self-emp-not-inc            12.63282172  3.92228739  7.80381438  8.660948e-59 -16.166704
race= White                            23.46491228 79.75317693 85.42735174  6.862918e-60 -16.322190
capital.gain=capital.gain (0,1]        11.68879056  3.87341153  8.32898253  1.735854e-73 -18.133430
education= Prof-school                  0.00000000  0.00000000  1.76898744  6.836459e-74 -18.184590
age=age:36 to 49                       18.87648950 24.38905181 32.47443260  2.386228e-75 -18.367626
education= Assoc-acdm                   0.00000000  0.00000000  3.27692638 1.761802e-137 -24.957694
occupation= Exec-managerial             9.73930152  4.83870968 12.48733147 1.016006e-152 -26.324108
hours.per.week=hours.per.week (45,99]  13.18111780 11.49804497 21.92500230 2.691090e-169 -27.734353
education= Assoc-voc                    0.00000000  0.00000000  4.24434139 6.897346e-179 -28.518556
workclass=level_NA                      0.65359477  0.14662757  5.63864746 3.106546e-214 -31.239437
education= Masters                      0.23215322  0.04887586  5.29160652 1.813750e-214 -31.256640
occupation=level_NA                     0.65111232  0.14662757  5.66014557 3.732362e-215 -31.307127
occupation= Prof-specialty              4.32367150  2.18719453 12.71459722 1.279227e-316 -38.039598
sex= Male                              17.98072510 47.87390029 66.92054912  0.000000e+00       -Inf
relationship= Husband                   0.29561131  0.47653959 40.51779736  0.000000e+00       -Inf
marital.status= Married-civ-spouse      4.18669872  7.66129032 45.99367341  0.000000e+00       -Inf
education.num=education.num (12,16]     0.90492128  0.89198436 24.77503762  0.000000e+00       -Inf
education.num=education.num (10,12]     0.00000000  0.00000000  7.52126777  0.000000e+00       -Inf
education.num=education.num (9,10]      0.02743108  0.02443793 22.39181843  0.000000e+00       -Inf
education= Some-college                 0.02743108  0.02443793 22.39181843  0.000000e+00       -Inf
education= Bachelors                    1.28851541  0.84310850 16.44605510  0.000000e+00       -Inf
```

```
$`5`
                                          Cla/Mod     Mod/Cla      Global       p.value     v.test
sex= Female                            30.3314455 56.79763561 33.07945088  0.000000e+00        Inf
relationship= Own-child                38.7726914 34.16203060 15.56463254  0.000000e+00        Inf
occupation=level_NA                    68.5295714 21.95757997  5.66014557  0.000000e+00        Inf
marital.status= Never-married          30.8714780 57.33657858 32.80918891  0.000000e+00        Inf
education.num=education.num (9,10]     68.6462762 87.01321280 22.39181843  0.000000e+00        Inf
education= Some-college                68.6462762 87.01321280 22.39181843  0.000000e+00        Inf
workclass=level_NA                     68.4095861 21.83588317  5.63864746  0.000000e+00        Inf
age=age:25 and under                   38.3091561 42.69819193 19.68919873  0.000000e+00        Inf
hours.per.week=hours.per.week [1,40)   31.0704625 41.93324061 23.84140536 1.126558e-250  33.816549
occupation= Adm-clerical               31.0875332 20.37552156 11.57826848 1.091346e-102  21.516482
relationship= Not-in-family            25.1053582 36.24826147 25.50597340  5.294236e-89  20.001967
marital.status= Divorced               26.9187486 20.79276773 13.64515832  6.831623e-62  16.601196
relationship= Unmarried                26.2333140 15.71627260 10.58321305  1.988412e-40  13.311355
capital.gain=capital.gain 0            18.4361285 95.67107093 91.67101747  1.252738e-38  12.998190
```

```
occupation= Machine-op-inspct              8.8411588  3.07719054  6.14845981  8.116645e-31 -11.541844
occupation= Exec-managerial               11.3871126  8.04937413 12.48733147  6.427546e-32 -11.757951
education= Doctorate                       0.4842615  0.03477051  1.26838856  3.276087e-32 -11.814728
workclass= Self-emp-inc                    5.8243728  1.13004172  3.42741316  1.331827e-32 -11.890140
occupation= Transport-moving               6.9505322  1.92976356  4.90464052  3.263567e-37 -12.746429
capital.gain=capital.gain (0,1]            9.1814159  4.32892907  8.32898253  1.252738e-38 -12.998190
workclass= Self-emp-not-inc                8.4218811  3.72044506  7.80381438  2.778885e-43 -13.793760
education= Prof-school                     0.5208333  0.05215577  1.76898744  2.751401e-44 -13.959557
workclass= Private                        15.6150864 61.61335188 69.70301895  1.909209e-47 -14.468746
age=age:50 to 64                           9.9022005  9.85744089 17.58545499  4.233744e-72 -17.956959
occupation= Craft-repair                   8.1727251  5.82406120 12.58867971  5.043468e-76 -18.451802
education= Assoc-acdm                      0.0000000  0.00000000  3.27692638  1.815492e-92 -20.395978
age=age:36 to 49                          11.6039342 21.33171071 32.47443260  2.716802e-93 -20.488676
occupation= Prof-specialty                 6.9565217  5.00695410 12.71459722 3.942803e-100 -21.241594
hours.per.week=hours.per.week (45,99]      9.1049167 11.30041725 21.92500230 5.727857e-115 -22.790257
education= Assoc-voc                       0.0000000  0.00000000  4.24434139 3.281160e-120 -23.312888
education= Masters                         0.5803831  0.17385257  5.29160652 3.498200e-131 -24.370702
education.num=education.num (10,12]        0.0000000  0.00000000  7.52126777 1.500190e-216 -31.409517
sex= Male                                 11.4043139 43.20236439 66.92054912  0.000000e+00       -Inf
relationship= Husband                      1.5386948  3.52920723 40.51779736  0.000000e+00       -Inf
marital.status= Married-civ-spouse         4.0464744 10.53546592 45.99367341  0.000000e+00       -Inf
education.num=education.num (12,16]        0.9049213  1.26912378 24.77503762  0.000000e+00       -Inf
education.num=education.num [1,9]          4.5682527 11.71766342 45.31187617  0.000000e+00       -Inf
education= HS-grad                         3.3425388  6.10222531 32.25023801  0.000000e+00       -Inf
education= Bachelors                       1.0830999  1.00834492 16.44605510  0.000000e+00       -Inf
```

2. Prediction errors with different processing of outliers

- Initial results (without taking care of outliers in C4.5 and using weights in CART)

C4.5: train:12.04%   test:17.04%

CART: train 14.65% Test: 15.94%

- Results when training without the outliers

C4.5

Training error without outliers: 11.75% error

Testing error without outliers: 17.35%

Testing error with outliers: 17.05%

CART:

Training error without outlier: 14.92%

Testing error without outlier: 16.22%

Testing error with outliers: 16.20%

3. R code

```
library(FactoMineR)
library(cluster)
library(class)
library(gtools)
library(xtable)

########################################
#        Read the data                 #
########################################
set.seed(10062014)
potec <-
  read.table("Adult.txt", header=FALSE, sep=",", na.strings="NA", dec=".")
names(potec)<-
c('age','workclass','fnlwgt','education','education.num','marital.status','occupation'
,'relationship','race','sex','capital.gain','capital.loss','hours.per.week','native.co
untry','target')

########################################
#        Pre-Processing                #
########################################
############## NA Values
#originally ' ?' is a level, so assigned to NA
for (i in 1:15) { potec[potec[,i]==' ?',i]<-NA}
summary(potec)

##############  Outliers detection
#this function is used to compute the initial mahalanobis distance
computeDistances <- function(x,G,V)
{
  lx <- x
  lg <- G
  lv <- V
  s <- svd(lv)
  D <- diag(1/s$d)
  linv <- s$v %*% D %*% t(s$u)
  distances <- seq(0,by=0, length = nrow(lx))
  for(i in 1:nrow(lx))
  {
    xi_minus_g <- as.matrix(lx[i,] - lg)
    maha_dist <- (xi_minus_g %*% linv) %*% t(xi_minus_g)
    distances[i] <- sqrt(maha_dist)
  }
  distances
}

# This function is used to compute the robust mahalanobis distance
loop.mahalanobis <- function (Dataset) {
  Bool <- FALSE
  s<-svd(cov(Dataset))
  D<-diag(1/s$d)
  Cov_inv <- s$v%*%D%*%t(s$u)
  Dm <- rep(0, nrow(Dataset))
  means <- colMeans(Dataset)



 n <- length(Dm)
  h <- round(0.75*n)
```

```
  Matrix <- Dataset

  while (n>2 && Bool == FALSE)
  {

    for (i in 1:n)
    { centralised  <- as.matrix(Matrix[i,] - means)
      mahasq <- centralised %*%  Cov_inv %*% t(centralised)
      Dm[i] <- sqrt(abs(mahasq))
    }
    Sorted_Dm <- sort.int(Dm, decreasing=TRUE,index.return=TRUE)
    New_index <- Sorted_Dm$ix[1:h]
    New_matrix <- Dataset[New_index,]

    s <- svd(cov(New_matrix))
    D <- diag(1/s$d)
    Cov_inv_new <- s$v %*% D %*% t(s$u)

    means_new <- colMeans(New_matrix)
    n <- h
    h <- round(0.75*n)

    if (Cov_inv_new == Cov_inv && means_new == means)
    {Bool <- TRUE}
    else {
      Cov_inv <- Cov_inv_new
      means <- means_new
      Matrix <- New_matrix
    }
  }
  return(Dm)
}

x <- potec[,c(1,3,5,11,12,13)]   #get only numeric columns
G <- as.matrix(colMeans(x))
V <- cov(x)

initial.distances <- computeDistances(x,G,V)
DMahalanobis.robust <- loop.mahalanobis(x)

#plot with outlier detection
plot(initial.distances, DMahalanobis.robust)
h = qchisq(.975,df=5)
abline(h = h, lty = 2, col = "red")
abline(v = h, lty = 2, col = "red")
outliers <- which(DMahalanobis.robust > h)

# those are the outliers
outliers<-
c(24511,24639,24674,24851,24984,25179,25373,25612,25634,25842,26084,26415,26443,26594,
26826,27078,27222,27359,27414,27636,27641,28055,28215,28265,28295,28319,28350,29636,29
807,30245,30497,30914,31112,31829,31973,32091,32239,32519)

# the weights are changed accordingly
w<-rep(1,dim(potec)[1])
w[outliers]<-0.00001
```

```
#Summary of outliers
summary(potec[outliers,c(1,3,5,11,12,13)])
#Summary of non outliers
summary(potec[-outliers,c(1,3,5,11,12,13)])
#capital gains of outliers
median(potec[outliers,c(11)])
#capital gains of nonoutliers
median(potec[-outliers,c(11)])

#######################  Factorization

cont<-c(3,5,11,12,13)  #the continuous variables that will be split into quartiles
for (i in cont){potec[,i]<-quantcut(potec[,i])}
for(i in cont) levels(potec[,i]) <- paste(colnames(potec)[i],levels(potec[,i]))

# Age variable
potec[,1]<- cut(potec$age, breaks = c(0,25,35,49,64,90))
levels(potec[,1])<-c("age:25 and under", "age:26 to 35", "age:36 to 49", "age:50 to 
64", "age:65 and up")


##################### Dealing with NA values
naval<-c(2,7,14)  # the variables containing NA values

# will set NA as a level: level_NA
for (i in naval){
  potec[,i]<- factor(potec[,i], levels = c("level_NA",levels(potec[,i])[-1]))
  potec[is.na(potec[,i]),i]<-'level_NA'
}
summary(potec)

########################################
#              MCA                     #
########################################

illus=c(15) #all the variables (except the target) are active ones
res.mca <- MCA(potec, quali.sup=illus,row.w=w)  # MCA with weights (outliers)

#####################  Plots
#plot of everything
plot(res.mca,label=c("var","quali.sup","quanti.sup"))   # too loaded

plot(res.mca,invisible=c("ind","quanti.sup","quali.sup"),autoLab="y",cex=0.7) # plot 
active variables
plot(res.mca,invisible=c("ind"),cex=0.7, selectMod="contrib 20", unselect="grey70") # 
20 variables contributed most
plot(res.mca,invisible=c("ind"),autoLab="y",cex=0.7,selectMod="cos2 
20",unselect="grey70") # 20 Variables most correlated
plot(res.mca, invisible=c("ind","var")) # illustrative variable (target modalities)

#plot of individuals  (not in the report)
plot(res.mca,invisible=c("var","quali.sup"),autoLab="y",cex=0.7) # individus




##################### Description of dimensions
```

```
dimdesc(res.mca)

##################### Eigenvalues
plot(res.mca$eig$eigenvalue, type="l") #plot of eigenvalues according to dimensioins
abline(v = 51, lty = 2, col = "grey70")

res.mca$eig[res.mca$eig[1]>1/14,,]
      #51 dimensions  ! keep eig > 1/(Number actives variables)..  # so nd=51

sum(res.mca$eig[1])/109  #109 is the total should be the total number of dimensions
res.mca$eig[res.mca$eig[1]>0.07142857,,] # also 51 dimensions with this rule  (mean)

########################################
#           Clustering                 #
########################################
res.mca2 <- MCA(potec, ncp=51, quali.sup=illus,row.w=w) # redo MCA with 51 dimensionns
kept
Psi<-res.mca2$ind$coord[,1:51] # Projetion of individuals on 51 kept dimensions

# CLUSTERING OF LARGE DATA SETS
#################### FIRST 2 KMEANS WITH K=12
n1 = 12   # arbitrary: can be changed
k1 <- kmeans(Psi,n1)
k2 <- kmeans(Psi,n1)
table(k2$cluster,k1$cluster)
clas <- (k2$cluster-1)*n1+k1$cluster
summary(clas) # 144 clusters (cross table of k1 and k2)
freq <- table(clas) # number of elts in each cluster
cdclas <- aggregate(as.data.frame(Psi),list(clas),mean)[,2:52] #52=nd+1, center of
gravity of cells of the cross table

##################### SECOND HIERARCHICAL CLUSTERING UPON THE CENTROIDS OF CROSSING
THE 2 KMEANS PARTITIONS
d2 <- dist(cdclas)
h2 <- hclust(d2,method="ward",members=freq) # Tree with ward criteria
plot(h2)
barplot(h2$height[(nrow(cdclas)-50):(nrow(cdclas)-1)]) # plot of the last 50
aggregations
nc = 5  # cut after 4th jump, will keep 5 clusters
c2 <- cutree(h2,5)  # cut the tree accordingly
cdg <- aggregate((diag(freq/sum(freq)) %*% as.matrix(cdclas)),list(c2),sum)[,2:52] #
final center of gravity of clusters

##################### Plot of clustering
plot(Psi[,1],Psi[,2],type="n",main="Clustering of individuals in 5 classes")
text(Psi[,1],Psi[,2],col=c2,cex = 0.6)
abline(h=0,v=0,col="gray")
legend("topright",c("c1","c2","c3","c4","c5"),pch=20,col=c(1:5))

# to help vizualising (not in the report) plot of the individuals colored according to
the target.
plot(Psi[,1],Psi[,2],type="n",main="target distribution")
text(Psi[,1],Psi[,2],col=unclass(potec[,15]),cex = 0.6)
legend("topright",levels(pote[,15]),pch=20,col=c(1:2)); abline(h=0,v=0,col="gray")
```

```
##################### CONSOLIDATION
k6 <- kmeans(Psi,centers=cdg)
k6$size  #size of the clusters 8259 2449 7917 8184 5752

##################### plot of the consolidatedclustering
plot(Psi[,1],Psi[,2],type="n",main="Clustering of individuals in 5 classes")
text(Psi[,1],Psi[,2],col=unclass(k6$cluster),cex = 0.6)
abline(h=0,v=0,col="gray")
legend("topright",c("c1","c2","c3","c4","c5"),pch=20,col=c(1:5))

##################### Description of clusters
potec.comp = cbind.data.frame(potec,k6$cluster) # A dataset with the cluster
assignment
potec.comp[,16]<-as.factor(potec.comp[,16])
pot<-potec.comp[,-c(15)] # don't want to describe the target with our clustering
(predicted variable)
CatDes<-catdes(pot,num.var=15)
CatDes$category

########################################
#           Prediction                 #
########################################
library(party)
library(RWeka)
library(partykit)
library(caret)
library(e1071)
library(rpart)
################### Split data into Training/Testing set
N<-dim(potec)[1]
learn<-sample(1:N,round(2*N/3))
nlearn<-length(learn)
ntest<-N-nlearn
################### Parameter optimization C4.5
c_sample <-seq(0.05,0.50,by=0.05)
#length(c_sample)

# create fixed sampling scheme (10-folds)
train <- createFolds(potec$target, k=10)

### (Prediction Tree)  the fitting of parameters will be done on train set, using 10
fold CV
C45Fit <- train(potec[learn,-15], potec[learn,15], "J48",
                tuneLength = 10,
                tuneGrid=expand.grid(.C=c_sample),
                trControl = trainControl(
                  method = "cv", indexOut = train, repeats=10))

plot(C45Fit$results[,1],C45Fit$results[,2],type='l') # acuracy according to tested
parameters
C45Fit  # accuracy keep increasing with C, so final model kept: 0.5
C45Fit$results   # table accuracy/parameters
C45Fit$bestTune  # best parameter
C45Fit$finalModel # can see whole description of tree (it is quite long)



#################### Build the model C4.5
```

```
# at this point we use the best parameter (c=0.5) to make a prediction with all
training data (no more CV)
model.tree<-J48(target~., data=potec, subset=learn, control=Weka_control(C=0.5),
na.action=NULL)

#################### Errors C4.5
# Training sample
pred.learn<-predict(model.tree, data=potec[learn])
tab<-table(pred.learn,potec$target[learn]) # contingency matrix
(error.learn<-100*(1-sum(diag(tab))/nlearn))  #12%  => This model is selected

# Test sample
pred.test<-predict(model.tree, newdata=potec[-learn,])#subset=-learn
tab<-table(pred.test,potec$target[-learn]) # contingency matrix
(error.test<-100*(1-sum(diag(tab))/ntest))  #17%

####################### Parameter optimization with CART
learndata <- potec[learn,]
cp_sample <-seq(0.001,0.01,by=0.0005)
#length(cp_sample)

train <- createFolds(potec$target, k=10)
rpart.fitted <- train(potec[learn,-15], potec[learn,15], "rpart", weights=w[learn],
                      tuneLength = 19,
                      tuneGrid=expand.grid(.cp=cp_sample),
                      trControl = trainControl(method = "cv", indexOut = train,
repeats=10))

rpart.fitted$results  # parameters/accuracy table
rpart.fitted$bestTune  #best parameter cp = 0.001
rpart.fitted$finalModel # can see whole description of tree

######################## Build model CART
p1 <- rpart(target ~ ., data=learndata,control=rpart.control(cp=0.001),
weights=w[learn])

######################## Plot a tree CART (no same complexity parameter because the
tree is too big)
#p1 <- rpart(target ~ ., data=learndata,control=rpart.control(cp=0.05),
weights=w[learn])
plot(rpart.fitted$results[,1],rpart.fitted$results[,2],type='l')
plot(as.party.rpart(p1),type="extended")

######################## Errors CART
#Training sample
pred.learn<-predict(p1, data=learndata, type="class")
tab<-table(pred.learn,learndata$target)
(error.learn<-100*(1-sum(diag(tab))/nlearn))  #15%  => This model isn't selected

# Testing sample
pred.test<-predict(p1, newdata=potec[-learn,], type="class")
tab.test<-table(pred.test,potec$target[-learn])
(error.test<-100*(1-sum(diag(tab.test))/ntest))  #16%
```