Journal of Intelligent Information Systems (2022) 59:455–477
https://doi.org/10.1007/s10844-022-00703-x

Obtaining synthetic indications and sorting relevant structures from complex hierarchical clusters of multivariate data

Damiano Fustioni[1] · Federica Vignati[1] · Alfonso Niro[1]

Received: 28 June 2021 / Revised: 3 February 2022 / Accepted: 20 March 2022 /
Published online: 21 May 2022

Abstract

Hierarchical clustering of multivariate data usually provide useful information on the similarity among elements. Unfortunately, the clustering does not immediately suggest the data-governing structure. Moreover, the number of information retrieved by the data clustering can be sometimes so large to make the results little interpretable. This work presents two tools to derive relevant information from a large number of quantitative multivariate data, simply by post-processing the dendrograms resulting from hierarchical clustering. The first tool helps gaining a good insight in the physical relevance of the obtained clusters, i.e. whether the detected families of elements result from true or spurious similarities due to, e.g., experimental uncertainty. The second tool provides a deeper knowledge of the factors governing the distribution of the elements in the multivariate space, that is the determination of the most relevant parameters which affect the similarities among the configurations. These tools are, in particular, suitable to process experimental results to cope with related uncertainties, or to analyse multivariate data resulting from the study of complex or chaotic systems.

Keywords Hierarchical clustering · Synthetic methodology · Graph theory · Experimental database · Experimental uncertainty

1 Introduction

Several physical phenomena and engineering applications are governed by a large number of factors. In the most fortunate case, the influential parameters are known, analytical

✉ Alfonso Niro
alfonso.niro@polimi.it

Damiano Fustioni
damiano.fustinoni@polimi.it

Federica Vignati
federica.vignati@polimi.it

[1] Dipartimento di Energia, Politecnico di Milano, Milan, Italy

models are available to describe the behavior of the system under scrutiny, and the obtained equations can be solved or, at least, interpreted. Conversely, in a large number of cases, one of the aforementioned conditions cannot be fulfilled, leading to an *a priori* impossibility to understand the phenomenon. To overcome this limitation, numerical simulations or experiments can be carried out, and they usually provide correlations between the output data and the input parameters. Unfortunately, three obstacles may still arise, to undermine the comprehension of the phenomena. The first problem is the selection of the influential parameters, which usually results from previous expertise and, therefore, can be easily solved. The second one consists in the interpretation of the resulting correlation, if any, since classical statistical tools, e.g., interpolation or regression, simply provide *analytic* functions, but no information on the *physical* link between the controlled parameters and the results. The third one arises when experimental or numerical results point to strange patterns: in particular, it is sometimes observed that small variations in the operational conditions may lead to different results, whereas farther configurations result in more similar output. The first step to approach all three of these obstacles may consist in the determination of the groups of configurations which lead to comparable results. This goal can be achieved by means of statistical clustering of multivariate data, i.e., a classification of the latter, based on some similarities among the members of the same set or *cluster*.

Two clustering approaches can be identified: flat and hierarchical. The first ones simply allocate each configuration to one of the defined sets. The computational of flat clustering is, in general, very efficient, but it presents several drawbacks, mainly the need to define *a priori* the number of sets and, to a first degree of approximation, their location. Hierarchical clustering methods, on the contrary, overcome some of these limitations, since the sets are built step-by-step: each cluster is recursively agglomerated with other ones or partitioned into sub-clusters. The deeper the cluster partitioning level, the larger the similarity among its elements.

Hierarchical clustering methods, therefore, require a slightly larger computational cost, but provide more information than flat ones (Sokal & Sneath, 1963; Johnson & Wichern, 1990; Anderson, 1984). When complex phenomena are analyzed, hierarchical clustering methods are therefore preferred. Their continuous improvements benefits from the advancements in computer science, which has allowed to develop smart, accurate and fast algorithms, which proved themselves useful in a large number of applications (Jain et al., 1999; Murthy, 1998; Hormiga, 1994; Campbell, 1996; Pampalk et al., 2003). However, also hierarchical clustering methods suffer from a number of limitations, mainly associated to non-unique results, due to the criterion adopted to define the similarity among the elements (James Rohlf & Sokal, 1962; Day & Edlesbrunner, 1985; Margot, 2015), the scaling of data (Jain, 2010; Kleinberg, 2002; Fortunato, 2010; Friedman & Rubin, 1967; Mahalanobis, 1936; Knorr et al., 2001), the efficiency (Jolion et al., 1991; Kumar & Orlin, 2008; Davé & Krishnapuram, 1997), the sensitivity to uncertainty (Jiang et al., 2013; Kriegel & Pfeifle, 2005; Aggarwal & Yu, 2009) and additional problems associated with the treatment of peculiar configurations (Fernández & Gómez, 2008). A previous work from the authors (Vignati et al., 2018) defines a novel algorithm which allows to partially fill these gaps, as it produces clusters more robust with respect to the data scaling, the experimental uncertainty and the ties in proximity problem (MacCuish et al., 2001; Jain & Dubes, 1998), at the same computational cost of classical techniques (Bouguettaya et al., 2015; Day & Edelsbrunner, 1984; Hruschka et al., 2009).

However, a question is still open: how to extract the most relevant information among the large number provided by a hierarchical clustering? The resulting dendrograms, indeed, indicate which configurations are *far*, *close*, *closer*, *closer-and-closer*, etc., but they are not helpful in uniquely identifying a partition among the elements. The larger the dendrogram size, indeed, the harder the definition of a unique, robust classification (Dunlop et al., 2015).

The goal of this work, therefore, is to obtain synthetic information —in the manner of flat clustering results— from the results of a robust clustering procedure —typical of hierarchical clustering techniques. This will allow to sort the most relevant structures, e.g., the influence of the different parameters, or to separate different levels of information. The latter goal, in particular, has been a hot topic for decades (Holton et al., 1993), and still to day represents an interesting field of research (Lu et al., 2020). For this reason, a procedure has been formulated, which provides two tools: the first one aims at summarizing the huge information, and to define some families of *reasonably close* elements. The second one considers the parameters which define the elements of the different families and automatically detects the most relevant ones, if any, which affect the similarity of multivariate data.

Both tools are very important when complex physical phenomena are investigated, like heat transfer by forced convection in ribbed channels. The results of the long-term study carried out at ThermALab of Politecnico di Milano —aimed at determining the Nusselt number and the friction factor for diverse-rib configurations in a large-aspect ratio duct with intermediate-Reynolds flows— which motivates this research, indeed, show that different rib geometries may result in comparable thermo-fluid-dynamic performances or, conversely, different results are obtained if small variations occur in the rib configuration. Due to the complex physics of turbulence and heat transfer, and to the lack of an evident underlying structure in multivariate data, i.e., the Nusselt number and the friction factor at different Reynolds numbers, a first analysis of the similarity appears a good strategy, followed by the novel synthesis step described in this work.

The paper is structured as follows: Section 2 summarizes the adopted methodology, with a focus on the more advisable choices to be adopted when clustering experimental data. The description of the first tool follows. The second tool is illustrated and discussed in Section 3, with an insight into its links with classical graph theory. A toy problem will be defined, to better understand the procedure. The estimation of the computational cost is discussed in Section 4, which is followed by a brief description of the procedure and dataset adopted to test the proposed method, i.e., Section 5. Conclusions and final remarks are eventually summarized in Section 6.

2 First tool: identification of groups from multivariate data

In this section, the identification of some groups of configurations will be performed. The procedure consists in three steps:

1. computation of the dendrogram resulting from hierarchical clustering of data;
2. cut-off of the dendrogram in correspondence of a physically meaningful similarity level;
3. refinement of the procedure.

2.1 Step I: the computation of the hierarchical clustering

The first step appears to be the simplest one, since it can be achieved by means of classical, well known techniques. Actually, the choice of the most suitable method turns out to be very helpful in the definition of the groups. When the definition of a number of families is the goal of the clustering, indeed, classical flat clustering could be directly applied. Unfortunately, it is mandatory to define a priori some parameters, such as the number of families, or the distance range among the centroids of the resulting sets, etc. This is usually possible when the problem under scrutiny is well known, or simple, e.g., the partition of a region into districts (Kalantari, 2013). When these data are not known, and hierarchical clustering must be therefore adopted, some constraints may sometimes apply to obtain simpler results: for example, in systematic biology, the *number* of branching occurring in the dendrogram is fixed (Michener et al., 1970), and therefore it is immediate to retrieve the required information even in large sets of data, e.g., the *families* within the Araneae *order*, in the Arachnida *class* of the Arthropoda *phylum* which belongs to the *kingdom* of all Animals (https://wsc.nmbe.ch/families).

When experimental results are analyzed, usually no constraint applies a priori, and moreover data are affected by measurement uncertainty. For this reason, the suggested clustering technique, whose mathematical details are duly described in Vignati et al. (2018), is based on the so-called “weighted asymmetric ∞-pseudometric” d^*_∞, which provides a unique clustering, regardless of the data scaling and experimental uncertainty (Vignati et al., 2018). This metrics, specifically designed for quantitative data, is defined as

$$\begin{aligned} d^*_\infty(P,Q) &= \min_{R=P,Q}\left(\max_i\left(\lim_{w_i\to R_i}\left|\frac{P_i-Q_i}{w_i}\right|\right)\right) \\ &= \min\left(\max_i\left(\lim_{w_i\to P_i}\left|\frac{P_i-Q_i}{w_i}\right|\right),\max_i\left(\lim_{w_i\to Q_i}\left|\frac{P_i-Q_i}{w_i}\right|\right)\right) \end{aligned} \tag{1}$$

where P_i, Q_i are the values of the i-th variable making up the multivariate data, and w_i the corresponding value of a weight. The fractions are computed by means of a limit operator to account for the possibility of zero-value variables.

This method, termed “dynamic”, is an agglomerative, bottom-up technique, based on a similarity criterion which disengages the results from the data scaling. This possibility comes from a proper metrics definition, which is automatically and dynamically updated during the clustering procedure. To detect the similarity between the configurations, an area of influence is identified around each configuration, termed “bounding box”, whose size is dynamically increased during the clustering procedure: if two bounding boxes interact with each other, then the two configurations are merged together to form a new cluster, which is represented by the centroid computed on all its elements. With the adopted d^*_∞ metrics, each side of the bounding box is proportional to the coordinate value of the central element along the considered direction, i.e., a percentage of it, and it will be therefore termed “percentage distance”. The interaction between the bounding boxes of two configurations occurs when one configuration is included in the bounding box built around the other one. Details on the novel proximity measure which substitutes classical metrics, on the algorithm computational cost and on its geometric meaning are duly discussed in previous papers by the Authors Vignati et al. (2018), Niro et al. (2016), and Fustinoni et al. (2019). Figure 1 shows two possible implementation strategies for the described clustering algorithm: the first one, i.e., the geometric approach, is more straighforward and flexible, albeit less efficient. The second one, which introduces the d^*_∞ pseudometric, disengaged the procedure from its geometric interpretation, and therefore it reduces its computational cost, at the expense of a lower flexibility.

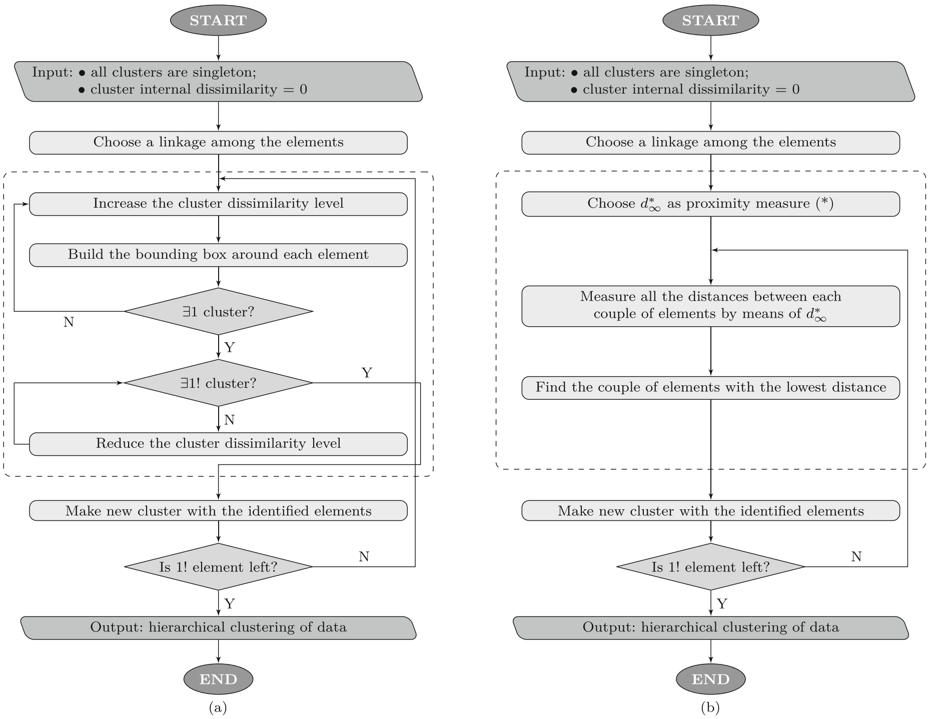

START
Input: • all clusters are singleton;
• cluster internal dissimilarity = 0
Choose a linkage among the elements
Increase the cluster dissimilarity level
Build the bounding box around each element
∃1 cluster?
N
Y
∃1! cluster?
Y
N
Reduce the cluster dissimilarity level
Make new cluster with the identified elements
Is 1! element left?
N
Y
Output: hierarchical clustering of data
END
(a)
START
Input: • all clusters are singleton;
• cluster internal dissimilarity = 0
Choose a linkage among the elements
Choose d^*_∞ as proximity measure (*)
Measure all the distances between each couple of elements by means of d^*_∞
Find the couple of elements with the lowest distance
Make new cluster with the identified elements
Is 1! element left?
N
Y
Output: hierarchical clustering of data
END
(b)

Fig. 1 Flowchart of two possible implementations of the dynamic hierarchical clustering, based on (b) the geometric interpretation of the bounding boxes and (a) the novel d^*_∞ pseudometric, respectively. The dashed boxes in the background highlight the difference between the two enforcements. The step marked by (*) in subfig. (b) can be substituted with the choice of a different metrics, as in other clustering methods

It is mandatory to stress that the proximity measure can be, in general, quantified by any metrics, since the synthesis of the results applies directly to the resulting dendrogram. However, due to the geometric interpretation of the described pseudometric, the implementation of the procedure is more straightforward and results are more robust when d^*_∞ is adopted.

A toy problem is now defined to understand the novel procedure described in the following. The configurations are represented by a generic set of 58 points, depicted in Fig. 2, identified by two coordinates, albeit the method can be generalized to spaces of any dimension. Similarly, since the adopted dynamic clustering does not depend on the data scaling, the numerical values are not relevant. Each coordinate represents the outcome of an experiment, after a proper averaging and uncertainty quantification. Experimental results depend on a number of controlled, i.e., independent variables and on a set of parameters: in the proposed procedure, one value of the controlled variable is considered, and the effect of the parameters on the outcomes is observed. In general, any kind of parameter, not only geometric ones, is valid: in the following example, the parameters are assumed to be

- state (s), with three possible values: *on*, *off*, *idle*;
- painting (p), over four values: *cyan*, *magenta*, *yellow*, *black*;
- direction (d), with six values: *north*, *south*, *west*, *east*, *zenith*, *nadir*
- figure (f), over four values: *square*, *round*, *triangular*, *irregular*.

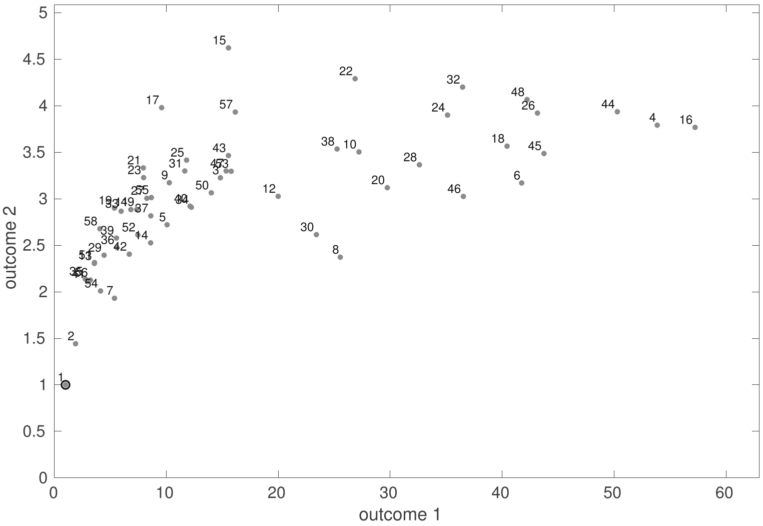


Fig. 2 Set of 58 data generated to define the discussed toy-problem. The coordinates of the two-dimensional, multivariate plane represent the experimental results. Each configuration is identified by four geometrical parameters

Therefore, when experiments are carried out in, e.g., configuration 7, one has $s = on$, $p = cyan$, $d = east$, $f = square$.

The results of the clustering is usually represented by means of a dendrogram. The dendrogram ordinate represents the increasing size of the bounding box, necessary to cluster the diverse configurations reported in abscissa. The ordinate starts from zero (for fully-identical configurations) and it increases to the maximum value (for the least similar configurations), which depends on each case.

Figure 3 depicts an the dendrogram resulting from the dynamic clustering of the points of the toy-problem. To fit in the page, it is rotated of 90° counterclockwise: for this reason, the ordinate, representing the percentage distance, is the horizontal axis. Similarly, all the original configurations are listed along the vertical axis, and represent experimental outcomes. The dendrogram in Fig. 3, as well as all other results presented in this paper and in Ref. Vignati et al. (2018), is generated by means of a code developed by the Authors and implemented in MATLAB (MATLAB and Statistics Toolbox Release, 2012b). For all the details on the clustering algorithm based on the novel d^*_∞ proximity measure, its implementation and discussion, the reader is addressed to Ref. Vignati et al. (2018).

2.2 Step II: derivation of groups

The second step processes the results of the clustering and provides a first indication of the possible groups. Since this tool summarizes the hierarchical clustering, it will be termed "reduced" clustering.

These groups are defines as groups of configurations which share a "reasonable" level of similarity. For this reason, a threshold value τ is selected, which allows to partition the hierarchical clustering: the configurations with a percentage distance lesser than or equal to the threshold are merged together to form groups. The solid vertical line in Fig. 3 represents

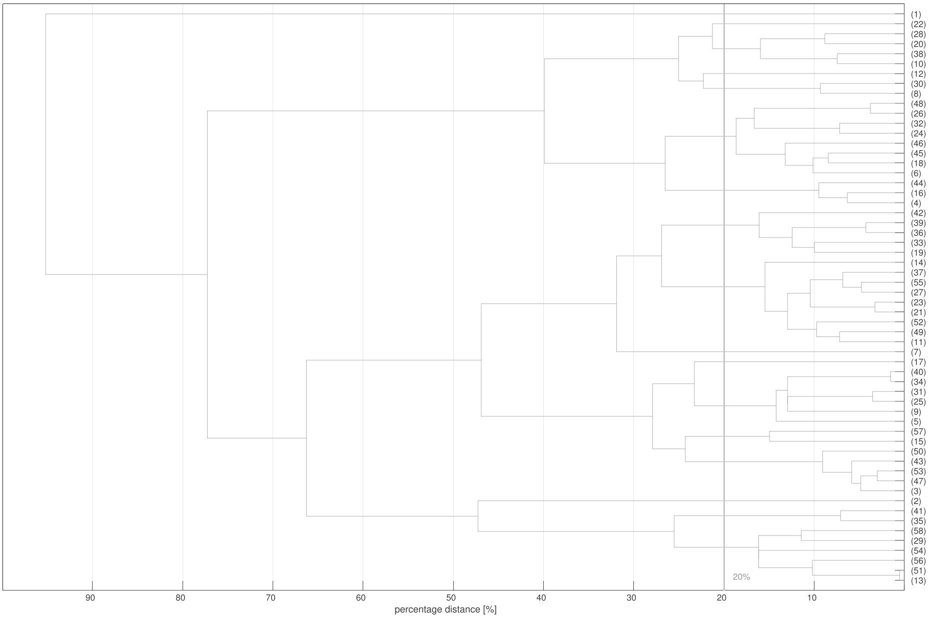


Fig. 3 Dendrogram resulting from the dynamic clustering of the 58 points adopted to define the toy problem depicted in Fig. 2. The investigated configurations are listed along the vertical axis, whereas the percentage distance among each other increases from right to left along the horizontal axis since the dendrogram was rotated of 90° counterclockwise to fit in the page

an example of possible threshold, i.e., the percentage distance of 20% among the configurations, which summarizes in one parameter all the effects of the multivariate data. Therefore, all the configurations that belong to the same group have a reciprocal distance smaller than 20% (in the example, or the selected threshold in general), and they are indeed linked by a branch intersecting the solid line.

2.2.1 Determination of τ

The selection of τ is non-trivial, as it is, in general, arbitrary; however, some guidelines are available to compute a range of possible values.

Due to the definition of the adopted pseudo-metric, defined in ref. Vignati et al. (2018), both the proximity measure and the experimental error have the same geometric interpretation, i.e., the area of the bounding boxes. For this reason, the selected threshold must be selected accounting for the experimental error, since it is mandatory for the threshold to recognize real distances between configuration pairs. In other words, the threshold must be sufficiently larger than the measurement error, which depends on the considered dataset and can be computed in accordance to standard techniques (Moffat, 1988). This workaround ensures that the identified groups of configurations are disengaged from the measurement errors. The uncertainty, indeed, implies that the experimental results are not fully deterministic, but each outcome should be, in general, a range of possible values, i.e., a band. When multivariate data are considered, configurations are identified by a number of outcomes, and therefore the whole element coordinates in the multivariate space actually represents a (multidimensional, in general) area of possible values, i.e., a box surrounding the configuration.

For this reason, since the d^*_∞ proximity measure is related to the area of the bounding box and the experimental uncertainty represents the area of the possible values for the considered configuration, the first one must be significantly larger than the second one, encompassing all the possible values. The lowest acceptable value of τ, therefore, must detect which configurations are certainly *far*, i.e., belonging to different groups, even considering possible deviations from the nominal values due to the experimental uncertainty.

Conversely, the upper limit is neither defined a-priori, nor, actually, mandatory. The latter can only be selected by means of a trial-and-error procedure, whose initial guess depends on the application, the required accuracy, etc. The larger the threshold value, indeed, the lower the selectivity of the procedure. If no information is available to the user for the selection of the upper value, this can be simply put equal to the maximum value of the percentage distance of each dendrogram. The definition of an upper limit for τ is indeed adopted only to reduce the computational cost of the following second tool, whose goal includes also the computation of the optimal value of the threshold. On the contrary, if the user has some criteria available (e.g., the approximate number of resulting families, or the maximum number of elements in each one, etc) the threshold upper limit can be immediately defined.

Other considerations depending on the user experience, investigation goal and possibility (in case of industrial components) of operating in off-design conditions may apply, further restricting the range of candidate threshold values. To uniquely define the optimal value for τ, however, an iterative procedure should be performed, including the (next) refinement step second tool (presented in Section 3). In both cases, therefore, i.e., whether the upper limit is defined or not, the refinement step included in the second tool is the core of the calculation of τ, since, if τ were too large, excessively different configurations would belong to the same group.

For the case presented in Fig. 3 (assuming the uncertainty to be 7%), it is observed that the lowest threshold value should be 15% whereas the largest one is set to 96%. The latter value was not selected a-priori, since the presented case is a simple toy-problem, and therefore there are no general guidelines to select the initial guess for the range of τ. For this reason, the largest investigated value for τ is simply the maximum value of the percentage distance in the dendrogram.

In the following, albeit different values of threshold are considered, results are shown for $\tau = 20\%$, resulting in groups including, e.g., configurations 28-20-38-10 or 44-16-4.

2.3 Step III: refinement of the procedure

When analyzing multivariate data, the relevance of each outcome may be different, to the purposes of the considered investigation. This is particularly true when experimental data are considered, since specific physical quantities are associated to the different directions in the multivariate space.

By means of the adopted proximity measure d^*_∞, the assignation of different relevance to each direction is straightforward. The aspect ratio AR of the bounding boxes must be simply modified: if the bounding boxes are enlarged in one or more directions, the clustering becomes more inclusive and therefore the physical quantity identified by the considered direction in the multivariate space is less influential on the partition into groups. Conversely, if the bounding boxes aspect ratio is shrunk along one or more directions, it becomes less likely for configurations to merge, and therefore the clustering procedure becomes more selective with respect to the considered physical quantity (or outcome). Clearly, a unit-value AR corresponds to the base case, where all the physical quantities have the same relevance and, therefore, all directions in the multivariate space are similarly accounted in the computation of d^*_∞.

The aspect ratio affects the size of the bounding boxes, and therefore the percentage distance among the configurations. For this reason, the threshold value may be modified with respect to the unit-value case, to account for the different selectivity of the bounding boxes. Figure 4 depicts the dendrogram resulting from the clustering performed with $AR = 5$ (along one direction in the multivariate space). Since the aspect ratio is larger than one, at each step of the clustering the bounding boxes are larger than the corresponding ones adopted for the standard clustering (resulting in the dendrogram in Fig. 3), and therefore the configurations are merged when their percentage distance (still computed by means of d^*_∞) is lower. The solid line in Fig. 4, indeed, is still in correspondence of a threshold value of 20% for comparison with Fig. 3, but it intersects only three branches, and therefore only three groups are identified, whereas seventeen were obtained by cutting the dendrogram in Fig. 3.

It is mandatory to specify that there is no way to decide a priori the optimal combination of τ and AR, because it depends on the specific problem and goal of the investigation. However, the difference between the two cases (in Figs. 3 and 4) is clear, and therefore special care must be paid if different aspect ratios or data treatments are adopted (Langfelder et al., 2008), since only $\tau = 1$ corresponds to a physical situation, where all directions in the multivariate space are equally weighted. To make the results more robust, a preliminary analysis is suggested, to investigate the effect of different bounding box aspect ratios and thresholds on the clustering of a specific dataset, in terms of, e.g., number of groups, selectivity, robustness, etc. Eventually, some indications may be retrieved by the application of the second tool (presented and discussed in Section 3), to recursively correct the values of AR and τ.

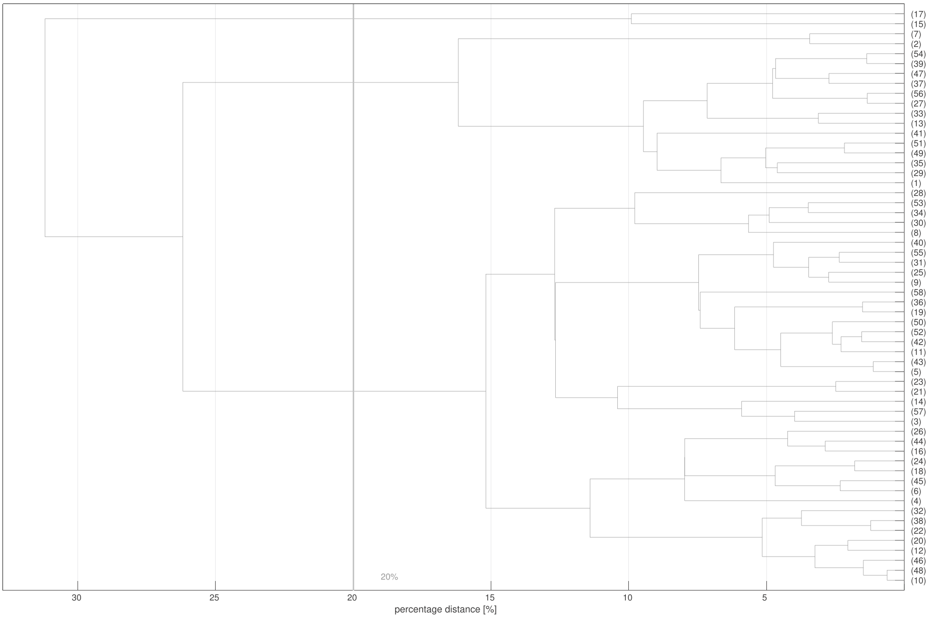


Fig. 4 Dendrogram obtained with a bounding box aspect ratio of 5 (dendrogram rotated, as in Fig. 3)

2.4 Comparison with flat clustering

The described cluster reduction results in some groups of configurations, in analogy with flat clustering. However, the proposed procedure presents three main advantages:

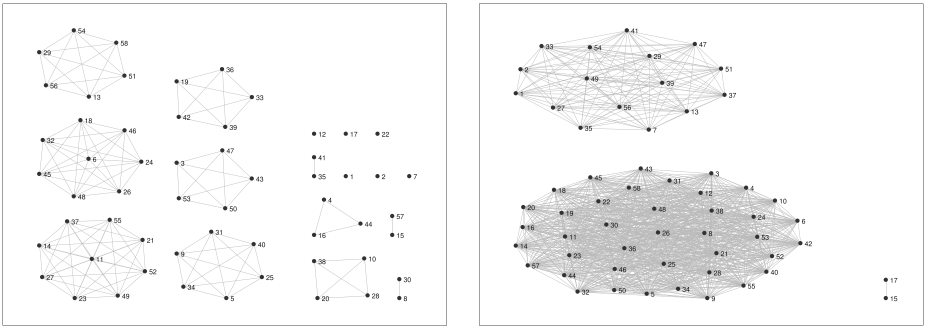
1. flat clustering algorithms usually require to define the so-called “seeds”, i.e., some reference points to compute the distance of all the configurations. These seeds usually represents the centroids of the clusters, and, therefore, the data partitioning depends on both the number and the initial position of the seeds. Conversely, no initial seed needs defining when hierarchical clustering is adopted, and therefore the configurations are merged together depending on their mutual similarity only;
2. even after the cluster reduction, the internal hierarchy of each group is preserved: for this reason, information retrieved by the procedure are more complete, and the mutual similarity among the configurations is available for further, more detailed analysis. On the contrary, with flat clustering, there is no difference among all the configurations belonging to the same cluster, and therefore it is not possible to retrieve additional information or perform sensitivity and uncertainty analysis;
3. although the computation of hierarchical clustering is less efficient than the corresponding one for flat clustering, and moreover the retrieved groups depend on the choice of the threshold, once the cluster reduction has been performed no additional operation is required. Conversely, a sensitivity analysis to determine the optimal seed number and arrangement is usually strongly recommended, which increases the computational cost of the flat clustering procedure.

3 Second tool: robust results synthesis

The second tool analyzes the groups of configurations obtained from the reduced clustering by means of graph theory. The developed tool, termed “aggregated matrix”, provides two relevant information: on the first hand, sets of configurations which systematically merge into groups, also for different values of AR or τ, and over a range of operating conditions, are identified and termed “families”. Families are more meaningful than groups, since they do not depend on the choice of the bounding box aspect ratio, threshold and experimental conditions, if any. On the second hand, the structure of the aggregated matrix automatically sorts which parameters, identifying the different configurations, are more relevant in determining their coalescence into clusters.

3.1 Identification of stable families of configurations

After the reduction of the hierarchical clustering described in Section 2, groups of configurations are identified, which share the same internal similarity. Each group, therefore, represents a so-called clique (Luce & Perry, 1949), the set of vertexes of a complete, undirected graph (Thulasiraman et al., 2016). For this reason, the problem of obtaining synthetic indications from hierarchical clustering can be approached in accordance with graph theory. In this perspective, the cut dendrograms resulting from the cluster reduction are disconnected graphs: each element is connected to all those elements, and those only, which belong to the same group. The graphs obtained after the cluster reduction are depicted in Fig. 5 for the two aspect ratio values of 1 and 5, showing the resulting 17 and 3 cliques, respectively.



(a) $AR = 1$

(b) $AR = 5$

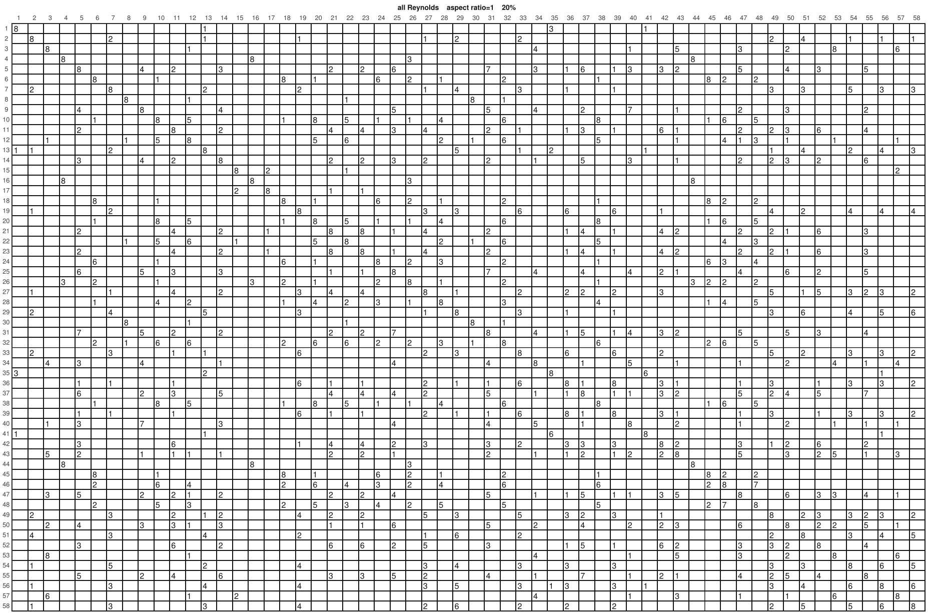
Fig. 5 Graphs associated to the reduced clustering, for the cases of (a) aspect ratio of 1 and (b) aspect ratio of 5

For every reduced clustering, the so-called "adjacency matrix" Λ (of size $n \times n$) is then computed, that is, a table with each row and column identifying one of the n configurations. If the two configurations associated to the $j-th$ row and $k-th$ column belong to the same group, then the cell value, hereafter termed "similarity coefficient" λ_{jk}, is 1, otherwise it is zero (Biggs, 1993). In the present work, a modified definition is adopted, since $\lambda_{jj} = 1$, whereas it is commonly 0 or ≤ 2 in undirected graphs (for simple and multigraphs, respectively). Therefore, all the adjacency matrices are symmetric and have unit values along the main diagonal. For a given dataset, Λ depends on both the bounding box aspect ratio and on the threshold value, for an overall number of possible adjacency matrices of $AR \times \tau$. Moreover, experimental investigations are usually designed with one or more quantities varying over multiple levels, i.e., the controlled or independent variables. Each set of points in the multivariate space, therefore, represents the outcomes at a given combination of independent variables. For this reason, Λ depends also on the value of the controlled variables. Unexpectedly, thanks to the following procedure, this large number of available adjacency matrices is the key to sort the most relevant information from the clustering. On the first hand, indeed, some sets, termed "families", are retrieved, containing configurations which tend to cluster together even for different values of AR and τ (disengaging the results from the arbitrary choice of the bounding box aspect ratio and threshold) or of the controlled variables (pointing to a physical and thus meaningful robustness of the results). On the second hand, it automatically identifies the parameters, e.g., geometrical or operational, identifying each configuration which mostly affect the clustering. This second result, in particular, is not usually permitted by simple cluster analysis, since it only retrieves *which* configurations are similar, but not *why* or *to what extent*. To achieve this result, a procedure is devised from the so-called weighted correlation network analysis and fuzzy clustering (Ross, 2004). In this approach, adjacency matrix elements may assume other values, in addition to 1 and 0, and therefore they do not simply identify a logical correlation, but a robustness of the link between the elements: the larger the value of a cell, the more robust or persistent the similarity between the two associated configurations. A number of fuzzy clustering algorithms are available (Bezdek, 1981; Dunn, 1973; Gustafson & Kessel, 1979), which require to perform a whole new clustering. To avoid this additional, memory consuming step, different adjacency matrices resulting from the reduced clustering step (for different AR or τ or operating conditions) are added. The result is a single matrix Λ^{agg}, of size $n \times n$, termed "aggregated". Boolean adjacency matrices are therefore added to build non-boolean aggregated matrices, whose values are still positive integers, but depart from 0 or 1. The analysis of aggregated matrices, therefore, allows to retrieve not only a boolean information, but a spectrum of similarities, ranging from 0 (none) to the diagonal values (used to set the reference of the maximum similarity). This approach is therefore not boolean, but fuzzy.

Different aggregated matrices can be defined. If the hierarchical clustering of a given dataset is performed with a constant bounding box aspect ratio, only the threshold affects the resulting adjacency matrices, and therefore their sum results in the "threshold-aggregated" matrix Λ^{τ}. Similarly, if τ and the operating conditions are kept constant, only the bounding boxes aspect ratio effect is observed on the clustering, and therefore the resulting matrix Λ^{AR} is termed "aspect ratio-aggregated", and its elements λ_{jk}^{AR} range from 0 to the number of explored aspect ratios. In the same way, both clustering parameters, i.e., AR and τ can be fixed, and hierarchical clustering applied to a number of dataset, representing outcomes of experiments carried out in different operating conditions: the matrix addition retrieves an "experimental-aggregated" matrix Λ^{e}, with λ_{jk}^{e} ranging from zero to the number of levels for the considered controlled variable. It is remarkable that only the last two operations are meaningful, since τ does not influence the clustering, and it is not linked to the physics of the investigated problem. Λ^{τ}, therefore, should be computed and studied only when a priori considerations on reasonable values of the threshold are not available, or when the range of possible τ is too wide. Eventually, combined effects may be considered, by, e.g., adding all Λ^{e} obtained for different aspect ratios or threshold, into the so-called "overall-aggregated" Λ^{O}.

As an example of the described technique, experiments are carried out in eight operating conditions. For each of the eight levels of the controlled variable, the 58 configurations adopted to compute the dendrogram in Fig. 3 are tested, and therefore eight dendrograms are built, by means of the dynamic procedure described in Vignati et al. (2018) and with $AR = 1$. The obtained hierarchical clusters are then reduced in accordance with the method described in Section 2, adjacency matrices are computed and summed, resulting in Λ^{e}, which is displayed in Fig. 6. Albeit different threshold values were tested, too, in the following only the cases with $\tau = 20\%$ will be shown, for brevity. The reading of adjacency and aggregated matrices is clearly not friendly, due to their large size; however, the computation and subsequent use of all matrices does not require an active role of the user, and therefore this limitation is naturally overcome by the automation of the method. Therefore, all the displayed matrices in this paper are for the purpose of a better understanding of the procedure steps.

During the experiments performed to compute the matrix in Fig. 6, the controlled variable is varied over 8 levels, and therefore λ_{jk}^{e} range from 0 to 8. The first case implies no similarity at all between the $j-th$ and the $k-th$ configurations, like between, e.g., 26 and 40. On the contrary, the second occurrence points to the maximum similarity between configurations j and k, i.e., the two configurations cluster for all the considered levels of the controlled variable like, e.g., 4, 16 and 44. These three configurations, therefore, make up a "family", i.e., the smallest group of configurations which result in similar outcomes (that is, which cluster together) even in different operating conditions and regardless of the selectivity of the procedure (which depends on AR and τ). All other values of λ_{jk}^{e} provide good indications, since they allow to draw considerations on the distribution of data, whether, e.g., there are families of configurations which cluster together *almost always* or *almost never*. The pair 41 and 35, for example, belong to the same group resulting from the reduction of the dendrogram reported in Fig. 3, but $\lambda_{35,41}^{e} = 6$ instead of 8, implying that, in two



all Reynolds aspect ratio=1 20%

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58
1	8												1																						3						1																	
2		8					2						1						1								1		2				2																2		4			1		1		1
3			8									1																						4						1			5				3			2			8				6	
4				8												8										3																		8														
5					8				4		2			3							2		2		6						7			3		1	6		1	3		3	2				5			4		3			5			
6						8				1								8		1				6		2		1				2						1							8	2		2										
7		2					8						2						2								1		4				3			1			1										3		3			5		3		3
8								8				1										1								8		1																										
9					4				8					4											5						5			4			2			7			1				2			3					2			
10						1				8		5						1		8		5		1		1		4				6						8							1	6		5										
11					2						8			2							4		4		3		4				2		1			1	3		1			6	1				2		2	3		6			4			
12			1					1		5		8								5		6						2		1		6						5					1			4	1	3		1			1				1	
13	1	1					2						8																5				1		2						1								1		4			2		4		3
14					3				4		2			8							2		2		3		2				2			1			5			3			1				2		2	3		2			6			
15															8		2					1																																			2	
16				8												8										3																		8														
17															2		8				1		1																																			
18						8				1								8		1				6		2		1				2						1							8	2		2										
19		1					2												8								3		3				6			6			6			1							4		2			4		4		4
20						1				8		5						1		8		5		1		1		4				6						8							1	6		5										
21					2						4			2			1				8		8		1		4				2					1	4		1			4	2				2		2	1		6			3			
22								1		5		6			1					5		8						2		1		6						5								4		3										
23					2						4			2			1				8		8		1		4				2					1	4		1			4	2				2		2	1		6			3			
24						6				1								6		1				8		2		3				2						1							6	3		4										
25					6				5		3			3							1		1		8						7			4			4			4		2	1				4			6		2			5			
26				3		2				1						3		2		1				2		8		1				2						1						3	2	2		2										
27		1					1				4			2					3		4		4				8		1				2			2	2		2			3							5		1	5		3	2	3		2
28						1				4		2						1		4		2		3		1		8				3						4							1	4		5										
29		2					4						5						3								1		8				3			1			1										3		6			4		5		6
30								8				1										1								8		1																										
31					7				5		2			2							2		2		7						8			4		1	5		1	4		3	2				5			5		3			4			
32						2		1		6		6						2		6		6		2		2		3		1		8						6							2	6		5										
33		2					3				1		1						6								2		3				8			6			6			2							5		2			3		3		2
34			4		3				4					1											4						4			8			1			5			1				1			2			4		1		4	
35	3												2																						8						6															1		
36					1		1				1								6		1		1				2		1		1		6			8	1		8			3	1				1		3			1		3		3		2
37					6				2		3			5							4		4		4		2				5			1		1	8		1	1		3	2				5		2	4		5			7			
38						1				8		5						1		8		5		1		1		4				6						8							1	6		5										
39					1		1				1								6		1		1				2		1		1		6			8	1		8			3	1				1		3			1		3		3		2
40			1		3				7					3											4						4			5			1			8			2				1			2			1		1		1	
41	1												1																						6						8															1		
42					3						6								1		4		4		2		3				3		2			3	3		3			8	2				3		1	2		6			2			
43			5		2				1		1	1		1							2		2		1						2			1		1	2		1	2		2	8				5			3		2	5		1		3	
44				8												8										3																		8														
45						8				1								8		1				6		2		1				2						1							8	2		2										
46						2				6		4						2		6		4		3		2		4				6						6							2	8		7										
47			3		5				2		2	1		2							2		2		4						5			1		1	5		1	1		3	5				8			6		3	3		4		1	
48						2				5		3						2		5		3		4		2		5				5						5							2	7		8										
49		2					3				2		1	2					4		2		2				5		3				5			3	2		3			1							8		2	3		3	2	3		2
50			2		4				3		3	1		3							1		1		6						5			2			4			2		2	3				6			8		2	2		5		1	
51		4					3						4						2								1		6				2																2		8			3		4		5
52					3						6			2							6		6		2		5				3					1	5		1			6	2				3		3	2		8			4			
53			8									1																						4						1			5				3			2			8				6	
54		1					5						2						4								3		4				3			3			3										3		3			8		6		5
55					5				2		4			6							3		3		5		2				4			1			7			1		2	1				4		2	5		4			8			
56		1					3						4						4								3		5				3		1	3			3		1								3		4			6		8		6
57			6									1			2																			4						1			3				1			1			6				8	
58		1					3						3						4								2		6				2			2			2										2		5			5		6		8

Fig. 6 Λ^e with $AR = 1$ and τ =20%. Matrix built by adding 8 adjacency matrices, each resulting from the clustering reduction of the dendrogram computed over a dataset. Each dataset considers 58 configurations, i.e., points in the multivariate space. During the experiments, the controlled variable is varied over 8 levels, and therefore λ^e_{jk} range from 0 (no similarity) to 8 (the two configurations cluster for all the considered levels of the controlled variable). Numbers in each cell of figure can be read in the digital copy, with a minimum suggested 2× enlargement

operating conditions, the two configurations do not result in near outcomes, an therefore that their similarity is not due to stable, physical reasons.

Moreover, if in an aggregated table there were too many or too few occurrences of the minimum or maximum value (outside of the main diagonal, which is maximum by definition), the user could reconsider the selected values of AR or τ, being the clustering or its reduction too selective or inclusive.

The identification of the families can be speeded up and made error-proof thanks to a proper permutation of the rows (and, similarly, of the columns, to preserve the symmetry) of the aggregated matrices. The order of rows and columns within each adjacency or aggregated matrix is indeed arbitrary, since it represents a run order associated to each configuration, and therefore it can be modified. All possible $n!$ permutations can be applied, preserving the meaning of the adjacency matrices, but some highlight a specific structure governing the values of the similarity coefficients, and therefore points to a possible criterion governing the experimental outcomes. Figure 7 shows a permuted version of the aggregated matrix depicted in Fig. 6, with similar rows sorted together. If any two or more rows are equal, they surely belong to the same family. It is worth stressing that this is ensured by the use of modified the adjacency matrices, with unit-values along the main diagonal. Equal groups of rows can be immediately spotted, when the number of configurations is low, like in the example. However, a large number of algorithms and tools are available to detect groups of equal rows and, therefore, of families.

all Reynolds aspect ratio=1 20% - sorted by repetitions

	20	38	10	3	53	16	44	4	18	45	6	21	23	8	30	36	39	1	2	5	7	9	11	12	13	14	15	17	19	22	24	25	26	27	28	29	31	32	33	34	35	37	40	41	42	43	46	47	48	49	50	51	52	54	55	56	57	58
20	8	8	8						1	1	1													5						5	1		1		4			6									6		5									
38	8	8	8						1	1	1													5						5	1		1		4			6									6		5									
10	8	8	8						1	1	1													5						5	1		1		4			6									6		5									
3				8	8																			1																4			1			5		3			2						6	
53				8	8																			1																4			1			5		3			2						6	
16						8	8	8																									3																									
44						8	8	8																									3																									
4						8	8	8																									3																									
18	1	1	1						8	8	8																				6		2		1			2									2		2									
45	1	1	1						8	8	8																				6		2		1			2									2		2									
6	1	1	1						8	8	8																				6		2		1			2									2		2									
21												8	8			1	1			2			4			2		1				1		4			2					4			4	2		2		2	1		6		3			
23												8	8			1	1			2			4			2		1				1		4			2					4			4	2		2		2	1		6		3			
8														8	8									1						1								1																				
30														8	8									1						1								1																				
36												1	1			8	8			1	1		1						6					2		1	1		6			1			3	1		1		3			1	3		3		2
39												1	1			8	8			1	1		1						6					2		1	1		6			1			3	1		1		3			1	3		3		2
1																		8							1															3			1														
2																			8		2				1				1					1		2			2												2		4		1		1		1
5												2	2			1	1			8		4	2			3					6					7			3		6	3		3	2		5			4		3		5			
7																1	1		2		8				2				2				1		4			3											3		3		5		3		3
9																				4		8				4					5					5			4		2	7			1		2			3				2			
11												4	4			1	1			2			8			2					3		4			2		1			3			6	1		2		2	3		6		4			
12	5	5	5	1	1									1	1									8						6					2			6								1	4	1	3		1						1	
13																		1	1		2				8										5			1		2			1						1		4		2		4		3
14												2	2							3		4	2			8					3		2			2			1		5	3			1		2		2	3		2		6			
15																											8	2		1																										2	
17												1	1														2	8																													
19																6	6		1		2								8					3		3			6						1					4		2		4		4		4
22	5	5	5											1	1									6			1			8					2			6									4		3									
24	1	1	1						6	6	6																				8		2		3			2									3		4									
25												1	1							6		5	3			3					8					7			4		4	4		2	1		4			6		2		5			
26	1	1	1			3	3	3	2	2	2																				2		8		1			2									2		2									
27												4	4			2	2		1		1		4			2			3					8		1			2			2			3					5		1	5	3	2	3		2
28	4	4	4						1	1	1													2						2	3		1		8			3									4		5									
29																1	1		2		4				5				3				1		8			3											3		6		4		5		6
31												2	2			1	1			7		5	2			2					7					8			4		5	4		3	2		5			5		3		4			
32	6	6	6						2	2	2			1	1									6						6	2		2		3			8									6		5									
33																6	6		2		3		1		1				6					2		3			8						2					5		2		3		3		2
34				4	4															3		4				1					4					4			8		1	5			1		1			2				1		4	
35																		3							2															8			6												1		
37												4	4			1	1			6		2	3			5					4		2			5			1		8	1		3	2		5		2	4		5		7			
40				1	1															3		7				3					4					4			5		1	8			2		1			2				1		1	
41																		1							1															6			8												1		
42												4	4			3	3			3			6						1		2		3			3		2			3			8	2		3		1	2		6		2			
43				5	5							2	2			1	1			2		1	1	1		1					1					2			1		2	2		2	8		5			3		2		1		3	
46	6	6	6						2	2	2													4						4	3		2		4			6									8		7									
47				3	3							2	2			1	1			5		2	2	1		2					4					5			1		5	1		3	5		8			6		3		4		1	
48	5	5	5						2	2	2													3						3	4		2		5			5									7		8									
49												2	2			3	3		2		3		2		1	2			4					5		3			5			2			1					8		2	3	3	2	3		2
50				2	2							1	1							4		3	3	1		3					6					5			2		4	2		2	3		6			8		2		5		1	
51																			4		3				4				2					1		6			2											2		8		3		4		5
52												6	6			1	1			3			6			2					2		5			3					5			6	2		3		3	2		8		4			
54																3	3		1		5				2				4					3		4			3											3		3		8		6		5
55												3	3							5		2	4			6					5		2			4			1		7	1		2	1		4		2	5		4		8			
56																3	3		1		3				4				4				3		5			3		1			1						3		4		6		8		6
57				6	6																			1			2												4			1			3		1			1						8	
58																2	2		1		3				3				4				2		6			2											2		5		5		6		8

Fig. 7 Permuted versions of the aggregated matrix Λ^{agg} reported in Fig. 6, with equal rows sorted together. If two or more configurations present the same coefficients in the corresponding rows, they surely belong to the same family

3.2 Hierarchy of relevance of the geometrical parameters

The configurations are defined by means of a collection of parameters, which are usually well defined and quantified, in particular when they represent experimental features. Some parameters are, in general, more relevant in determining which configurations belong to the same groups or families. In the following, three indicators are defined and adopted, to determine the parameters hierarchy of relevance.

3.2.1 Matrices permutation

A first indication is obtained by means of a proper permutation of the rows (and columns) of the aggregated matrices.

In the proposed procedure, the rows of a given aggregated Λ^{agg} are sorted to gather together, in horizontal bands, all the experimental configurations sharing the same value of each geometric parameter. Thanks to the matrix symmetry, also the columns are automatically sorted in the same way, identifying vertical bands of configurations identified by the same parameter. In the discussed case, four parameters are used to define the configurations, and therefore four permuted matrices are obtained, i.e., Λ_s^{agg}, Λ_p^{agg}, Λ_d^{agg} and Λ_f^{agg}. Each of them presents square blocks of cells along the matrix main diagonal: Λ_s^{agg}, for example, shows two blocks (one with all elements in state *on* and one with all elements in state *off*). All the cells out of these two blocks represent the similarity coefficients between configurations with different states.

The block-structure of a permuted matrix is more evident when the parameter which drives the sorting is more influential on the clustering. Therefore, the stronger one parameter relevance, the larger the average of λ_{jk}^{agg} within the blocks and the smaller outside. The limit case is obtained when a parameter is so decisive that is completely segregates all the configurations which do not share the same value of the considered parameters: in this case, the average value of λ_{jk}^{agg} outside of the blocks is ideally zero. On the contrary, when a parameter has a weak influence, it is not able to separate families with configurations characterized by the same value from other ones, and therefore the coefficients of the aggregated matrix are more distributed.

It is worth mentioning that, to a first degree of approximation, the internal order of the bands is not relevant, and therefore several possible permuted matrices exist for each parameter. For this reason, any permuted matrix should provide similar indications.

Two permutations of the aggregated matrix in Fig. 6, driven by *state* and *force*, respectively are shown in Fig. 8. The block structure is more evident in picture Fig. 8a than in picture Fig. 8b, pointing to the *state* as the most relevant parameter.

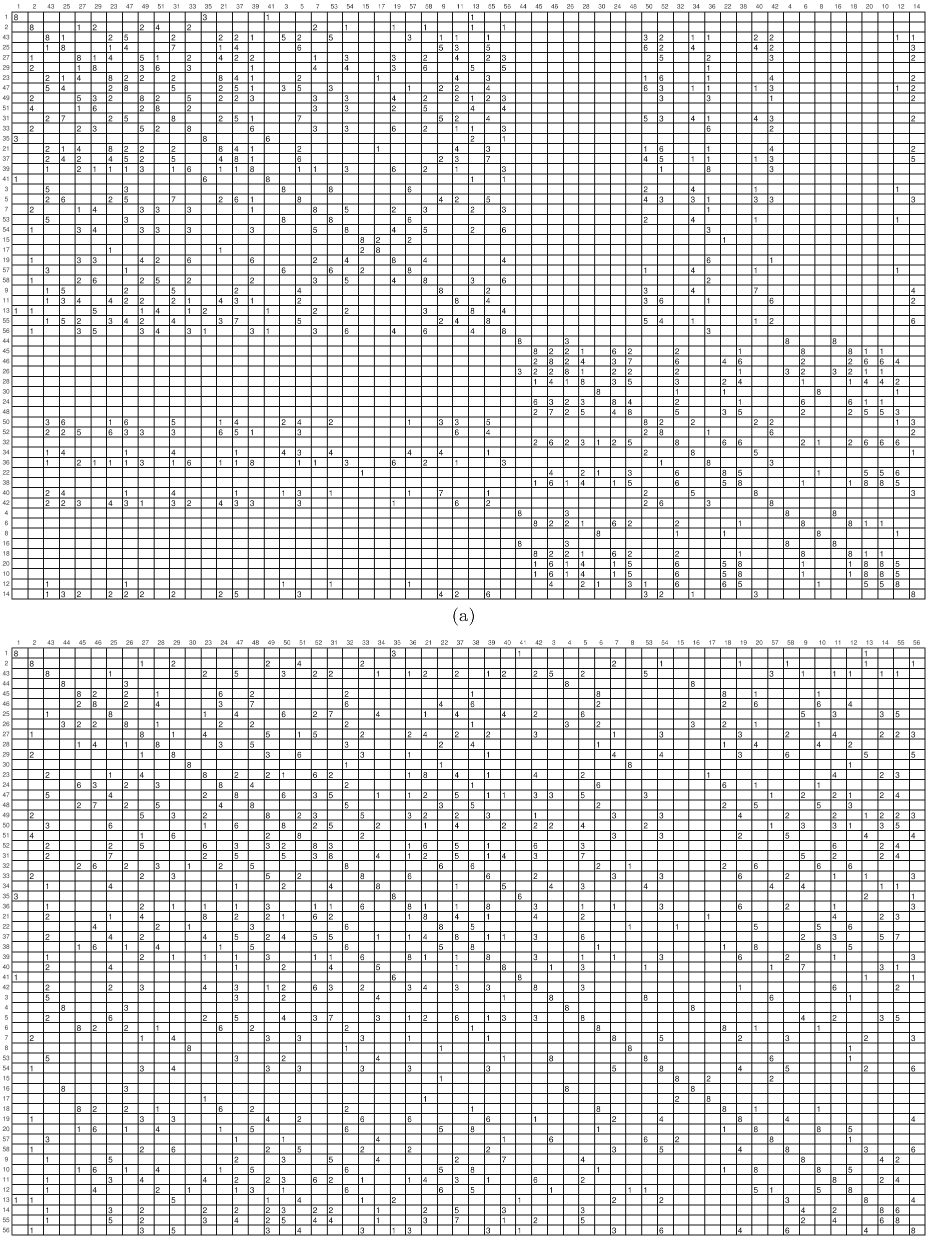
3.2.2 Relevance coefficients

The permutation of the matrices is an immediate indicator of the relevance of each parameter, but it cannot provide exact, quantitative results. For this reason, a new coefficient α is defined and termed "relevance coefficient", which summarizes the effect of the permutation on a given aggregated matrix. Therefore, there are as many relevance coefficients as the parameters are (four in the present study: α_s, α_p, α_d, α_f). Let x be a generic parameter, which can assume X values (e.g., the *painting* p with 3 possible occurrences), then the relevance coefficient of x is defined as

$$\alpha_x^{agg} = \frac{\left[\sum_{i=1}^{X}\Big(\sum_{j,k=1}^{n}\lambda_{jk}^{agg}\cdot\hat{\delta}_j^i\hat{\delta}_k^i\Big)\Big/\sum_{i=1}^{X}\Big(\sum_{j,k=1}^{n}\hat{\delta}_j^i\hat{\delta}_k^i\Big)\right]}{\left[\sum_{i=1}^{X}\Big(\sum_{j,k=1}^{n}\lambda_{jk}^{agg}\cdot\hat{\delta}_j^i(1-\hat{\delta}_k^i)\Big)\Big/\sum_{i=1}^{X}\Big(\sum_{j,k=1}^{n}\hat{\delta}_j^i(1-\hat{\delta}_k^i)\Big)\right]} \tag{2}$$

$\hat{\delta}_*^i$ is a logical parameter which has unit value if, for the $* - th$ configuration, $x = x_i$, and zero otherwise. The numerator computes the average λ_{jk}^{agg} within the blocks of the permuted matrix (hereafter termed block "density"), whereas the denominator is the average λ_{jk}^{agg} outside of the blocks. For this reason, the stronger the influence on the clustering of a parameter, the more evident the block structure, and therefore the larger the value of α_x^{agg}.

All relevance coefficients must be computed, i.e., one for each parameter, and their values compared, resulting in a hierarchy of relevance of the parameters. α_x^{agg} represents a good indicator of the parameters influence, since it confronts the average λ_{jk}^{agg} inside and outside of the blocks, so as to disengage the resulting parameter hierarchy from the choice of AR and τ. If only the block internal density were considered, indeed, it would depend on the selectivity of the procedure. Conversely, with the adopted definition, the resulting coefficient are normalized with respect to the average trend of the configurations to cluster together. According with the definition, moreover, the relevance coefficients do not depend on the internal order of the blocks, but simply on the driving parameter. Eventually, it is quite efficient, since it does not actually require to permute the aggregated matrix, because



(a)

(b)

Fig. 8 Permuted versions of the aggregated matrix Λ^{agg} reported in Fig. 6, by (a) *state* (Λ_s^{agg}) and (b) *force* (Λ_f^{agg}), showing the higher influence of s on the clustering, with respect to f. Due to the small size of the figures, numbers in each cell of figure can be read in the digital copy, with a suggested 4× enlargement

$\hat{\delta}^i_*$ automatically selects the elements of the same block directly from the original, non permuted aggregated matrix. The consideration on the block structure of the permuted matrices of Fig. 8 are confirmed, since $\alpha_s^{agg} = 4.0155$ and $\alpha_f^{agg} = 2.0717$, pointing to the *state* as a more influential parameter than the *force*.

3.2.3 Modified relevance coefficients

A third indicator is the set of coefficients, introduced here and named "modified relevance coefficient", $\tilde{\alpha}_x^{agg}$. All three coefficients are derived from α_x^{agg}, since they compute the density inside and outside of the boxes in the permuted matrices. The difference is that they do not only compare the average λ_{jk}^{agg} of configurations sharing the same parameter with all the other ones, but also confront specific values of the considered parameter.

The first group of X coefficients $\tilde{\alpha}^{agg}_{x,i-i}$, desribed in (3a), indicate the densities within each block, and identify the trend to aggregate of the configurations characterized by the same value of $x = x_i$. The sum of all $\tilde{\alpha}^{agg}_{x,i-i}$ over X represents the numerator of α_x^{agg}, and the double subscript means that it considers only the $i - th$ value of the parameter. For example, for the parameter *painting*, one has $\tilde{\alpha}_{p,C-C}$, $\tilde{\alpha}_{p,M-M}$ and $\tilde{\alpha}_{p,Y-Y}$, being C, M and Y cyan, magenta and yellow, respectively. For each parameter, therefore, one has as many $\tilde{\alpha}^{agg}_{x,i-i}$ as X, and the total number of coefficients of the first type is the same as the whole number of levels of all the parameters (13, in the example).

Conversely, the second group of X elements contains the densities computed over the band of the $i - th$ value of the parameter x but outside of the block, namely $\tilde{\alpha}^{agg}_{x,\bar{i}-\bar{i}}$. With reference to (3b), they indicate the average trend of the configurations with $x = x_i$ to aggregate with all the other ones, i.e., configurations with $x \neq x_i$. Also in this case, the sum of all $\tilde{\alpha}^{agg}_{x,\bar{i}-\bar{i}}$ over X returns the denominator of α_x^{agg}. Also in this case, every parameter is characterized by X coefficients, for an overall number of $\tilde{\alpha}^{agg}_{x,\bar{i}-\bar{i}}$ equal to the total number of levels.

Eventually, the third array of coefficients provide a punctual indication of the average cross-correlation between configurations associated to $x = x_i$ and the ones associated to each value of $x \neq x_i$. They are indicated by $\tilde{\alpha}^{agg}_{x,i-\bar{i}}$ and defined in (3c). The number of coefficients of the third type is larger than the previous two, since it is $X(X-1)/2$ for each paramter.

$$\tilde{\alpha}^{agg}_{x,i-i} = \Big(\sum_{\substack{j,k=1 \\ k\neq j}}^{n} \lambda^{agg}_{jk} \cdot \hat{\delta}^i_j \hat{\delta}^i_k \Big) / \Big(\sum_{\substack{j,k=1 \\ k\neq j}}^{n} \hat{\delta}^i_j \hat{\delta}^i_k \Big) \quad \text{for } i = 1, \dots X \tag{3a}$$

$$\tilde{\alpha}^{agg}_{x,\bar{i}-\bar{i}} = \Big(\sum_{j,k=1}^{n} \lambda^{agg}_{jk} \cdot \hat{\delta}^i_j (1 - \hat{\delta}^i_k) \Big) / \Big(\sum_{j,k=1}^{n} \hat{\delta}^i_j (1 - \hat{\delta}^i_k) \Big) \quad \text{for } i = 1, \dots X \tag{3b}$$

$$\tilde{\alpha}^{agg}_{x,i-\bar{i}} = \Big(\sum_{j,k=1}^{n} \lambda^{agg}_{jk} \cdot \hat{\delta}^i_j \hat{\delta}^{\bar{i}}_k \Big) / \Big(\sum_{j,k=1}^{n} \hat{\delta}^i_j \hat{\delta}^{\bar{i}}_k \Big) \quad \text{for } i = 1, \dots X; \quad \bar{i} = 1, \dots X; \quad \bar{i} \neq i \tag{3c}$$

The densities within each block are computed without considering the diagonal elements, which contain, by definition, the maximum value, as they indicate the self-correlation of each configuration.

Due to the large number of results, modified relevance coefficients are less synthetic than α_x^{agg}. However, they are useful to account for:

- whether a parameter maintains its relevance over all its levels. It is indeed possible that a parameter affects the physics of the problem, and therefore the distribution of experimental results in the multivariate space, only within a range of values. The numerator of α_x^{agg} accounts for the average influence of x, over the whole range of operating conditions and regardless of which x_i produce the most populated blocks in the permuted aggregated matrices. $\tilde{\alpha}^{agg}_{x,i-i}$, on the contrary, shows the relevance of the parameters in correspondence of each value, and therefore allows to detect also possible trends in the parameter relevance, if any. A similar, albeit weaker indication, is provided by $\tilde{\alpha}^{agg}_{x,\bar{i}-\bar{i}}$ as opposed to the denominator of α_x^{agg};
- the possible existence of cross correlations between pairs of parameters, by means of the third array of modified coefficients, i.e., $\tilde{\alpha}^{agg}_{x,i-\bar{i}}$.

For these two reasons, the analysis of the modified relevance coefficient can highlight, if any, isolated cases going against the general observation on the hierarchy of relevance of the geometrical parameters.

4 Estimation of the overall computational cost

The overall computational cost of the procedure depends on both the effort required by the clustering and the post-processing of the resulting dendrograms.

4.1 Computation of the hierarchical clustering

For the first part of the procedure, i.e., the hierarchical clustering of experimental data, some cost estimations are available, as presented in Section 1 (Sokal & Sneath, 1963; Johnson & Wichern, 1990; Anderson, 1984; Bouguettaya et al., 2015; Day & Edelsbrunner, 1984; Hruschka et al., 2009). For a dataset of n data, indeed, the computational cost for the hierarchical clustering is the same as classical ones, i.e., n^2, when the novel, d^*_∞ proximity measure is adopted. Conversely, it increases by a factor n if the geometric approach, based on the bounding boxes construction, is chosen. The refore, the latter is not convenient, since it results in the same dendrogram, but its definition is useful since:

1. it provideds a physical interpretation of the novel proximity measure;
2. it is easily implemented, and therefore it provides a benchmark;
3. it allows to define the control parameter AR, as the bounding boxes aspect ratio, which can make the procedure more or less selective with respect to the considered parameter.

4.2 Synthesis of the data: dendrograms reduction, derivation of stable families and retrieval of relevant parameters

The effort required by the second part of the procedure, i.e., the post-processing of the dendrograms, depends on the goal of the investigation. Each dataset, indeed, may result in different clusters, depending on AR, and therefore the computational cost of the clustering should be multiplied by the number of explored aspect ratios, yielding $AR \cdot n^2$ operations. Albeit the increased number of operations due to the different AR values is performed during the first part of the work, i.e., the dendrogram derivation, its effect falls back on the second part, since the investigation of the aspect ratios effect makes sense only if a post-processing applies to the clustering.

The AR dendrograms are then partitioned at the level τ, retrieving the "reduced" clustering, i.e., a set of groups summarized in the boolean "adjacency matrices" Λ. The number of obtained matrices is not exactly $\tau \cdot AR$, since the number of investigated values of τ depends on AR, as discussed in Section 2.2: the larger the boxes, indeed, the lower the upper value of τ, yielding a computational cost fpr the definition of Λ of $\sum_{AR}(n^2 \cdot \tau_{AR})$, where τ_{AR} is the number of investigated threshold values for a given aspect ratio.

To make the resulting families of configurations robust against the numerical parameters AR and τ, the aggregated matrices are computed as sum of adjacency matrices, retrieving as many Λ^{τ} as the investigated aspect ratios (from the sum of all Λ related to the partitioning of dendrograms built with the same aspect ratio), and a number of Λ^{AR}, similarly defined, depending on the number of values assumed by τ in the range common to all the clustering computed with different AR. To compute Λ^{AR}, therefore, it is mandatory that the investigated values of τ are the same for all the dendrograms, at least in the common range, albeit its upper value depends on AR. The computational effort for these operations is added to the previous one, and it depends on the actual enforcement. If the algorithm is implemented in a high-level language, which is equipped with libraries for matrix algebra, the cost is lower than n^2. However, the amount of dataset elements n can be arbitrarily large, whereas only a reasonable, much lower number of AR and τ levels are commonly investigated. For this reason, they do not significantly affect the magnitude order of the computational cost, which remains controlled by n^2.

Λ^{τ} and Λ^{AR} point to families of configurations which aggregate with each other regardless of the procedure selectivity (controlled by τ) or of the relevance of each variable in the multivariate space, which depends on the application (and is quantified by AR). Therefore, these aggregated matrices indicate the *robustness* of the found families against numerical parameters introduced by the mathematical enforcement of the algorithm. On the contrary, a different approach is devised to investigate the *stability* of the resulting families, if the initial dataset may vary. Indeed, if experimental results are obtained in different operating conditions, different dataset must be considered, and the resulting dendrograms are increased by a factor equal to the number of levels for the considered controlled variables, albeit this does not depend on the algorithm efficiency, but, clearly, only on the number of the experimental data available to the user, as for n. However, also in this case, a number of adjacency matrices can be computed, in the number of the values of the controlled variables times the aspect ratios selected during the clustering (assuming AR values fixed). The latter are added, retrieving the experimental-aggregated Λ^e, at a cost which depends on the actual algebraic enforcement of the operation, but which is, in general, lower than n^2.

Λ^e not only points to stable families of configurations, but also determines a hierarchy of relevance among the X parameters defining the different configurations, i.e., which ones are more influential on the retrieved families, thanks to the novel "(modified) relevance coefficients". Their calculation requires to permute rows and columns of the matrices Λ^e: if only the direct effect of the parameters must be accounted, X permutations are required, while they become $2^X - 1$ if second-order effects, due to the parameters interactions, are of interest, accounting for all the possible combinations among them. For arch Λ^e, therefore, the 2^X permutations are first performed, and then the relative coefficients are computed and ranked. The cost for this operation is added to the previous ones; however, unlike for AR and τ, the number of permutations is not necessarily negligible, if compared to n^2. X and n are, indeed, independent, with $X \leq n$ by definition, and $X << n$ in experimental investigations; however the exponential function (such as 2^X) grows faster than a power-law (like n^2). For this reason, the cost of the latter operation could be comparable to n^2, even for X smaller than n, but only if second-order effects are to be considered, and also in this case the overall effort would not exceed the order of n^2.

For the proposed toy problem, with $n = 58$, no interaction is considered among the parameters, and therefore 4 permutations apply, and 8 relevance coefficients (including the modified ones) are computed.

5 Validation of the procedure

Albeit all the results in this paper refer to the described toy problems, the proposed method was tested on 45 dataset: 8 of them include real experimental data, of about 450 elements, and 37 were generated to observe the robustness and efficiency of the procedure also in different situations. For the generated datasets, the number and the distribution of elements in the multivariate space is variable (in the order of tens or hundreds), but they are designed to mimic the structure of experimental data, i.e., depending on parameters and providing a number of outputs —the directions in the multivariate space.

In the first tests, some centroids were first defined in a two-dimensional space, and some points were arranged in proximity of each one. The distance between each point and one specific centroid was defined to be much smaller than the distance between the same point and all other centroids: this guaranteed the possibility to know *a priori* a reasonable allocation of all the test points, and, therefore, whether the resulting reduced clustering was correct. In the following tests, some uncertainty (up to 15%) was introduced, to explore the robustness, and the dimension of the multivariate space was increased up to 4. In the subsequent tests, data were increasingly spread in the multivariate space, to make the proximity less trivial, and eventually normally distributed data were adopted. Even in the last case, the algorithm was stable, provided the limitations on τ were observed, and a meaningful uncertainly applied. Moreover, in the final tests, the approach based on the reduced hierarchical clustering proved its better performances against a simple flat one, since it was very difficult, if not impossible, to define the seeds.

6 Conclusions and final remarks

Two novel tools to analyze the results of hierarchical clustering of experimental data were presented and discussed. Both tools stem from the consideration that hierarchical clustering results in a large number of indications, but with no indications on the most relevant features, criteria or parameters governing the distribution of the multivariate data. Synthetic indications, on the contrary, are very important when the structure underlying the data is unknown. This is the case of, e.g., results of experimental investigations of complex physical systems like the one motivating this work, carried on at ThermALab of Politecnico di Milano.

The first tool, i.e., the reduced clustering, suggests how to cut the dendrogram resulting from the hierarchical clustering. Moreover, it helps testing the robustness of the resulting groups of configurations against the so-called “bounding box aspect ratio”, that is a possible modification of the clustering selectivity with respect to one or more directions in the multivariate space.

The second, most important tool is derived from graph theory and requires to define some modified adjacency and aggregated matrices resulting from the reduced clustering.

The analysis of the matrices structure and maxima, by means of novel coefficients, identifies the parameters which mostly affect the clustering, and determines the so-called "families" of configurations, i.e., sets of elements with persistent similarity. Moreover, it indicates to what extent these results are *stable* over different operating conditions, and *robust* in the range of variability of the parameters and numerical coefficients adopted during the clustering, if any.

A toy problem was set-up to better understand the role of the different operations and defined quantities.

An estimation of the computational cost of the procedure is eventually added, for completeness.

Both tools, especially the first one, are particularly suitable to the processing of clusters obtained when the so-called d_{∞}^{*} proximity measure is adopted, since it automatically accounts for data uncertainty. However, they apply to any hierarchical clustering.

Author Contributions All authors contributed to the study conception and design. In detail: D. Fustinoni mainly contributed in terms of Conceptualization, Methodology and Investigation. F. Vignati contributed in Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing (original draft, review & editing). A. Niro contributed also in Conceptualization, Project administration, Resources and Supervision.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Availability of data and material Toy problem used to show the method

Code Availability Custom code

References

Aggarwal, C. C., & Yu, P. S. (2009). A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), 609–623.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function*. New York: Plenum Press.

Biggs, N. (1993). Algebraic Graph Theory, Cambridge Mathematical Library (2nd ed.), Cambridge University Press.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, *42*, 2785–2797.

Campbell, J. F. (1996). Hub location and the p-hub median problem. *Operations Research*, *44*(6), 923–935.

Davé, R. N., & Krishnapuram, R. (1997). Robust clustering methods: A unified view. *IEEE Transaction on Fuzzy Systems*, *5*(2), 270–293.

Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*, 7–24.

Day, W. H. E., & Edlesbrunner, H. (1985). Investigation of proportional link linkage clustering methods. *Journal of Classification*, *2*, 239–254.

Dunlop, J. A., Penney, D., & Jekel, D. (2015). A summary list of fossil spiders and their relatives, World Spider Catalog Natural History Museum Bern.

Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact, well separated clusters. *J. of Cybernetics*, *3*(3), 32–57.

Fernández, A., & Gómez, S. (2008). Solving Non-Uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, *25*, 43–65.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174.

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. American Statistical Association Journal, pp 1159–78.

Fustinoni, D., Vignati, F., Gramazio, P., Vitali, L., & Niro, A. (2019). Insight in thermal and fluid-dynamic properties of ribbed ducts by means of a novel clustering method, 37-th UIT Conference Padova.

Gustafson, D., & Kessel, W. (1979). Fuzzy clustering with a fuzzy covariance matrix. Proc. IEEE CDC, 761–766, San Diego USA.

Holton, D., May, R. M., & noise, Distinguishing chaos from (1993). In The Nature of Chaos, Chap. 7 Oxford University Press.

Hormiga, G. (1994). Cladistics and the comparative morphology of linyphiid spiders and their relatives (Arneae, Araneoidea, Linyphiidae). *Zoological Journal of the Linnean Society*, *111*(1), 1–71.

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & de Carvalho, A.C.P.L.F. (2009). A survey of evolutionary algorithms for clustering. *IEEE Trans. on Systems, Man and Cybernetics Part C: Applications and ReviewsOpen 39*, *2*, 133–155.

https://wsc.nmbe.ch/families.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*, 651–666.

Jain, A. K., & Dubes, R. C. (1998). Algorithms for clustering data, prentice hall advanced reference series: Englewood Cliffs NJ.

Jain, A. K., Murty, M. N., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323.

James Rohlf, F., & Sokal, R. R. (1962). The description of taxonomic relationships by factor analysis. *Systematic Zoology*, *11*(1), 1–16.

Jiang, B., Pei, J., Tao, Y., & Lin, X. (2013). Clustering uncertain data based on probability distribution similarity. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 751–763.

Johnson, R. A., & Wichern, D. W. (1990). *Applied Multivariate Statistical Analysis*. New York: Pearson Education.

Jolion, J., Meer, P., & Bataouche, S. (1991). Robust clustering with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(8), 791–802.

Kalantari, B. (2013). The State of the Art of Voronoi Diagram Research. In *Transactions on Computational Science XX, Lecture Notes in Computer Science 8110*. Berlin: Springer. https://doi.org/10.1007/978-3-642-41905-8_1.

Kleinberg, J. (2002). *An Impossibility Theorem for Clustering, Advances in Neural Information Processing Systems 15*, (pp. 446–453). Boston: MIT Press.

Knorr, E. M., Ng, R. T., & Zamar, R.H. (2001). Robust space transformations for distance-based operations. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining* (pp. 126–35).

Kriegel, H. P., & Pfeifle, M. (2005). Density-Based Clustering of uncertain data. In *Proceedings of the 11th ACM KDD Conference pn Knowledge Discovery in Data Mining* (pp. 672–677).

Kumar, M., & Orlin, J. B. (2008). Scale-invariant clustering with minimum volume ellipsoids. *Computers & Operations Research*, *35*, 1017–29.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720.

Lu, Z., Kim, J. Z., & Bassett, D.S. (2020). Supervised chaotic source separation by a tank of water. *Chaos*, *30*, 021101. https://doi.org/10.1063/1.5142462.

Luce, R. D., & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, *14*, 95–116. https://doi.org/10.1007/BF02289146.

MATLAB and Statistics Toolbox Release (2012b). The MathWorks, Inc., Natick, Massachusetts, United States.

MacCuish, J., Nicolaou, C., & MacCuish, N.E. (2001). Ties in proximity and clustering compounds. *J. Chem. Inform. Comput. Sci.*, *41*, 134–146.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, *2*, 49–55.

Margot, J. L. (2015). A quantitative criterion for defining planets. *The Astronomical Journal*, *150*(6), 185–191.

Michener, C. D., Corliss, J. O., Cowan, R. S., Raven, P. H., Sabrosky, C. W., Squires, D. S., & Wharton, G.W. (1970). Systematics In Support of Biological Research, tech. report of Division of Biology and Agriculture, National Research Council, Washington D.C.

Moffat, R. J. (1988). Describing the uncertainties in experimental results. *Experimental Thermal and Fluid Science*, *1*, 3–17.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A Multi-Disciplinary survey. *Data Mining and Knowledge Discovery*, *2*, 345–389.

Niro, A., Fustinoni, D., Vignati, F., Gramazio, P., & Ciminà, S. (2016). Considerations on the thermal performances of ribbed channels by means of a novel dynamic method for hierarchical clustering, 7-th Eurotherm kraków.

Pampalk, E., Dixon, S., & Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. In *Proc. Sixth Internat, Conf. on Digital Audio Effects (DAFx-03)* (pp. 7–12).

Ross, T. J. (2004). *Fuzzy Logic With Engineering Applications*. UK: John wiley & sons ltd.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco: W.H. Freeman and Company.

Thulasiraman, K. K. T., Arumugam, S., Brandstädt, A., & Nishizeki, T. (2016). Handbook of graph theory, Combinatorial Optimization, and Algorithms, Chapman & Hall/CRC Computer and Information Science Series.

Vignati, F., Fustinoni, D., & Niro, A. (2018). A novel scale-invariant, dynamic method for hierarchical clustering of data affected by measurement uncertainty. *Journal of Computational and Applied Mathematics*, *334*, 521–531.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.