Robust distances for mixed-type data

Aurea Grané

Universidad Carlos III de Madrid

DaSSWeb

Data Science and Statistics Webinar

University of Porto, May 20th 2025

Outline

1. Motivation
2. Tailoring a metric with Related Metric Scaling
3. Clustering algorithms for large datasets
4. Application: Mental well-being profiles in older adults

Motivation

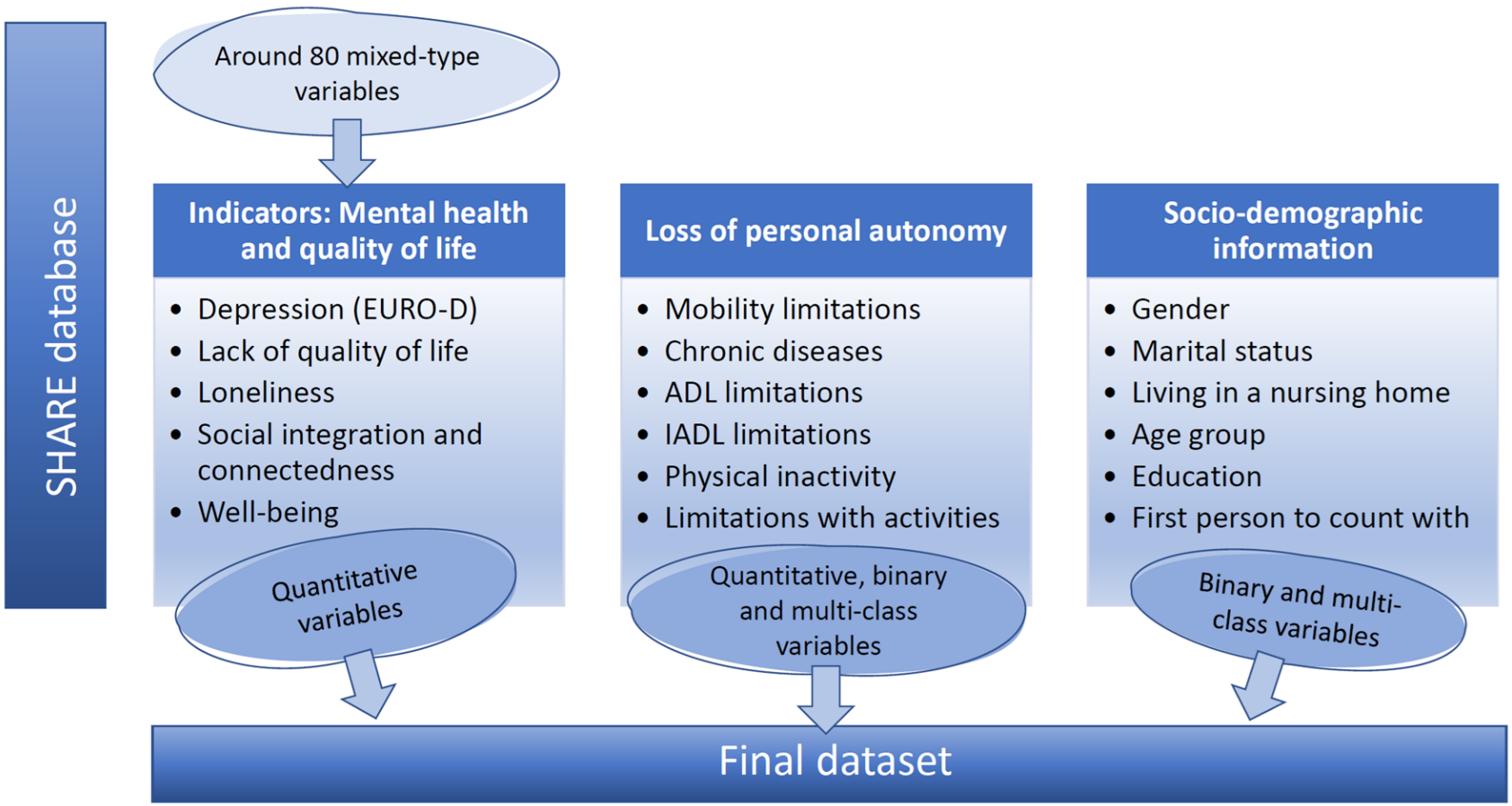
- The nature of data is, more than ever, of mixed type: quantitative and qualitative variables, textual, functional data, etc.
- Challenges arise in clustering mixed-type data with correlations, redundancies, and outliers.
- A common approach (when working with quantitative and qualitative variables) is to compute Gower distance between units, obtain a Euclidean representation (orthogonal axes) and finally cluster the units via k-means, k-medians. Other approaches skip the Euclidean representation by applying k-medoids directly to Gower's distance matrix.
- These strategies can lead to sub-optimal results since
  - Classical Gower distance is neither robust nor able to incorporate redundancy among variables,
  - k-medoids turns out to be computationally unfeasible for moderately large sample sizes.

Main goals

1. To develop a package with new distances for mixed-type data able able to deal with redundancy and outliers
   - `PyDistances`, `https://pypi.org/project/PyDistances/`
2. To develop computationally efficient clustering algorithms for large datasets based on these new robust distances
   - Fast k-medoids and q-Fold Fast k-medoids, `https://pypi.org/project/FastKmedoids/`
3. To evaluate their performance through an extensive simulation study, and compare them to a wide range of existing clustering alternatives in terms of both predictive power and computational efficiency.
4. Application: To create profiles of older adults to better understand the differing levels of emotional well-being across Europe (Survey of Health, Ageing and Retirement in Europe).

Let's start with the dataset

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a rich panel database of individuals aged 50 or over in 26 European countries and Israel. We took wave 9 (2021/2022).



Tailoring a metric with Related Metric Scaling (RelMS)

My contributions to the design of robust metrics for mixed-type data started with the work Albarrán et al. (2015) and rely in a general technique developed by Cuadras and Fortiana (1995).

Since then, this methodology has been applied to:

- Visualization of mixed-type data (R. Romera, UC3M),
- Cluster weighted mixed-type data (I. Albarrán, UC3M),
- Detect multivariate outlying units (S. Salini, UNIMI),
- Robustify distance-based predictive models (E. Boj, UB).

Distance measures for mixed-type data via RelMS

Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_m]$ be a data matrix corresponding to a set of variables of m different types measured on a sample of n units.

- In this presentation, we use $m = 3$ for quantitative, binary and multi-class variables.
- The procedure that follows is general enough to be applied to any type of variable/information (functional data, textual data, images, manifolds, etc.) provided that a distance measure can be computed between pairs of units (See Cuadras and Fortiana 1995, Grané and Romera 2018, for the proofs and properties of Related Metric Scaling).

1. Split $\mathbf{X}$ in m matrices $\mathbf{X}_k$, $k = 1, \ldots, m$, regarding each variable type.
2. For each $\mathbf{X}_k$ consider a proper distance measure between units, according to the characteristics of the data. Next, compute the matrix of squared pairwise distances (conveniently standardized) by its geometric variability (Cuadras and Fortiana 1995):

$$\mathbf{\Delta}_k = \frac{1}{V_{\mathbf{\Delta}_k}} \left(\delta_k^2(\mathbf{x}_{k,i}, \mathbf{x}_{k,j}) \right)_{\{1 \leq i,j \leq n\}}, \tag{1}$$

   where $\mathbf{x}_{k,i}$ is the i-th row of $\mathbf{X}_k$, $V_{\mathbf{\Delta}_k} = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_k^2(\mathbf{x}_{k,i}, \mathbf{x}_{k,j})$.
3. For each $\mathbf{\Delta}_k$ compute its Gram matrix $\mathbf{G}_k = -\frac{1}{2} \mathbf{H} \mathbf{\Delta}_k \mathbf{H}$, where $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}'$, $\mathbf{I}$ $n \times n$ identity matrix and $\mathbf{1}$ $n \times 1$ vector of ones.
4. Check for *Euclideanity*: $\mathbf{G}_k$ must be positive definite; If this is not the case, transform $\mathbf{\Delta}_k$ so that $\mathbf{G}_k > 0$ (Borg and Groenen 2005).
5. Combine all Gram matrices to get the joint Gram matrix (RelMS):

$$\mathbf{G} = \sum_{k=1}^{m} \mathbf{G}_k - \frac{1}{m} \sum_{k \neq l} \mathbf{G}_k^{1/2} \mathbf{G}_l^{1/2}, \tag{2}$$

   where $\mathbf{G}_k^{1/2}$ is the square root of $\mathbf{G}_k$.
6. Recover the joint distance: $\mathbf{\Delta} = \mathbf{g}\,\mathbf{1}' + \mathbf{1}\,\mathbf{g}' - 2\,\mathbf{G}$, where $\mathbf{g} = \mathrm{diag}(\mathbf{G})$ column vector containing the diagonal of $\mathbf{G}$.

- We call the first addend of formula (2) generalized Gower distance (G-Gower), because it mimics classical Gower distance by adding different distances, although here the addition is done through the matrices of square distances.
- The second addend of formula (2) is responsible of discarding redundant information coming from different sources. This second part equals to zero when for each $k \neq l$ the Euclidean configurations associated to each $\mathbf{\Delta}_k$ and $\mathbf{\Delta}_l$ generate orthogonal subspaces on $\mathbb{R}^n$. This second part can be computationally expensive.
- A simplified version is obtained by considering G-Gower (square) distance:

$$\mathbf{\Delta}_{GG} = \sum_{k=1}^{m} \mathbf{\Delta}_k, \tag{3}$$

where each $\mathbf{\Delta}_k$ is defined as in (1). Equivalentely, (square) G-Gower is obtained from the first addend of formula (2).

Distances included in PyDistances

Joint distances for mixed-type data can be obtained via RelMS or G-Gower by combining:

- Distances for numerical data: Euclidean (ℓ^2 distance), Manhattan (ℓ^1 distance), Canberra, Pearson (standardized ℓ^2 distance), Mahalanobis, robust Mahalanobis (MAD, trimmed, winsozired),
- Distances for binary data: Associated to Jaccard (ignores doble zeros) and Sokal-Michener (takes into account double zeros) similarity coefficients. The general transformation given in Gower (1966) is applied to obtain a distance from a similarity coefficient.
- Distances for multi-class data: Hamming (associated to simple matching coefficient).

Our recommendations

- In the presence of an underlying correlation structure or outlying units, combinations including robust Mahalanobis distance for quantitative data are recommended.
- In such cases, compared to classical Gower distance, we get:
  - More stability and robustness in MDS representations (Albarrán et al. 2015, Grané and Romera 2018),
  - More stability in clustering results (Grané et al. 2021, Grané and Scielzo-Ortiz 2025),
  - More efficiency in distance-based predictive models (Boj and Grané 2024, 2025).

Clustering algorithms for large datasets

Two proposals:

- Fast k-medoids
- q-Fold Fast k-medoids

To be used in combination with the previous distances specially designed for mixed-type data

Part of the PhD Thesis of F. Scielzo-Ortiz (on-going work)

Fast k-medoids

$\mathbf{X}$ data matrix corresponding to a set of variables measured on a sample of n units, $\delta(\mathbf{x}_i, \mathbf{x}_j)$ distance between units i, j.

Fast k-medoids algorithm:

1. Partition-step: Partition at random the sample units in two disjoint sets, S and $\overline{S}$. Let $\mathbf{X}_S$ be the matrix with the measurements of the n_S selected units and $\mathbf{X}_{\overline{S}}$ the corresponding one for the not selected units.
2. Cluster-step: Apply classic k-medoids on $\mathbf{X}_S$, obtain the sample clusters $C_1, \ldots, C_k$, and the medoids $\overline{\mathbf{x}}_{C_1}, \ldots, \overline{\mathbf{x}}_{C_k}$.
3. Assignment-step: Assign the not selected units to the nearest cluster. For each unit $i \in \overline{S}$, assign unit i to cluster C_{r^*} where $r^* = \arg \min_{r \in \{1,\ldots,k\}} \delta(\mathbf{x}_i, \overline{\mathbf{x}}_{C_r})$, and $\mathbf{x}_i$ is the i-th row of $\mathbf{X}_{\overline{S}}$ and $\overline{\mathbf{x}}_{C_r}$ the medoid of cluster C_r.

Key parameters of Fast k-medoids: δ, k (as in k-medoids), n_S.

q-Fold Fast k-medoids

Fast k-medoids may loose accuracy as the sample size increases.
q-Fold Fast k-medoids algorithm is a hierarquical strategy:

1. Split $\mathbf{X}$ in q disjoint folds $\mathbf{X}_{F_1}, \ldots, \mathbf{X}_{F_q}$
2. For each fold $\mathbf{X}_{F_j}$ $(j = 1, \ldots, q)$, apply Fast k-medoids, obtain the clusters $C_1^{F_j}, \ldots, C_k^{F_j}$, and the medoids $\overline{\mathbf{x}}_{1^{F_j}}, \ldots, \overline{\mathbf{x}}_{k^{F_j}}$.
3. Build the medoid matrix $\mathbf{X_M}$ by concatenating the medoids of each fold by rows.
4. Apply Fast k-medoids to $\mathbf{X_M}$ and obtain the clusters $C_1^{\mathbf{X}_M}, \ldots, C_k^{\mathbf{X}_M}$.
5. Final clusters rule: unit i is assigned to the 4-step cluster that contains the medoid of the 2-step cluster where i belongs. If $i \in C_h^{F_j}$, then i is assigned to $C_r^{\mathbf{X}_M} \Leftrightarrow \overline{\mathbf{x}}_{h^{F_j}} \in C_r^{\mathbf{X}_M} \quad \forall i = 1, \ldots, n,\ j = 1, \ldots, q,\ h, r = 1, \ldots, k.$

Key parameters: those already key in Fast k-medoids and q.

Empirical evaluation

These algorithms are implemented in `FastKmedoids` Python package, and here we evaluate them using G-Gower distance:

Distances in G-Gower's family are built as a combination of three distances (conveniently standardized by their geometric variability):

Num. Euclidean, Manhattan, Canberra, Mahalanobis, robust Mahalanobis (trimmed, winsorized, MAD),

Bin. Sokal-Michener, Jaccard,

Cat. Hamming.

Adjusted Accuracy$^{(1)}$, Adjusted Rand Index and computation time are used to evaluate the goodness of the clustering as well as the efficiency of the algorithms.

(1) Defined as the accuracy of the optimal encoding of the clusters. All possible encoding combinations for the clusters are explored in order to compute it.

Scenarios considered

All datasets were initially generated with $p = 8$ numerical variables with underlying correlation structure. Next, four of them were categorized in order to end with a mixed-type dataset with 4 numerical, 2 binary and 2 multi-class variables.

	n	Clusters	Correlation structure L: lower than 0.3 - H: higher than 0.6 (in absolute value)	Outlier contamination (in numerical variables)
1	35k	4	71 % L - 7 % H	5 % in 2 variables
2	35k	4	61 % L - 4 % H	no outliers
3	100k	4	71 % L - 7 % H	5 % in 2 variables
4	300k	3	64 % L - 4 % H	5 % in 2 variables
5	600k	3	64 % L - 4 % H	5 % in 2 variables
6	1M	3	64 % L - 4 % H	5 % in 2 variables
7	300k	3	54 % L - 4 % H	no outliers
8	300k	3	64 % L - 14 % H	7 %–14 % in 4 variables

Computational experiments

I Comparison between k-medoids and Fast k-medoids (accuracy and computing time),

II Stability of Fast k-medoids (accuracy and adjusted Rand index),

III Accuracy and computation time for Fast k-medoids,

IV Accuracy and computation time for q-Fold Fast k-medoids,

V Empirical comparison to other clustering algorithms.

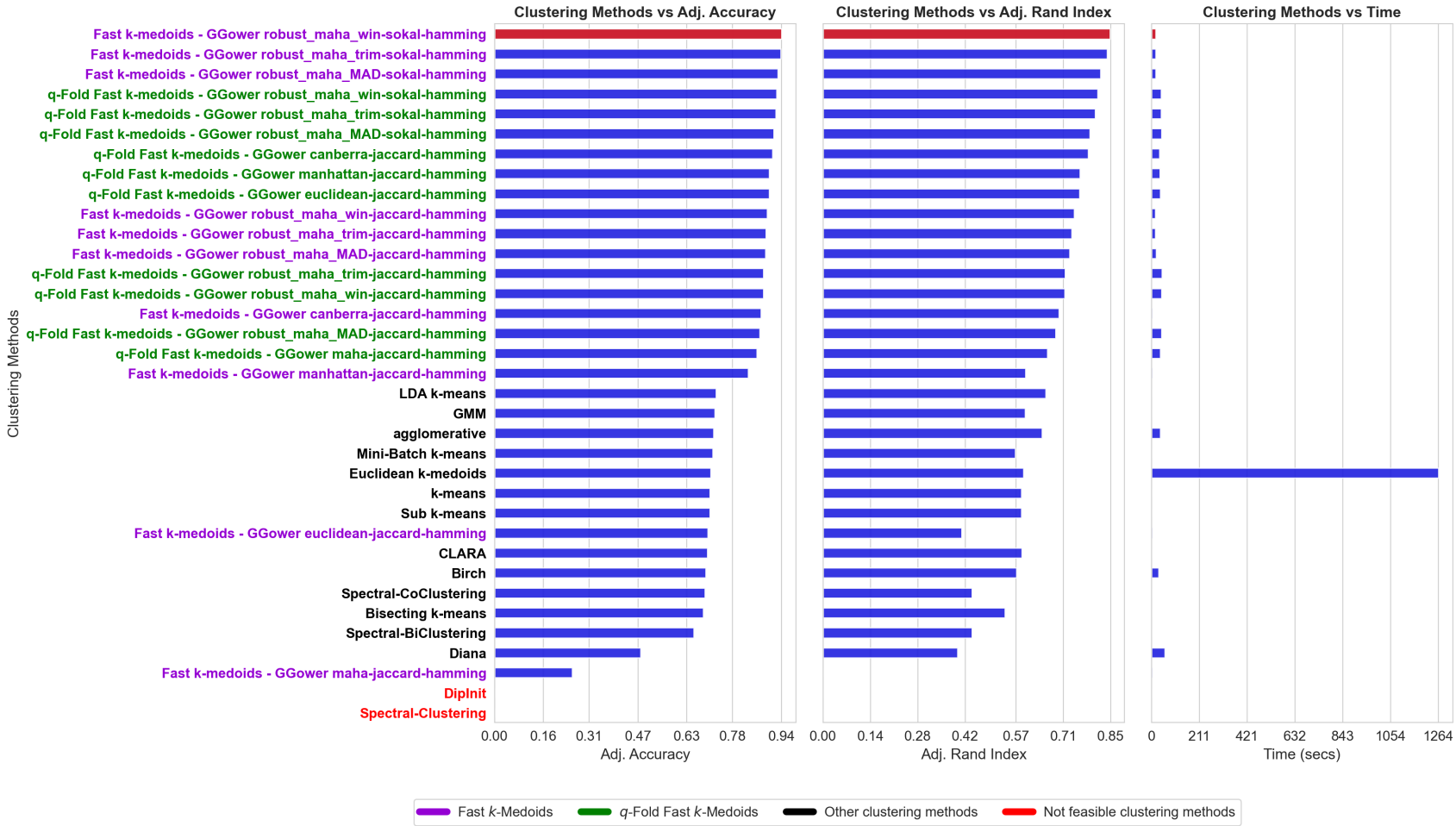
(See Grané and Scielzo-Ortiz 2025 for a complete study).

V. Empirical comparison to other clustering algorithms

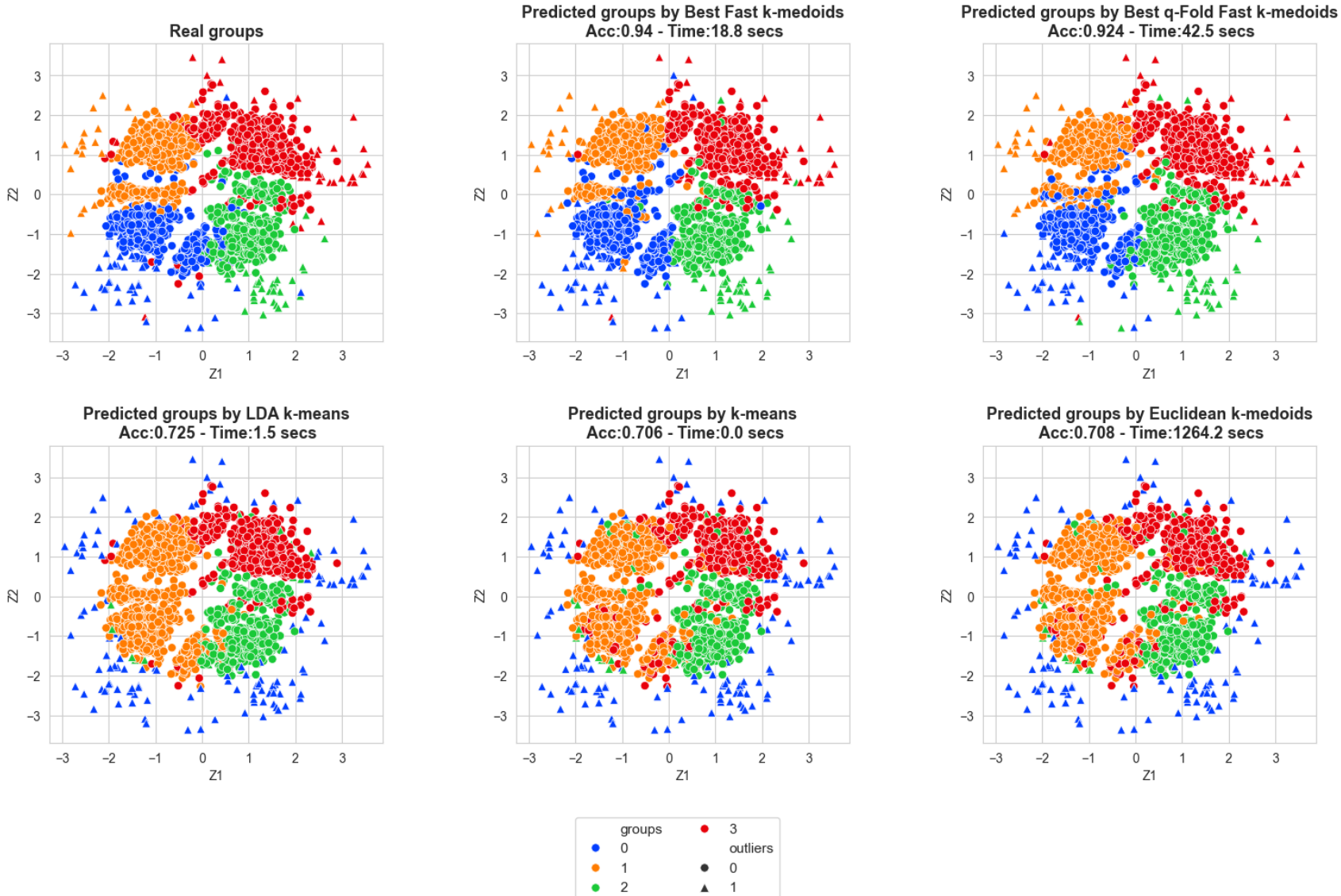
- Agglomerative,
- Birch,
- CLARA,
- Diana,
- DipInit,
- GMM,
- k-medoids,
- k-means, LDA k-means, MiniBatch k-means, Bisecting k-means, Sub k-means,
- Spectral clustering, Spectral bi-clustering, Spectral co-clustering,

All implemented in `scikit-learn` and `scikit-learn extra` Python packages.

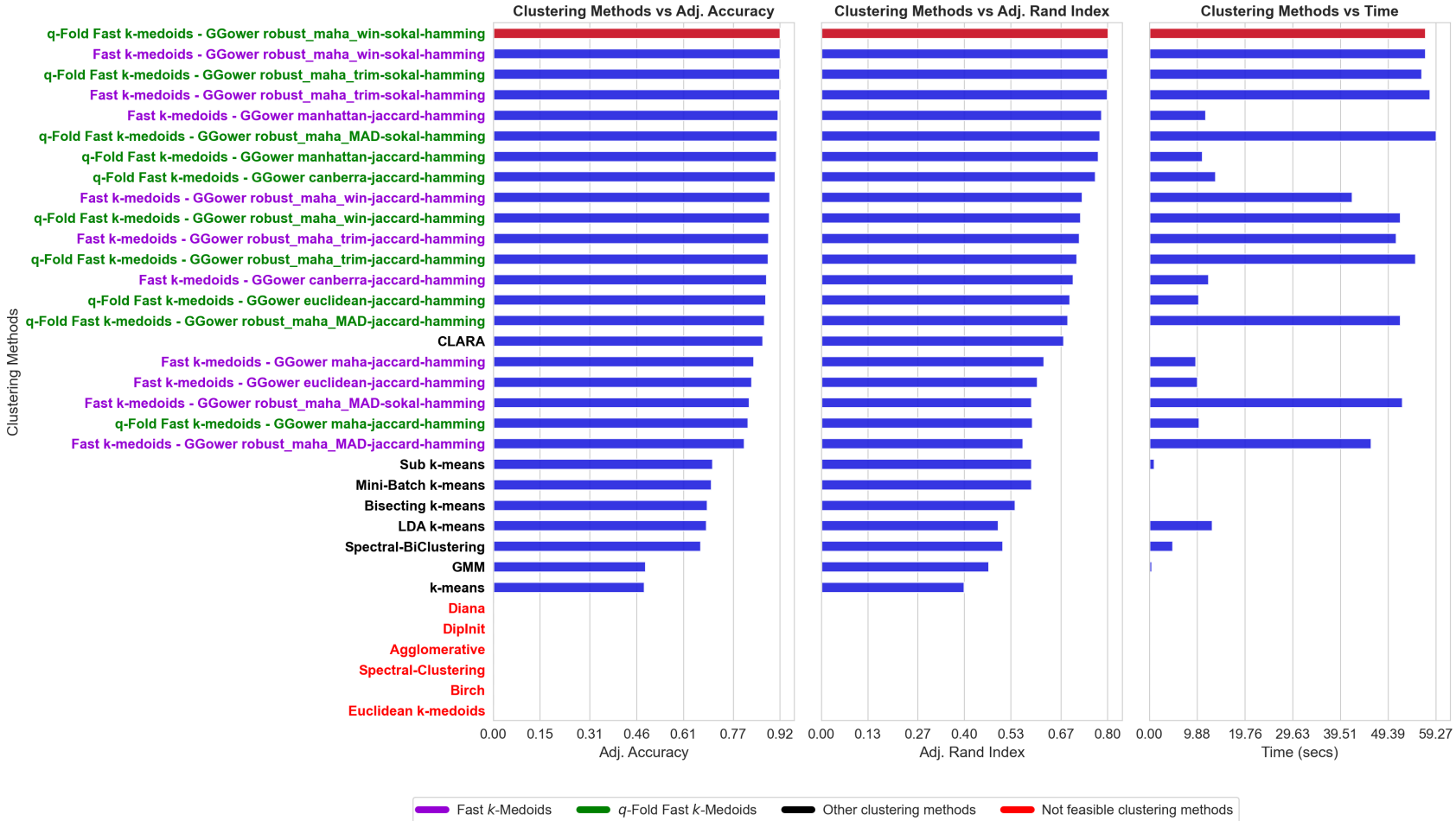
Dataset 1 (35k)



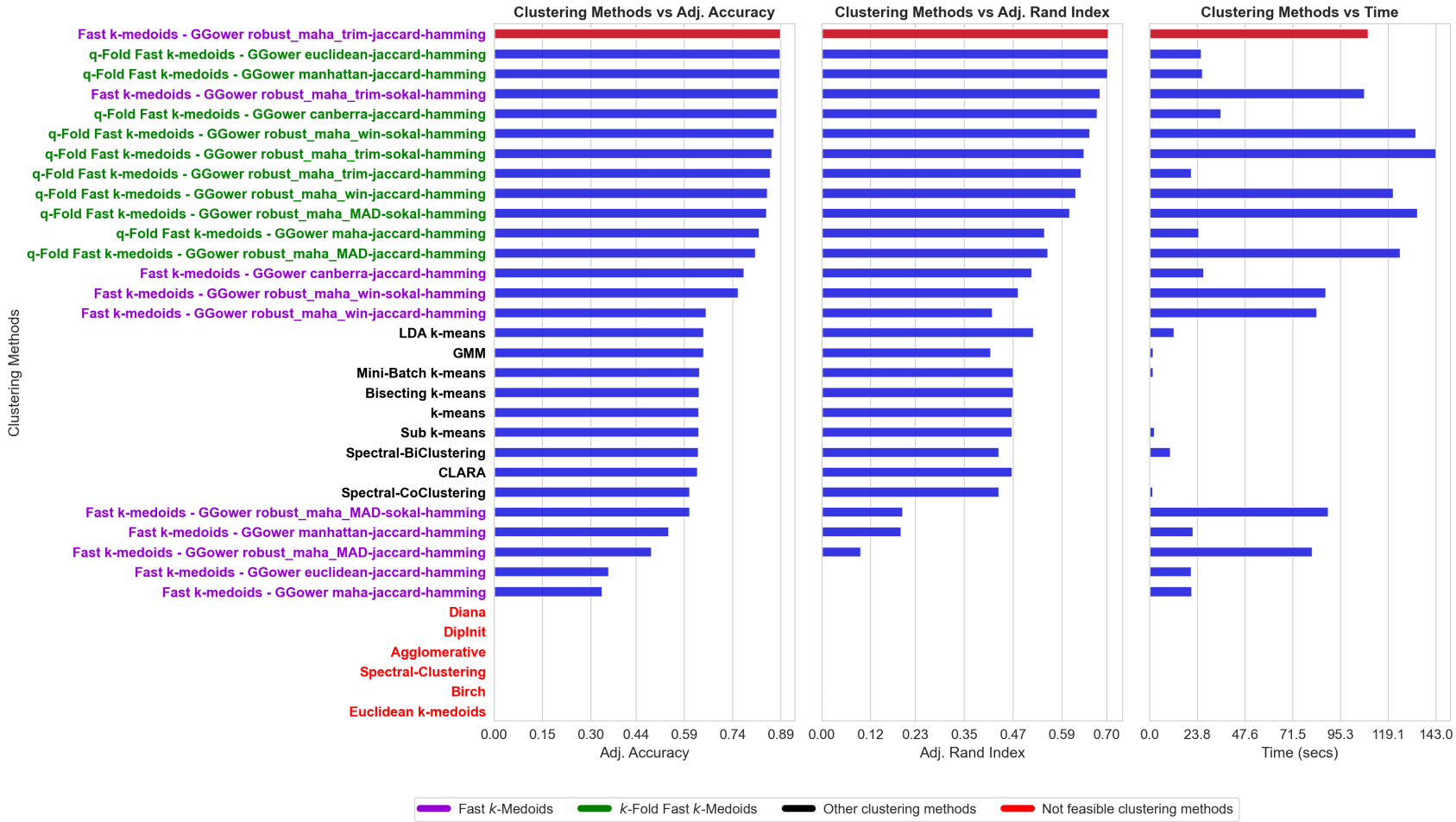
Dataset 1 (35k). Clustering visualization with MDS computed from G-Gower with robust Mahalanobis (winsorized), Sokal and Hamming



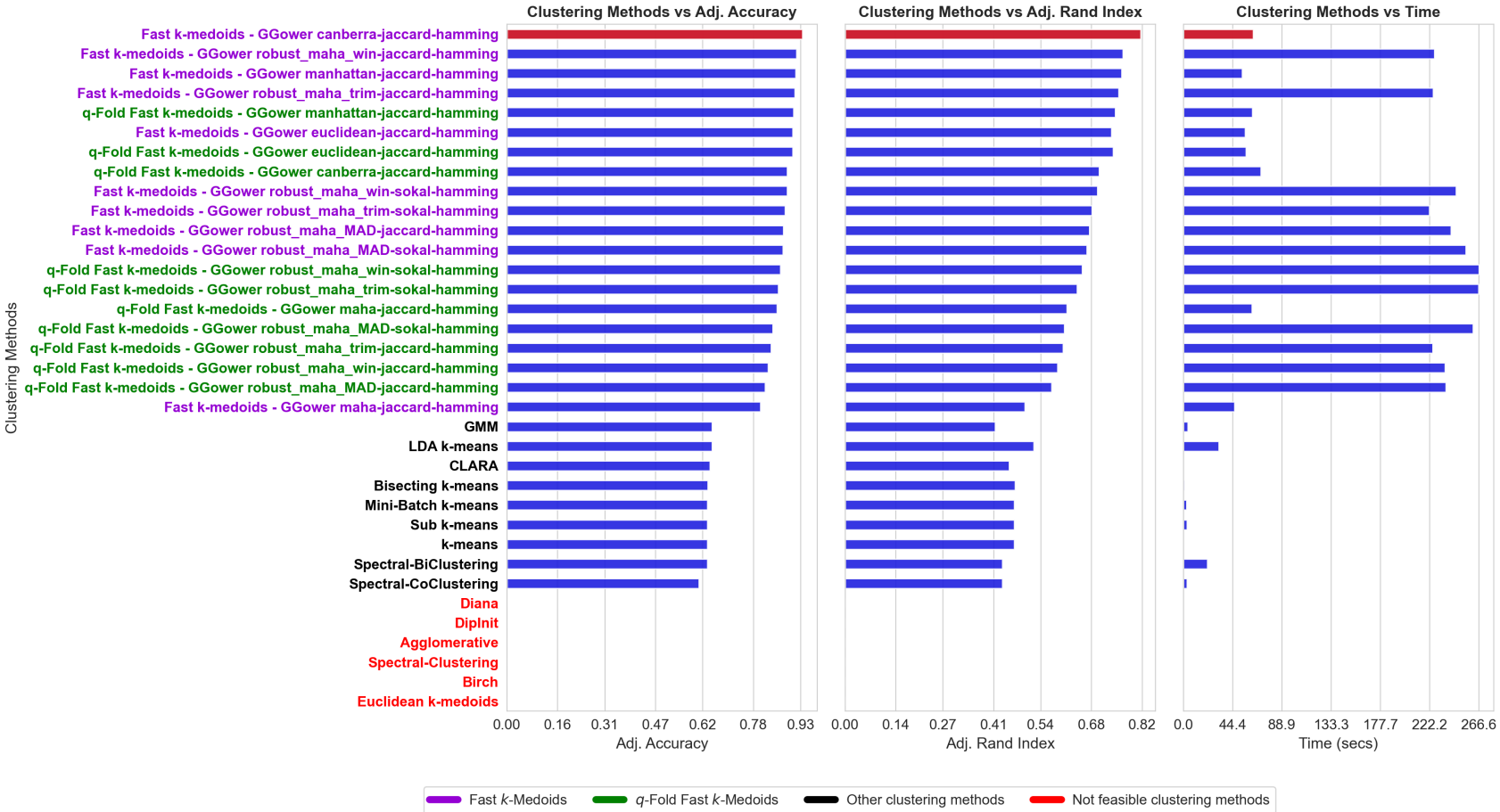
Dataset 3 (100k)



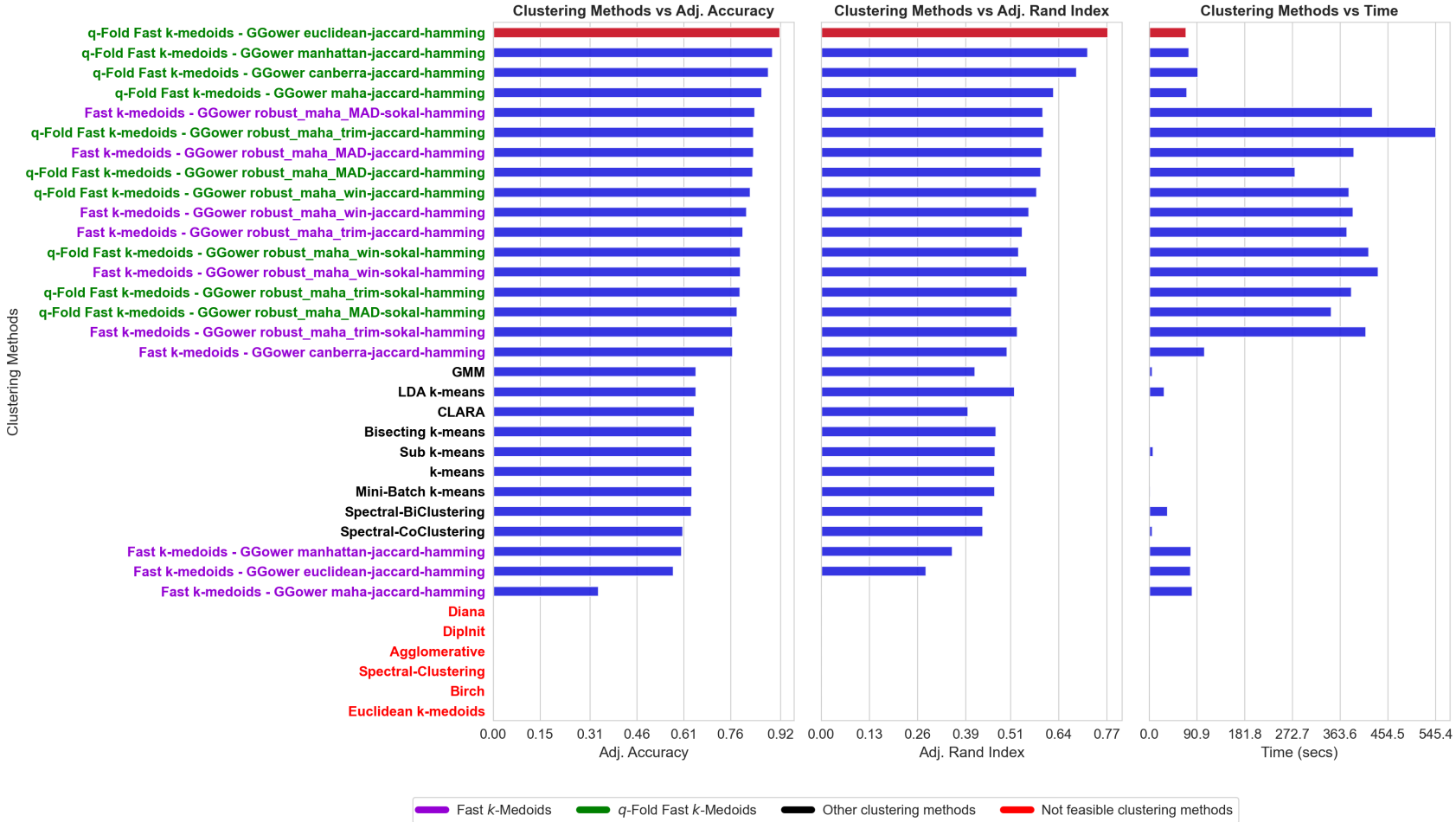
Dataset 4 (300k)



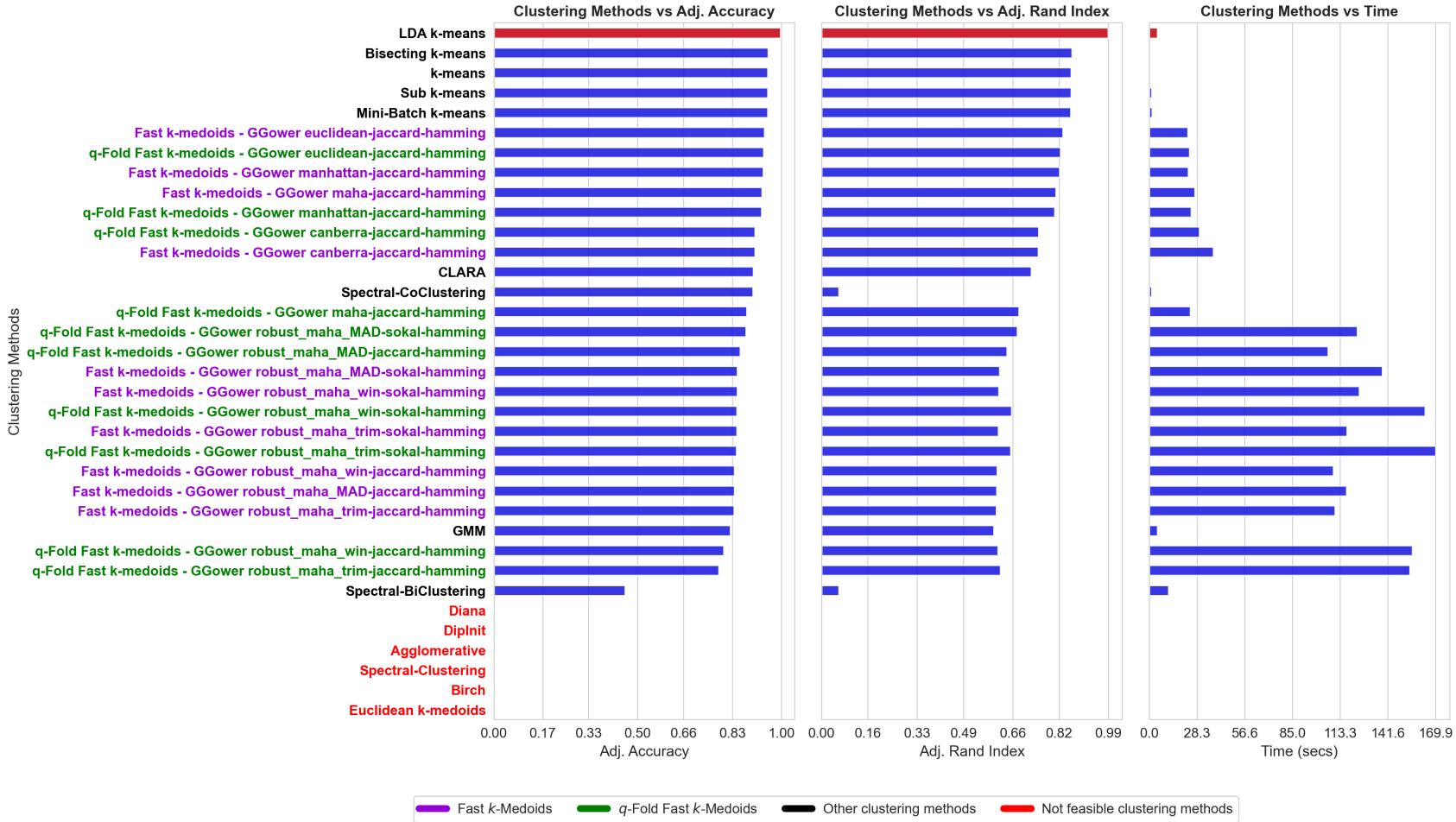
Dataset 5 (600k)



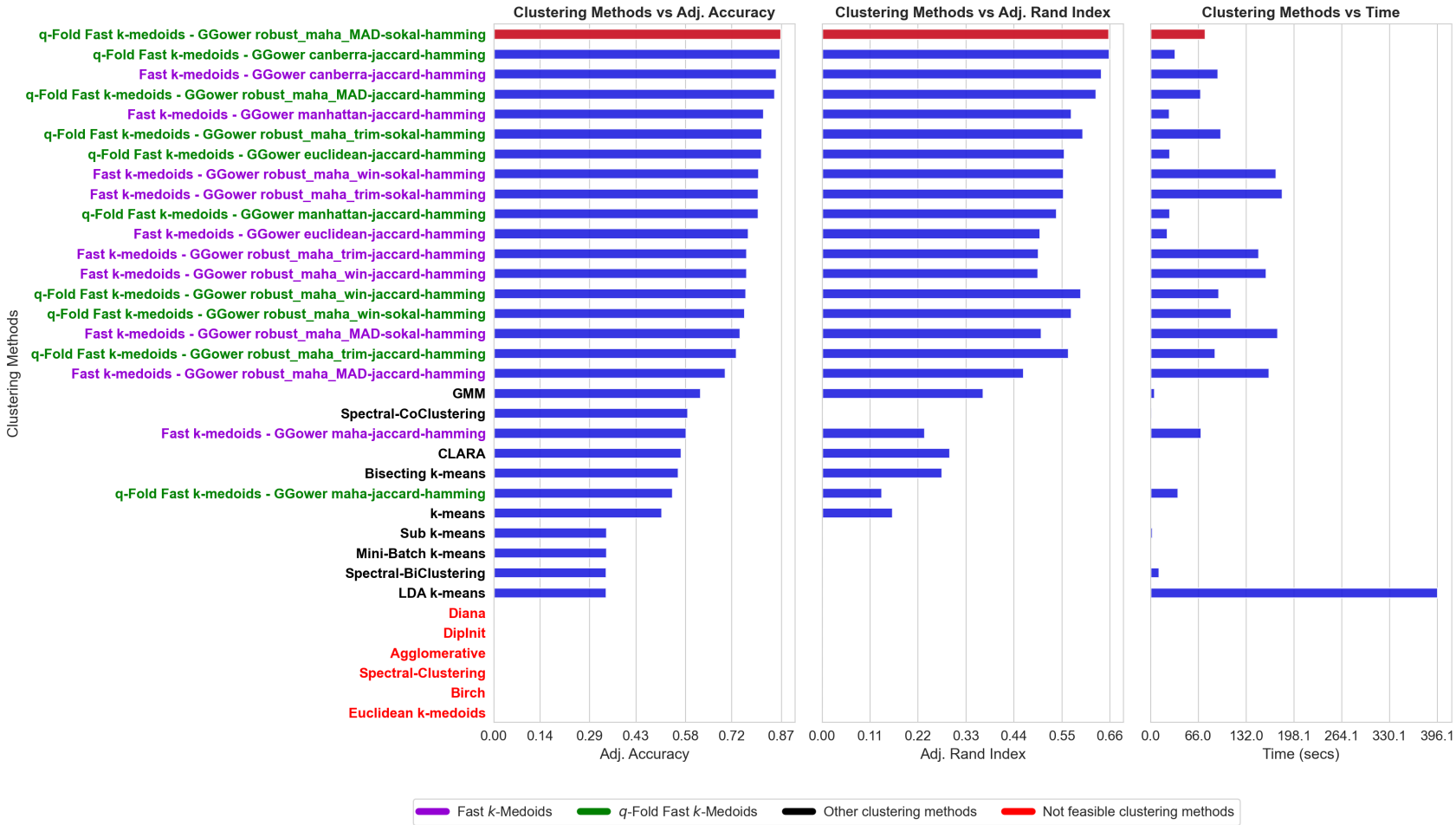
Dataset 6 (1M)


Clustering Methods vs Adj. Accuracy
Clustering Methods vs Adj. Rand Index
Clustering Methods vs Time
Clustering Methods
q-Fold Fast k-medoids - GGower euclidean-jaccard-hamming
q-Fold Fast k-medoids - GGower manhattan-jaccard-hamming
q-Fold Fast k-medoids - GGower canberra-jaccard-hamming
q-Fold Fast k-medoids - GGower maha-jaccard-hamming
Fast k-medoids - GGower robust_maha_MAD-sokal-hamming
q-Fold Fast k-medoids - GGower robust_maha_trim-jaccard-hamming
Fast k-medoids - GGower robust_maha_MAD-jaccard-hamming
q-Fold Fast k-medoids - GGower robust_maha_MAD-jaccard-hamming
q-Fold Fast k-medoids - GGower robust_maha_win-jaccard-hamming
Fast k-medoids - GGower robust_maha_win-jaccard-hamming
Fast k-medoids - GGower robust_maha_trim-jaccard-hamming
q-Fold Fast k-medoids - GGower robust_maha_win-sokal-hamming
Fast k-medoids - GGower robust_maha_win-sokal-hamming
q-Fold Fast k-medoids - GGower robust_maha_trim-sokal-hamming
q-Fold Fast k-medoids - GGower robust_maha_MAD-sokal-hamming
Fast k-medoids - GGower robust_maha_trim-sokal-hamming
Fast k-medoids - GGower canberra-jaccard-hamming
GMM
LDA k-means
CLARA
Bisecting k-means
Sub k-means
k-means
Mini-Batch k-means
Spectral-BiClustering
Spectral-CoClustering
Fast k-medoids - GGower manhattan-jaccard-hamming
Fast k-medoids - GGower euclidean-jaccard-hamming
Fast k-medoids - GGower maha-jaccard-hamming
Diana
DipInit
Agglomerative
Spectral-Clustering
Birch
Euclidean k-medoids
0.00 0.15 0.31 0.46 0.61 0.76 0.92
Adj. Accuracy
0.00 0.13 0.26 0.39 0.51 0.64 0.77
Adj. Rand Index
0.0 90.9 181.8 272.7 363.6 454.5 545.4
Time (secs)
Fast k-Medoids
q-Fold Fast k-Medoids
Other clustering methods
Not feasible clustering methods

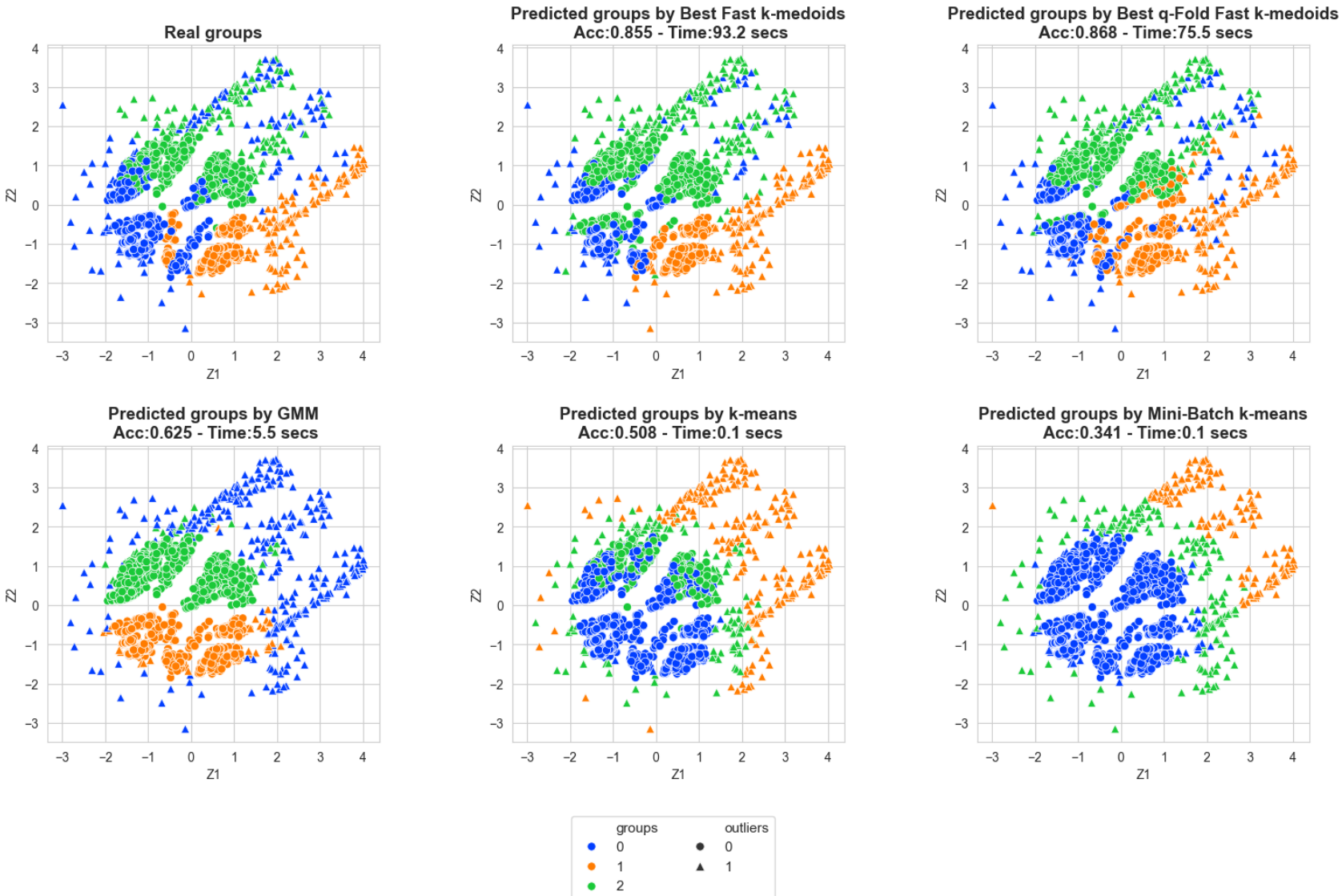
Dataset 7 (300k, no outliers)



Dataset 8 (300k, more outliers, more correlation)



Dataset 8 (300k, more outliers, more correlation). Clustering visualization with MDS computed from G-Gower with robust Mahalanobis (MAD), Sokal and Hamming

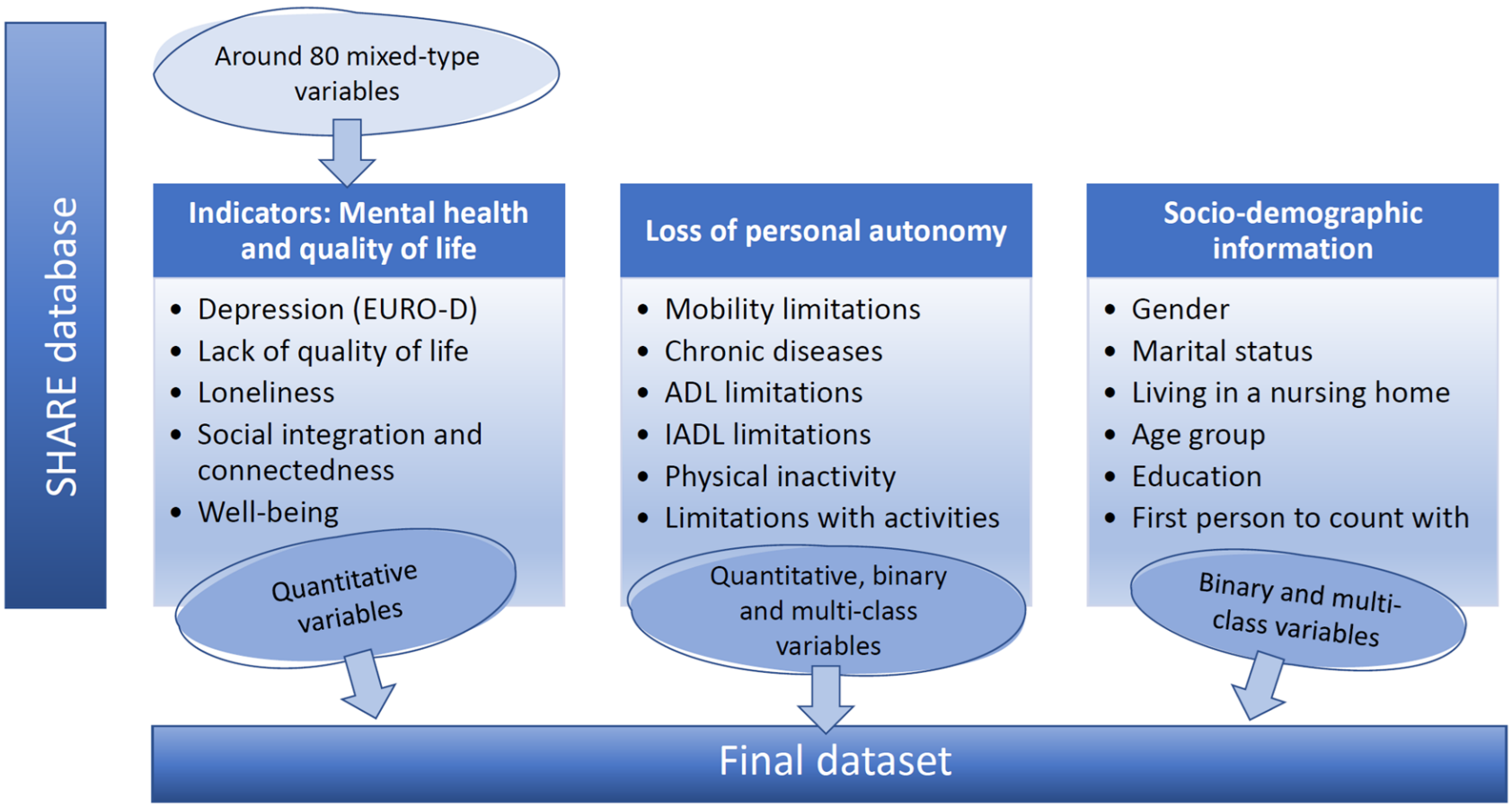


Mental well-being profiles in older adults

On-going work with I. Albarrán and F. Scielzo-Ortiz
(preliminary results)

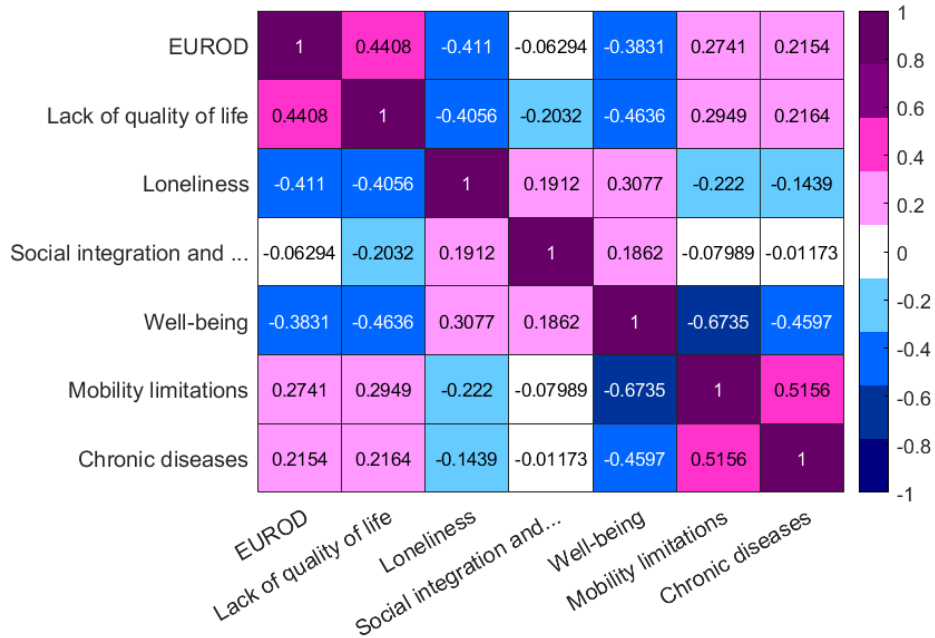
Going back to the data

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a rich panel database of individuals aged 50 or over in 26 European countries and Israel. We took wave 9 (2021/2022).



- 40,000 units, representing over 181 million aged individuals in EU,
- 17 variables of mixed-type,

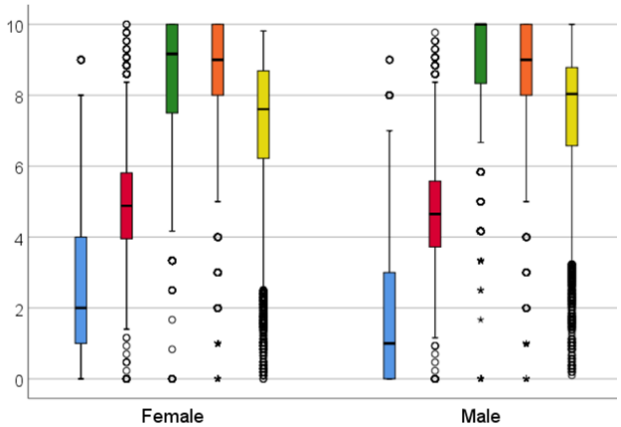
Pearson's correlation coefficient



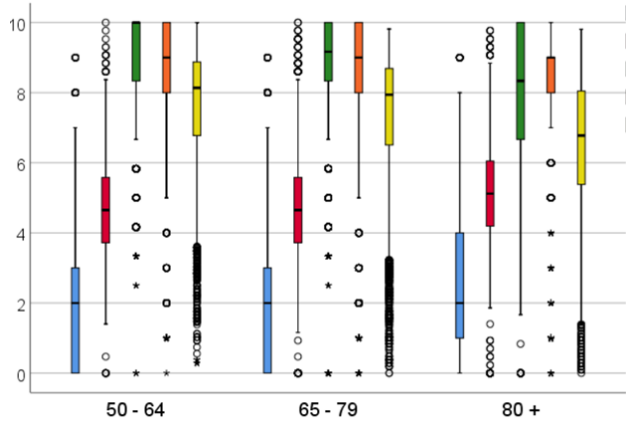
- G-Gower distance with robust Mahalanobis for quantitative data, Hamming distance for categorical multi-class, and a distance associated to Jaccard similarity coefficient for binary data.
- Fast k-medoids algorithm, with $k = 5$ (best silhouette results).

Mental health and quality of life indicators

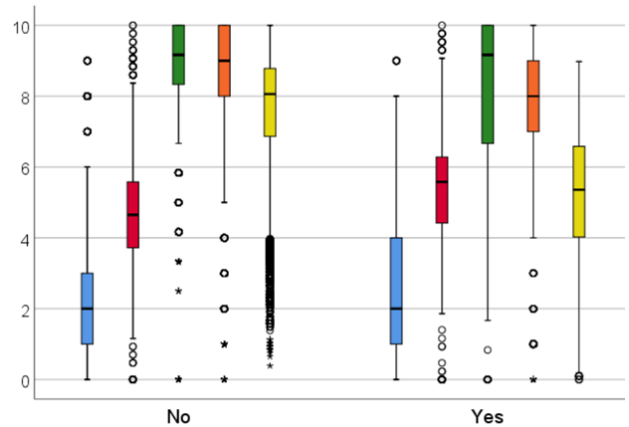
By gender



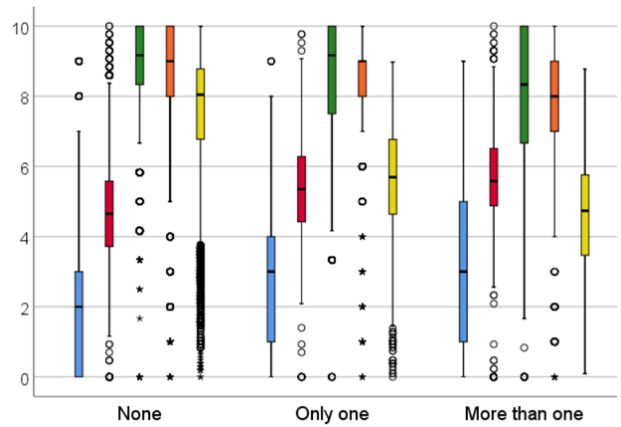
By age group



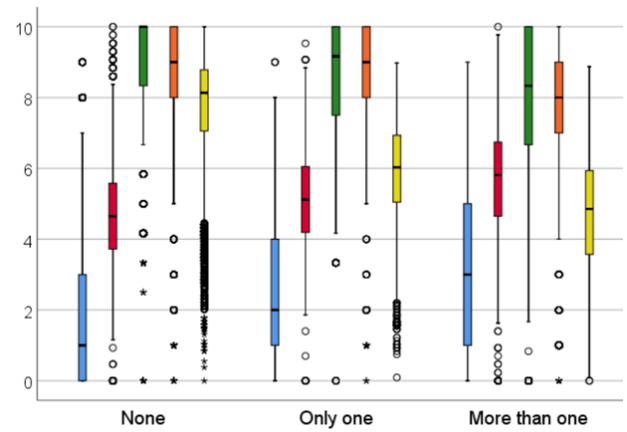
By physical inactivity



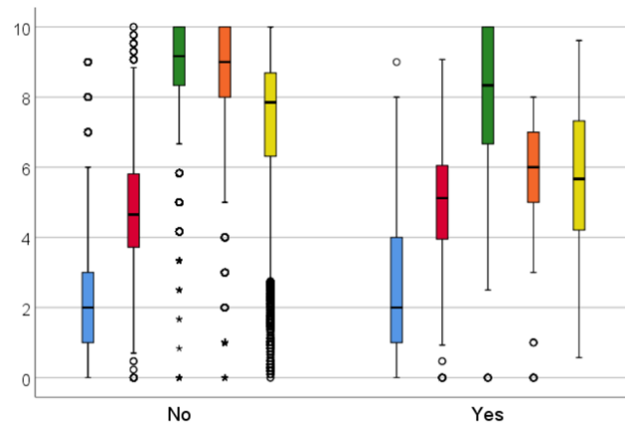
By ADL limitations



By IADL limitations



By living in a nursing home



EURO-D
Lack of quality of life in older age
Loneliness
Social integration and connectedness
Well-being

Visualizing profiles

Profile 1 – 16.3M

- 16% respondents
- 65-80+ years old
- Western EU
- High EUROD
- Feeling lonely although having social network
- Moderate well-being
- Few limitations to ADL or IADL or mobility
- Few chronic diseases

Profile 2 – 29.4M

- 30% respondents
- 50-79 years old
- Northern EU
- Secondary-University education
- Feeling lonely although having a social network
- High well-being
- Very unlikely to have limitations to ADL or IADL
- Few chronic diseases

Profile 3 – 18.7M

- 19% respondents
- 80+ years old (oldest group)
- Southern EU
- Widowed/Divorced
- Primary education
- Living in a nursing home
- High EUROD
- Feeling lonely
- No social network
- Lowest well-being
- Many limitations to ADL, IADL and mobility
- More than 3 chronic diseases

Profile 4 – 15.5M

- 16% respondents
- Never married
- Southern & Eastern EU
- Some mobility limitations
- 1-4 chronic diseases

Profile 5 – 19.0M

- 19% respondents
- 50-64 years old (youngest group)
- Secondary-University education
- North & Eastern EU
- Lowest EUROD
- Feeling lonely although having a social network
- Highest well-being
- No physical inactivity
- No ADL/IADL limitations

Profiles of special tracking

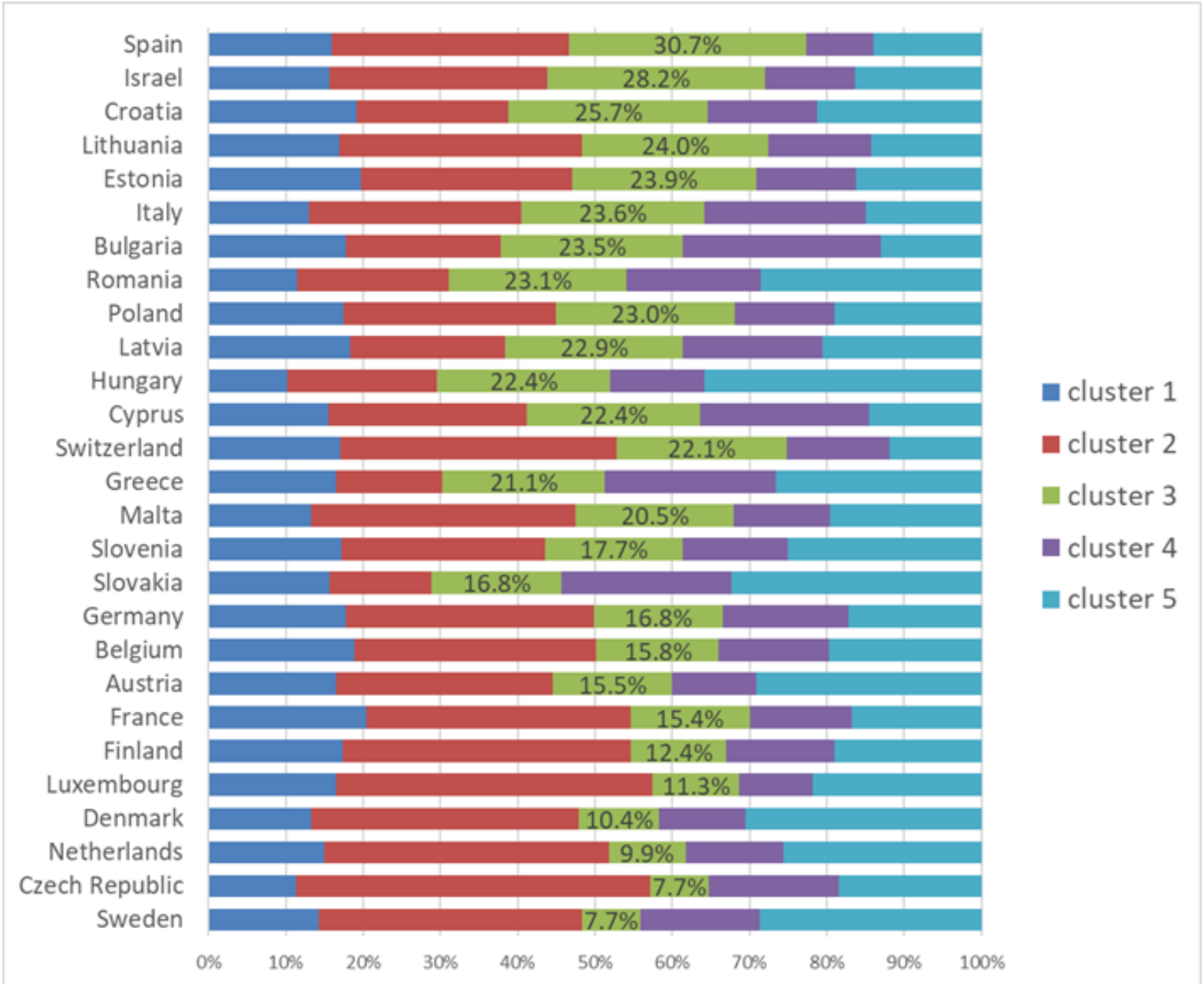
Profile 5 - Healthy ageing group. Will they be able to maintain good physical, mental and social health and well-being as they age?

- The youngest Europeans, mostly women, not living in nursing homes, in good physical and mental condition (no chronic diseases, no ADL-IADL limitations, with an active social network and a good quality of life).

Profile 3 - Elderly people to be cared for.

- The oldest Europeans living in residential care homes, mostly women with physical health problems (mobility and ADL-IADL limitations and several chronic diseases). Their basic needs are covered but they feel depressed and perceive their quality of life as poor.

EU State Members and Israel, ranked by percentage of Profile 3



Conclusions

- Two clustering algorithms based on distances specially designed for mixed-type data are proposed and studied,
- They are highly effective, particularly when using robust G-Gower's for datasets with outliers,
- Competitor methods fail in identifying outlying units, and classify them as a separate group,
- In case of Big data (more than 100k), our methods are still the best when outliers and moderate to high correlations are involved,
- Although our methods are not the most computationally efficient, they are significantly more affordable than k-medoids.

References

- Albarrán, I., Alonso, P. and Grané, A.: Profile identification via weighted related metric scaling: an application to dependent Spanish children. JRSS-A 178, 1–26 (2015)
- Boj, E. and Grané, A.: The robustification of distance-based linear models: some proposals. Socio-Economic Planning Sciences 95, 11992 (2024)
- Borg, I. and Groenen, P.: Modern multidimensional scaling: theory and applications. New York: Springer (2005)
- Cuadras, C.M. and Fortiana, J.: A continuous metric scaling solution for a random variable. Journal of Multivariate Analysis 52, 1–14 (1995)
- Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–338 (1966)
- Grané, A. and Romera. R.: On visualizing mixed-type data: A joint metric approach to profile construction and outlier detection. Sociol Methods Res 47(2), 207–39 (2018)
- Grané, A., Salini, S. and Verdolini, E.: Robust multivariate analysis for mixedtype data: Novel algorithm and its practical application in socio-economic research. Socio-Economic Planning Sciences 73, 100907 (2021)
- Grané, A. and Scielzo-Ortiz, F.: On generalized Gower distance for mixed-type data: Extensive simulation study and new software tools. Submitted to SORT (2024)
- Grané, A. and Scielzo-Ortiz, F.: Fast k-medoids and q-Fold Fast k-medoids: New distance-based clustering algorithms for large mixed-type data. Working Paper 2025-02, Statistics and Econometrics Series, University Carlos III of Madrid (2025) `https://hdl.handle.net/10016/46673`

Developed software

- **PyDistances** documentation
- **FastKmedoids** documentation