

# FLEXBEAN

Flexibility for  
Energy Transition

## PROJECT

D.2.2.2

### Household's energy literacy and flexibility survey

-clustering of household profiles

Mohamed Laib, Eric Roseren.



LUXEMBOURG  
INSTITUTE OF SCIENCE  
AND TECHNOLOGY



---

**Project acronym** FlexBeAn

**Full title** Flexibility potentials and user Behaviour Analysis

**Project start date** 2022-05-02

**Project duration** 41 months

---

**Work package 2** Customer Incentive mapping

**Deliverable lead organisation** LIST

**Authors** Mohamed Laib (LIST)  
Eric Roseren (LIST)

**Version** 1.0

**Status** Final

**Dissemination level** PU: Public

---

**How to cite this report** Laib, M. and Roseren, E. (2025). Deliverable D2.2.2 - Household's energy literacy and flexibility survey: clustering of household profiles. FlexBeAn Project.

This report details the data analysis performed on the dataset obtained from the energy literacy and flexibility survey conducted in the WP2 of the FlexBeAn project, with the goal of finding clusters of households having different characteristics. It highlights the obtained clusters and their main characteristics regarding the household model (see D2.1.4), through the survey's questionnaire.

**Table of Contents**

---

- 1. First investigations..... 5
  - 1.1. Analysis of the full dataset..... 5
    - 1.1.1. PCA analysis of the continuous variables ..... 5
    - 1.1.2. MCA analysis for the categorical variables..... 6
    - 1.1.3. Factor analysis of mixed data (FAMD) ..... 7
    - 1.1.4. Clustering using Gower distance and Partitioning around Medoid (PAM) ..... 11
    - 1.1.5. K-means clustering using Euclidean distance..... 13
    - 1.1.6. Clustering based on personality test..... 15
- 2. Refined analysis..... 17
  - 2.1. Medoid-based clustering approach..... 17
    - 2.1.1. Main steps ..... 17
    - 2.1.2. Application..... 17
  - 2.2. Results and discussion..... 19
    - 2.2.1. Clustering Performance and Visualization..... 19
    - 2.2.2. Detailed Cluster Analysis ..... 20
    - 2.2.3. Influence of Variables on Cluster Assignment..... 21
    - 2.2.4. Distribution of Survey Scores within Clusters..... 22
  - 2.3. Discussion ..... 23
  - 2.4. Conclusion ..... 23
  - 2.5. Relevant references ..... 24
- 3. Appendix: Overview of used techniques and measures ..... 25
  - 3.1. Appendix A. Medoid-based clustering..... 25
  - 3.2. Appendix B. Gower distance ..... 26
  - 3.3. Appendix C. Silhouette Measure ..... 27
  - 3.4. Appendix D. TSNE..... 28

List of Figures

Figure 1. Correlation plot of the continuous variables of the survey dataset. Regarding continuous variables, no significant correlation can be observed in the dataset. ....5

Figure 2. Principal Component analysis of the continuous variables.....6

Figure 3. Multiple correspondence analysis results applied to all categorical variables of the survey data.....6

Figure 4. Graph of individuals color-segmented by education level. ....8

Figure 5. Graph of individuals color-segmented by gender. ....9

Figure 6. Graph of individuals color-segmented by heating system. ....9

Figure 7. Graph of individuals color-segmented by household type. ....10

Figure 8. The representation of variables (relationship square) of both continuous and categorical variables on the same figure..... 11

Figure 9. Average Silhouette width index to help in determining the optimal number of cluster..... 12

Figure 10. Results of the PAM clustering projected on the first two principal components..... 13

Figure 11. K-means clustering results. .... 14

Figure 12. Description of the categories. .... 15

Figure 13. Overview of the personality and literacy scores obtained by every participant. Each row is a participant survey answers to all tests. Each column corresponds to a question. .... 16

Figure 14. Result of the Hierarchical Clustering approach with Jaccard distance. Both variables and respondents are clustered. .... 16

Figure 15. Medoid based clustering. 2 clusters are obtained based on the values of silhouette..... 19

Figure 16. visualisation of these two cluster in 2-dimensional space using TSNE... 20

Figure 17. Boxplot showing 2 clusters: red vs green (median, 1st and 3rd quantile for each variable). .... 20

Figure 18. Radar visualisation of median of scaled variables vs obtained clusters.. 21

Figure 19. List of variables that influence the most GLM model..... 22

Figure 20. Visualisation of scores of survey within the detected clusters (scaled between 0 and 1). .... 23

## 1. First investigations

The aim of the work reported in this section was to compare respondent answers to the Energy literacy and flexibility study and detect any answering pattern. As a first step, an overall clustering algorithm was implemented on all categorical, ordinal and numerical features to detect any potential clusters. Then, the clustering analysis was focused on the features with Likert scale data (a Likert scale is a rating scale used to measure survey participants' opinions, attitudes, motivations, and more.).

### 1.1. Analysis of the full dataset

In this section, an overall analysis on all features was performed without focusing on any particular variable. We proceeded as follows:

- Performed simple exploratory data analysis on the data and detect trends manually.
- PCA and MDA are dimension reduction techniques that are useful to visualize correlated features as well as their contribution in filling the variance space.
- A detailed factorial analysis of mixed data was performed on the overall dataset to detect and highlight patterns.
- Apply various clustering methods using an appropriate similarity measure (e.g., Gower, or Euclidean Manhattan distance) and then apply hierarchical clustering on the distance matrix.
- Performed regression and identify the importance of the features.

#### 1.1.1. PCA analysis of the continuous variables

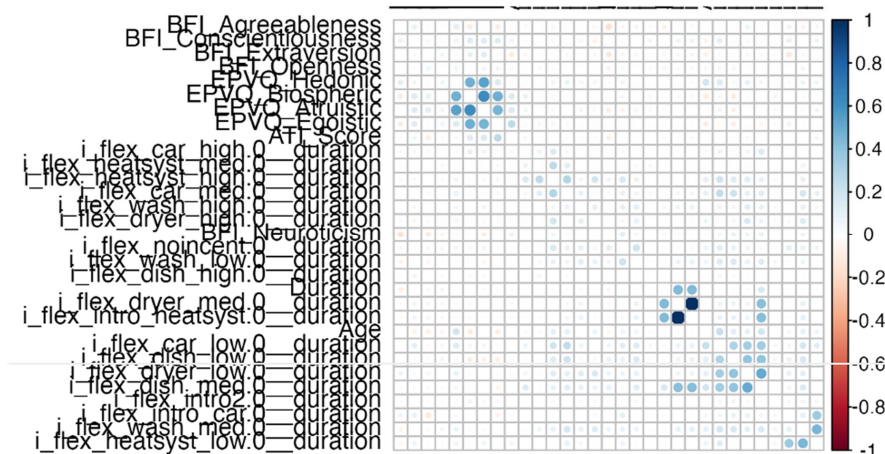


Figure 1. Correlation plot of the continuous variables of the survey dataset. Regarding continuous variables, no significant correlation can be observed in the dataset.

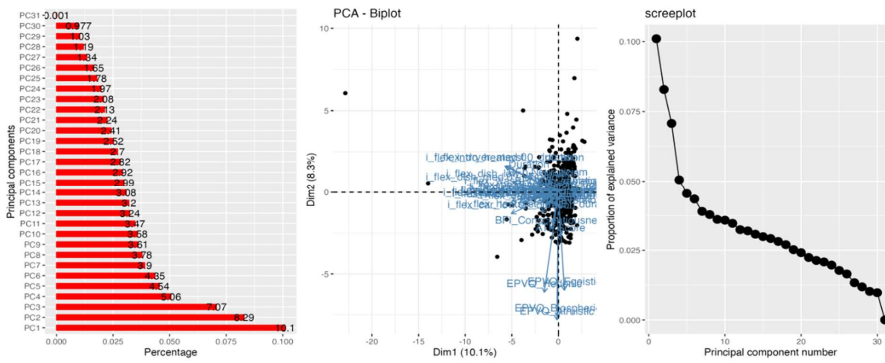


Figure 2. Principal Component analysis of the continuous variables.

Conclusion of the PCA analysis:

- no significant correlation can be observed in the dataset
- The first three PCs capture only roughly 25% of the variance in the dataset.
- The EPVQ features mainly contribute to the first PC whereas flexibility scores mainly contribute to the second dimension.
- The screen plot is not informative since only 25% of the variation in the dataset is explained by the three first principal components.

1.1.2. MCA analysis for the categorical variables

In Multiple Correspondence Analysis (MCA), discovering associations among categorical variables involves computing the chi-square distance across various categories within the variables and among the individuals (or respondents). We obtain a plot as shown in Figure 3.

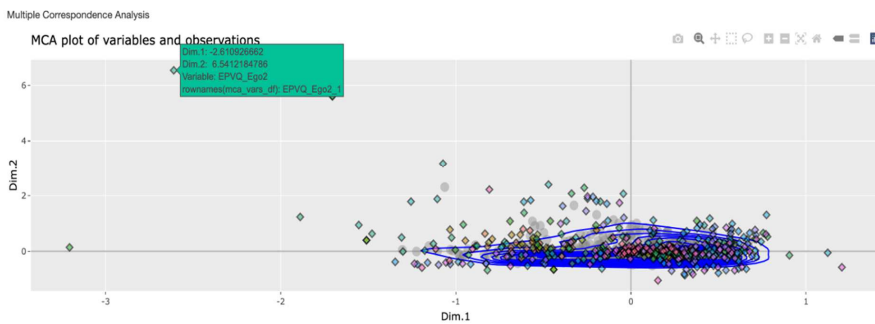


Figure 3. Multiple correspondence analysis results applied to all categorical variables of the survey data.

The MCA shows that the first two dimensions together explain only about 18 % of the total inertia, as is common with survey-scale categorical data. Dimension 1 is dominated by tenure status (owner vs. renter) and heating system type, while Dimension 2 separates respondents primarily by age group and education level. Although the inertia captured is modest, these axes reveal interpretable associations that hint at mild clustering among key socio-demographic factors.

### 1.1.3. Factor analysis of mixed data (FAMD)

FAMD functions akin to principal components analysis (PCA) for quantitative variables and multiple correspondence analysis (MCA) for qualitative variables. This allows us to analyze the importance of each feature in the datasets and the individual data points. The following graphs were obtained (see figures 4,5,6,7).

**Graph of individuals.** Both Figure 4 and Figure 5 give examples of a FAMD analysis. Here are the key interpretations coming from these plots:

- The scatter plot of individuals does not show a particular trend when plotted vs the two main PCs. The points seemed randomly dispersed along both axes.
- With regards to education levels (Figure 4), most individuals appear to have a master level education and tend to be slightly more present in the positive part of the second dimension.
- There is a clear trend however in Figure 5 where the first dimension clearly separates the gender categories. The positive part of the first dimension contains mainly female individuals whereas the negative part contains mainly the male individuals.

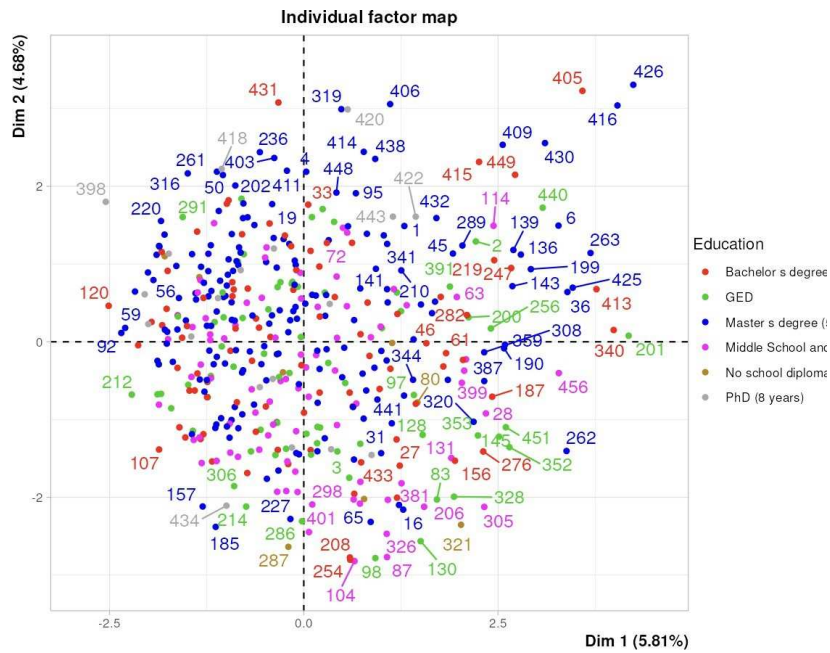


Figure 4. Graph of individuals color-segmented by education level.

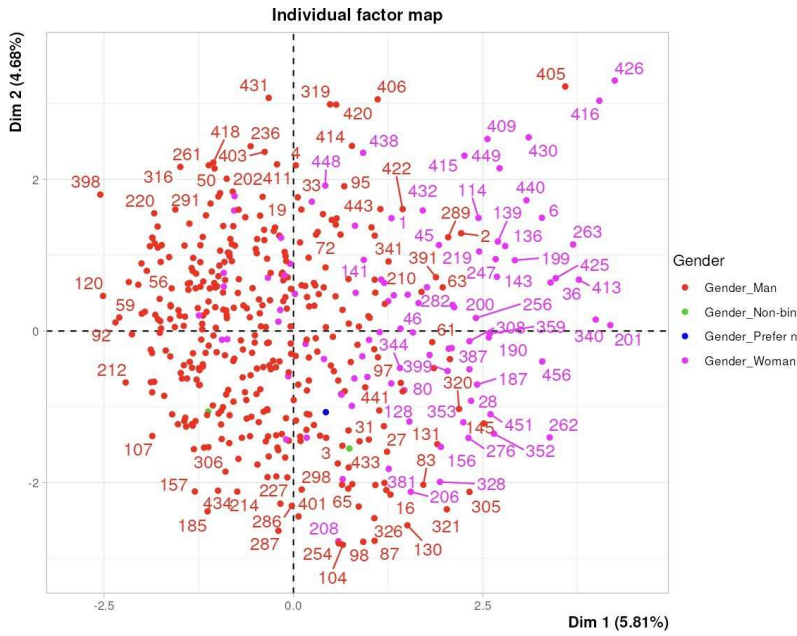


Figure 5. Graph of individuals color-segmented by gender.

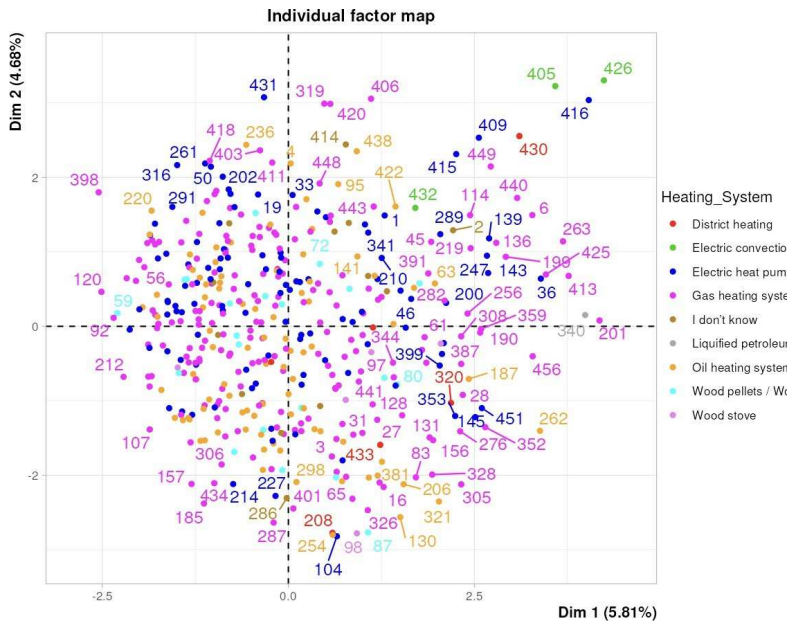


Figure 6. Graph of individuals color-segmented by heating system.



value and personality traits form a secondary, largely independent dimension.

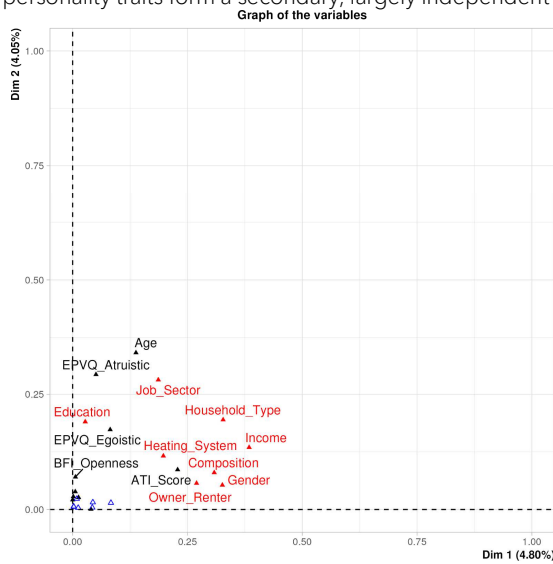


Figure 8. The representation of variables (relationship square) of both continuous and categorical variables on the same figure.

#### 1.1.4. Clustering using Gower distance and Partitioning around Medoid (PAM)

Gower’s (1971) proposal is the most popular way of measuring the similarity/dissimilarity between observations in the presence of mixed-type variables. The Gower’s distance can be defined as the complement to one of the Gower’s similarity coefficients:

$$s_{\{ij\}} = \frac{\sum_{k=1}^p w_k s_{ijk}}{\sum_{k=1}^p w_k}, \quad d_{\{ij\}} = 1 - s_{\{ij\}}$$

where

- $p$  = total number of variables,
- $w_k$  = weight for variable kkk (1 if both observations are non-missing, 0 otherwise),
- $s_{ijk}$  = partial similarity between observations iii and jjj on variable kkk (defined according to that variable’s measurement level),
- $s_{ij}$  = overall Gower similarity, and
- $d_{ij}$  = Gower distance (the complement to one of the similarity).

It serves as a dissimilarity or distance metric between unit  $i$  and unit  $j$ , denoted  $d^t = 1 - s^t$  where  $d^t$  represents the distance computed based on the  $t^{th}$  variable.  $s^t$  denotes the similarity between units  $i$  and  $j$  concerning the  $t^{th}$  variable. Its value is contingent on the type of variable in question.

This method can be used to construct the dissimilarity distance matrix and measure how different two records are. The distance was therefore a good option to deal with the mixed-data type of the survey's variables. Indeed, the records contained a combination of logical, numerical, categorical and text data. The Gower distance is always a positive number and is bound in the [0, 1] interval where a value of 0 indicates a significant similarity (identical) whereas a value of 1 indicates that the two records are completely different (maximally dissimilar).

**Conclusion from PAM clustering:**

Using the Gower distance and based on both the total within sum of square and the silhouette we obtained an optimal number of 3 clusters using Partitioning around Medoid (PAM) clustering.

Figure 9 illustrates the optimal number of clusters when dealing with all mixed variables and the Gower distance as the dissimilarity between two records. For visualization purposes, the clustering results were projected on the first two principal components. From Figure 10, we can see that no clear clusters can be visualized on the two first principal components (which is normal since they both account for less than 20% of the variation).

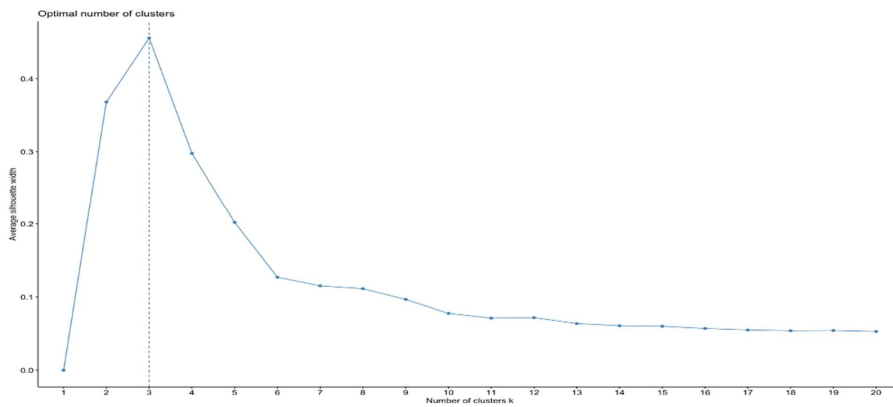


Figure 9. Average Silhouette width index to help in determining the optimal number of cluster.

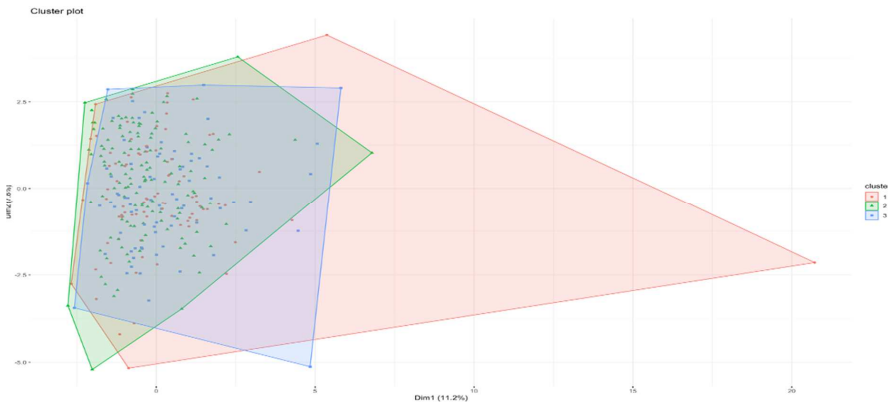


Figure 10. Results of the PAM clustering projected on the first two principal components.

#### 1.1.5. K-means clustering using Euclidean distance

Given that the results obtained so far were not as satisfactory as hoped, an alternative clustering procedure using the Euclidean distance and K-means was implemented. As K-means is an unsupervised machine learning algorithm that requires also the number of partitions as input, we tested several indices such as the silhouette and the total within-cluster sum of squares. The number of optimal clusters was three, which was similar to our previous result. Figure 11 and Figure 12 illustrate the results obtained for the K-mean clustering applied to the whole survey dataset.

#### Conclusion from K-means clustering:

- The value in the table corresponds to the v-test value. The v-test here corresponds to the quantile of the normal distribution which is associated with p-value. The sign indicates an over- or under- representation.
- A v-test lower than -2 means that the individuals in this cluster obtain a value significantly different. For continuous variables, a v-test lower than -2 indicates that the variable in the cluster has a value significantly lower than the population.

	Cluster 1 ↕	Cluster 2 ↕	Cluster 3 ↕
Income=3000-4999 €	-8.15	8.34	-0.331
Owner_Renter=Renter	-8.1	7.88	1.1
Household_Type=Apartment	-7.64	6.98	2.65
Composition=One person household	-7.4	7.15	1.17
Job_Sector=Transportation and Logistics	-5.32	5.45	-0.207
Heating_System=District heating	-4.26	4.33	-0.0646
Gender=Woman	-4.03	3.77	1.23
Composition=Prefer not to disclose	-3.83	3.9	-0.054
Job_Sector=Retail	-3.83	3.91	-0.0752
Lit_Benef_PV	-3.16	2.87	1.41
Lit_EV_Charge	-3.12	3.07	0.479
Heating_System=Electric convection heater	-2.82	-0.541	5.16
Income=1000-2999 €	-2.36	1.1	3.02
Lit_Delay_DW	-2.33	1.59	3.12
Income=< 1000 €	-2.19	2.23	-0.0218
Job_Sector=Hospitality	-2.19	2.23	-0.0218
Job_Sector=Student	-2.19	-0.382	4.02
Job_Sector=Education	-2.17	2.34	-0.431
BFI_Conscientiousness	-1.97	1.97	0.164
Composition=Multiple adults in cohabitation	0.198	-1.74	2.8

Figure 11. K-means clustering results.

Description of each cluster by the categories

```

=====
`$ 1`

```

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Owner_Renter=Owner	87.58782	98.4210526	93.2314410	5.545902e-16	8.098899
Income>= 9000 €	93.63057	38.6842105	34.2794760	3.806435e-06	4.621681
Composition=Couple with child(ren)	89.76378	60.0000000	55.4585153	1.807911e-05	4.287382
Gender=Man	86.70360	82.3684211	78.8209607	1.152102e-04	3.856100
Household_Type=Semi-detached house	91.26984	30.2631579	27.5109170	2.517212e-03	3.021265
Job_Sector=Information Technology	97.67442	11.0526316	9.3886463	2.518337e-03	3.021129
Heating_System=Wood pellets / Wood chips	100.00000	5.5263158	4.5851528	1.798910e-02	2.365842
Composition=Single-parent household	100.00000	5.2631579	4.3668122	2.188674e-02	2.292328
Job_Sector=Finance	92.98246	13.9473684	12.4454148	2.385523e-02	2.259453
Income=Prefer not to disclose	90.32258	22.1052632	20.3056769	2.968758e-02	2.174234
Household_Type=Detached house	86.91589	48.9473684	46.7248908	3.578491e-02	2.099363
Job_Sector=Education	70.45455	8.1578947	9.6069869	3.036455e-02	-2.165302
Income< 1000 €	0.00000	0.0000000	0.4366812	2.869483e-02	-2.187654
Job_Sector=Student	0.00000	0.0000000	0.4366812	2.869483e-02	-2.187654
Job_Sector=Hospitality	0.00000	0.0000000	0.4366812	2.869483e-02	-2.187654
Income=1000-2999 €	50.00000	1.3157895	2.1834061	1.823323e-02	-2.360847
Heating_System=Electric convection heater	0.00000	0.0000000	0.6550218	4.782472e-03	-2.821332
Composition=Prefer not to disclose	0.00000	0.0000000	1.0917031	1.284926e-04	-3.829330
Job_Sector=Retail	14.28571	0.2631579	1.5283843	1.284524e-04	-3.829407
Gender=Woman	68.08511	16.8421053	20.5240175	5.520869e-05	-4.032406
Heating_System=District heating	0.00000	0.0000000	1.3100437	2.070631e-05	-4.257135
Job_Sector=Transportation and Logistics	30.00000	1.5789474	4.3668122	1.057633e-07	-5.316533
Composition=One person household	21.42857	1.5789474	6.1135371	1.364016e-13	-7.399789
Household_Type=Apartment	43.33333	6.8421053	13.1004367	2.248514e-14	-7.635554
Owner_Renter=Renter	19.35484	1.5789474	6.7685590	5.545902e-16	-8.098899
Income=3000-4999 €	21.21212	1.8421053	7.2052402	3.530530e-16	-8.153667

Figure 12. Description of the categories.

### 1.1.6. Clustering based on personality test

In this section, rather than performing a clustering on all features, a clustering analysis was performed on the result of the personality and electricity literacy test only. Figure 13 shows the different scores obtained by each participant with regard to each personality and literacy question. The questions are variables with values ranging from 1 to 7 and represent Likert scale scores. For this reason, we chose the Jaccard similarity coefficient. The Jaccard coefficient quantifies similarity between finite sample sets by comparing the size of their intersection to the size of their union. It is defined by  $J(A, B) = |A \cap B| / |A \cup B|$ , and the Jaccard distance is expressed as  $d(A, B) = 1 - J(A, B)$ .

By performing hierarchical clustering with an agglomerative approach and using Jaccard's distance as the distance matrix between respondents we obtained the following clusters shown in Figure 14 with an optimal cut-off at four clusters for the variables and five for the respondents.

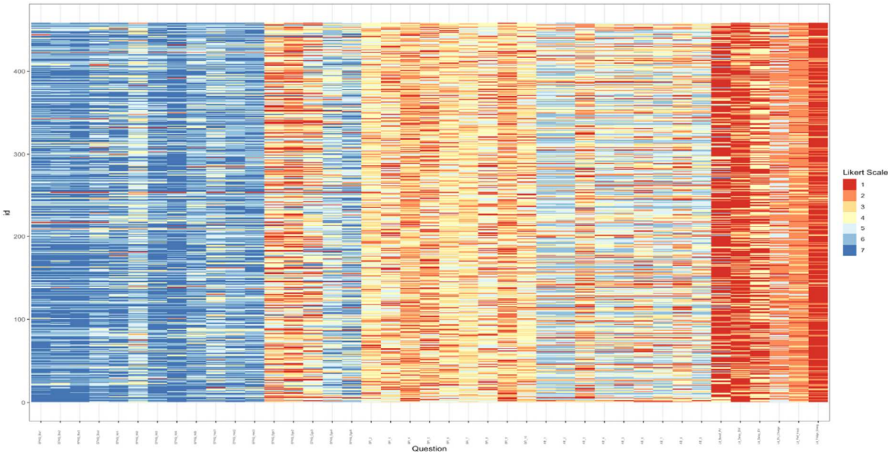


Figure 13. Overview of the personality and literacy scores obtained by every participant. Each row is a participant survey answers to all tests. Each column corresponds to a question.

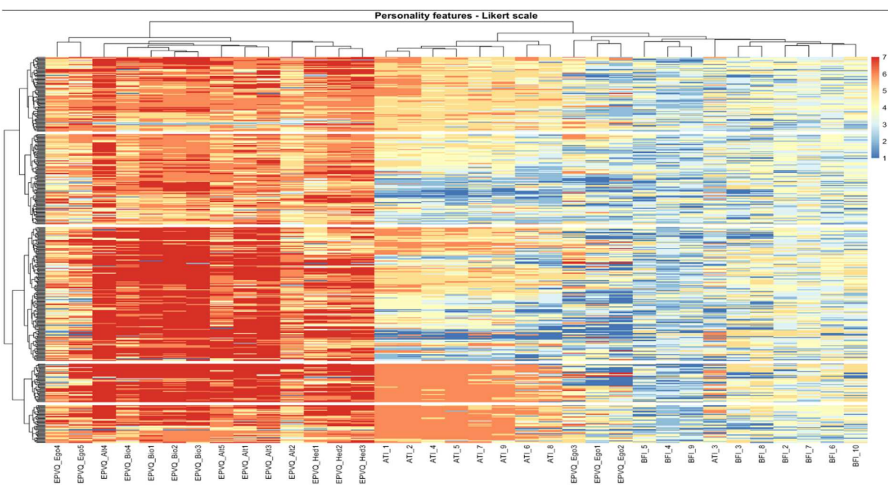


Figure 14. Result of the Hierarchical Clustering approach with Jaccard' distance. Both variables and respondents are clustered.

## 2. Refined analysis

---

This section reports on an advanced approach that was explored as an alternative to the classical ones detailed in the previous section, to get more significant and exploitable results. The goal was to uncover meaningful clusters among respondents based on their responses, using a medoid-based clustering approach. The methodology provides several advantages, enhancing the analysis and delivering valuable insights.

The survey contained multiple questions, some of which were correlated. Therefore, the analysis prudently focused on reducing the number of features and highlighting key variables based on expert knowledge. This method aimed to uncover meaningful clusters within the data, which would be beneficial for understanding the survey results.

### 2.1. Medoid-based clustering approach

#### 2.1.1. Main steps

The main steps of the approach are:

1. Start with an initial dataset based on expert knowledge.
2. Complete the input space (if needed).
3. Reduce redundancy in the data.
4. Perform medoid-based clustering using Gower distance as dissimilarity measure.
5. Visualize and describe the obtained clusters.
6. Use models to infer relationships between input variables and clusters.

#### 2.1.2. Application

The proposed approach offers an efficient way to analyse the data, ensuring finding meaningful clusters with the smallest number of features.

##### Step 1: Initial Dataset

After thorough discussions, key features were selected based on their relevance to the analysis:

- "BF1\_1", "BFI\_6", "BFI\_2", "BFI\_7", "BFI\_3", "BFI\_8", "BFI\_4", "BFI\_9"
- "EPVQ\_Bio1", "EPVQ\_Bio2", "EPVQ\_Bio3", "EPVQ\_Bio4"
- "Owner\_Renter", "Dongle", "Heating\_System", "Household\_EV", "Household\_PV".

This initial selection ensured that the most important features were included from the outset.

##### Step 2: Adding Features

To complete the input space, additional features were added to ensure no relevant information was lost. This step ensured that all potentially relevant features were included, thereby maximizing the information available for clustering.

### Step 3: Reducing Redundancy

Using an unsupervised feature selection algorithm, redundant features were removed, keeping only the most informative features. This step improved the clarity of the analysis and enhanced the quality of the clusters.

The final dataset was:

- "ATI\_8", "i\_flex\_dish\_low.0\_\_RESPONSE", "i\_flex\_dish\_med.0\_\_RESPONSE", "Age"
- "ATI\_3", "EPVQ\_Ego1", "Job\_Sector", "i\_flex\_wash\_low.0\_\_RESPONSE"
- "Education", "BFI\_5", "Household\_Type", "EPVQ\_Hed1"
- "i\_flex\_dish\_high.0\_\_RESPONSE", "ATI\_9", "i\_flex\_car\_low.0\_\_RESPONSE", "Income".

### Step 4: Clustering

#### Step 4.1: Medoid-Based Clustering

Two clusters were obtained based on silhouette scores.

#### Step 4.2: Cluster Description

The clusters were described through various visualizations, which helped in understanding the distinguishing features of each cluster.

Main variables distinguishing the clusters included:

- "Age", "Flex\_Car\_low", "ATI\_8", "BFI\_3", "Flex\_dish\_med", "Flex\_dish\_low", "Flex\_dish\_high", "Flex\_wash\_low", "EPVQ\_Hed1".

These detailed descriptions provide clarity and enhance the understanding of the clusters.

### Step 5: Model-Based Inference

An additional step was performed to understand better the results. Various models were used to infer relationships between the input variables and the clusters, as example we applied:

1. **General Linear Model (GLM):** The variables influencing the model the most were highlighted.
2. **Random Forest (RF):** Important features in the random forest model were identified.

These inference models provide valuable insights into the characteristics and behaviour of each cluster, enhancing the analysis' depth.

## 2.2. Results and discussion

In this section, we present the results of the clustering analysis performed on the survey data. The clustering aimed to identify distinct groups within the dataset to understand better the variability and commonalities in respondent behaviours and preferences.

### 2.2.1. Clustering Performance and Visualization

Figure 15 displays the results of the medoid-based clustering analysis, where two distinct clusters were identified. The clustering validity was assessed using silhouette scores, a measure of how similar an object is to its own cluster compared to other clusters. Higher silhouette scores indicate a better-defined clustering structure, and as shown in Figure 15, the silhouette scores suggest that the two clusters are indeed well-separated and coherent.

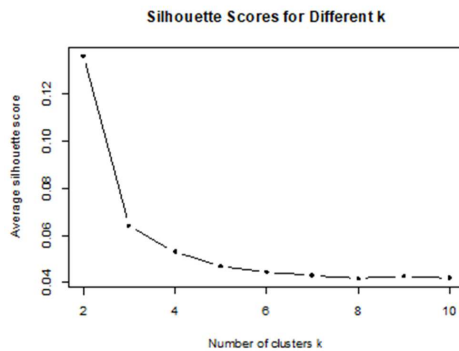


Figure 15. Medoid based clustering. 2 clusters are obtained based on the values of silhouette.

The visualization of these clusters in a two-dimensional space is presented in Figure 16 using TSNE (t-distributed Stochastic Neighbor Embedding). This method is particularly useful for visualizing high-dimensional data in lower dimensions while preserving the local structure of the data. Figure 16 illustrates how data points grouped into clusters are distinctly set apart, highlighting the effectiveness of the clustering process in distinguishing between different respondent types based on their survey responses.

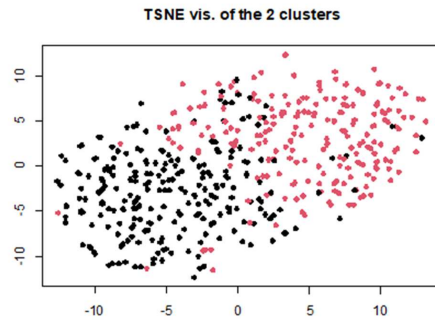


Figure 16. visualisation of these two cluster in 2-dimensional space using TSNE.

2.2.2. Detailed Cluster Analysis

Figure 17 presents a boxplot visualization, where the internal distribution of variables within each of the two identified clusters is shown. This plot, distinguishing the clusters in red and green, provides insights into the median, first, and third quartiles for each variable. The clear differentiation in the interquartile ranges and medians across many variables supports the existence of meaningful distinctions between the two clusters.

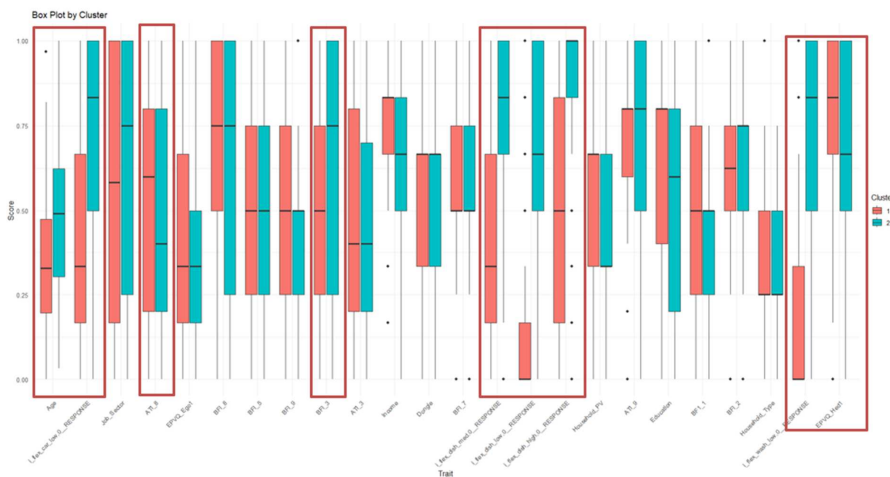


Figure 17. Boxplot showing 2 clusters: red vs green (median, 1st and 3rd quantile for each variable).

In Figure 18, a radar chart compares the median values of scaled variables across the obtained clusters. This visualization method allows for an overview of multiple

variables simultaneously, facilitating a direct comparison of how each cluster's characteristics differ. The radar chart in Figure 18 effectively underscores the distinct profiles captured by each cluster, reflecting diverse respondent characteristics that might be related to their preferences or demographic backgrounds.

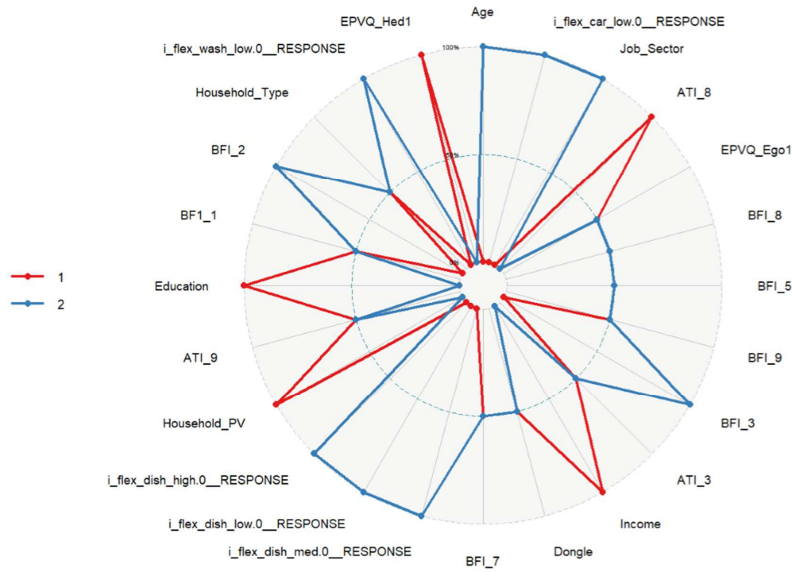


Figure 18. Radar visualisation of median of scaled variables vs obtained clusters.

### 2.2.3. Influence of Variables on Cluster Assignment

Figure 19 ranks and visualizes the variables that most significantly influence the General Linear Model (GLM). This figure highlights the predictive power of each variable concerning the cluster assignments, illustrating which features are most instrumental in defining the clusters. Understanding these key drivers is crucial for interpreting the clustering results and can guide further analysis or interventions based on these insights.

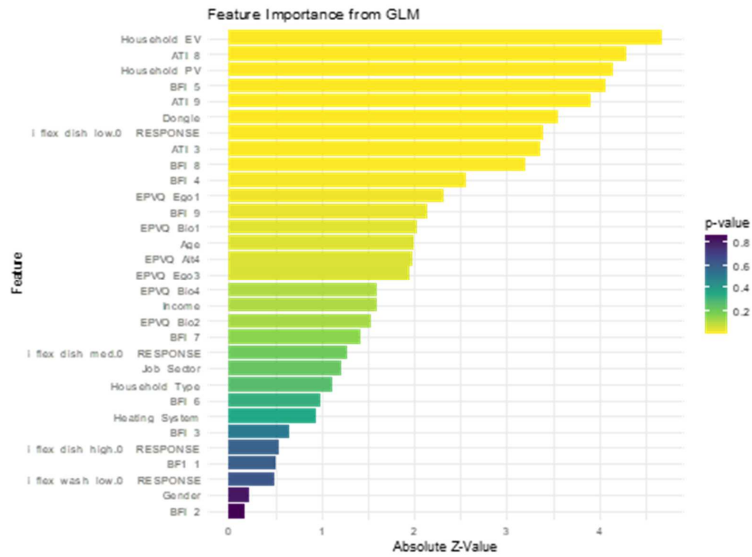


Figure 19. List of variables that influence the most GLM model.

2.2.4. Distribution of Survey Scores within Clusters

Figure 20 explores the distribution of survey scores within the clusters, with scores scaled between 0 and 1. This visualization helps compare how different survey responses are distributed across the identified clusters, showcasing the extent of variation or consistency in responses within each cluster. This figure provides a nuanced view of the clusters' composition, revealing potential patterns or anomalies in how different groups perceive or respond to survey items.

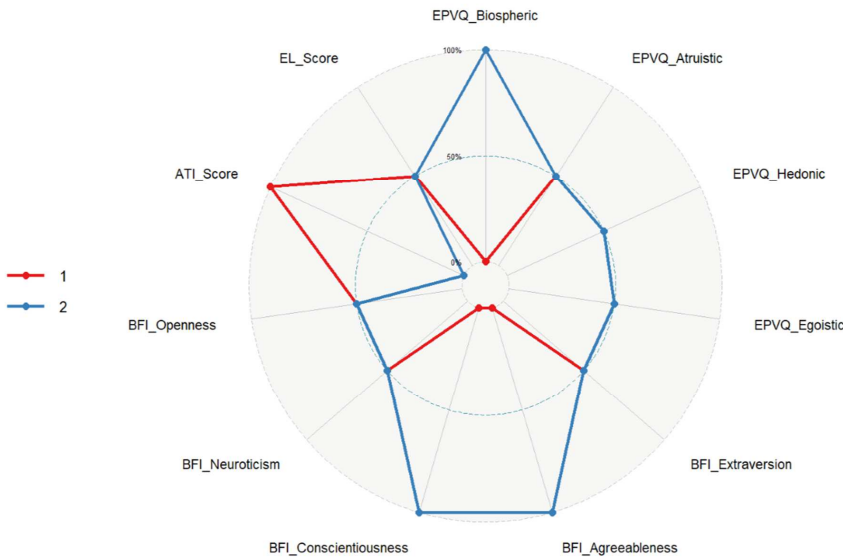


Figure 20. Visualisation of scores of survey within the detected clusters (scaled between 0 and 1).

### 2.3. Discussion

**Group 1** has the same characteristics as in previous analysis, if we consider the other scores with medium scores are the same as in the other group and we do not seek to analyse them. In addition, since the ATI score remains low, we can now say people in group 1 mostly care about the environment, they are agreeable and conscientious and are mostly not technology friends.

**Group 2** keeps only a high ATI score and would be characterised in addition by its low scores in biospheric, conscientiousness and agreeableness. People in group 2 do not really care about the environment, are technology friends, are mostly not agreeable, nor conscientious. They tend to be more literate than in group 1 but the difference is not significant. They might also have a higher openness to experience than people in group 1 have, but this remains to be confirmed.

### 2.4. Conclusion

The analysis presented here, supported by statistical and visual tools, confirms the presence of distinct clusters within the energy literacy and flexibility survey data. These results provide valuable insights into the diverse characteristics and preferences of the survey respondents, enhancing our understanding of the underlying patterns in the data.

From their answer to the survey, the participants sample could be grouped into two separated clusters of similar households having each their own unique characteristics. One category of households is represented by individuals with a higher emphasis on environmental issues, who are agreeable and conscientious, and most are not technology friends. The other category represents individuals who are less sensitive to environment, with a lower agreeableness and conscientiousness, but who are technology friends. They tend to be more literate and have a higher openness to experience (linked to their personality) than people from the first group, but the difference is not significant, and this remains to be confirmed. The other characteristics are mostly similar among the two groups, not statistically determinant. Overall, this analysis confirms the relationship between sensitivity to environment and conscientiousness, highlighting in addition the influence of agreeableness (which we know however is biased), and the affinity to technology interaction, which might also be a determinant.

## 2.5. Relevant references

### 1. Medoid-Based Clustering:

- Kaufman, L., & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. This book discusses various clustering techniques including medoid-based methods such as Partitioning Around Medoids (PAM), which are less sensitive to outliers compared to k-means.

### 2. Silhouette Measure:

- Rousseeuw, P.J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*, 20, 53-65. This paper introduces the silhouette measure as a means to assess the quality of clustering and determine the optimal number of clusters.

### 3. Gower Distance:

- Gower, J.C. (1971). "A General Coefficient of Similarity and Some of Its Properties". *Biometrics*, 27(4), 857-871. This paper details the Gower distance metric, ideal for handling mixed data types in clustering analysis.

### 4. t-SNE (t-Distributed Stochastic Neighbor Embedding):

- van der Maaten, L., & Hinton, G. (2008). "Visualizing Data using t-SNE". *Journal of Machine Learning Research*, 9(Nov), 2579-2605. This paper provides a comprehensive overview of t-SNE, a technique for dimensionality reduction particularly suited for visualizing high-dimensional datasets.

### 3. Appendix: Overview of used techniques and measures

---

#### 3.1. Appendix A. Medoid-based clustering

Medoid-based clustering is a type of partitioning algorithm within cluster analysis that is particularly useful when a measure of the "centre" or most representative point of a cluster is needed. Unlike k-means clustering, which minimizes the sum of squared deviations from the mean (making it sensitive to outliers), medoid-based clustering aims to minimize the sum of dissimilarities between points labelled to belong to a cluster and a point designated as the centre of that cluster, known as the medoid.

##### Key Concepts

- A medoid is the most centrally located object in a cluster, with respect to sum of distances or dissimilarities to all other objects within the cluster. It can be thought of as the most representative or most typical point of a cluster.
- One of the most common medoid-based clustering algorithms is the Partitioning Around Medoids (PAM). The PAM algorithm works as follows:
  - **Initialization:** Start by randomly selecting k points as the medoids.
  - **Assignment Step:** Assign each object to the nearest medoid, based on the chosen distance metric (e.g., Euclidean, Manhattan).
  - **Update Step:** For each medoid and each non-medoid pair, calculate the total cost of swapping the medoid with a non-medoid. If any swap reduces the cost, the swap is made. This process iterates until no further improvements can be made.
- Another variation is the k-medoids algorithm, which is more computationally efficient and effective for large datasets.
- Medoid-based methods are more robust to noise and outliers compared to centroid-based methods like k-means. Since medoids are actual data points, their position is less influenced by extreme values.
- Medoid-based clustering is suitable for applications where a typical exemplar is needed for each cluster. It's widely used in fields like finance (identifying typical market behaviors), biology (finding representative genes or proteins), and image processing (segmentation based on typical pixel values).
- The computational cost can be higher than k-means, especially for large datasets, because of the need to continually compute and compare the cost of swapping medoids with non-medoids.
- Choosing the appropriate number of clusters (k) and the distance metric still requires domain knowledge and possibly heuristic approaches, as with other clustering methods.

Medoid-based clustering is particularly advantageous in scenarios where the dataset contains outliers or anomalies that might skew the mean of a cluster, as in k-means. It's also beneficial when the data involves non-numeric attributes since medoids require only a defined distance or dissimilarity metric rather than means or averages

that are not always meaningful for categorical data. Thus, medoid-based methods offer a versatile and robust approach to clustering, making them a valuable tool in the data analysis toolkit.

### 3.2. Appendix B. Gower distance

Gower distance is a metric used primarily in statistics for measuring similarity or dissimilarity between data points when the variables involved are of mixed types—such as ordinal, nominal, and interval variables. This measure is particularly useful in clustering and other multivariate analyses where data do not conform to purely numerical or categorical types. The Gower distance metric is versatile and robust, accommodating datasets with varied variable types without requiring transformation to a single common scale.

#### How Gower Distance is Computed

The computation of Gower distance involves the following steps:

1. **Standardization of Variables:** Each variable type is treated differently to ensure comparability:
  - **Quantitative variables:** Typically scaled by dividing by the range (maximum minus minimum) to normalize them between 0 and 1.
  - **Ordinal variables:** Converted to ranks if not already in rank form, then scaled similarly to quantitative variables.
  - **Nominal (categorical) variables:** Converted into a binary dissimilarity measure (0 if the same, 1 if different).
2. **Pairwise Comparisons:** For each pair of objects, the dissimilarity is calculated for each variable. The Gower distance is the average of these dissimilarities. The formula for a pair of objects  $x$  and  $y$  across  $n$  is:

$$D(x, y) = \frac{\sum_{j=1}^n w_j \cdot d_j(x, y)}{\sum_{j=1}^n w_j}$$

where  $d_j(x, y)$  is the dissimilarity score for variable  $j$  (scaled to  $[0, 1]$ ), and  $w_j$  is the weight for variable  $j$ , which can be adjusted based on the importance or reliability of each variable.

3. **Handling Missing Data:** Gower distance can easily handle missing data by adjusting the denominator in the calculation for each pair of objects. Weights  $w_j$  are only summed for variables that are not missing for either object in the pair, allowing comparisons even in incomplete datasets.
  - **Flexibility:** Gower distance can handle mixed data types effectively, making it highly versatile for real-world data scenarios.
  - **Robustness to Missing Data:** It naturally accommodates missing values by adjusting the base of comparison, which is crucial in datasets where missing data are common.

- **No Need for Data Transformation:** Unlike other metrics that require conversion of all data to a single type (e.g., all numeric), Gower distance can use data in their original forms.
- **Performance Issues:** Calculating Gower distance can be computationally intensive, especially for large datasets, because it requires pairwise comparisons across all variables for each pair of objects.

### 3.3. Appendix C. Silhouette Measure

The silhouette measure, or silhouette coefficient, is a metric used to evaluate the quality of clusters in cluster analysis. Introduced by Peter J. Rousseeuw in 1987, it quantifies how well each object has been classified in the clustering process. The silhouette measure provides a concise graphical representation of how closely each point in one cluster is to points in the neighbouring clusters. This method is widely used because it provides a clear and interpretable measure of cluster tightness and separation.

#### Calculation of the Silhouette Measure

The silhouette measure is calculated for each individual data point in a dataset. For each point  $i$ :

1. **Calculate  $a(i)$ :** This is the average distance between  $i$  and all other data points in the same cluster.  $a(i)$  represents how well  $i$  is matched to its own cluster, considered as a measure of cohesion.
2. **Calculate  $b(i)$ :** This is the minimum average distance from  $i$  to all points in any other cluster, of which  $i$  is not a member.  $b(i)$  quantifies how well  $i$  is separated from the nearest cluster to which it does not belong, considered as a measure of separation.
3. **Calculate the silhouette coefficient ( $s(i)$ )** for each individual data point:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- $s(i)$  ranges from  $-1$  to  $1$ .
- A value of  $1$  indicates that the point is very well matched to its own cluster and poorly matched to neighbouring clusters.
- A value of  $0$  indicates that the point is on or very close to the decision boundary between two neighbouring clusters.
- A value of  $-1$  indicates that the point might have been placed in the wrong cluster.

#### Uses of the Silhouette Measure

- **Cluster Validation:** The silhouette measure is used to validate the consistency within clusters in a dataset. A high average silhouette width across all data points indicates good clustering.
- **Determining the Number of Clusters:** The silhouette measure can be used to determine the optimal number of clusters by calculating it for various numbers of clusters and choosing the number that maximizes the average silhouette coefficient.
- **Comparative Assessment:** It allows for the comparison of the effectiveness of different clustering algorithms or configurations, providing a visual way to compare how tightly grouped the clusters in different models are.

#### Advantages

- **Interpretability:** The silhouette measure is straightforward and provides an intuitive graphical method to assess clustering quality.
- **Applicability:** It does not assume clusters to be of any specific shape and size, which makes it applicable to a wide variety of data types and clustering algorithms.
- **Versatility:** It can be used with any distance metric and is not limited to Euclidean spaces.

### 3.4. Appendix D. TSNE

t-Distributed Stochastic Neighbour Embedding (t-SNE) is a powerful machine learning algorithm primarily used for the visualization of high-dimensional data. It was developed by Laurens van der Maaten and Geoffrey Hinton in 2008. t-SNE is particularly well-suited for the visualization of complex datasets by reducing the number of dimensions to two or three, making it easier to plot and interpret visually.

#### How t-SNE Works

1. **Similarity Computation in High-Dimensional Space:** t-SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. These probabilities denote the likelihood that one data point would pick another as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at that point.
2. **Similarity Computation in Low-Dimensional Space:** t-SNE then defines a similar probability distribution over the points in the low-dimensional map, but it uses a Student-t distribution to measure similarities in the lower-dimensional space. This distribution has heavier tails, which allows distant points to be modeled more accurately in the reduced space.
3. **Minimization of Kullback-Leibler Divergence:** The algorithm iteratively adjusts the positions of the points in the low-dimensional map to minimize the Kullback-Leibler divergence between the two distributions (high-dimensional and low-dimensional). This divergence is a measure of how one probability distribution diverges from a second, expected probability distribution.

### Advantages of t-SNE

- **Cluster Visibility:** t-SNE is very effective at creating a map that reveals structures within the data, such as clusters of similar data points.
- **Handling Non-Linearity:** Unlike PCA which is a linear dimension reduction technique, t-SNE can capture much more complex nonlinear structures.
- **Flexibility:** It can be used on a wide range of data types and is particularly good at handling data that lie on several different, but related, manifolds.

### Limitations

- **Computational Complexity:** t-SNE can be quite slow on very large datasets because of its non-linear nature and the optimization process it uses.
- **Hyperparameter Sensitivity:** The performance of t-SNE is highly dependent on the choice of perplexity and learning rate. These parameters can significantly affect the outcome and need careful tuning.
- **Global vs. Local Trade-Off:** While t-SNE excels at capturing the local structure of data, it sometimes exaggerates clusters and can lose the global picture.

Despite these limitations, t-SNE remains a popular tool for data exploration and visualization in many fields such as bioinformatics, finance, and computer vision, where high-dimensional data is common.