An evaluation of the IFCS Cluster Benchmarking Data Analysis Challenge

Christian Hennig

Abstract Eight clusterings of a lower back pain dataset were submitted to the IFCS Benchmarking Cluster Analysis Challenge. The aim of the challenge was to find clusterings of the 112 baseline variables that help with predicting 9 outcome variables. These clusterings are compared here, using data visualisation (multidimensional scaling and discriminant coordinates on both baseline and outcome variables), outcome means and uncertainty intervals, and four cluster validation indices, namely the Average Silhouette Width, the Pearson correlation version of Hubert's Γ, the Calinski/Harabasz index, and the Adjusted Rand Index. The different comparison approaches give quite different assessments of the clustering quality.

Christian Hennig
Department of Statistical Science, University College London
Gower St., London WC1E 6BT, United Kingdom
Tel.: +44-207-6791698

✉ c.hennig@ucl.ac.uk

Archives of Data Science, Series B
(Online First)
KIT Scientific Publishing
Vol. 1, No. 1, 2019

DOI 10.5445/KSP/1000085952/08
ISSN 2510-0564

1 Introduction

This paper presents an evaluation of the eight clusterings that were submitted to the cluster analysis challenge of the IFCS Cluster Benchmarking Task Force. The challenge was to carry out a cluster analysis of a dataset that had won an earlier dataset submission challenge. The dataset was provided by Werner Vach and colleagues. The dataset contains a baseline assessment and outcome measures from a longitudinal study of adult low back pain patients who consulted chiropractors. The research question was rooted in a need for a better understanding of the mechanisms underlying the very heterogeneous conditions of low back pain. This was translated into a search of a clinically useful grouping of patients (based on their 112 variables of baseline characteristics only, "baseline variables" in the following) that could help to predict the development of low back pain through therapy. Furthermore there were three different outcome measurements, namely global perceived improvement, a Roland Morris disability score, and a lower back pain intensity score (LBP), taken after 2 weeks, 3 months and 12 months. These measurements could be used to assess the predictive quality of a clustering. In this sense, the clustering task is semi-supervised. There were many missing values, including in the outcome variables, and the baseline variables were of mixed types, i.e., nominal (including binary), ordinal and interval scaled. Of the outcome variables, the Roland Morris scores can be seen as interval scaled (which may be controversial), whereas the other two measures are ordinal. More information on the dataset and the challenge is given in van Mechelen and Vach (2018). The dataset, along with some documentation and a questionnaire giving information about the background and characteristics of a desirable clustering, is available on `https://ifcs.boku.ac.at/repository/`.

This dataset was to be analysed in a second challenge. There were eight teams of contributors to this challenge who analysed the dataset in question, namely Hanneke van der Hoef, Yordan Raykov & Reham Badawy, Mario Fordellone, Fengmei Liu & Suchara Gupta & Cristina Tortora, Michael Greenacre, Le Phan & Hongzhe Liu & Cristina Tortora, Vladimir Makarenkov & Alexandre Gondeau, Joseph Fitch & Nazia Khan & Cristina Tortora. Most of these contributors provided submissions for the current Special Issue of Archives of Data Science, in which their clusterings are explained in detail (Hennig et al (2018)). In the present paper, the final clusterings from each contribution are evaluated in

various ways. I will not focus on ranking the contributions here. There was actually a prize winner in the challenge, but a central criterion for the prize was the justification of the proposed clustering, and its connection to the background information, and the prize was not awarded based on formal rankings. The aim of the present paper is rather a less formal exploration of how the clusterings differ, and to what extent they grasp various features of the data, although some of the evaluation given here pertains to the predictive task and can therefore be taken as basis for ranking the clusterings, if required. One issue with this is that there are nine outcome measurements (or rather three measurements over three time points each) with different observations missing, so any ranking would either be based on a single measurement only, or on aggregating measurements, for which there are many ways conceivable.

There were 928 patients in the dataset. A clustering would normally have assigned each patient to a cluster, although it was permissible to classify some observations as "outliers". Furthermore, some observations were discarded by contributors because of too many missing values.

Table 1: Number of clusters K and number of observations in all clusters for the eight clusterings. Cluster column "0" refers to observations classified as outliers. The last column is the number of observations discarded (because of too many missing values).

		Observations in cluster									
	K	1	2	3	4	5	6	7	8	outliers	discarded
van der Hoef	5	108	165	377	106	169				0	3
Raykov/Badawy*	8	168	208	66	149	57	50	70	97	63	0
Fordellone	3	277	360	291						0	0
Liu/Gupta/Tortora	8	219	166	138	118	114	75	62	36	0	0
Greenacre	6	90	189	191	177	47	234			0	0
Liu/Phan/Tortora	3	240	463	212						0	13
Gondeau/Makarenkov	5	73	198	440	59	151				7	0
Fitch/Khan/Tortora	4	225	257	168	278					0	0

(*) The original clustering of Raykov/Badawy had 17 clusters, but they singled out the eight biggest ones as potentially meaningful, so observations in the smaller ones (fewer than 20 observations per cluster) were declared outliers.

The number of clusters K was not given, so all contributors had to decide this number. Table 1 gives information on K, the cluster sizes, and discarded/outlying

observations in the eight clusterings. Here is a list of decisions that had to be made by the contributors:

- What to do with missing values?
- How to handle the mixed types of variables?
- Transformation and standardisation
- Variable selection and/or dimension reduction
- What clustering method to choose?
- How to select the final solution (including K)?
- How to interpret the solution?
- How to validate the solution externally?

Contributors were invited to submit to this Special Issue, and six of them explained their clustering in detail, see above. Raykov and Badawy used a Bayesian Dirichlet process approach for clustering, in which missing values were estimated from the same Bayesian model, see Raykov et al (2016). Fordellone used multiple imputation, a factorial analysis for mixed type data (Pagès (2004)) and K-means clustering. The latter two did not submit papers for the Special Issue.

In Section 2 I discuss some issues with the evaluation. In Section 3, the clusterings are visualised in three different ways, looking at both baseline and outcome variables. In Section 4 the clusterings are compared by use of four cluster validation indices. Section 5 provides a conclusion.

2 Some evaluation issues

The IFCS Cluster Benchmarking Task Force states as their philosophy (IFCS Task Force for Benchmarking (2016)): "In cluster analysis (...) different aims of clustering may lead to different clusterings on the same dataset that could be optimal according to different criteria (e.g., overall low within-cluster distances,

or optimal representation of every object by the centroid object of the cluster to which it is assigned, or optimal fit by a mixture probability model). (...) It is therefore particularly important for benchmarking clustering to define properly the clustering problem that a method aims to solve, by specifying as precisely as possible what kinds of clusters are of interest." (See also Hennig (2015)).

The evaluation of clusterings therefore should take into account the aim of clustering and the available background information. There are certain specifications given in the questionnaire accompanying the dataset, such as

- "to ensure clinical acceptance, it is desirable to have between 3 and about 12 clusters/groups",
- " a small group of patients classified as 'unclassifiable' may be acceptable",
- "clusters can vary in size. A large number of small clusters ($< 2\,\%$) would limit the clinical acceptability",
- "a sufficient degree of similarity, which allows a conceptual labelling" is required for observations in a cluster or
- ultimately the aim of clustering is the prediction of the future outcomes, in a clinically interpretable way.

These specifications can be used for evaluation in different ways. Some can just be checked to see whether they are fulfilled, e.g., whether the number of clusters is in the required range. Some are essentially informal: various contributions discussed clinical interpretability, but it is hard to define a quantitative measurement of this aspect; one could look at cluster purity regarding certain variables but selecting these would benefit from collaboration with a practitioner.

There is some other background information, e.g., the fact that the baseline variables include both summary scores for a number of a patient's characteristics, and the detailed scores of which the summary scores are made up, or the meaning of the different categories of nominal and ordinal variables, with implications on whether for example it could be appropriate to treat the ordinal variables as continuous. How these aspects have been taken into account can hardly be evaluated from the clustering alone.

The predictive power of the clustering for the outcome variables can in principle be measured, although there are various ways of doing it, particularly because of the non-trivial structure of the outcome variables. Another aspect that can be measured is that within-cluster homogeneity (similarity regarding the baseline variables) is required and far more important than separation between clusters. There is not much reason to believe that there are clusters in this dataset that are clearly and meaningfully separated (splitting clusters along any discrete variable technically introduces separation, but this would be an artefact).

A general issue with formal performance evaluation in a competition in the absence of a given "true" clustering unknown to the contributors is the following. If there was a transparent evaluation criterion, contributors could just try to optimise it directly, rather than applying and adapting existing clustering methods, which therefore would not be "benchmarked". Also, one would expect clustering methods to perform better, the closer their rationale matches the evaluation criterion.

In the present situation this issue plays out in the following way. In order to optimise prediction performance, the outcome variables could be used to choose the final clustering, for example using cross-validation or even some kind of within-sample prediction error. This could either be somehow implemented in the clustering method, or could serve at least to pick one out of several candidate clusterings, generated by possibly different methods, in the end. Contributors were explicitly asked not to use the outcome variables for producing the actual clustering but could use them for validation, which most of them did. Actually, on the one hand, better clusterings in terms of prediction performance could probably be achieved by using the outcome variables for finding the clustering, but on the other hand, there is a certain danger that this may lead to over-optimism regarding the final (optimised) prediction performance.

3 Visualisation of clusterings

In this section, the clusterings are visualised in ways that allow comparing them. There are three visualisations. The first one uses discriminant coordinates and multidimensional scaling on the variables to be clustered. The second one shows the clusterwise means of the outcome variables with uncertainty intervals. The third one shows discriminant coordinate plots of the outcome variables.

3.1 Discriminant coordinates of baseline variables

The first visualisation shows the clusterings on discriminant coordinate plots of a multidimensional scaling representation of the baseline variables. Discriminant coordinates are connected to linear discriminant analysis; the first two discriminant coordinates are the two dimensions (orthogonal with respect to the pooled within-cluster covariance matrix) along which the ratio of the projected pooled within-cluster variance and the projected pooled between-cluster variance is minimum, see Rao (1952) (where they are called "canonical variates") and other standard textbooks on multivariate analysis. They serve to find a two-dimensional linear projection of the data along which a given clustering is most clearly expressed in the sense explained above.

This should give some ideas about the homogeneity of clusters, to what extent there are separated clusters in the data, and to what extent they coincide, or do not coincide, with the clusterings. Because the data are of mixed format and there are many missing values which cannot be handled by standard discriminant coordinates, multidimensional scaling was performed first in order to achieve a Euclidean representation of the data.

For the distance measure and particularly the treatment of mixed type data, I followed Hennig and Liao (2013). The nominal variables are coded as dummy variables (one for each category). The ordinal variables are used as Likert coded (i.e., subsequent numbering of categories). The continuous variables are scaled to unit standard deviation, and the ordinal and dummy variables are standardised in order to make their contributions to the overall distance comparable, see Hennig and Liao (2013) for details and justification. A Euclidean distance between these standardised variables is then computed. In case of missing values, variables with missing values between a pair of observations are ignored for that pair, and variables without any missing value between the pair are scaled up accordingly. I do not claim that this is the optimal distance that can be defined on these data; in particular, shared information between variables and dependence could be taken into account as well as variable importance, which however would require more complex decisions than I wanted to make for this evaluation.

Multidimensional scaling was carried out using ratio MDS in the R-package "smacof" (de Leeuw and Mair (2009)). This approach preserves the quantitative information in the distances (which is advisable if one's interest is to look

for “gaps” in the data corresponding to clusters) whereas the distances to approximate are not squared as in classical MDS, giving less weight to the largest distances and potentially outlying observations (see Borg et al (2012) for more discussion). A 20-dimensional MDS output (stress 5.8 %) was used as basis of computing the discriminant coordinates.

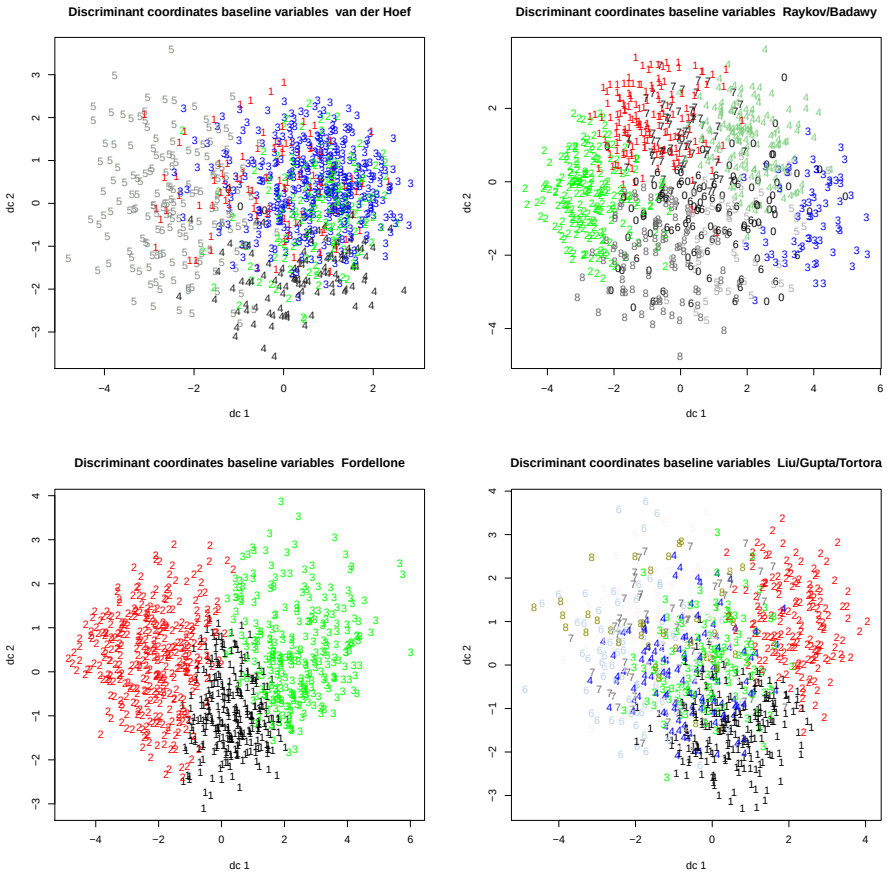

Figure 1a: Discriminant coordinates/ratio MDS plots of baseline variables, clusterings by van der Hoef, Raykov/Badawy, Fordellone, Liu/Gupta/Tortora (symbol “0” refers to observations not assigned to any cluster).

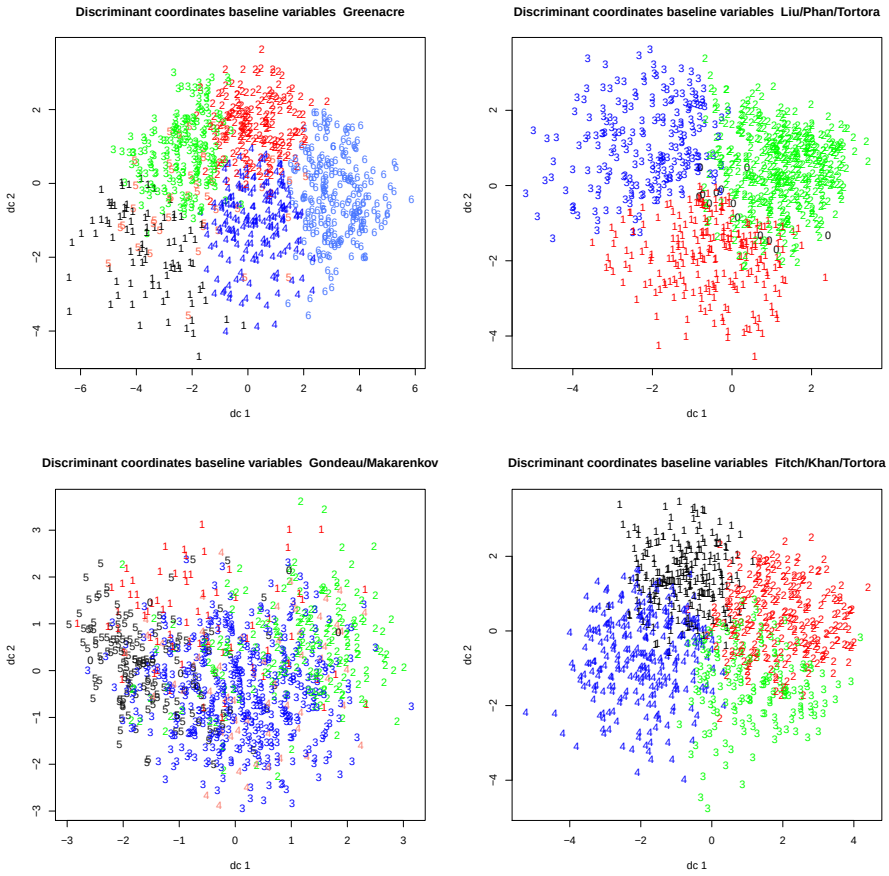

Figure 1b: Discriminant coordinates/ratio MDS plots of baseline variables, clusterings by Greenacre, Liu/Phan/Tortora, Gondeau/Makarenkov, Fitch/Khan/Tortora (symbol "0" refers to observations not assigned to any cluster).

The resulting plots are shown in Figures 1a and 1b. None of these projections shows clearly separated clusters in the sense of density gaps (neither do some other plots not shown such as a 2-dimensional MDS configuration, looking at further dimensions, "rotating" through the data etc.), therefore, it seems that such clusters do not exist in the dataset and consequently the contributors cannot be expected to find such clusters.

From these plots, the clusterings of Fordellone, Liu/Phan/Tortora and Fitch/Khan/Tortora show the clearest homogeneity in terms of the distance measure used here; clearly, clusters bring together most similar points and within-cluster distances are smaller. These clusterings have small K (3 or 4); with a higher number of clusters homogeneity seems to be more difficult to achieve for all clusters, although most clusters still look homogeneous in the clusterings of Greenacre and Raykov/Badawy with larger K. The remaining clusterings do not show much within-cluster homogeneity regarding the given distance measure. This does not necessarily mean that they are bad, as long as one can argue that they are homogeneous regarding other potentially more appropriate distance measures. The main issue here seems to be variable selection and dimension reduction. Van der Hoef, Liu/Gupta/Tortora and Gondeau/Makarenkov apparently selected information that characterised the similarity structure quite differently from the distance measure used here. Greenacre and Gondeau/Makarenkov have one or two apparently quite heterogeneous clusters besides some more homogeneous ones, and also in van der Hoef's clustering there seem to be differences in within-cluster variation. This may indicate that the selected variables for these clusterings did not represent all the variation, which is not necessarily a problem as long as the represented information is still suitable for prediction.

3.2 Means and uncertainty of outcome variables

Some of the original contributions to the challenge showed the within-cluster means of the outcome variables over time, which look reasonably different for most clusterings. But this does not take into account how much uncertainty there is in these means. In Figures 2a-c (which contain a lot of sometimes overplotted information and may therefore be hard to decipher initially) the within-cluster means are shown over time along with uncertainty intervals (thin lines of the same colour below and above the fat mean line). The uncertainty interval borders are defined as mean plus/minus 1.96 times their estimated standard error, based on the number of non-missing observations in the cluster. Technically these are invalid as confidence intervals, because this standard formula does not take into account that data dependent clustering was carried out first. However, the outcome variables were not used for clustering, and the intervals may therefore

be reasonable approximations. In any case they give an idea about the relative uncertainty compared over clusters and clusterings.

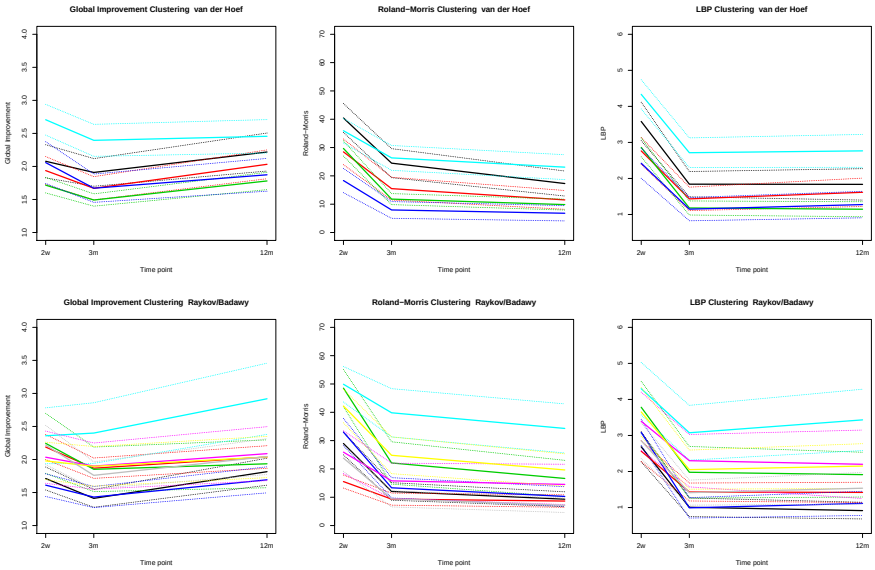

Figure 2a: Cluster means (fat) and uncertainty intervals of outcome variables for van der Hoef and Raykow/Badawy. The time points are 2 weeks, 3 months and 12 months.

The plots show that there is a quite strong amount of overlap between the uncertainty intervals, indicating that often the differences between within-cluster mean curves may not be statistically significant, although the differences between the most extreme clusters seem significant in most cases. The clusters are best distinguished regarding the Roland-Morris score and worst regarding global improvement. Fordellone's clusters show the clearest differences, particularly regarding the Roland-Morris score, followed by Liu/Phan/Tortora. It does not look coincidental that these are the clusterings with $K = 3$, the lowest number of clusters. Clusterings with more clusters predict a larger range of outcome values, but this comes with larger uncertainty, which is in all likelihood caused by lower numbers of observations per cluster. For Gondeau/Makarenkov's clustering, all intervals overlap for the LBP score. Fitch/Khan/Tortora's clustering looks as if three out of four cluster mean lines are well separated; the mean lines of

two of their clusters cross for the Roland-Morris score, which is the clearest crossing of within-cluster developments over time. Van der Hoef's cluster with largest means (light blue) seems to predict higher values of global improvement and LBP clearly separated from the other clusters' means, but not regarding the Roland-Morris score, which is remarkable given that generally Roland-Morris seems to be the easiest score to predict from clusters.

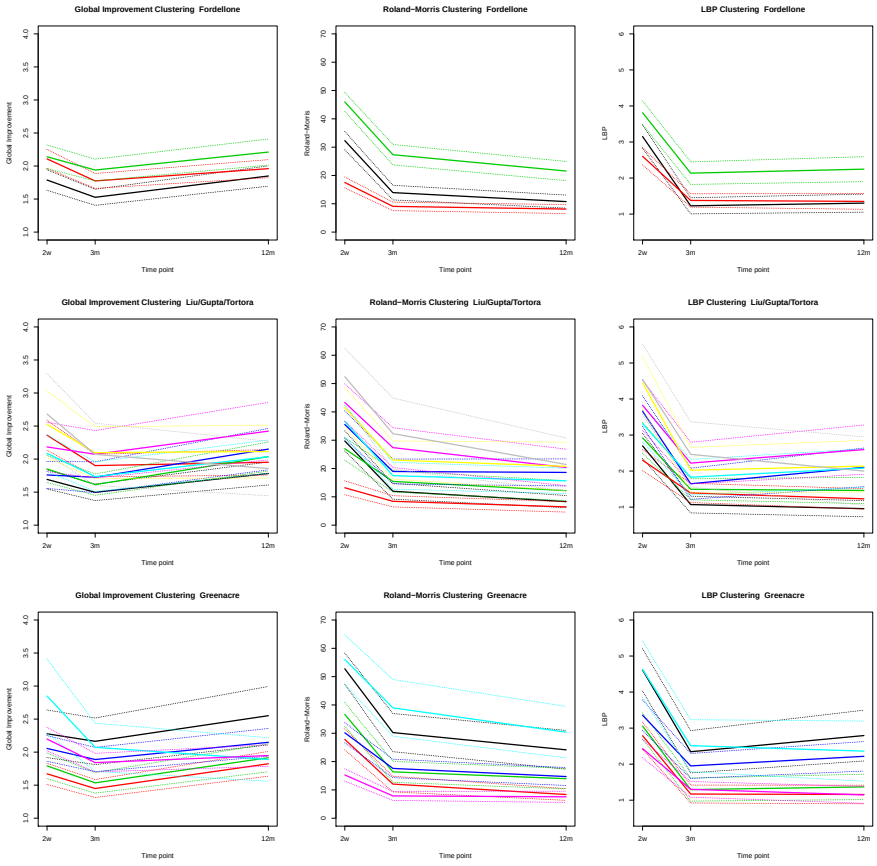

Figure 2b: Cluster means (fat) and uncertainty intervals of outcome variables for Fordellone, Liu/Gupta/Tortora and Greenacre. The time points are 2 weeks, 3 months and 12 months.

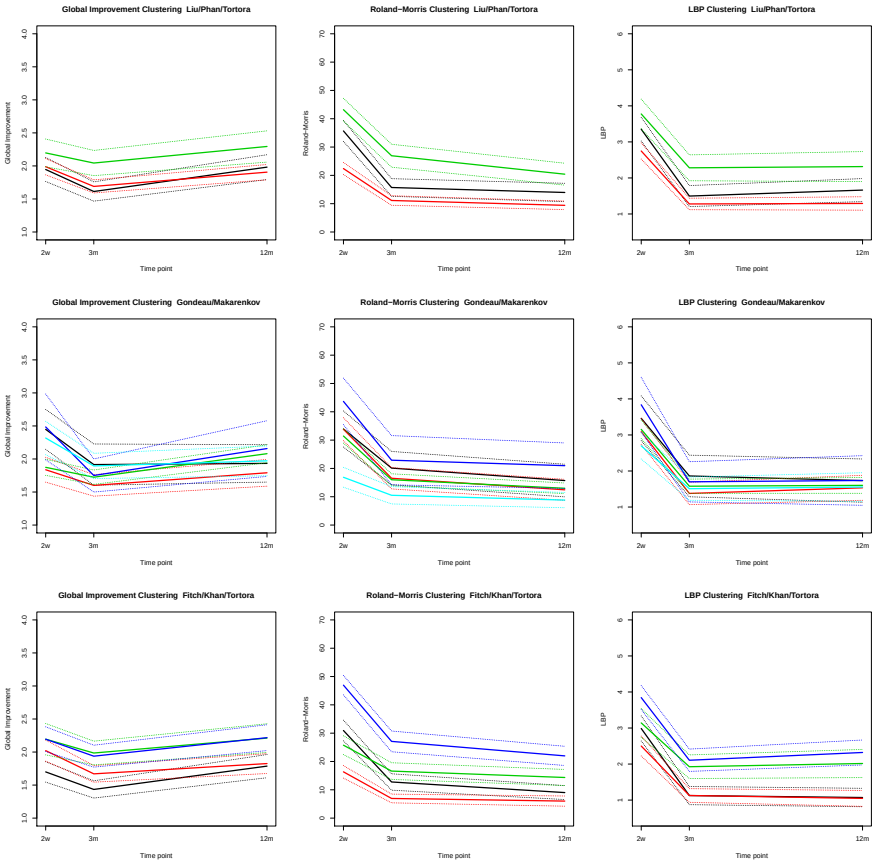

Figure 2c: Cluster means (fat) and uncertainty intervals of outcome variables for Liu/Phan/Tortora, Gondeau/Makarenkov and Fitch/Khan/Tortora. The time points are 2 weeks, 3 months and 12 months.

I also produced similar plots (not shown) but with uncertainty intervals based on standard deviations instead of standard errors to see to what extent the actual values of the outcome variables overlap between clusters (rather than the uncertainty in the estimated means). This overlap is generally very strong.

3.3 Discriminant coordinates of outcome variables

The uncertainty intervals from Section 3.2 illustrate the uncertainty in the within-cluster means. They do not show the within-cluster variation of the observations on the outcome variables. The homogeneity, or heterogeneity, respectively, of the clusters regarding the outcome variables can be visualised in the same way as in Section 3.1. There are only nine outcome variables with a considerable number of missing values. I excluded all observations with five or more missing values (i.e., more than half) on the nine outcome variables. The resulting sample size is 742. Distances, ratio MDS (9 dimensions, stress 8.7 %) and discriminant coordinates have then been applied as in Section 3.1. The resulting plots are shown in Figures 3a and 3c.

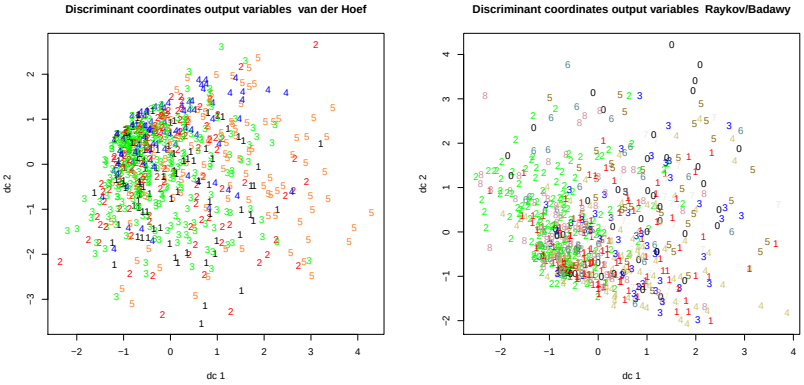

Figure 3a: Discriminant coordinates/ratio MDS plots of outcome variables, clusterings by van der Hoef, Raykov/Badawy (symbol "0" refers to observations not assigned to any cluster).

As could be expected because the outcome variables were not used for clustering, the clusters look much more heterogeneous than in Figures 1a and 1b. In some of these plots one needs to look quite hard for systematic differences between clusters. The clusterings with more clusters seem to have an advantage here in the sense that at least the differences between the more extreme clusters can be clearly seen (e.g., clusters 3/4 vs. 6/8 in Raykow/Badawy). In practice, it may be useful to know that in the same clustering different clusters may be more or less informative when it comes to the prediction of the outcomes. One message of

the plots is that the ability of the clusters to "locate" patients in outcome space and thus predict the patient's process is quite limited. Obviously in these plots the global improvement, Roland-Morris and LBP scores have been aggregated, and prediction of the Roland-Morris score alone may work somewhat better. The correlation between the vectors of distances (stacking the distances for all pairs of patients) in baseline space and outcome space is 0.289, showing that the baseline variables should have an existing though limited predictive power.

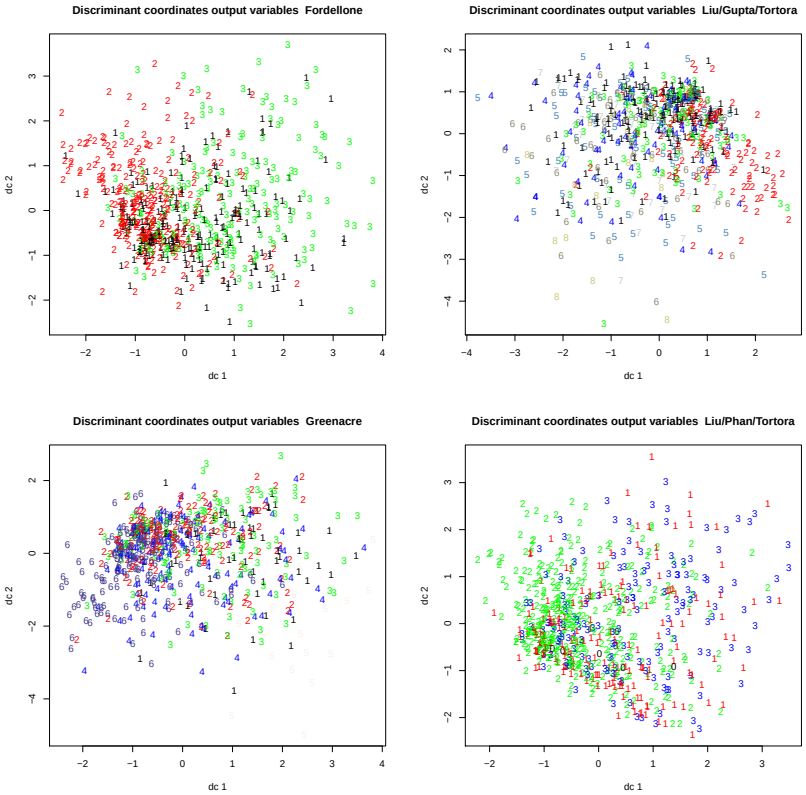

Figure 3b: Discriminant coordinates/ratio MDS plots of outcome variables, clusterings by Fordellone, Liu/Gupta/Tortora, Greenacre, Liu/Phan/Tortora (symbol "0" refers to observations not assigned to any cluster).

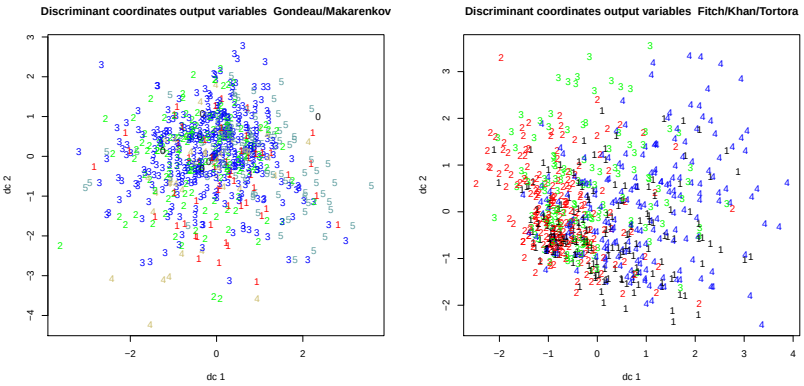

Figure 3c: Discriminant coordinates/ratio MDS plots of outcome variables, clusterings by Gondeau/Makarenkov, Fitch/Khan/Tortora (symbol "0" refers to observations not assigned to any cluster).

4 Validation indices

Many indices have been proposed in the literature to measure the "validity" or quality of a clustering, see, e.g., Halkidi et al (2015). "Internal" cluster validity in the literature usually refers to measurement of the clustering quality just based on the clustering and the clustered data, without any reference to some underlying (known) "true clustering". In contrast, "external" validation uses information external to the data that was clustered. A number of indices is available for these tasks. For the lower back pain data analysed here, computing such indices can be of interest for both the baseline data and the outcome data (with in both cases using the clusters that were derived from the baseline variables). Note that computing such indices on the outcome data constitutes an "external" validation of the clusterings on the baseline data. Obviously, in the latter case one would expect a lower clustering quality.

I use four different validation indices here. Note that most competition participants used one or more of these indices directly or indirectly (for fixed K, K-means is equivalent to optimising the CH-index, see below) for finding their clustering. In principle one should expect that a clustering does better on an index if the index was used for clustering, particularly if the clustering was

found by optimising the index. However, keeping in mind that the distances used by the participants were all different from mine, and that some of the indices were used in a very marginal way (for example, van der Hoef used a voting scheme out of 30 indices including three of those given below for finding the number of clusters), it would be hard to interpret the presence or absence of any connection between the index values given here, and which indices were in some way used by the participants, which I therefore will not consider.

4.1 Average Silhouette Width

The Average Silhouette Width (ASW, Kaufman and Rousseeuw (1990)) compares for every observation the average distance to observations in the same cluster with the average distance to observations in the closest different cluster and aggregates the resulting "silhouette widths". The resulting value is between -1 and 1. High values mean that the average distance to the closest different cluster for all or most observations is much higher than the average distance to observations in the same cluster. This is desirable in cluster analysis and means that clusters are homogeneous and separated from their closest neighbours. Values around zero mean that on average, the two average distances are very similar; negative values mean that many points would be better off in a neighbouring cluster than in the one to which they were actually assigned.

Average Silhouette Widths can be seen in Figure 4. The correlation between ASWs on the baseline and the outcome variables is 0.15, which means that the ASW on the baseline variables is somewhat but not very informative about the ASW on the outcome variables. Fordellone's clustering is best regarding both distances. Van der Hoef's clustering is the worst on the baseline variables but does better regarding the outcome variables. Liu/Phan/Tortora do not do so well regarding the ASW. All ASW values are quite low; actually on the outcome variables all are negative. This is not a good result, although it does not mean that the clusterings are all useless. In fact, the definition of the ASW implies a search of the best neighbouring cluster for all observations. When calculating ASW on the outcome variables, "best neighbouring" is based on a match in terms of the outcome variables (whereas the clusterings under study are derived from just the baseline variables). Nevertheless, as long as the cluster to which an observation is assigned still allows a prediction of the outcomes that is better

than prediction from a random cluster, the clustering still is of some use for prediction.

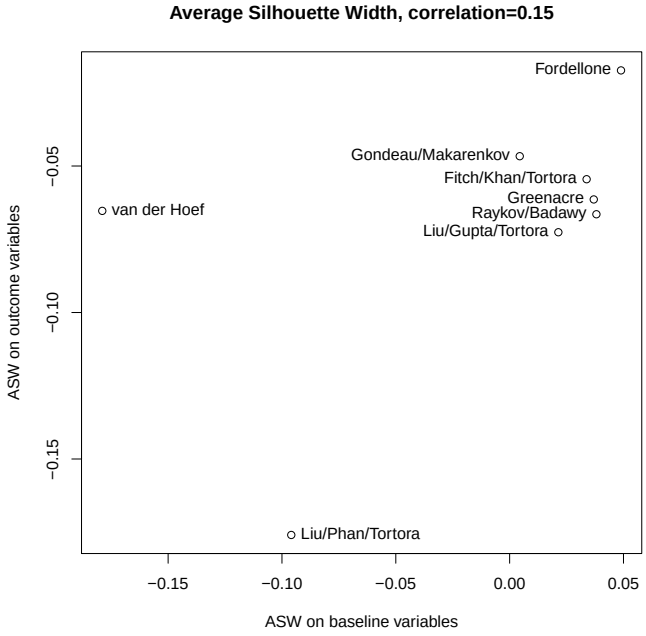

Figure 4: Average Silhouette Widths on baseline and outcome variables.

Although the ASW is one of the most popular validation indices, it has too much of a focus on separation between neighbouring clusters, which is not very relevant for the data analysed here.

4.2 Pearson correlation version of Hubert's Γ

Hubert and Schultz (1976) introduced a general principle for constructing validity indices, sometimes referred to as Hubert's Γ, one of which is the Pearson correlation between the vector of dissimilarities for all pairs of observations and a 0-1 vector which is 0 if the two observations are in the same cluster and 1 if they are in different clusters. I call this index Pearson-Γ, see Halkidi et al

(2015). This index measures to what extent the clustering information represents the information in the dissimilarities. This can be seen as relevant particularly for the outcome variables, because it addresses the question how strongly the clusters are informative about the outcomes.

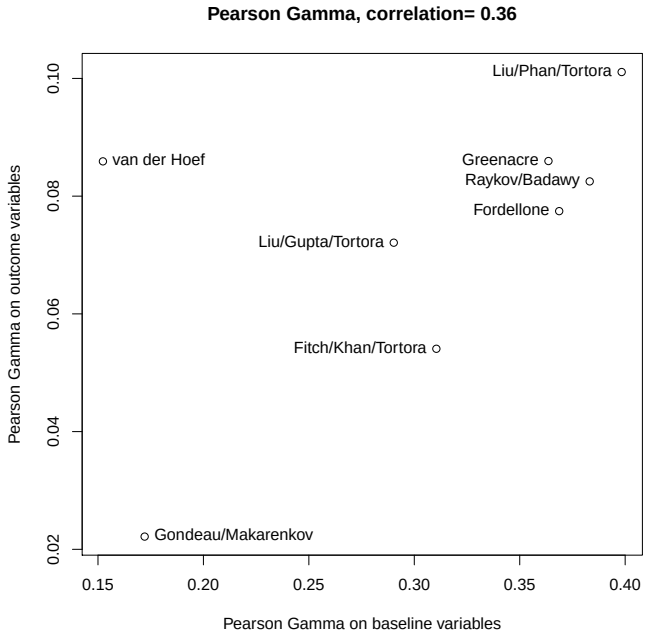

Figure 5: Pearson-Γ on baseline and outcome variables.

The Pearson-Γ values are shown in Figure 5. The correlation between Pearson-Γ values on the baseline and outcome variables is 0.36, much higher than for the ASW. It is remarkable that Liu/Phan/Tortora achieve the best value on the outcome variables after having scored lowest on the ASW. This shows how strongly differently these two indices assess the clusterings. Van der Hoef, Greenacre and Raykov/Badawy have the next highest values. Van der Hoef again scores lowest on the baseline variables while still achieving a competitive performance on the outcome variables. All correlations are positive (if rather low), so the clusterings at least give some information about the outcome variables.

4.3 Calinski and Harabasz index

The Calinski and Harabasz index (CH, Calinski and Harabasz (1974)) is another quite popular index. It is usually used for Euclidean data but can be defined for general dissimilarities, see Halkidi et al (2015). CH is based on the ratio of between-cluster and within-cluster variation. This is scaled appropriately to make clusterings with different numbers of clusters comparable. Larger values are better. On the other hand, as opposed to the ASW and Pearson-Γ, it is hard to attribute absolute meaning to CH values that could be compared across different datasets. Because the within-cluster variation does not look specifically at neighbouring clusters or smallest dissimilarities between clusters, CH can be interpreted as measuring within-cluster homogeneity, standardised by overall variation, whereas separation does not contribute strongly. I chose CH here because within-cluster homogeneity is much more relevant than between-cluster separation, and realistic to achieve.

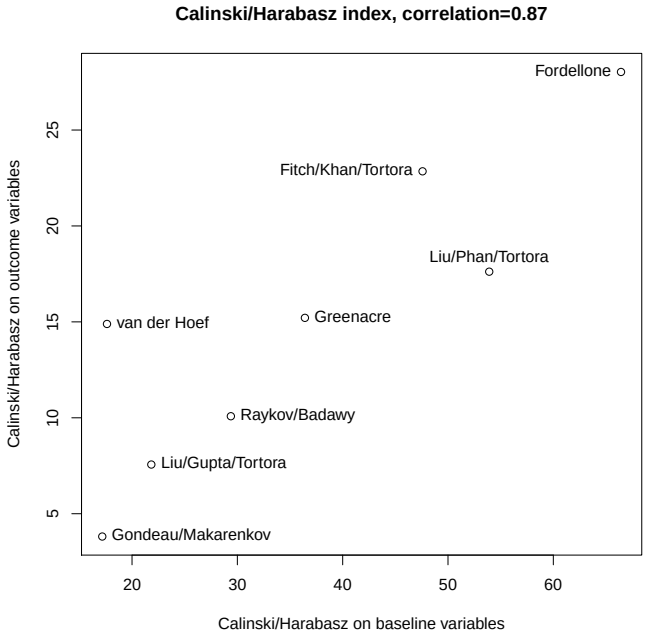

Figure 6: Calinski/Harabasz on baseline and outcome variables.

The values are shown in Figure 6. Interestingly, the correlation between the CH values on the baseline and outcome variables is very high at 0.87, so homogeneity on the baseline variables seems to be connected to homogeneous predictions (although the very high correlation can to some extent be an unstable product of the low number of clusterings). Fordellone has the best results on both distances. Van der Hoef again looks bad on the baseline variables but better on the outcome variables. Of course all these quality judgements rely on the appropriateness of the distances used here, which can be disputed.

4.4 Adjusted Rand Index

The Adjusted Rand Index (ARI; Hubert and Arabie (1985)) is an index that compares two different partitions. Its value range is between -1 and 1 with 1 indicating a perfect match and 0 being the expected value when comparing two random partitions with the same cluster sizes as the two partitions that are actually compared. Often it is used as a quality measurement comparing a computed partition to a known "true" clustering. Such a "true" clustering is not given here, but the ARI can be used to compare all pairs of the eight submitted clusterings.

This yields ARI values between 0.03 and 0.46. The ARI is a similarity measure, and I defined a dissimilarity by computing $0.5-\text{ARI}$, which is between 0 and 1 because no ARI here is larger than 0.5. On this dissimilarity, a ratio MDS was performed. The result can be seen in Figure 7. I also tried out $1-\text{ARI}$ but this distributes the clusterings more uniformly in the ratio MDS, giving a less interesting plot.

The pairs of clusterings of Greenacre and Raykov/Badawy, as well as Fordellone and Fitch/Khan/Tortora are the most similar. The other four clusterings are more "idiosyncratic". What is surprising here is that the similarity between Greenacre, Liu/Gupta/Tortora and Liu/Phan/Tortora is not that high, given that all of these submissions relied strongly on Correspondence Analysis. On the other hand, I would have expected Fitch/Khan/Tortora to be more outlying, given that they used Spectral Clustering, which is quite different and less focused on within-cluster homogeneity than the clustering methods used by most others, particularly k-means and clara. What this shows is that preprocessing decisions, probably particularly variable selection and dimension reduction, perhaps also

treatment of the mixed types of variables, had more impact on the clusterings than the finally used clustering algorithm.

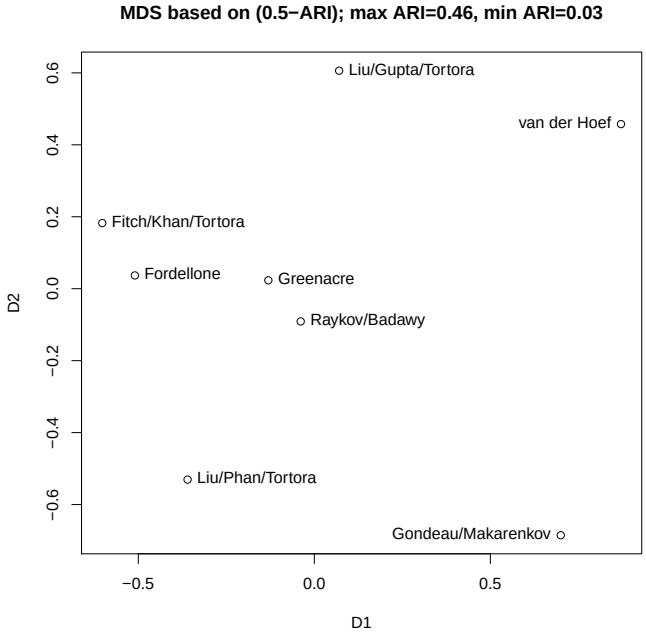

Figure 7: Ratio MDS of distances defined as 0.5−ARI between clusterings.

5 Conclusion

I compared the eight clusterings submitted to the IFCS cluster analysis challenge on the lower back pain dataset in a largely exploratory fashion, not aiming at generating a quality ranking in the first place, but rather at illustrating differences and similarities between the contributions.

Besides comparing the specific clusterings, this shows some general features of the problem, namely that overall the use of the clusters for predicting the outcome variables is rather limited; no baseline cluster in any of the clusterings

leads to really homogeneous outcomes, although there is certainly some limited amount of information about outcomes in the baseline variables.

Preprocessing decisions including variable selection and dimension reduction may have a stronger impact on the clustering than the clustering algorithm finally used. To what extent clustering quality on baseline and outcome variables is related depends strongly on the index used to measure this, and the different cluster validity indices gave quite different assessments of the clustering quality. Generally, the different analyses carried out here give quite different pictures of what the methods achieved and how good they were (if they would be used in this way), which illustrates the general difficulty of cluster benchmarking and the importance to carefully think about the used criteria.

Thinking of general cluster analysis benchmarking, I can well imagine that some of the techniques applied here can be informative when evaluating and comparing a set of different clusterings on a real benchmark dataset without given "true" clustering.

References

Borg I, Groenen PJ, Mair P (2012) Applied Multidimensional Scaling. Springer, New York. ISBN: 978-3-642318-47-4

Calinski T, Harabasz J (1974) A Dendrite Method for Cluster Analysis. Communications in Statistics - Theory and Methods 3:1–27

Halkidi M, Vazirgiannis M, Hennig C (2015) Method-Independent Indices for Cluster Validation and Estimating the Number of Clusters. In: Handbook of Cluster Analysis, Hennig C, Meila M, Murtagh F, Rocci R (eds), Chapman & Hall/CRC, Boca Raton FL, chap. 26, p. 595–618

Hennig C (2015) Clustering Strategy and Method Selection. In: Handbook of Cluster Analysis, Hennig C, Meila M, Murtagh F, Rocci R (eds), Chapman & Hall/CRC, Boca Raton FL, chap. 31, p. 703–730

Hennig C, Liao TF (2013) Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification (with discussion). Journal of the Royal Statistical Society, Series C 62:309–369

Hennig C, van Mechelen I, Dean N (eds) (2018) Special Issue: The IFCS Cluster Analysis of Target Data Set Competition, IFCS-2017, Tokyo. 1(1)Archives of Data Science, Series B , DOI 10.5445/KSP/1000085952/08

Hubert L, Arabie P (1985) Comparing Partitions. Journal of Classification 2(2):193–218, Springer

Hubert LJ, Schultz J (1976) Quadratic assignment as a general data analysis strategy. British Journal of Mathematical and Statistical Psychology 29:190–241

IFCS Task Force for Benchmarking (2016) Cluster Benchmark Data Repository - Philosophy, URL https://ifcs.boku.ac.at/repository/philosophy.html [accessed 2017-11-11]

Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data. Wiley, New York, DOI 10.1002/9780470316801

de Leeuw J, Mair P (2009) Multidimensional Scaling Using Majorization: SMACOF in R. 31(i03)Journal of Statistical Software , DOI 10.18637/jss.v031.i03

van Mechelen I, Vach W (2018) Cluster analyses of a target data set in the IFCS cluster benchmark data repository: Introduction to the special issue. Archives of Data Science, Series B 1(1):1–12, DOI 10.5445/KSP/1000085952/01

Pagès J (2004) Analyse factorielle de données mixtes. Revue de Statistique Appliquée, p. 93–111, Société française de statistique, URL http://www.numdam.org/item/RSA_2004__52_4_93_0

Rao CR (1952) Advanced statistical methods in biometric research. Wiley, New York, DOI 10.1002/ajpa.1330120224

Raykov YP, Boukouvalas A, Baig F, Little MA (2016) What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLOS ONE 11(9):1–28, Public Library of Science, San Francisco, DOI 10.1371/journal.pone.0162259