UNIVERSITÀ
DEGLI STUDI
DI MILANO

OECD Countries Happines Index Clustering

An Unsupervised Learning Approach

Filippo Menegatti*

*Data Science and Economics, University of Milan

Contents

1 Introduction **3**

2 Analysis **3**

2.1 Principal Component Analysis 3

2.2 Hierarchical Clustering 4

2.3 K-Means Clustering 4

3 Analysis and discussion **5**

4 Results **11**

5 Table **13**

6 Bibliography **13**

7 Code **14**

1 Introduction

In this essay we are going to analyze the numerical variables used to calculate the Better Life Index in the OECD countries, including also some relevant Non-OECD ones like Brazil, Russia and South Africa. The data set - available also on Kaggle website [1] - is composed by 39 individuals (including a general OECD Total row) and 24 variables which represent different social or environmental characteristics of each country, such as the percentage of housing expenditure, the average student skills, the air pollution in UGM3 and many others[1].

We are going to analyze the characteristics of the data set using different *Unsupervised Learning* methods like Principal Component Analysis, K-Means Clustering and Hierarchical Clustering. In this way we are going to detect patterns which connect the "more similar" countries in order to finally confront them with the real ranking made by OECD.

2 Analysis

2.1 Principal Component Analysis

The principal component analysis (PCA) [2] is a statistical method mainly used to reduce the dimension of a given matrix, mainly to be able to plot the the individuals and the variables using a 2-dimensional graph, controlling the loss of information. This is made possible by the calculation of the *principal components*: normalized linear combinations of the features, orthogonal one with the other. We consider specifically the two which are associated with the highest variance, namely the first and second principal components.

Assuming that all the variables are normalized, to find the principal components (for example the first one) it is necessary to solve the following optimization problem:

[1] The complete list is available at the end of the essay.

$$\max \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} xij \right)^2 \right\} \text{subject to} \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

The loadings ϕ_j are used to define the principal components:

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} x_{ip}$$

The loading vector ϕ_1 with all its elements can be used to describe the direction of maximum variation of the data in the feature space.

So thanks to PCA it is possible to visualize the data in a new coordinate set with the first two principal components on the axis, giving a nice overview of the correlation between them and the features.

2.2 Hierarchical Clustering

The Hierarchical Clustering [3], as the name suggests, is a clustering method which can be used to find patterns in the data and represent them in a tree-like plot called *dendrogram*. To obtain the result a dissimilarity measure like Euclidean distance is chosen and all the N observation are treated as single clusters. Iteratively, the two "most similar" clusters are merged together until there is just one cluster left. When there is more than one observation in a cluster a *linkage method* is needed to choose in what way the clusters have to be merged. There are many available procedures like the *complete*, *average*, *single*, *centroid.* The average linkage method is going to be used: it computes all the pairwise dissimilarities between the observations of two clusters and uses the mean.

2.3 K-Means Clustering

The K-Means Clustering is another famous unsupervised method useful to perform the clustering of data. Differently from the Hierarchical Clustering, here we have to choose the number of clusters (or cluster centers) we want to obtain. The initial centers are generally randomly chosen, then an optimization process is performed in order to optimize the position of the

centroids with respect to the data points. To choose the "optimal" K there are many methods available, and we are going to explicitly use three of them: the *elbow method*, the *silhouette method* and the *gap statistic* [4].

- Elbow Method: A graph with the calculated total within sum of squares for each K is created and the *kink*, so the point with the highest variation, is chosen
- Silhouette Method: The Silhouette score of the form $s(i) = \dfrac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, where $a(i)$ is the is the *average distance* between i and all the other data points in the cluster to which i belongs and $b(i)$ is the *minimum average distance* from i to all clusters to which i does not belong. The K correspondent to the highest average score is selected.
- Gap Statistic: The Gap statistic, of the form $Gap(k) = \dfrac{1}{B}\sum_{b=1}^{B} \log(W_{kb}) - \log(W_k)$, where W_k is the within intra-cluster variation for the k-th cluster and W_{kb} is the total within intra-cluster variation of an artificial set of data of cardinality B with a random uniform distribution. The optimal number of cluster is chosen in this way: $Gap(k) \geq Gap(k+1) - \sigma_{k+1}$, so the $Gap(k)$ is within one standard deviation with respect to $Gap(k+1)$.

3 Analysis and discussion

As a first step of the analysis, after the normalization process, the Pearson correlation is calculated in order inspect a first linear relation between the variables.

The correlation matrix represented follows a quite logical pattern, for example variables like long term unemployment are highly positively related with the net financial wealth while the life expectancy is negatively related with the lack of basic facilities.It is useful to show us the first superficial patterns which characterize the data.

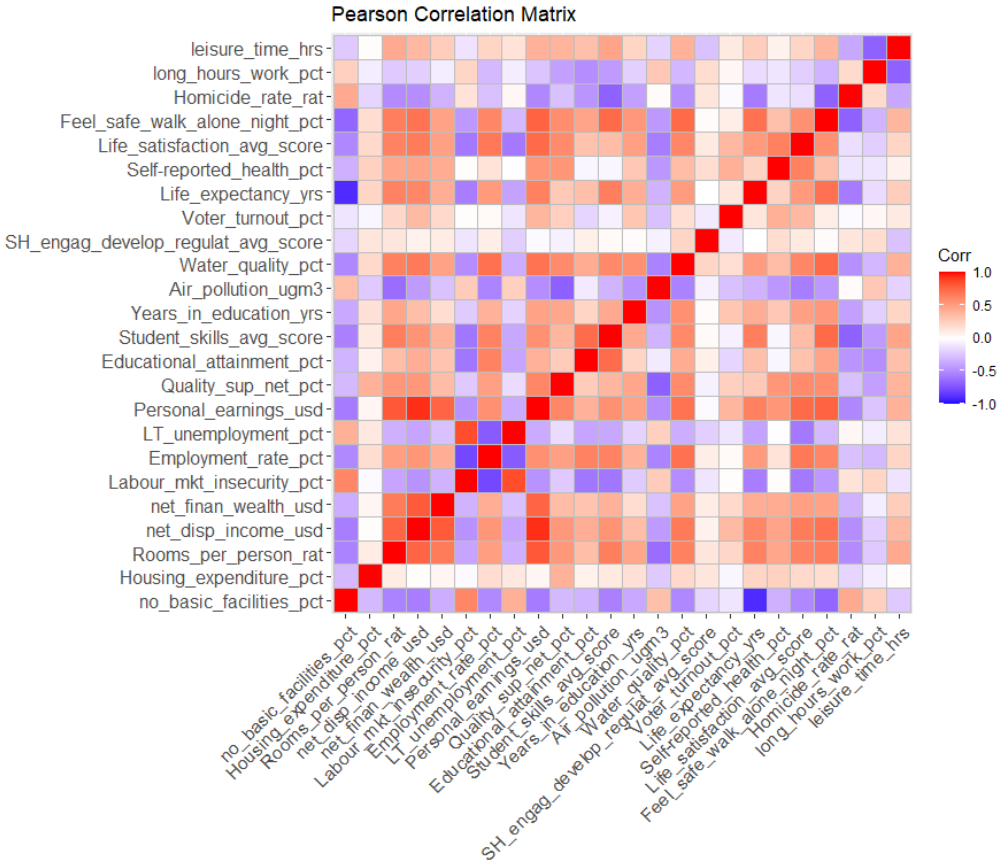


Figure 1: Matrix of the Pearson Correlation

Then a Principal Component Analysis is attempted in order to go deeper in the variables' analysis and to facilitate the graphical representation of our results.

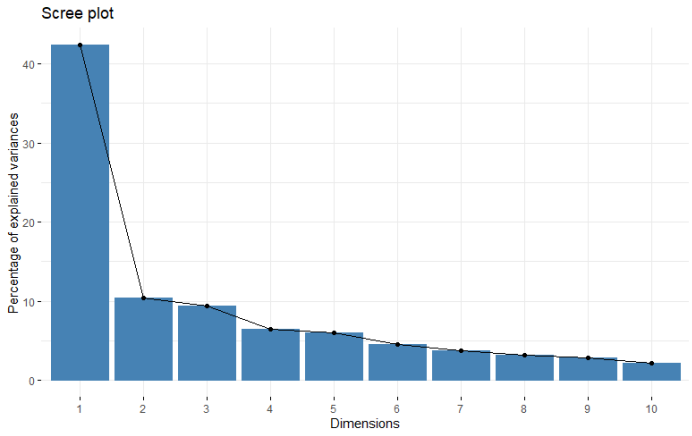


Figure 2: Variance Explained by PCA Dimensions

As we can see in Figure 2, where the variance explained by the first ten dimensions is represented, the first two dimensions together are able to describe about 53% of the total variance, so limiting our analysis to them we accept a loss of 47% of information.

In figure 3 and 4 we can see two representations of the individuals in a 2-dimensional plot with the Principal Components represented on the axes.

Analyzing the variables plot we can obtain the direction and the size of their contribution to the principal components. The "negative" variables like pollution, homicide rate, etc. have - as logic can suggest - an opposite direction with respect to the "positive" ones and are also negatively related with the first principal components. The second component seems to contain information more related with labor market and education of the population. From the individual plot we can easily identify South Africa as having different characteristics with respect to the majority of the other countries.

Next to the PCA dimensional reduction it is possible to apply a clustering algorithm to group countries relying on their characteristics. The Hierarchical Clustering algorithm described in the previous section permits to

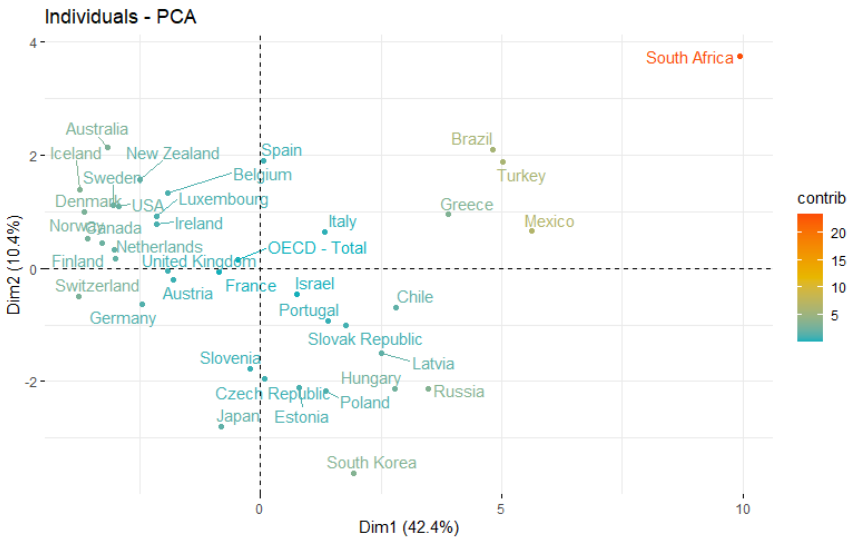


Figure 3: PCA Representation of Individuals

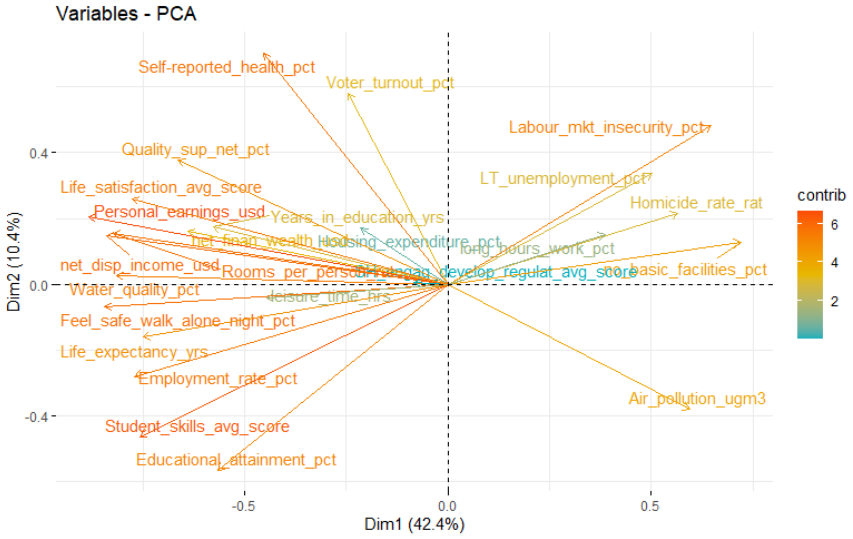
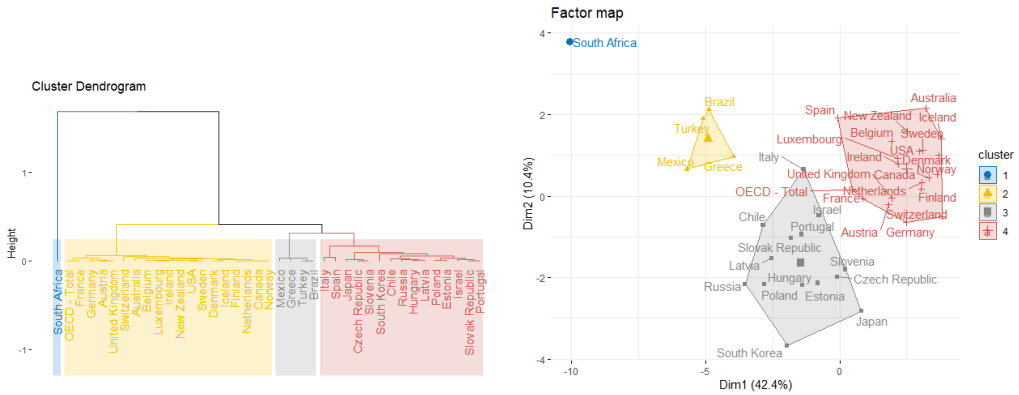


Figure 4: PCA Representation of Variables

automatically optimize the height of the dendrogram and consequently the number of clusters to maintain.



The two resulting plots show four main groups of countries including a singleton cluster containing only South Africa. The two big clusters are characterized by high values of the positive observations and low values of the negatives: high employment rates, low murder rates, high incomes, and so on. The other two clusters are instead characterized by the opposite results.

In order to control the result using another method, the K-Means clustering is going to be applied. An important step in the application of this algorithm is the process of choice of the number of clusters. In this essay we are going to apply the three most famous methods: Elbow method, Silhouette method and Gap method. Then to generalize more we are going to apply the `NbClust` function, which returns the result of 30 indexes and suggests the optimal number of clusters.

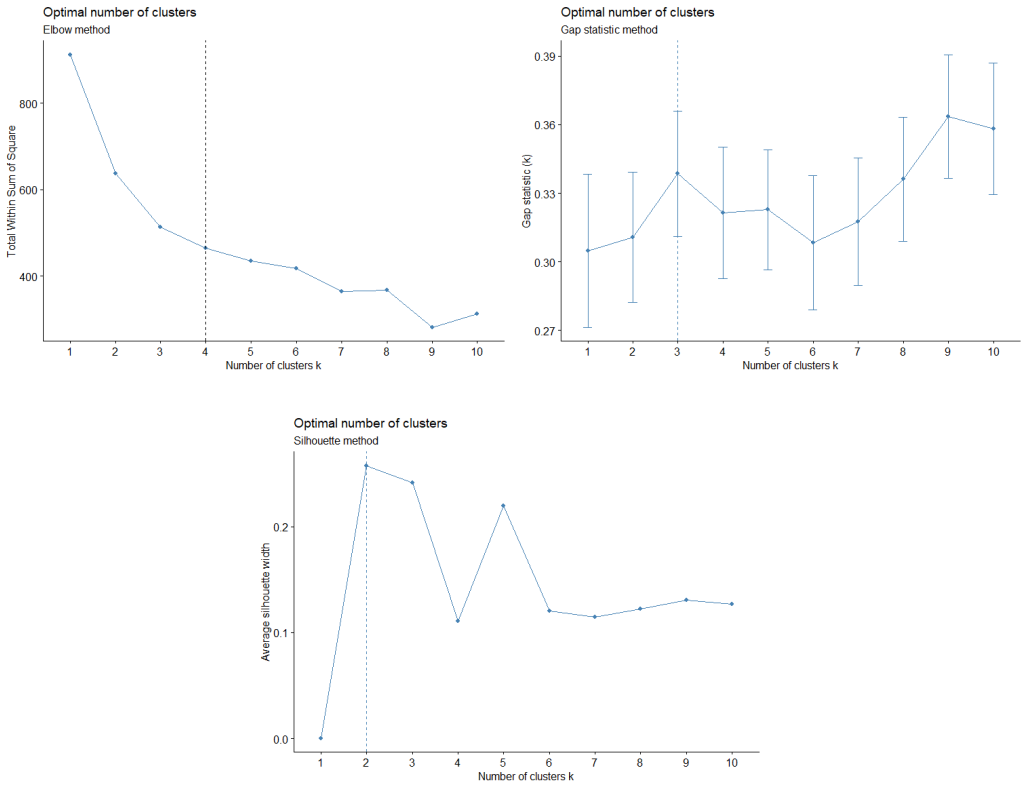


Figure 5: From upper left: Elbow Method, Gap Statistic and Silhouette Method

The resulting choice, based on the majority rule, suggests the use of 3 clusters. The algorithm includes South Africa in the first cluster, even if its results are more extreme with respect to the other members.

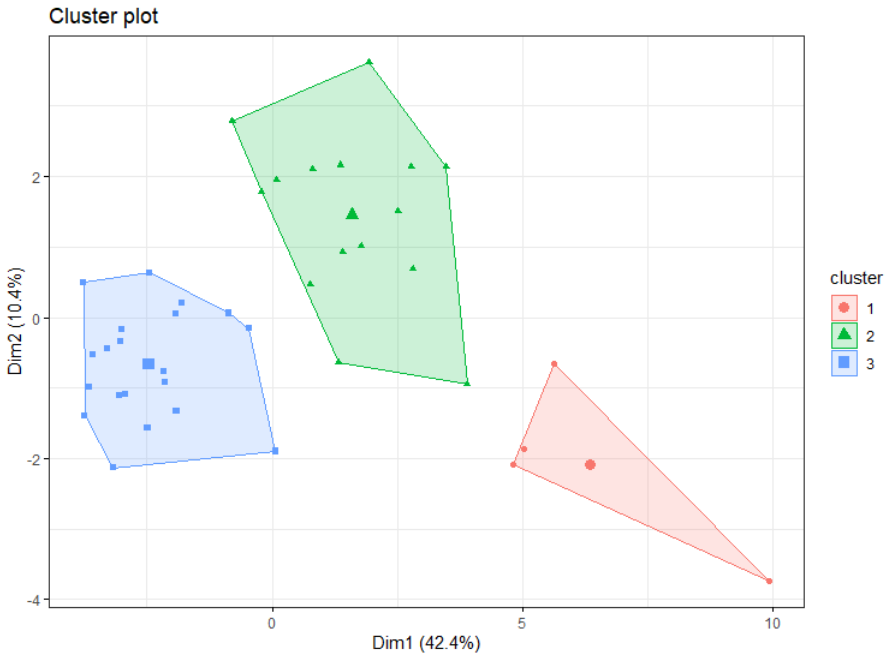


Figure 6: K-Means Clusters

4 Results

Using the different unsupervised learning methods, we are able to isolate the characteristics of the different countries available in the data set and to provide an optimal graphical representation of the clusters.

The majority of the countries are, as logic and data suggest, characterized by an optimal level of the variables needed to obtain an high Happiness Index. While Brazil, Turkey, Mexico and Greece have slightly lower scores, South Africa is identified as an extremely bad place to live in.

Principal Component Analysis, Hierarchical Clustering are so two excellent methods which can be combined in order to identify patterns and similarities in groups of individuals. Also the addition of the K-Means Clustering method confirmed the previous results, but suggesting only 3 clusters, including South Africa in the "unhappy cluster" with the other countries in the analysis. Our clustering division is robust with the ranking made by the OECD [5] and so the algorithms used permitted to have a consistent result and a good

representation of the most similar countries with just a few lines of code.

5 Table

Table 1: List of variables

Variables	
no_basic_facilities_pct	Air_pollution_ugm3
Housing_expenditure_pct	Water_quality_pct
Rooms_per_person_rat	SH_engag_develop_regulat_avg_score
net_disp_income_usd	Voter_turnout_pct
net_finan_wealth_usd	Life_expectancy_yrs
Labour_mkt_insecurity_pct	Self-reported_health_pct
Employment_rate_pct	Life_satisfaction_avg_score
LT_unemployment_pct	Feel_safe_walk_alone_night_pct
Personal_earnings_usd	Homicide_rate_rat
Quality_sup_net_pct	long_hours_work_pct
Educational_attainment_pct	leisure_time_hrs
Student_skills_avg_score	Years_in_education_yrs

6 Bibliography

[1] Kaggle Dataset

[2] "An Introduction to Statistical Learning" - G. James Et Al. (2017)

[3] "Elements of Statistical Learning" - T. Hastie Et Al. (2009)

[4] "Estimating the number of clusters in a data set via the gap statistic" - R. Tibshirani Et Al. (2000)

[5] Better Life Index Ranking

7 Code

```
library(factoextra)
library(NbClust)
library(readr)
library(FactoMineR)
library(ggplot2)
library(ggcorrplot)

OECD <- read_csv("OECDBLI2017cleanedcsv.csv")
names <- OECD$Country
OECD$Country <- NULL
rownames(OECD) <- names
colnames(OECD) <- gsub(" ", "_", colnames(OECD))
colnames(OECD) <- c("no_basic_facilities_pct" ,
                    "Housing_expenditure_pct",
                    "Rooms_per_person_rat",
                    "net_disp_income_usd",
                    "net_finan_wealth_usd",
                    "Labour_mkt_insecurity_pct",
                    "Employment_rate_pct",
                    "LT_unemployment_pct",
                    "Personal_earnings_usd",
                    "Quality_sup_net_pct",
                    "Educational_attainment_pct",
                    "Student_skills_avg_score",
                    "Years_in_education_yrs",
                    "Air_pollution_ugm3",
                    "Water_quality_pct",
                    "SH_engag_develop_regulat_avg_score",
                    "Voter_turnout_pct",
                    "Life_expectancy_yrs",
                    "Self-reported_health_pct",
                    "Life_satisfaction_avg_score",
                    "Feel_safe_walk_alone_night_pct",
```

```
                    "Homicide_rate_rat",
                    "long_hours_work_pct",
                    "leisure_time_hrs")

OECD <- scale(OECD)

ggcorrplot(cor(OECD), lab = TRUE,
  title = 'Pearson Correlation Matrix', lab_size = 2.5)

pca <- princomp(OECD)

fviz_eig(pca)

comp <- data.frame(pca$scores[,1:2])
# Plot
plot(comp, pch=16, col=rgb(0,0,0,0.5))

fviz_pca_ind(pca,
             col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
)

fviz_pca_var(pca,
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
)


# Eigenvalues
eig.val <- get_eigenvalue(pca)
eig.val

# Results for Variables
```

```
res.var <- get_pca_var(pca)
res.var$coord
res.var$contrib
res.var$cos2

# Results for individuals
res.ind <- get_pca_ind(pca)
res.ind$coord
res.ind$contrib
res.ind$cos2

fviz_nbclust(OECD, kmeans, method = "wss")+
  geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")

fviz_nbclust(OECD, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")

set.seed(33)

fviz_nbclust(OECD, kmeans, nstart = 25,
  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")

NbClust(data = OECD, distance = "euclidean",
        min.nc = 2, max.nc = 15, method = 'kmeans')

KM <- kmeans(OECD, centers = 3, iter.max = 10, nstart = 25)

res.pca <- PCA(OECD, ncp = 2, graph = FALSE)
# Compute hierarchical clustering on principal components
res.hcpc <- HCPC(res.pca, graph = FALSE,
  nb.clust = -1, method = 'average')

fviz_dend(res.hcpc,
```

```
          cex = 0.8,
          palette = "jco",
          rect = TRUE, rect_fill = TRUE,
          rect_border = "jco",
          labels_track_height = 0.8
)

fviz_cluster(res.hcpc,
             repel = TRUE,
             show.clust.cent = TRUE,
             palette = "jco",
             ggtheme = theme_minimal(),
             main = "Factor map"
)
```