Clustering and Dimension Reduction

Jordi Tarroch Mejón

***Abstract*—Hotels have multiple ways to increase profits using data science but is not so clear how. Targeting the right clients who will probably cancel their reservations less and who will fill most of the variety rooms the hotel offers are strategies that could be designed well using unsupervised learning techniques.**

- **To target bookings whose cancellation ratio is the minimum, multiple clustering techniques were developed to select the best one. They use the hotel's Property Management Systems data and train a clustering model to group clients in groups whose booking's cancellation ratio is different.**
- **To target clients' nationalities that will most likely want certain rooms and to know also distribution channels that will most likely target certain rooms, a correspondence analysis was developed. It finds the relationship between the two categories studied. This would help the hotel to target the right nationality and distribution channel to fill the rooms they need.**

Results show different cancellation ratios between the groups created showing that if a client's booking features fall into a certain group, the odds of cancellation increase or decrease. Results also show a relationship between each type of room and a nationality, as well as the relationship between a type of room and a distribution channel, helping the hotel decide which is its appropriate target based on its needs.

***Index Terms*—Bookings cancellation, clustering, prediction, modeling, dimension reduction, correspondence analysis, efficiency hotel room**

CONTENTS

Jordi Tarroch Mejón is with Data Science, CUNEF, e-mail: jordi.tarroch@cunef.edu

I. INTRODUCTION

- This study presents different unsupervised learning techniques for clustering in order to see, based on the variables a hotel has about their bookings:
  - What technique is the best in order to group bookings with the minimum cancellation ratio. Group that as a business point of view should be maximized in order to increase profits.
  - What technique is the best in order to group bookings in the most homogeneous way and find out the exact common features that minimize the cancellation ratio.
- This study also presents an unsupervised learning technique to reduce the dimension in order to see what is the relationship between:
  - Reserved Room Type vs Distribution Channel: relationship that as a business point of view can help to match hotel's needs making the most of the different types of rooms offered by it. Going to the appropriate Distribution Channel based on the Type of Room they need to fill.
  - Reserved Room Type vs Countries: relationship that would help to know what nationality its advertising should focus on to occupy the rooms they need to fill.

II. METHODOLOGY

The need to automatically group bookings based on its features leads the study to research between unsupervised hierarchical and non hierarchical clustering techniques. Clustering techniques try to minimize the variance within the group and maximize the variance (distance) between the groups.

First of all is necessary to see the quality of a possible clustering based on its underlying features with a Goodness Cluster Analysis. In case of being positive, the research can continue.

Then the optimal number of clusters is calculated. From a business point of view trying to group the bookings in only 2 groups makes sense, as there would be one group with a higher cancellation ratio and another with a smaller cancellation ratio. However, the best solution is to see what is the most optimal number of groups in order to target hotel's bookings better

and in the most heterogeneous way between clusters and in the most homogeneous way intracluster.

After calculating the optimal number of clusters, clustering techniques are used and for each one of them a few measures are performed:

- **Cluster Plot:** it helps to see the overlap or not between groups in a 2-dimensional space. But as clustering is done in multiple dimensions, although there is overlap, the clusters may be grouping well. Performance of good clustering is measured later.
- **Contingency Table:** through a contingency table that displays the (multivariate) frequency distribution of the variables is possible to extract the cancellation ratio. As the goal of this paper is to see the cancellation ratio of each group, the final results have been shown.
- **Silhouette Plot:** this graphic shows the following information.
  - Each line representing an observation.
  - The average similarity of each group.
  - Depending on the profile values:
    - Tending to 1 means a very good classification.
    - Tending to 0 means border observations between two groups
    - Negative values mean possible miss-classification

Finally, a correspondence analysis is performed to help the hotel target the right country or distribution channel to optimize hotel room efficiency.

III. Clustering

A. Goodness of the Cluster Analysis

The Hopkins statistician is used to advise on the existence of underlying groups in the data set, other than a mere random assignment. It contrasts through a hypothesis of uniform distribution of data space. Results support the presence of two or more clusters in the set of observations as it is close to 0.

Hopkins statistician: 0.06892939

B. Distances

Most of the clustering techniques can process mixed data (numerical and categorical data), whose dissimilarity matrices distances can be calculated using the Gower distance. Gower distance scales any type of data from 0 to 1 and then uses a linear combination of user-specific weights (in general, an average) to finally calculate the distance matrix.

However, CLARA clustering technique requires Euclidean or similar distances to calculate the dissimilarity matrix. That forces the study, with the CLARA clustering, to use numerical data for this model, forbidding the use of mixed data (categorical and numerical data).

Table I

TABLE I
Distance Measures based on Data Type

Clustering Technique	Data Type	Distance Measure
PAM	Mixed	Gower
CLARA	Numeric	Euclidean
FUZZY	Mixed	Gower
HCUT	Mixed	Gower

C. Optimal Number of Clusters

As a non-hierarchical clustering is being used, which requires a predetermined number of groups, and that means prior knowledge, or at least a previous administrative decision. Having an objective number of groups prior to introducing any domain knowledge introduces less bias to the study.

Having an excess groups can result in a granularity that does not add much interest and, above all, that can be overly expensive when implementing it in practice. Whilst having too few groups may not show the impact that is being pursued.

The method used to calculate the optimal numb of groups is called profile chart, which offers a measure the quality of segmentation. The optimal number of clusters will be given by the value of 'k' that maximizes the average profile. So a high average profile means a correct segmentation.

1) Non-Hierarchical Clustering: These techniques attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster.

PAM was prioritized over other methods like K-means as sensitivity to outliers is not desired, being PAM a more robust model. Because the study uses a large data set, CLARA is even a better solution as it internally computes PAM.

a) PAM (k-medoids or partitioning around medoids):

[Fig. 1 about here.]

The optimal number of clusters for PAM is 4.

b) CLARA:

[Fig. 2 about here.]

The optimal number of clusters for CLARA is 2.

c) FUZZY:

[Fig. 3 about here.]

The optimal number of clusters for FUZZY is 2.

2) Hierarchical Clustering: In the hierarchical clustering there is no need to set up a default number of groups and the result of the study just invites the research to discover how many groups are possible. However, it is also possible to get an objective number of groups based on the same study used for the non hierarchical clustering techniques.

a) HCUT:

[Fig. 4 about here.]

The optimal number of clusters is 2.

D. Non-Hierarchical Clustering

1) PAM: Main features:

- Centers: choose data points as centers (medoids).
- Sensitivity: solve the sensitivity to outliers of k-means. More robust in their presence than k-means as its centers are data points and not centroids (arithmetic mean of all the data points that belong to that cluster).

a) Cluster Plot:

[Fig. 5 about here.]

b) Contingency Table: Table II

TABLE II
CANCELLATION RATIO OF PAM

Groups	Cancellation Ratio
Group 1	25.89286
Group 2	42.85714
Group 3	21.2766
Group 4	11.76471

c) Silhouette Plot: The average silhouette width is 0.43.

[Fig. 6 about here.]

2) CLARA: It is used for large sets of several thousand observations as in this case. It applies the PAM algorithm in each new subset by choosing the k medoids in each one and assigning each observation of the original set to the nearest medoid.

a) Cluster Plot:

[Fig. 7 about here.]

b) Contingency Table: Table III

TABLE III
CANCELLATION RATIO OF CLARA

Groups	Cancellation Ratio
Group 1	23.49304
Group 2	37.96034

c) Silhouette Plot: The average silhouette width is 0.54.

[Fig. 8 about here.]

3) FUZZY: This is what is called a soft segmentation, compared to the hard segmentation used before. Observations close to the center of each group are more likely to belong to it than those found in the outer limits:
The degree of membership to a cluster is measured through a probability. This algorithm minimizes intraclusters dispersion.

a) Cluster Plot: The Dunn Coefficient is measured giving as a result 0.504429626, which means that the segmentation is neither too good nor too bad, as it may vary between values '1' and '0'.

[Fig. 9 about here.]

b) Contingency Table: Table IV

TABLE IV
CANCELLATION RATIO OF FUZZY ANALYSIS

Groups	Cancellation Ratio
Group 1	30.18 %
Group 2	26.19 %

c) Silhouette Plot: The average silhouette width is 0.37.

[Fig. 10 about here.]

E. Hierarchical Clustering

In the hierarchical clustering there is no default number of groups and the result of the study just invites the research to discover how many groups are possible. However, it is also possible to get an objective number of groups based on the same study used for the non hierarchical clustering techniques.

1) HCUT: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

a) Dendrogram:

[Fig. 11 about here.]

b) Cluster Plot:

[Fig. 12 about here.]

c) Contingency Table: Table V

TABLE V
CANCELLATION RATIO OF HCUT

Groups	Cancellation Ratio
Group 1	29.47368 %
Group 2	12 %

d) Silhouette Plot: The average silhouette width is 0.48.

IV. DIMENSION REDUCTION

A. Correspondence Analysis

CA is the proposed technique for reducing the dimension of categorical variables in order to summarize a set of data in two-dimensional graphical form, in a similar manner to PCA. However, overcoming the limitations of PCA and Factor Analysis.

These associated positions of the variables studied are related to the degree of association. By using a Simple Correspondence Analysis 2 features are studied simultaneously. In this study:

- Reserved Room vs Distribution Channel
- Reserved Room vs Country

After performing the Independent Test for both correspondence analysis a relationship between them is confirmed so the analysis can be done.

1) Reserved Room vs Distribution Channel: These plots are a visual help to show which distribution channel should be contacted in order to occupy the type of room of the hotel that needs to be filled. The minimum the angle between the two differents categories the highest the relationship between the distribution channel and the type of room.

[Fig. 13 about here.]

[Fig. 14 about here.]

2) Reserved Room vs Country: These plots are a visual help to show which country marketing of the hotel should be focused on in order to occupy the type of room of the hotel that needs to be filled. The minimum the angle between the two differents categories the highest the relationship between the distribution channel and the type of room.

[Fig. 15 about here.]

[Fig. 16 about here.]

V. Conclusion

A. Clustering - Discrimination according to groups

In order to increase profits and maximize the number of people that don't cancel their bookings the hotel should focus on the bookings that fall into the following group:

- **Group 4 of PAM clustering technique,** with a **Cancellation Ratio of 11.76%,** group that results from 4 groups. Followed by HCUT with a cancellation Ratio of 12%, on group 2, that results from 2 groups.

Table VI

TABLE VI
Cancellation Ratio based on Clustering Technique

Clustering Technique	Group	Cancellation Ratio
PAM	4	11.76%
CLARA	1	23.49%
FUZZY	2	26.19%
HCUT	2	12%

The Clustering Technique that is able to perform **the best classification of groups is CLARA**, with an **Average Silhouette width of 0.54**. Although its cluster plot shows overlap in two dimensions, this is clearly the technique that best classifies the bookings, as the clustering is performed in multiple dimensions that are not able to visually be seen. If visualization in multiple dimensions was possible, a good segmentation would be seen. Important things to consider about its segmentation are that this model is using numeric data and calculating euclidean distances instead of gower distances.

CLARA's good segmentation is followed by HCUT algorithm with an average silhouette width of 0.48.

Although a good segmentation is possible with **CLARA** clustering, its minimum cancellation ratio for Group 1 of 23.49% can not be compared with the lowest 11.76% offered by PAM. But in order to **easily segment** (as there are just a few 15 numeric variables compared with other models using all of them) customers with a low cancellation ratio (not the lowest), this technique would make sense.

Table VII

TABLE VII
Average Silhouette Width

Clustering Technique	Average Silhouette Width	Number of Groups
PAM	0.43	4
CLARA	0.54	2
FUZZY	0.37	2
HCUT	0.48	2

B. Dimension Reduction - Correspondence Analysis

Correspondence analysis end up being a good visual tool to help departments of the hotel take decisions. Using two plots for each relationship and depending on the rooms that are still not filled for certain days, the responsible of it can directly focus on certain Distribution Channels and Nationalities to reach the **maximum hotel room efficiency**.

VI. References

- https://www.r-bloggers.com/clustering-mixed-data-types-in-r/
- https://datascience.stackexchange.com/questions/22/k-means-clustering-for-mixed-numeric-and-categorical-data
- https://chrisalbon.com/machine_learning/preprocessing_structured_data/encoding_ordinal_categorical_features/

List of Figures

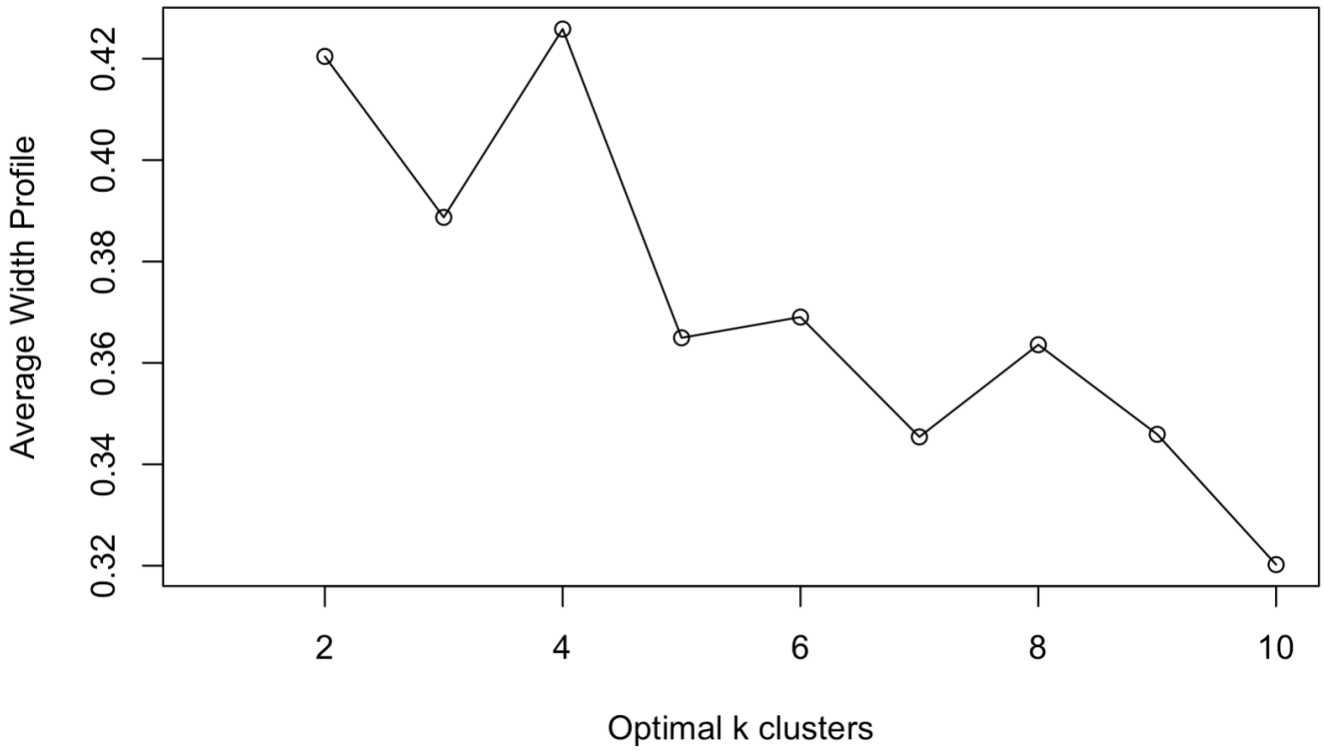

Fig. 1. Optimal number of clusters - PAM

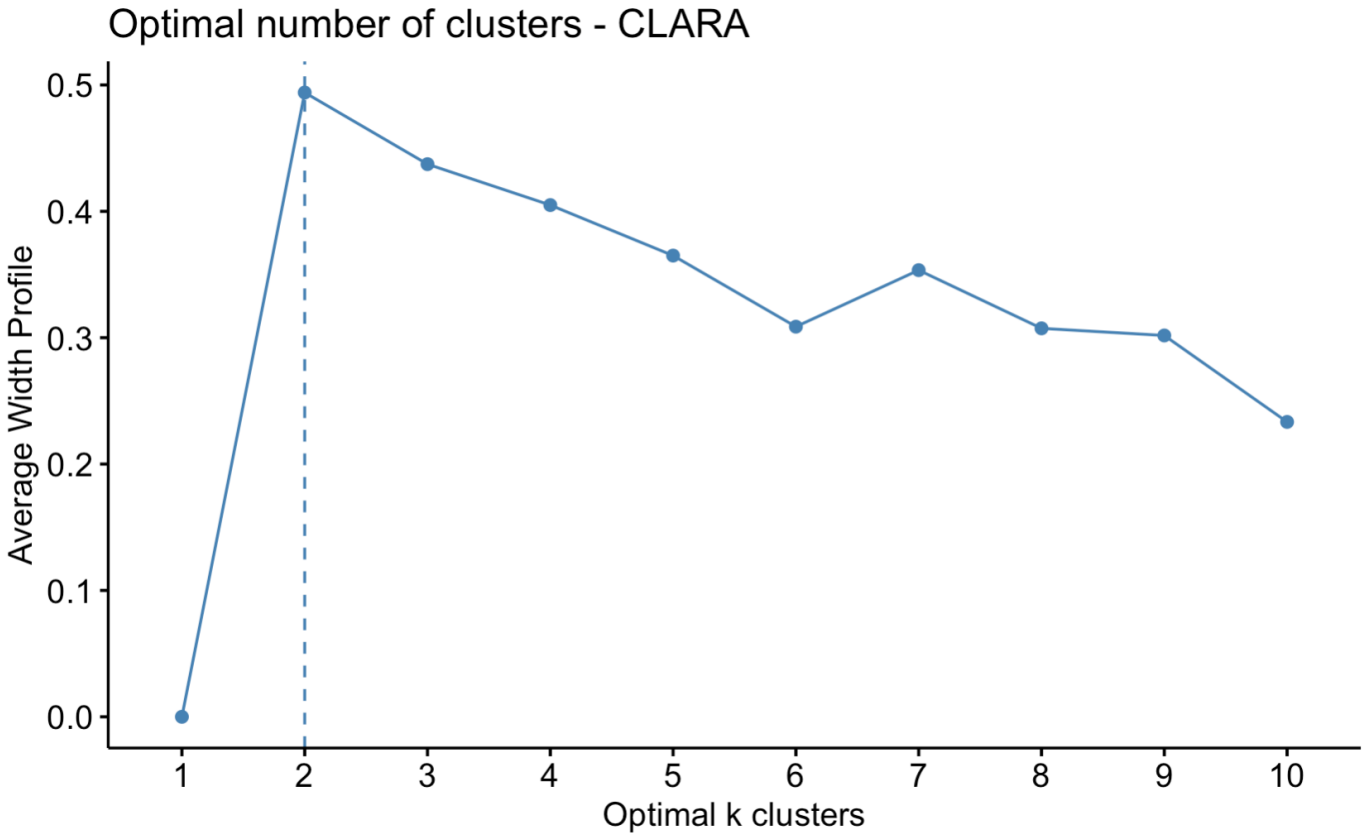

Fig. 2. Optimal number of clusters - CLARA

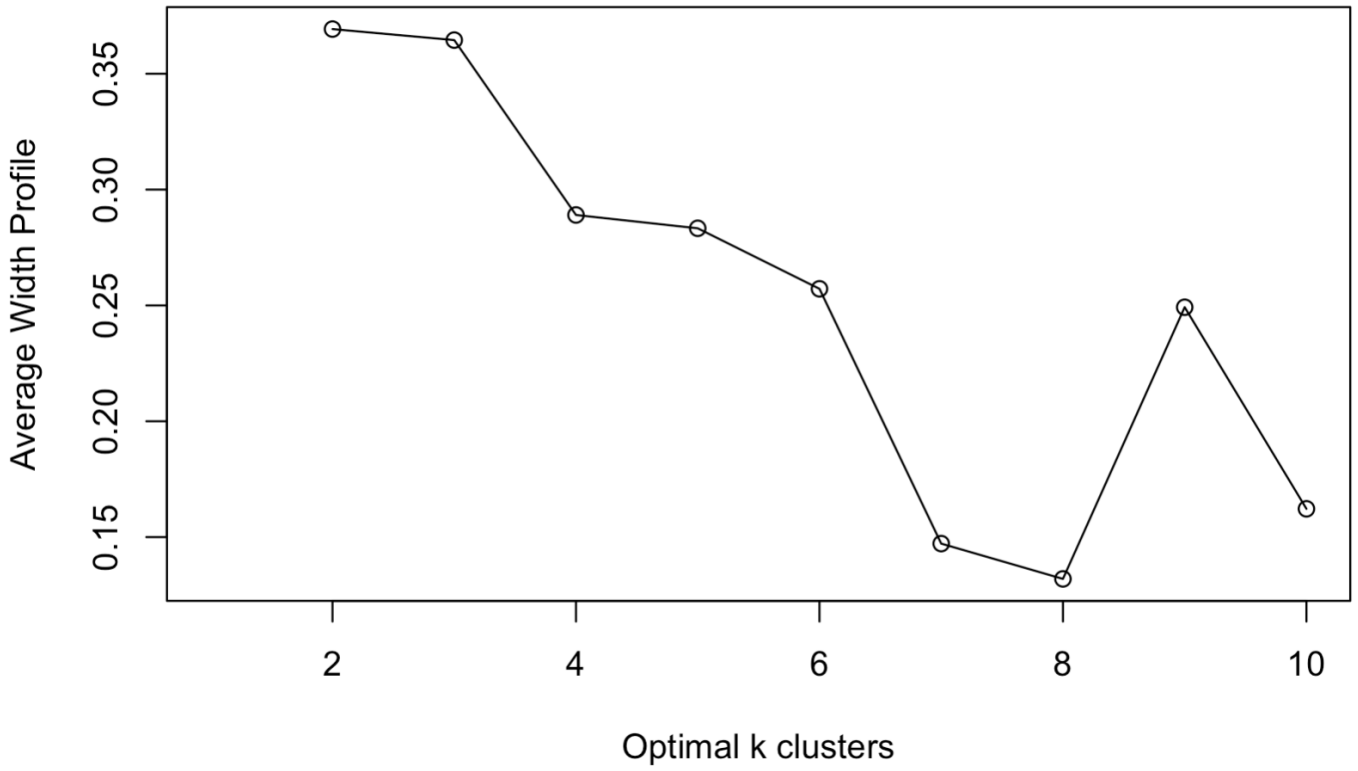

Fig. 3. Optimal number of clusters - FUZZY

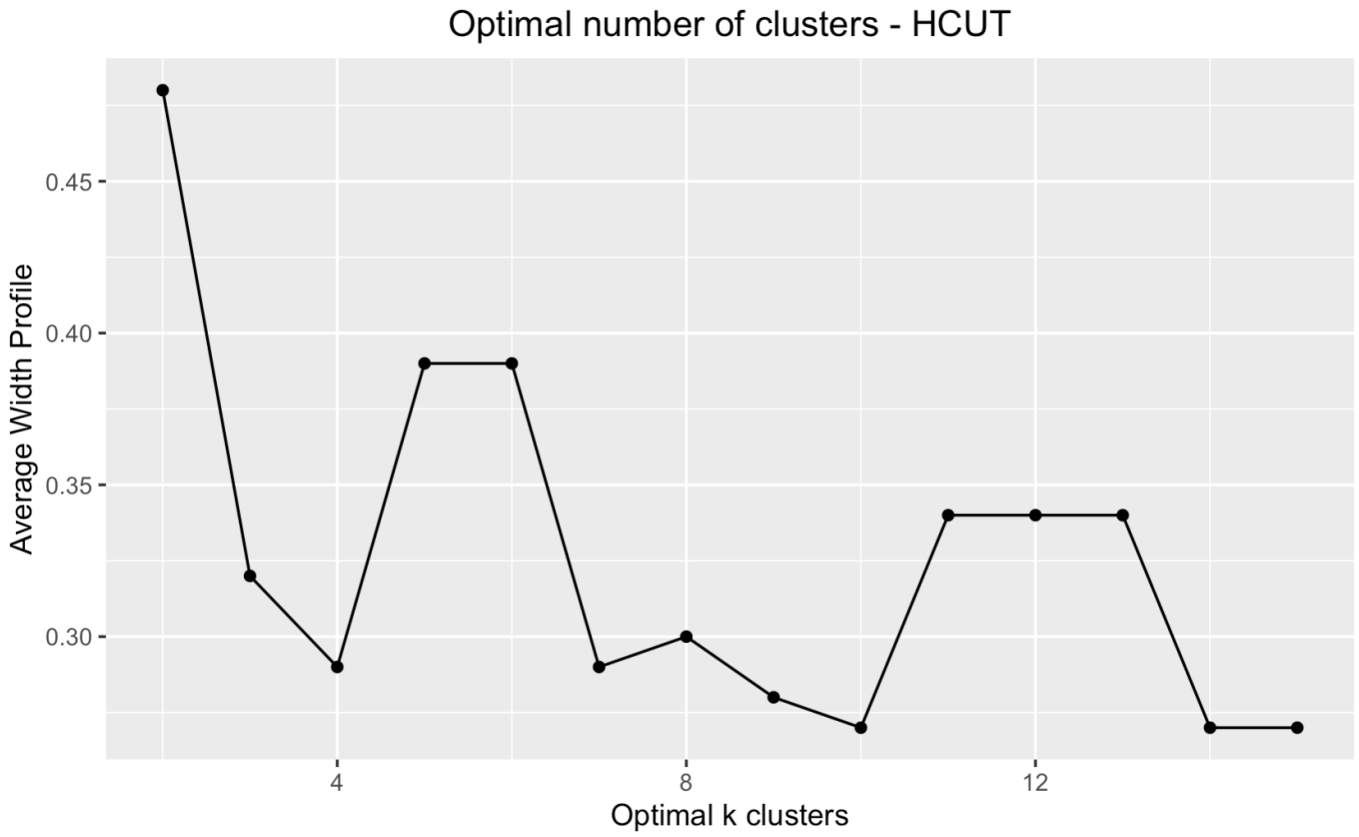

Fig. 4. Optimal number of clusters - Hierarchical

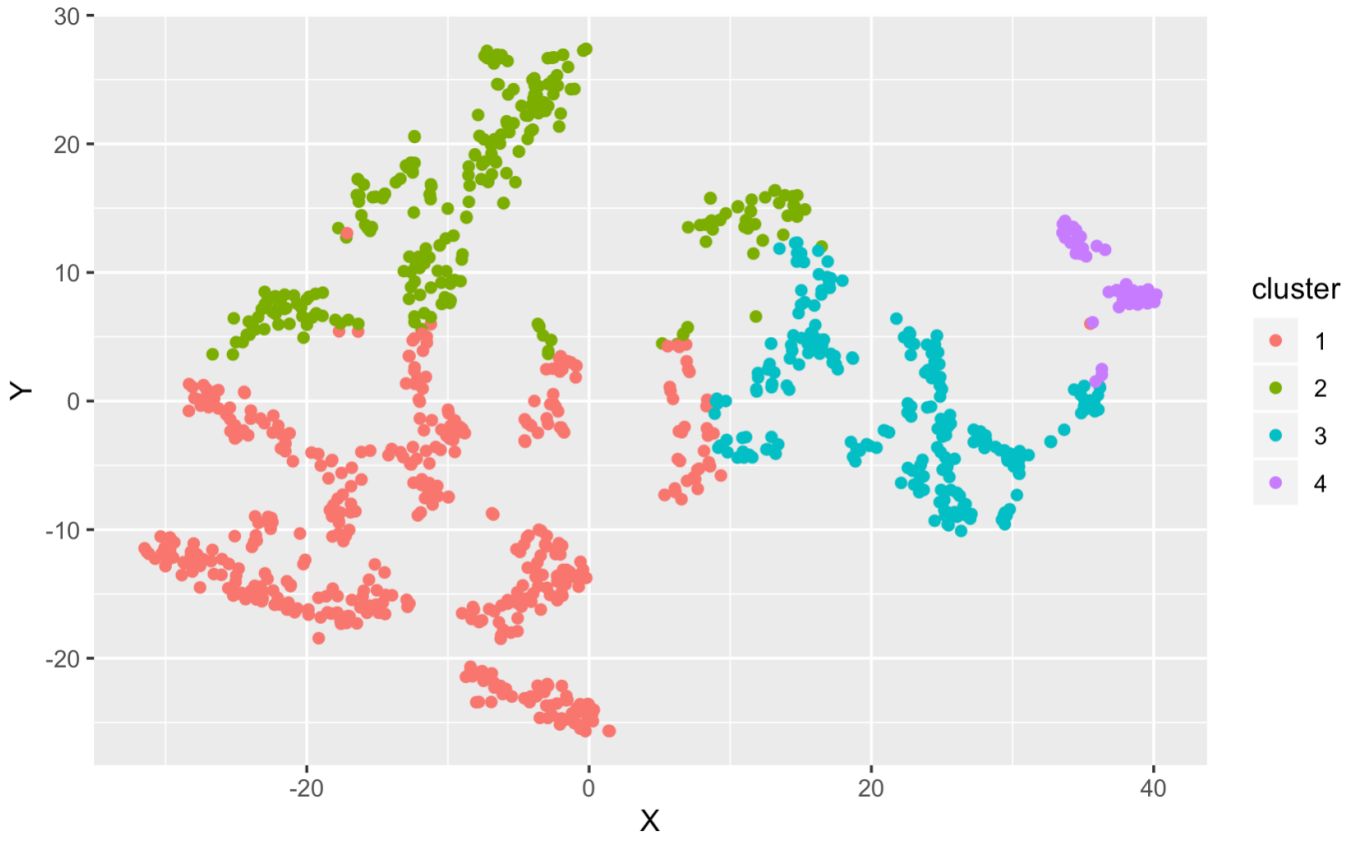

Fig. 5. Cluster Plot PAM

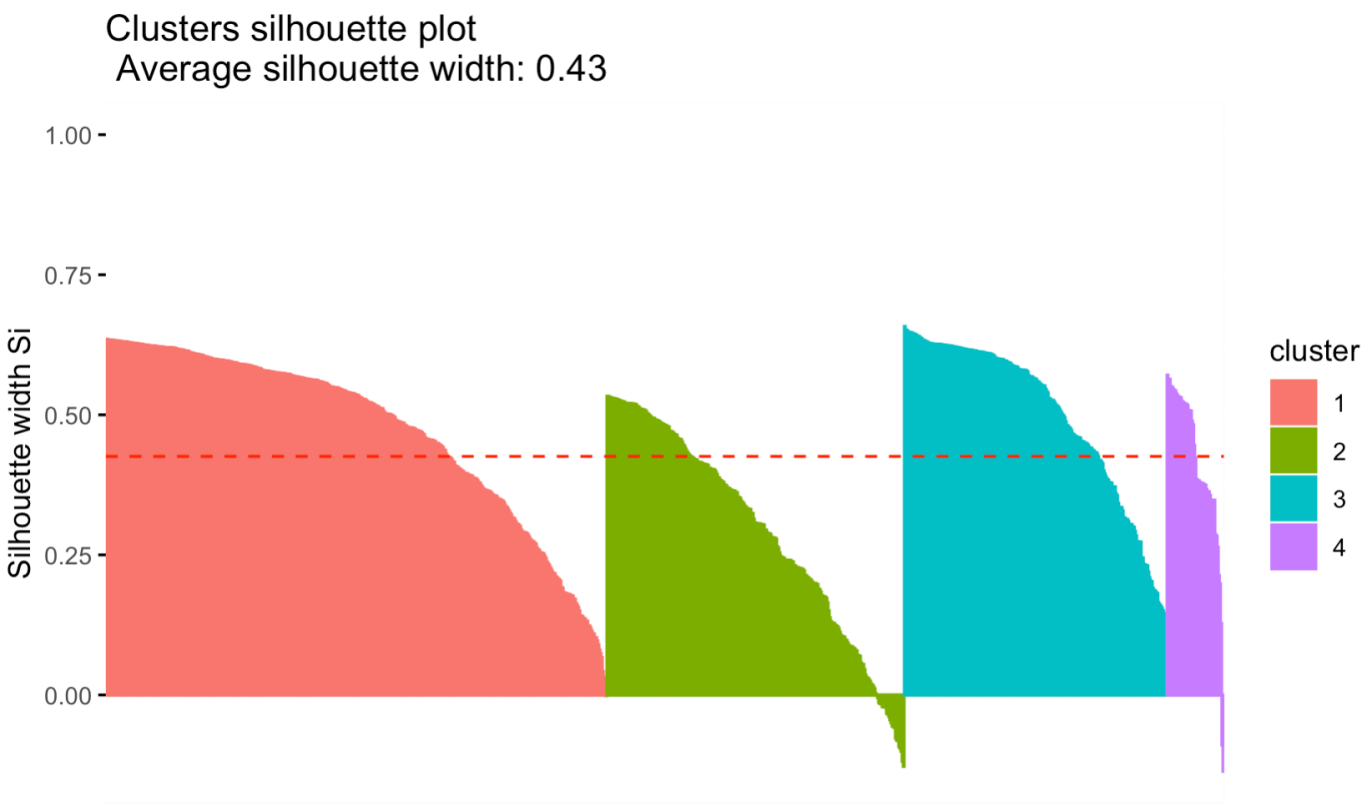

Fig. 6. Silhouette Plot PAM

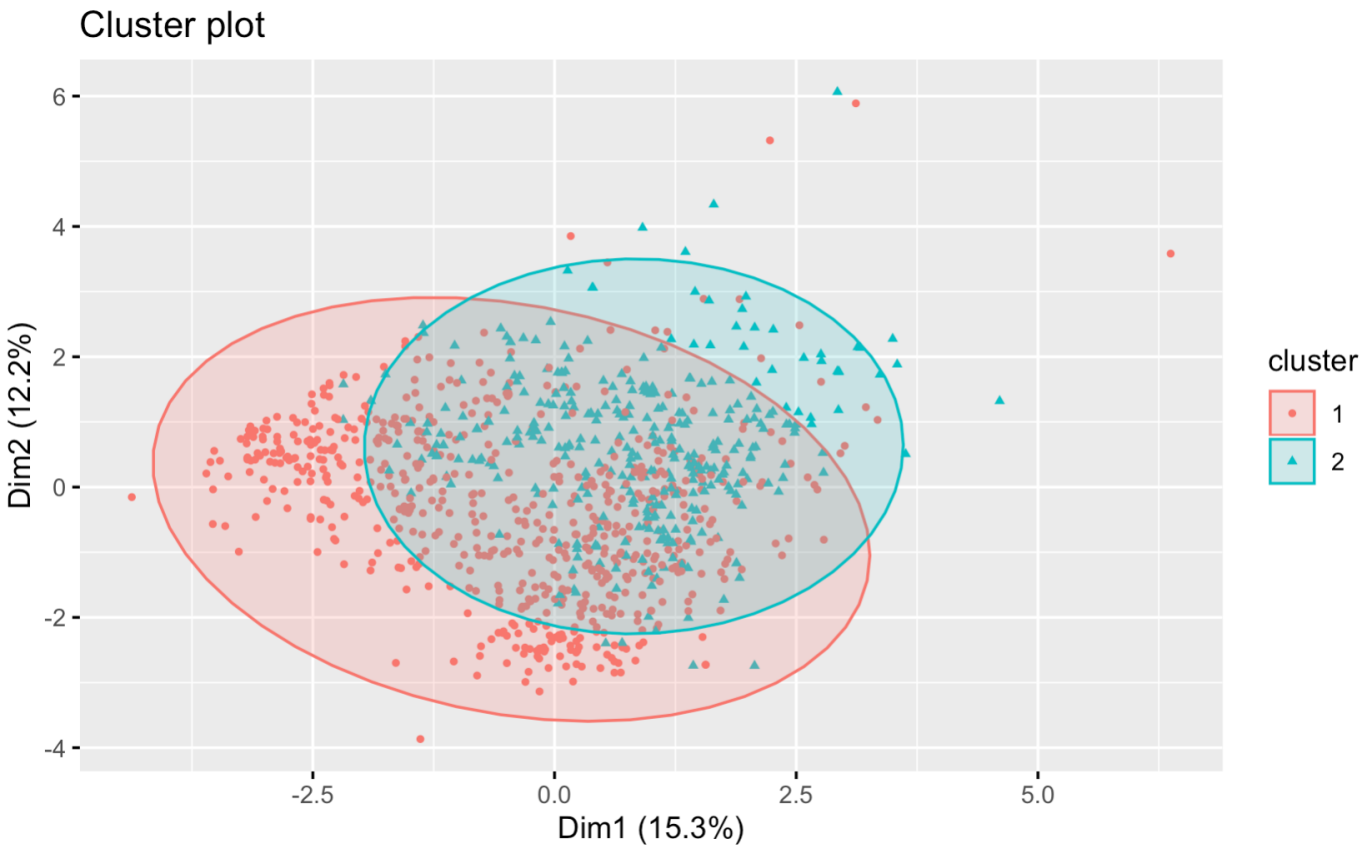

Fig. 7. Cluster Plot CLARA

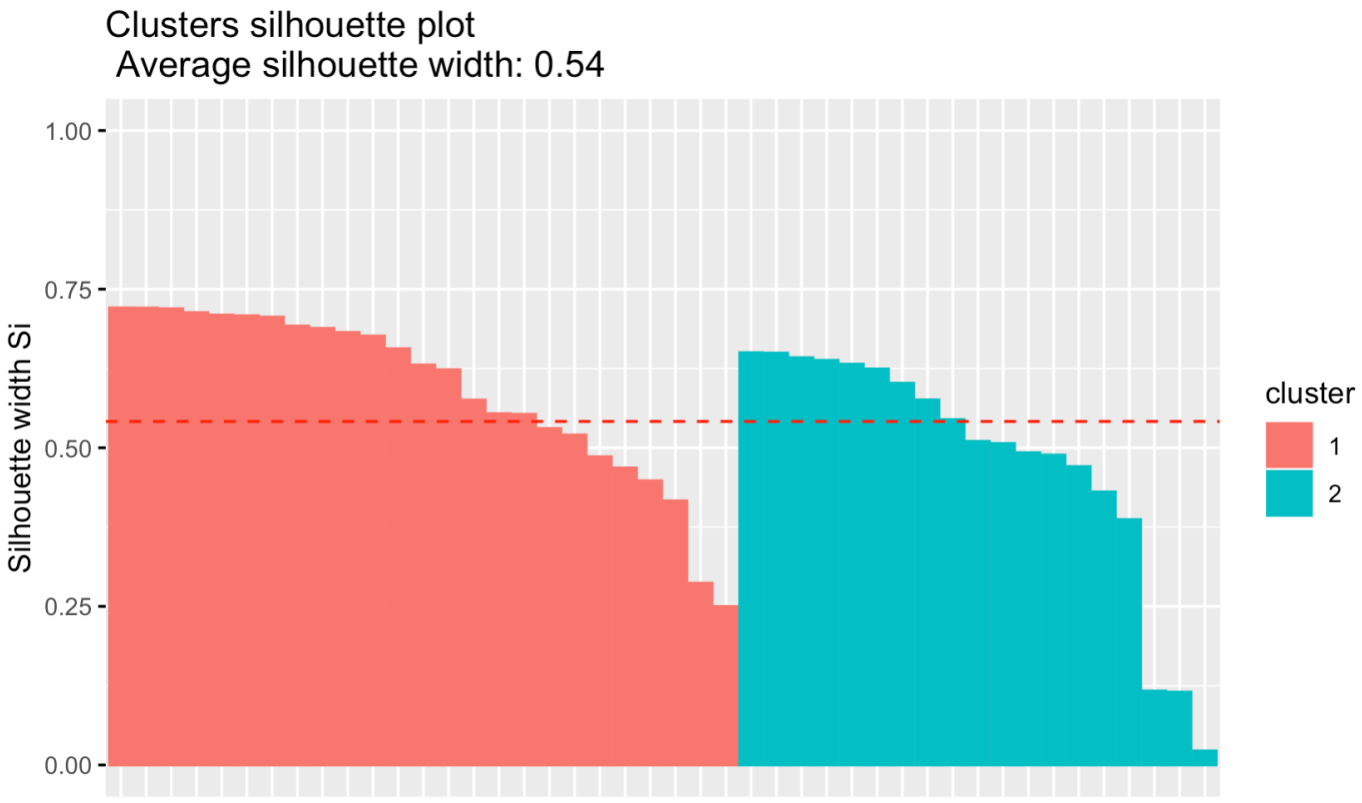

Fig. 8. Silhouette Plot PAM

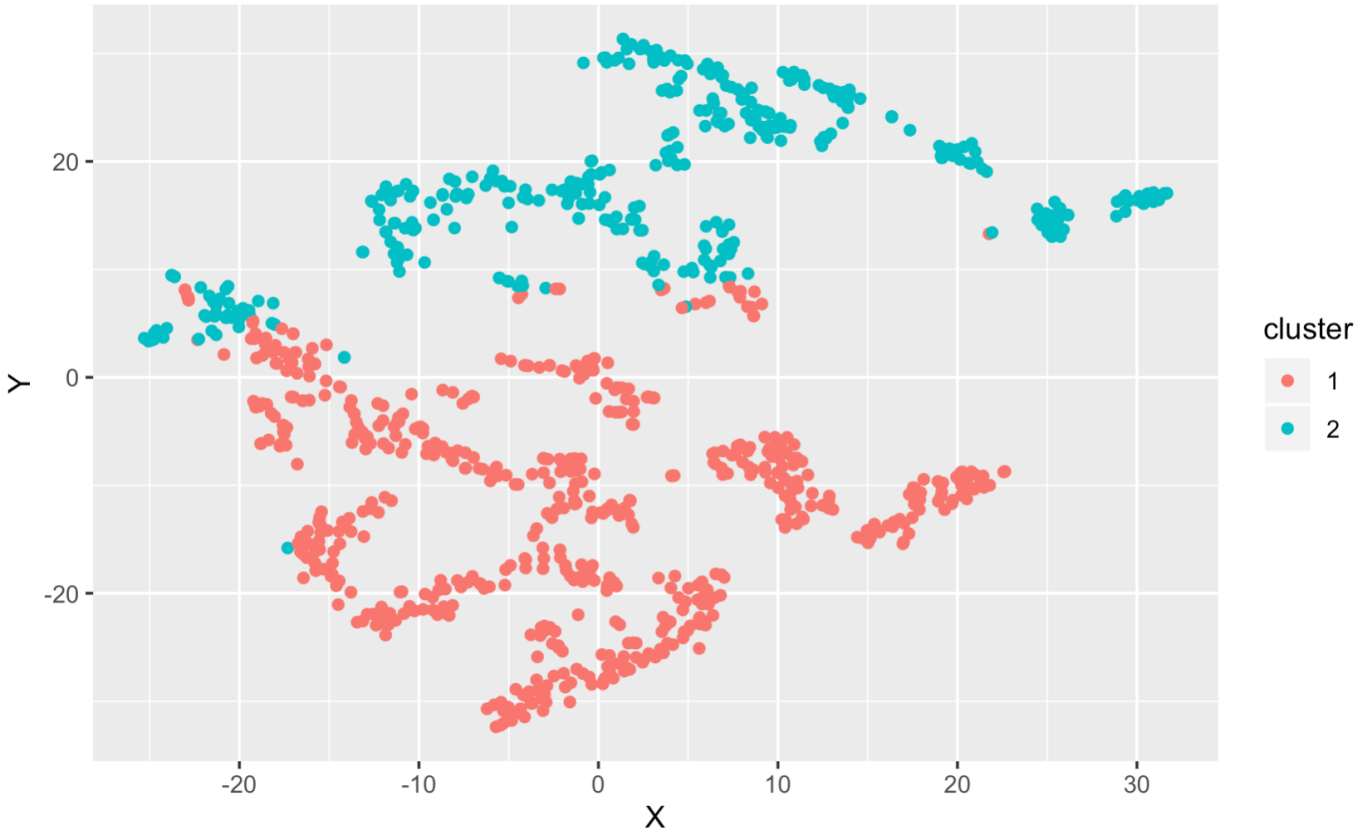

Fig. 9. Cluster Plot Fuzzy

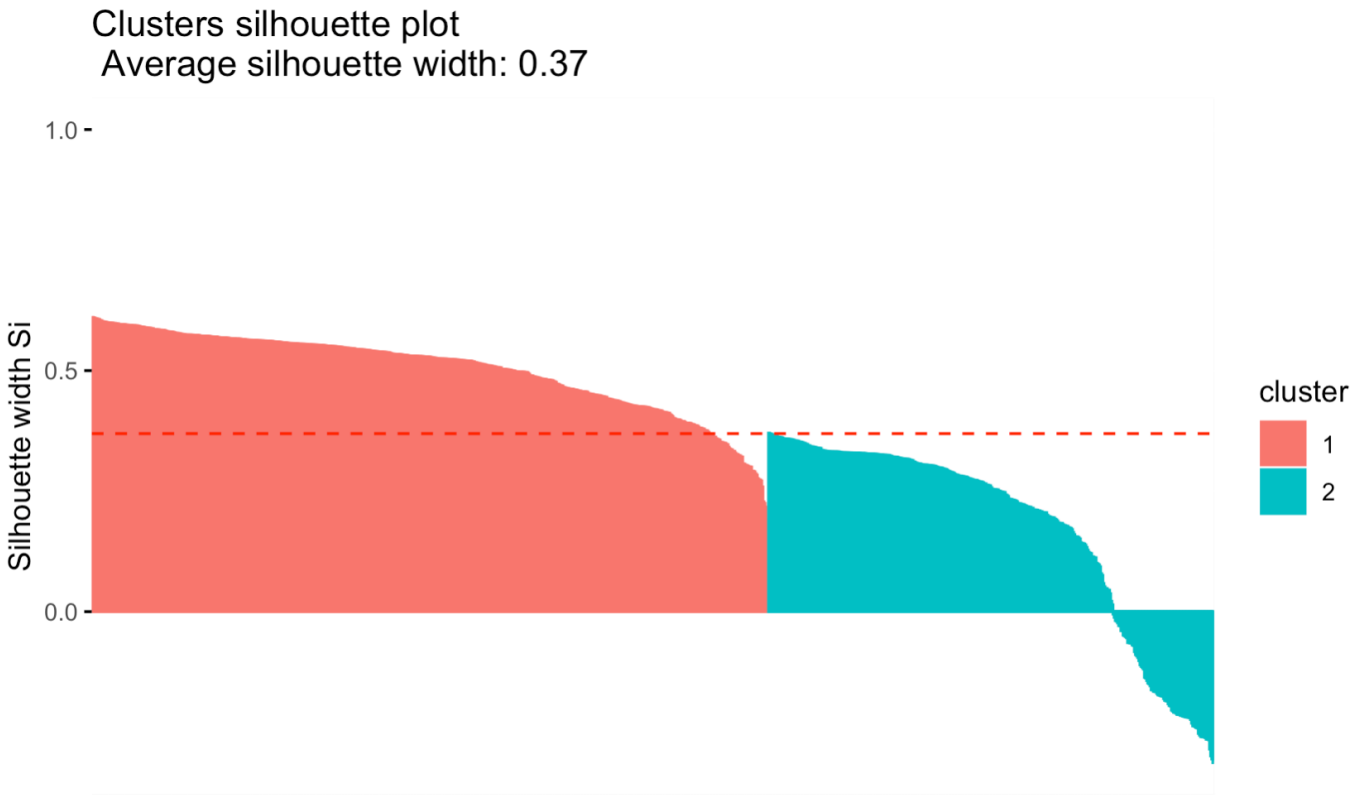

Fig. 10. Silhouette Plot FUZZY

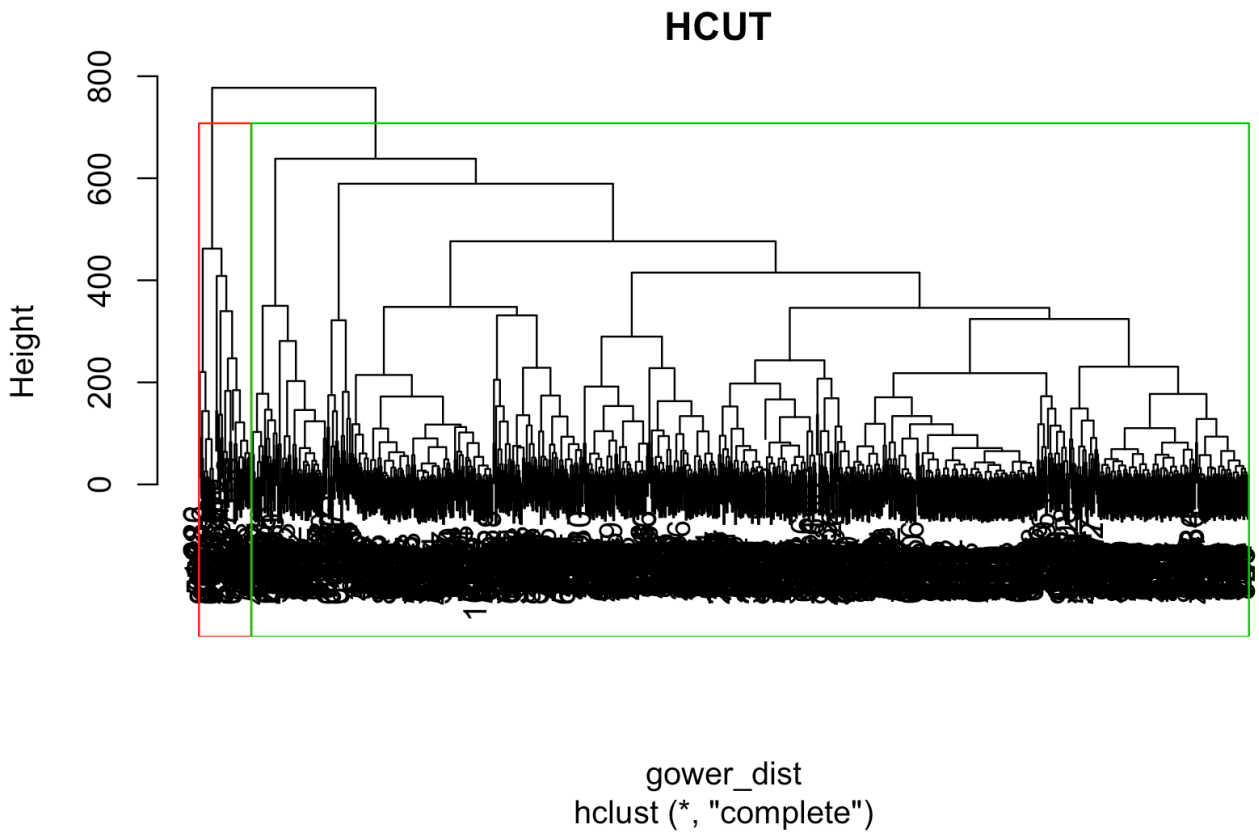

Fig. 11. Dendrogram of HCUT

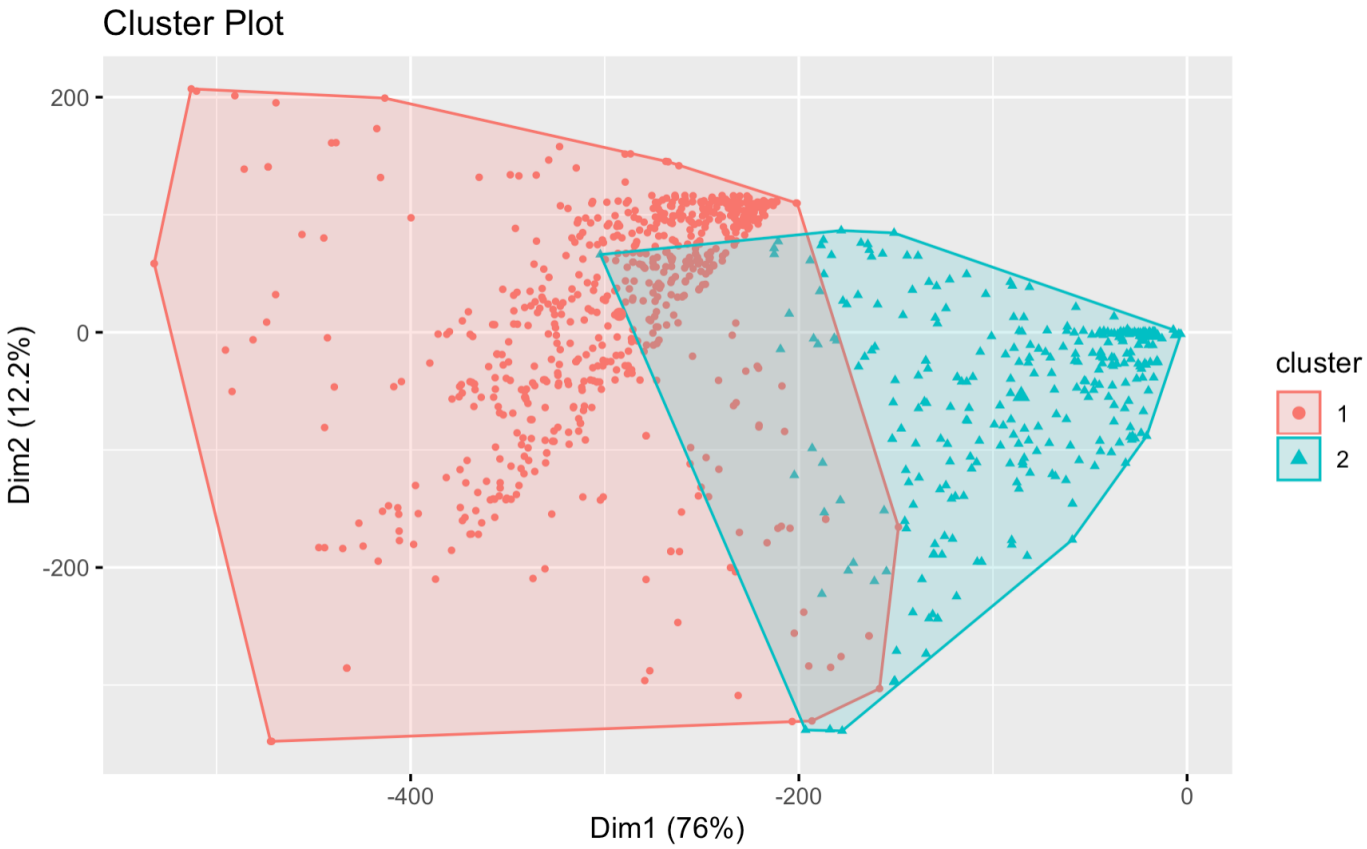

Fig. 12. Cluster Plot HCUT

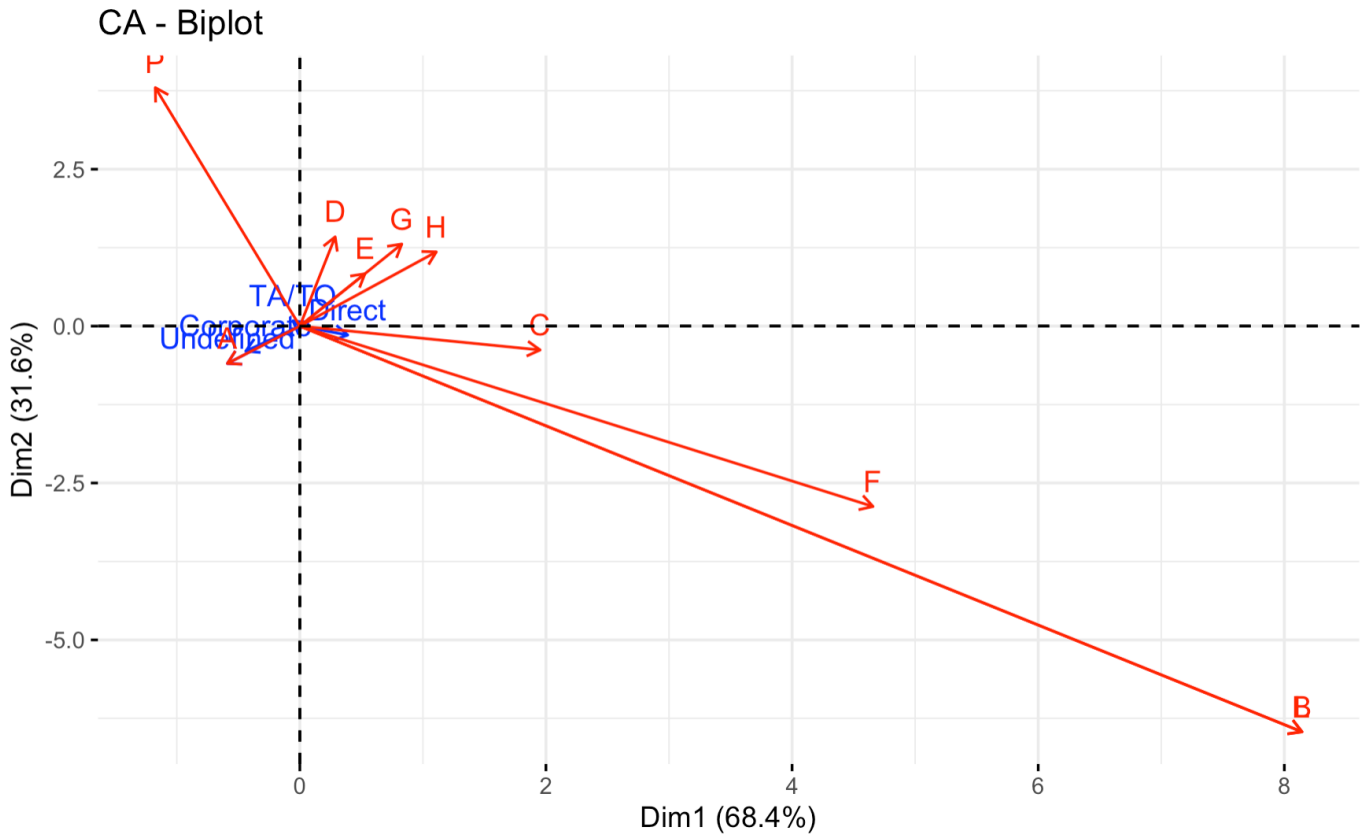

Fig. 13. Contribution of Room Types to dimensions

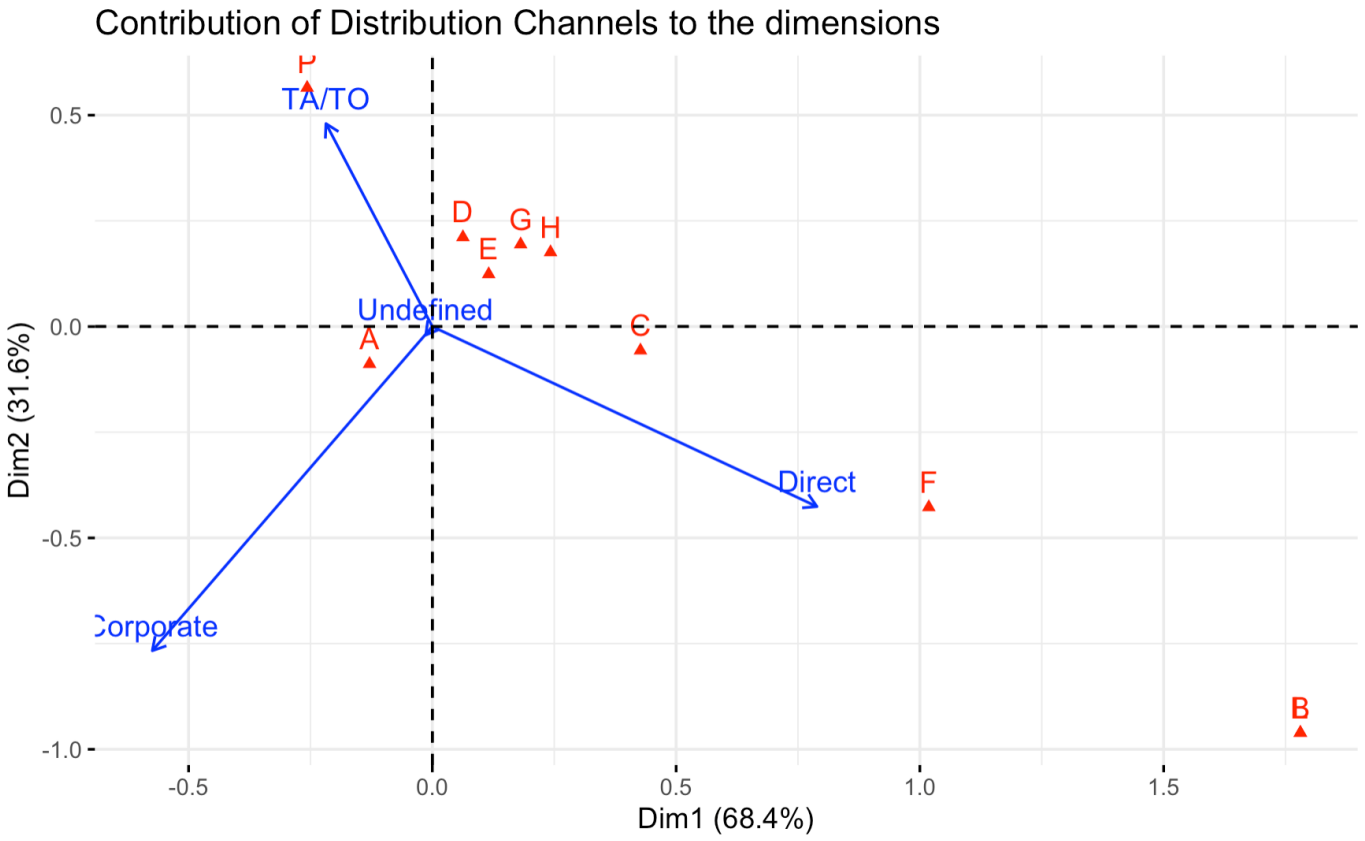

Fig. 14. Contribution of Distribution Channels to dimensions

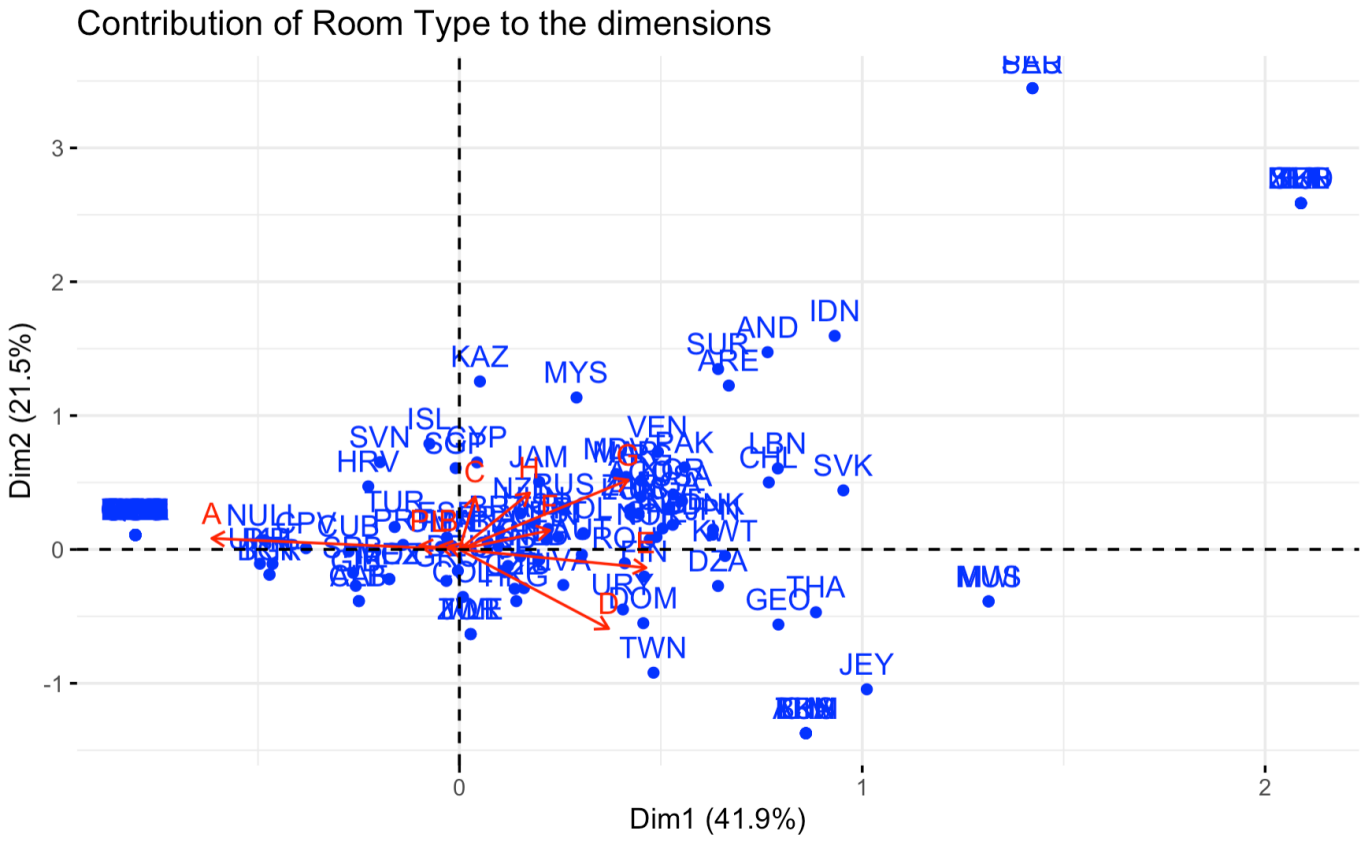

Fig. 15. Contribution of Room Type to the dimensions

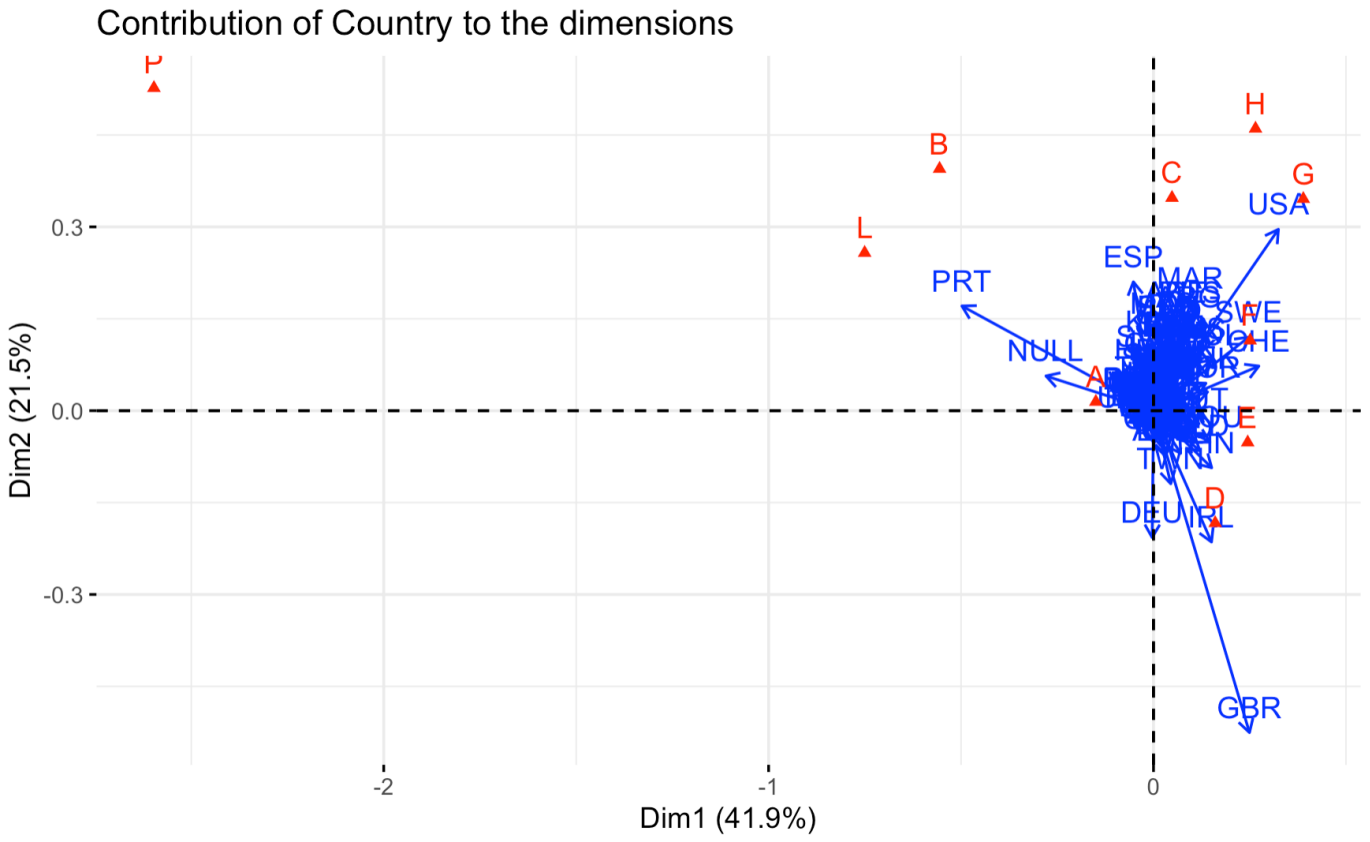

Fig. 16. Contribution of Country to the dimensions