UNIVERSITY OF CATANIA
DEPARTMENT OF ECONOMICS AND BUSINESS
MASTER'S DEGREE IN DATA SCIENCE FOR MANAGEMENT

Alessandro Platania

Clustering speed data via a new Ward's method based on the harmonic mean

MASTER'S THESIS

Supervisor:
Prof. Antonio Punzo
Co-supervisor:
Prof. Salvatore Ingrassia

Academic Year 2022/2023

A chi sarebbe accanto a me
se fosse ancora qui

Contents

Clustering speed data via a new Ward's method based on the harmonic mean

Abstract

In statistics and machine learning, cluster analysis plays a crucial role as a foundational method to discover knowledge in multidimensional data, aiding in the identification of common patterns and inherent structures within datasets. One of the most commonly employed clustering techniques is agglomerative hierarchical clustering which nests several methods. Among them, Ward's method represents one of the milestones in clustering literature since it generally guarantees a high degree of reliability in partitioning datasets into meaningful clusters. However, when working with real-world datasets that contain variables given by rates or ratios, the original Ward's method fails to work properly as it depends on the use of the arithmetic mean for determining the mean to compute the within sum of squares. In this paper, we present a novel approach founded on the original Ward's method enhanced by the incorporation of the harmonic mean. Indeed, it is the correct average to be used when dealing with quantities that are inversely proportional (such as rate-based values). The underlying idea is to implement a modification to the traditional Ward's method that considers the nature of the data analysed. The proposed method has been illustrated on a dataset about speeds.

Keywords: Hierarchical clustering, Ward's method, Harmonic mean, Speed dataset

1 Introduction

Agglomerative hierarchical clustering is an unsupervised technique of cluster analysis which aims to create a hierarchical structure of clusters. In contrast to partitioning methods (e.g. K-Means), hierarchical clustering methods do not require the number of clusters to be fixed in advance but produce a series of partitions in multiple stages (Giordani *et al.*, 2020). "Agglomerative" refers to a bottom-up approach, where initial observations form separate clusters at the lower level. As we move up the hierarchy, these clusters are merged based on their similarity, ultimately resulting in a single cluster at the top (Patel *et al.*, 2015). The process involves computing a dissimilarity (or distance) matrix, a square matrix containing

pairwise dissimilarities (or distances) between elements in a given set. With the use of a dendrogram it is possible to graphically represent the result of the clustering process (Zhao *et al.*, 2005).

The most common strategies of agglomerative hierarchical clustering are single, complete, average, centroid linkage and Ward's method. Their main distinction lies in how they choose the data points that influence the final similarity or distance criterion.

In the single linkage method, the distance between two clusters J_a and J_b is determined by the smallest distance between any two points $\mathbf{x}_a$ and $\mathbf{x}_b$, each belonging to different clusters in the pair (Florek *et al.*, 1951), namely $min_{\mathbf{x}_a,\mathbf{x}_b}\{d(\mathbf{x}_a,\mathbf{x}_b)|\mathbf{x}_a \in J_a, \mathbf{x}_b \in J_b\}$. This tends to create elongated clusters since it only considers the closest points.

On the other hand, complete linkage and average linkage methods take a different approach. Complete linkage considers the farthest distance $max_{\mathbf{x}_a,\mathbf{x}_b}\{d(\mathbf{x}_a,\mathbf{x}_b)|\mathbf{x}_a \in J_a, \mathbf{x}_b \in J_b\}$. This results in tight, compact clusters, as it prioritizes the farthest points. Average linkage calculates $(|J_a| \cdot |J_b|)^{-1} \sum_{\mathbf{x}_a \in J_a} \sum_{\mathbf{x}_b \in J_b} d(\mathbf{x}_a, \mathbf{x}_b)$, the average of all pairwise distances between points in different clusters to establish the distance between the clusters (Sokal *et al.*, 1958). As a result, average linkage strikes a balance between single and complete linkage, leading to more moderate cluster shapes (Ros and Guillaume, 2019).

Centroid linkage method offers a distinct approach to clustering. Instead of focusing on individual point-to-point distances, it calculates the distance between the centroids of two clusters $\bar{\mathbf{x}}_a$ and $\bar{\mathbf{x}}_b$, say $d(\bar{\mathbf{x}}_a, \bar{\mathbf{x}}_b)$.

Differently from the other approaches, Ward's method focuses its attention on the minimization of an objective function that could be "any functional relation that an investigator selects to reflect the relative desirability of grouping" (Ward, 1963). The function proposed by Ward is the within sum of squares (WSS). Any merger of two clusters leads to an increase of the WSS and the main goal is to minimize the worsening of the sum of all the WSS of the clusters. This results in the formation of groups where data points are generally in close proximity to their respective cluster means, indicating that clusters consist of similar elements.

Ward's method is one of the most performing methods for data clustering as demonstrated by different authors, e.g. Vijaya *et al.* (2019) and Mongi *et al.* (2019). However, it relies on the utilization of the arithmetic mean to compute the mean used in the formula

of the WSS, while we know that for some data types it is not the correct average to be computed and alternative means should be considered. We focus on the harmonic mean (HM) which represents one of the three Pythagorean means together with the arithmetic and geometric counterparts (Bakker, 2003 and Brown, 1975 and Huffman, 2005) and it is defined as the reciprocal of the arithmetic mean of the reciprocals. It is a measure of location to be used when data consist of a set of rates, such as prices, speeds, or productivity (Komić, 2011). Following these considerations, in this work we introduce a new method that we define *harmonic Ward's method* (HWM), based on the original Ward's method but shifted in a "harmonic space". Therefore, our work concentrates on the introduction of this novel approach with the aim of implementing it in datasets that involve variables derived from rate-based metrics.

The rest of the paper is organized as follows. An overview of related work is presented in Section 2. Section 3 explains the methodology adopted to describe how the algorithm works. Section 4 illustrates how the new method has been implemented on R (R Core Team, 2022). It also emphasizes a real case in which our approach has been applied. To conclude, Section 5 points out some final considerations.

2 Related works

Hierarchical clustering (HC) is a widely used cluster analysis method in unsupervised machine learning, aiming to group data points into clusters based on their distances. In multiple sectors HC techniques have been applied, such as industry (Maleki and Bingham, 2019), health (Ashton *et al.*, 2018), environment (Senthilnath *et al.*, 2019), geology (Unglert *et al.*, 2016) and for different purposes like anomaly detection (Shi *et al.*, 2020), sentiment analysis (Hamdani *et al.*, 2020), pattern recognition (Unglert *et al.*, 2016), image segmentation (Senthilnath *et al.*, 2019). A notable reference for our work is given by the paper of Fernández and Gómez (2019). Within this perspective, they presented a family of space-conserving strategies for agglomerative hierarchical clustering based on generalized mean, through which they introduced a new infinite system of agglomerative hierarchical clustering methods that leads to new linkage methods such as geometric linkage or harmonic linkage. The **mdendro** package provides the implementation of this methodology.

Related to Ward's method, different studies have explored its application. Krolak-Schwerdt *et al.* (1991), proposed a regression analytic modification of Ward's method, where within sums of squares are partitioned into the proportion accounted for by the cluster centers and the residual variation. The procedure consists of fusing the two clusters that minimize the residual variation not predicted by the centers. Vijaya *et al.* (2019), and Mongi *et al.* (2019), tested it on different case studies emphasizing the relevant performances of this method compared to other HC approaches. Ogasawara and Kon (2021), pointed out the need of clustering methods employing interval-valued dissimilarity measures and proposed two clustering approaches based on Ward's method to face the problem. Strauss and von Maltitz (2017), have argued that Ward's method cannot be used only with Euclidean distance. In their paper, they generalized the algorithm to use it with the l_1 norm providing proof of the theoretical correctness of this claim. Szekely and Rizzo (2005), have described a new agglomerative hierarchical clustering method that extends the original Ward's method by introducing a cluster distance and objective function that are defined using a power within the range of $(0, 2]$ applied to the Euclidean distance between cluster centers. Ward's method represents the limiting case, namely when the power is set to 2. Lee and Willcox (2014) considered several generalizations of Ward's method and analysed the effect on the clustering when the Minkowski distance has been adopted. Amorim (2015) proposed a new HC algorithm called $Ward_p$. It differs from the original Ward algorithm by introducing feature weights, which can be seen as feature re-scaling factors thanks to the incorporation of the L_p norm. Cluster-specific feature weights enable individual features to exhibit varying levels of importance across different clusters.

3 Methodology

In this section, we focus on the main elements that characterize our novel method. Section 3.1 presents the theoretical foundations of our statements, describing the use of the harmonic mean (Section 3.1.1) and outlining the overall idea of Ward's method (Section 3.1.2). The explication of the "harmonic transformation" of Ward's method and the demonstration of how it is possible to use the classical WSS as a way to compute our new objective function follows (Section 3.2).

3.1 Preliminaries

3.1.1 From generalized mean to harmonic mean

For our purposes, it is convenient to recap the role of generalized mean (Abramowitz and Stegun, 1964). Given a set of n positive numbers x_1, x_2, ..., x_n, and a real number p, the generalized mean M of order p is calculated as:

$$M_p(x_1, x_2, \ldots, x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}. \tag{1}$$

According to the value of p we can obtain different means:

- for $p = 1$ we get the arithmetic mean;
- for $p = 0$ we get the geometric mean;
- for $p = -1$ we get the harmonic mean.

Our primary focus is on the case $p = -1$. The harmonic mean, say $\tilde{x}$, can be defined as the reciprocal of the arithmetic mean of the reciprocals of a set of numbers. Limiting the generalized mean to the case of our interest, (1) becomes:

$$\tilde{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Example of use. The harmonic mean is the correct average to compute when rates are being averaged over a constant numerator (Komić, 2011). Consider a situation in which a person travels at 120 km/h for 10 km and 40 km/h for another 10 km. The actual average speed is given by the harmonic mean, as per Table 1.

Mean	Value
Arithmetic	80
Harmonic	60

Table 1: Comparison between arithmetic and harmonic mean

Indeed in 5 minutes 10 km are covered at 120 km/h, while 15 minutes are required for the same length at 40 km/h. The total time taken to cover a distance of 20 kilometres is 20 minutes, equivalent to an average speed of 60 km/h.

Objective function to be minimized. The use of the harmonic mean leads to the minimization of an objective function which is different from the function minimized by the more common arithmetic mean. In the latter case, the function being considered is the sum of squares (SS) and it is given by (Hogg and Craig, 1978):

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{2}$$

where $\bar{x}$ is the arithmetic mean of the values x_i. According to Berger and Casella, 1992, if we use the harmonic mean the equation to minimize becomes what we call harmonic sum of squares (HSS) and takes the following form:

$$HSS = \sum_{i=1}^{n} \left(\frac{1}{x_i} - \frac{1}{\tilde{x}} \right)^2. \tag{3}$$

Following this reasoning, if we set $y_i = \frac{1}{x_i}$, we can state that in this new domain, (3) is defined as:

$$HSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{4}$$

where $\bar{y}$ is the arithmetic mean of the values y_i.

3.1.2 Ward's method

Instead of focusing on distances directly, Ward's method aims to minimize the increase of the within sum of squares when attempting to combine clusters, where this minimization reflects the cohesion and compactness of clusters. Let k be the number of groups in a level of clustering and $\Omega_k = \{J_1, \ldots, J_k\}$ be a set of groups. From (2) which is defined

in a univariate space, we can define the WSS for a cluster $J \in \Omega_k$ as the sum of squared deviations between the observations $\mathbf{x}_i$ and the centroid $\bar{\mathbf{x}}_J$ as follows:

$$WSS(J) = \sum_{i=1}^{n_J} \|\mathbf{x}_i - \bar{\mathbf{x}}_J\|^2, \tag{5}$$

where $\|\cdot\|$ represents the Euclidean norm. To select the clusters to merge, every possible pair of combinations needs to be considered. Considering two clusters $J_a \in \Omega_k$ and $J_b \in \Omega_k$, the combination that has the lowest merging cost $\Delta WSS(J_a, J_b)$ represents the merge to be done, and it is given by:

$$\Delta WSS(J_a, J_b) = \sum_{i=1}^{n_{J_a \cup J_b}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{J_a \cup J_b}\|^2 - \sum_{i=1}^{n_{J_a}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{J_a}\|^2 - \sum_{i=1}^{n_{J_b}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{J_b}\|^2. \tag{6}$$

where $\bar{\mathbf{x}}_{J_a \cup J_b}$, $\bar{\mathbf{x}}_{J_a}$, $\bar{\mathbf{x}}_{J_b}$ are the centroid of the union of the clusters J_a and J_b, the centroid of J_a, and the centroid of J_b, respectively. With the help of (5), we can rewrite (6) as:

$$\Delta WSS(J_a, J_b) = WSS(J_a \cup J_b) - WSS(J_a) - WSS(J_b).$$

Algorithm 1 shows the steps to perform Ward's method.

Algorithm 1: Ward's method clustering algorithm

Input: Data matrix with n observations

Output: Clustering tree

1 Start with n. of clusters $k = n$, i.e. initialize each data point as its own cluster;

2 **repeat**

3 Calculate the $\binom{k}{2}$ merging costs for each pair of clusters (J_a, J_b), with $J_a, J_b \in \Omega_k$;

4 Find (J_a, J_b) with the lowest merging cost;

5 Combine J_a and J_b into a single new cluster, say $J_a \cup J_b$;

6 **until** *only one cluster remains*;

3.2 Harmonic transformation of Ward's method

Following our claims, for specific cluster analysis it is required to adopt an approach that uses the harmonic mean as the correct average. To align with this statement we make a "harmonic transformation" of our original data, computing the reciprocal. As outlined before, (3) is the function minimized by the HM. Therefore, we can modify (5) and (6) in such a way to establish the concepts of the within harmonic sum of squares

$$WHSS(J) = \sum_{i=1}^{n_J} \left\| {\mathbf{x}_i}^{-1} - \widetilde{\mathbf{x}}_J^{-1} \right\|^2 \tag{7}$$

and the corresponding merging cost function

$$\Delta WHSS(J_a, J_b) = \sum_{i=1}^{n_{J_a \cup J_b}} \|\mathbf{x}_i^{-1} - \tilde{\mathbf{x}}_{J_a \cup J_b}^{-1}\|^2 - \sum_{i=1}^{n_{J_a}} \|\mathbf{x}_i^{-1} - \tilde{\mathbf{x}}_{J_a}^{-1}\|^2 - \sum_{i=1}^{n_{J_b}} \|\mathbf{x}_i^{-1} - \tilde{\mathbf{x}}_{J_b}^{-1}\|^2 \tag{8}$$

that with the help of (7) can be rewritten as

$$\Delta WHSS(J_a, J_b) = WHSS(J_a \cup J_b) - WHSS(J_a) - WHSS(J_b).$$

The next step should be to modify Ward's algorithm 1 to take into account the new centroid (i.e. the harmonic mean vector) and the new merging cost function (8). Advantageously, by using the same reasoning leading from (3) to (4), it is easy to realize that harmonic Ward's method on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is equivalent to implement classical Ward's method on $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where $\mathbf{y}_i = \mathbf{x}_i^{-1}$, $i = 1, \ldots, n$, so letting us to write (7) as

$$WHSS(J) = \sum_{i=1}^{n_J} \|\mathbf{y}_i - \bar{\mathbf{y}}_J\|^2 . \tag{9}$$

To write (7) as in (9) we have used the following result

$$\tilde{\mathbf{x}}_J^{-1} = \bar{\mathbf{y}}_J. \tag{10}$$

The result in (10) can be straightforwardly shown as

$$\begin{aligned} \left(\frac{n}{\sum_{i=1}^{n} \mathbf{x}_i^{-1}} \right)^{-1} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \\ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^{-1} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^{-1}. \end{aligned}$$

So, no "cost" for the implementation of the new agglomerative hierarchical method is required.

4 Real data analysis

To evaluate and validate our methodology, the R language has been employed. The code of the analysis is available at the following link: https://github.com/alesplat/Harmonic-Ward-s-method. To perform the clustering operation we use the `hclust.vector` function of the **fastcluster** package, which allows to receive in input data and perform agglomerative clustering with memory-saving algorithms. To check the optimal number of clusters the functions `NbClust` and `fviz_nbclust` have been used. To compute the confusion matrix and find the accuracy we use the package **caret** with the `confusionMatrix` function. Moreover, a comparison with the linkage method introduced by Fernández and Goemz implemented in the **mdendro** package concludes the experimental tests.

4.1 Data

The analysed dataset has been taken from Kaggle and contains results from the 2019 Ironman World Championship in Kona, Hawaii. It consists of the category to which each athlete belongs with their finish time for the three activities of the championship (swimming, cycling, and running). According to our claims, knowing the length of each path these three variables have been converted to speed in m/s. Once a speed is obtained, we can perform our harmonic clustering strategy since we are now dealing with a space which is constant for each dimension. The four categories considered are "MPRO", "F35-39", "F65-69" and "M70-74", related to the different professional levels and age groups of the athletes. The scatter plot of the data is displayed in Figure 1.

4.2 Analysis

The goal of the analysis is to apply our harmonic Ward's method to evaluate the capability to recognize the categories of our dataset and consequently assess whether the observations are clustered accurately. Indeed, even though we are working on an unsupervised technique we know in advance how many categories we have and the actual classification of

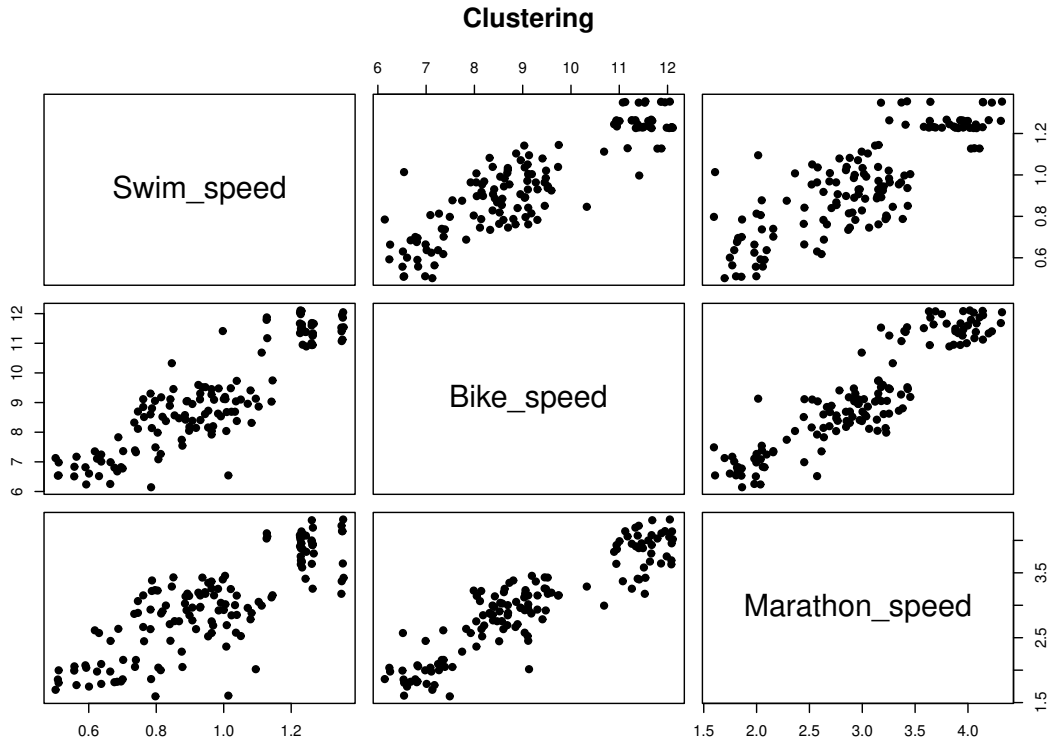

Figure 1: Scatter plot of the data without labels.

observations. This information is used as a reference for our analysis, to precisely quantify the accuracy of the method.

After having performed our HWM, the next step is to decide which is the best number of clusters to group our data. Determining this number is a challenging task since it doesn't follow some specific and always working rules, being an exploratory procedure. The interpretation of the resulting hierarchical structure is context-dependent and often several solutions are equally good from a theoretical point of view. However, there are different measurements that in literature have been presented to evaluate clustering results.

`NbClust` function (contained in the package of the same name) provides multiple indices for determining the relevant number of clusters. Among them, there are methods such as the silhouette, Hartigan, the Dunn index, the C-index, the Dindex, and the Beale index. Table 2 highlights the results obtained. As we can note, there is an *ex aequo* between the values $k = 3, 4, 5$. Therefore, we need an (exogenous) criterion to select such a best

number.

To investigate further, other tools can be evaluated when looking for the optimal number of clusters. One of the most popular is the elbow method (e.g. Pandey and Malviya, 2018, and Humaira and Rasyidah, 2020) which allows you to find the number of clusters for which the variation of the total within sum of squares (TWSS, given by the sum of all the WSS in a level of clustering) for different numbers of clusters starts to decrease. The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve. The elbow point is the number of clusters we can use for our clustering algorithm since choosing a value greater than what the elbow indicates could lead to over-fitting, "explaining" more of the variation. Our HWM focuses on minimizing this TWSS. Figure 2 shows that the best choice is 4 clusters.

Number of clusters (k)	Count
2	4
3	6
4	6
5	6
6	1

Table 2: Number of times each value of k is selected by the internal validation indexes included in the R function `NbClust`.

After all these considerations, although there is no certainty when we have to choose how many clusters we should consider, the most suitable choice seems to be $k = 4$ clusters. As previously affirmed, we actually know the categories we have and consequently, it is possible to confirm that it is the best choice. The dendrogram (Figure 3) shows the representation of hierarchical relationships between data points based on their dissimilarity. With different colours we can highlight the groups that have been identified.

Figure 4 shows the scatter plot of our data with the label associated to the observations by our HWM. After passing this step, with the help of the `confusionMatrix` function we calculate the accuracy in percentage.

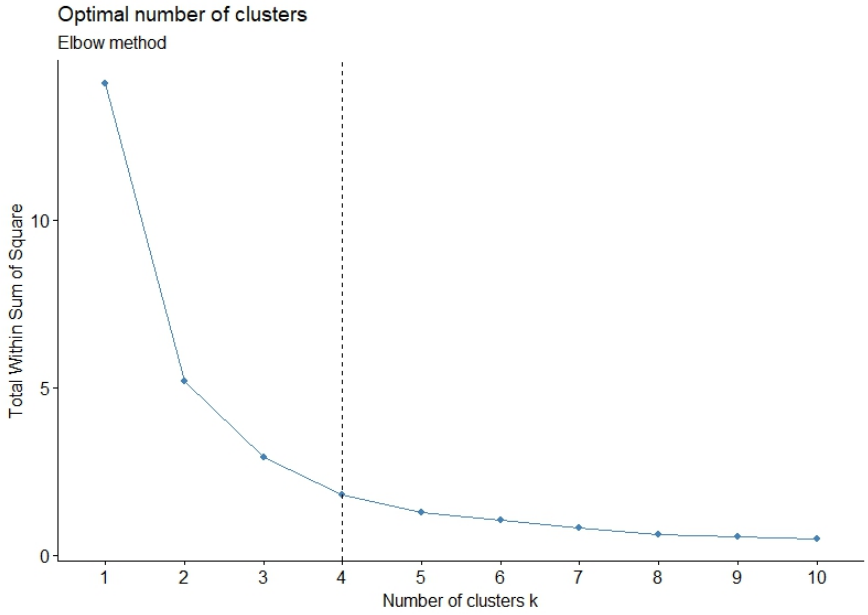

Figure 2: Elbow method.

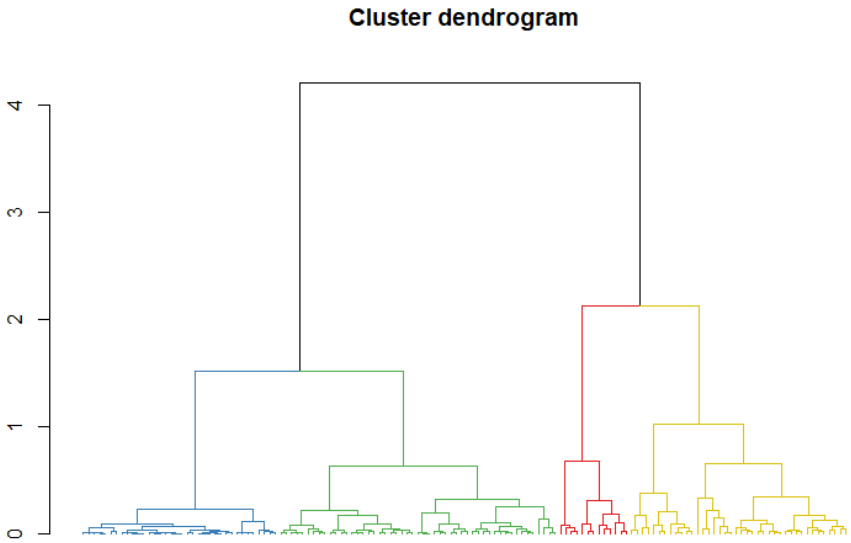

Figure 3: Dendrogram of our dataset with the HWM.

The confusion matrix provides valuable insights into the performance of a classification model. Even though in a real context clustering operations are performed in non-labeled data, having the ground truths allows us to understand with precision where the algorithm

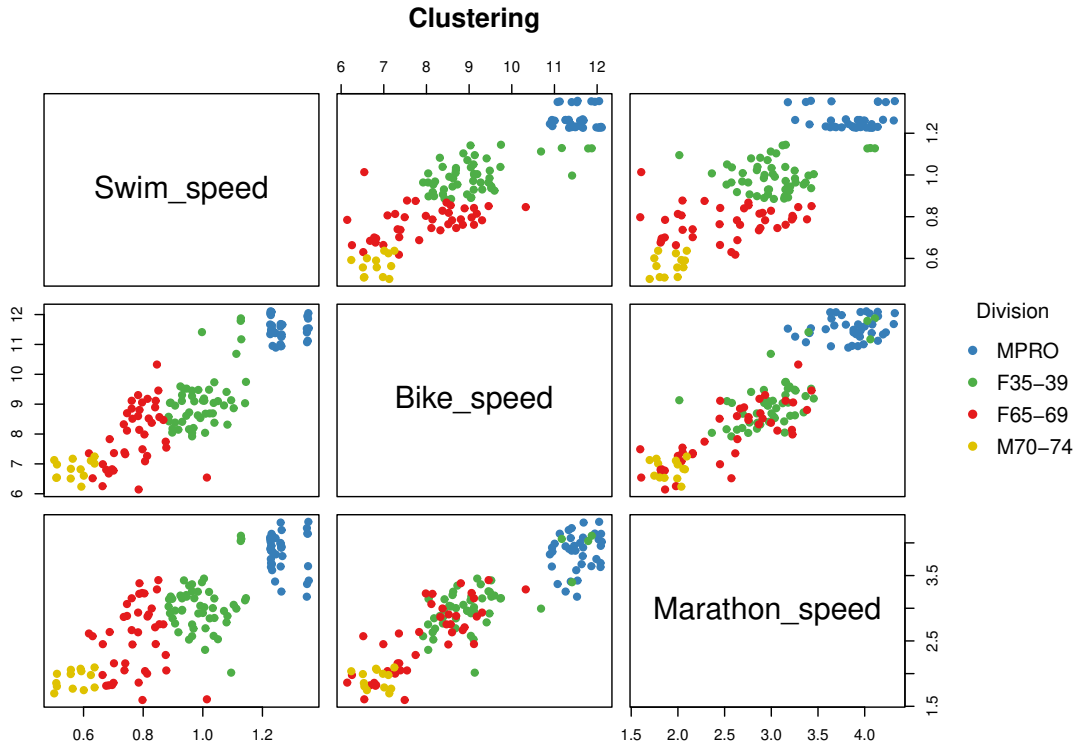

Figure 4: Scatter plot with clusters found by our method. On the graph related to *bike_speed* and *Marathon_speed* can be seen some classification problems.

misclassifies data. Table 3 points out the results we got. In this case, it's evident how the first class ("MPRO") is correctly associated with each observation most of the time, while the second class ("F35-39") and fourth one ("M70-74") are those where the algorithm makes the majority of mistakes. Overall, considering the complexity of the data, with an accuracy of 73.57% we can state that our method has been able to classify the observations with a noticeable level of confidence.

To make a comparison, we decided to test the **mdendro** package with its versatile linkage. Fernandez and Gomez based their agglomerative hierarchical clustering strategies on the use of the generalized mean to present new linkage methods that substitute the classical approaches. The generalized mean contains several well-known particular cases, depending on the value of the power p. When $p = -1$, the generalized mean is equal to the harmonic mean that leads to harmonic linkage (HL). The main idea is based on the

average linkage, where in this case the distance between clusters $d(J_a, J_b)$ is given by the harmonic mean (instead of the arithmetic mean) of the Euclidean distances between each pair of observations $\mathbf{x_i} \in J_a$ and $\mathbf{x_l} \in J_b$, as per formula

$$d(J_a, J_b) = \left(\frac{1}{|J_a| \cdot |J_b|} \sum_{i=1}^{n_{J_a}} \sum_{l=1}^{n_{J_b}} \|\mathbf{x_i} - \mathbf{x_l}\|^{-1} \right)^{-1}. \tag{11}$$

Our test uses this HL on the dataset chosen and verifies its ability to recognize clusters. Table 4 highlights the results obtained.

		Actual			
		MPRO	F35-39	F65-69	M70-74
Predicted	MPRO	36	0	0	0
	F35-39	5	46	0	0
	F65-69	0	21	12	7
	M70-74	0	0	4	9

Table 3: Confusion matrix with HWM.

		Actual			
		MPRO	F35-39	F65-69	M70-74
Predicted	MPRO	40	0	0	0
	F35-39	1	1	0	0
	F65-69	0	66	2	0
	M70-74	0	0	14	16

Table 4: Confusion matrix with versatile linkage.

Compared to our method, versatile linkage performs worse almost for all the categories. "F35-39" is always confused with "F65-69" and also the observations belonging to this category are most of the time misclassified with the category "M70-74". The computed accuracy is 42.14. It is a very low score: the method struggles to identify correctly the real clusters, and in this situation other methods (such as ours) should be preferred.

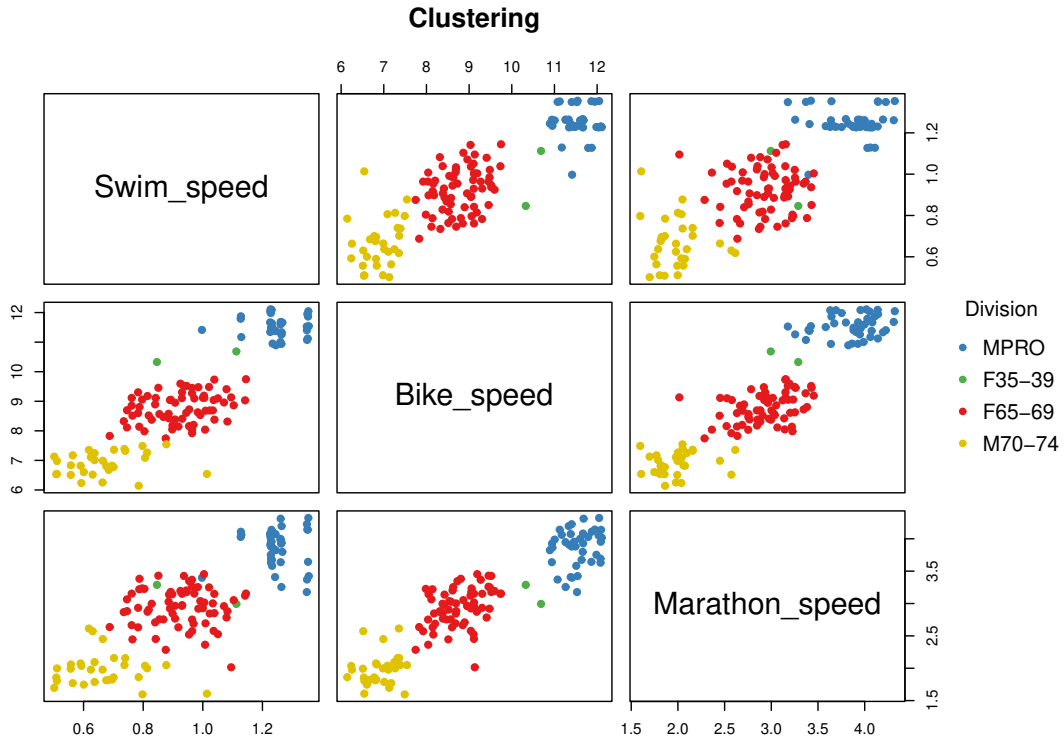

Figure 5: Scatter plot with clusters found by versatile linkage ($p = -1$).

5 Conclusions

Clustering algorithms allow the identification of patterns in data that may be used to gain important insights which at first glance may appear as hidden or at least not easily recognizable. In numerous scenarios involving rates or ratios, the harmonic mean emerges as the sole measure that accurately represents the average. As a consequence, when dealing with such data (like speed data in which the space considered is constant), using clustering methods that work with arithmetic mean is not suitable for specific purposes.

Instead, it becomes evident the need to explore alternative strategies. We showed that applying Ward's method with a harmonic transformation of the data could be seen as a valuable approach in such situations, with significant results. As an open point for further research, an intriguing avenue involves the development of novel clustering strategies tailored to the unique characteristics of the data being analyzed. Each mean

minimizes an objective function; consequently, when we realize that a particular variable needs a particular mean, then we can define new versions of Ward's method that use that mean and its related objective function.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

Amorim, R. (2015). Feature relevance in wardâs hierarchical clustering using the l p norm. *Journal of Classification*, **32**, 46–62.

Ashton, J., Borca, F., Mossotto, E., Phan, H., Ennis, S., and Beattie, R. (2018). Analysis and hierarchical clustering of blood results prior to diagnosis in paediatric inflammatory bowel disease. *Inflammatory Bowel Diseases*, **26**.

Bakker, A. (2003). The early history of average values and implications for education. *Journal of Statistics Education*, **11**.

Berger, R. L. and Casella, G. (1992). Deriving generalized means as least squares and maximum likelihood estimates. *The American Statistician*, **46**(4), 279–282.

Brown, M. (1975). Pappus, plato and the harmonic mean. *Phronesis*, **20**(2), 173–184.

Fernández, A. and Gómez, S. (2019). Versatile linkage: a family of space-conserving strategies for agglomerative hierarchical clustering. *Journal of Classification*, **37**(3), 584–597.

Florek, K., Åukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, **2**, 282–285.

Giordani, P., Ferrero, M. B., and Martella, F. (2020). *An Introduction to Clustering with R*, volume 1. Springer.

Hamdani, M., Aziz, W., Almarashi, M., Khan, I., Nadeem, M., and Habib, N. (2020). A cooperative binary-clustering framework based on majority voting for twitter sentiment analysis. *IEEE Access*, **PP**, 1–1.

Hogg, R. V. and Craig, A. T. (1978). *Introduction to mathematical statistics; 4th ed.* Macmillan, New York, NY.

Huffman, C. (2005). *Archytas of Tarentum: Pythagorean, Philosopher and Mathematician King*. Cambridge University Press.

Humaira, H. and Rasyidah, R. (2020). Determining the appropiate cluster number using elbow method for k-means algorithm. EAI.

Komić, J. (2011). *Harmonic Mean*, pages 622–624. Springer Berlin Heidelberg, Berlin, Heidelberg.

Krolak-Schwerdt, S., Orlik, P., and Kohler, A. (1991). *A Regression Analytic Modification of Wardâs Method: A Contribution to the Relation between Cluster Analysis and Factor Analysis*, pages 23–27.

Lee, A. and Willcox, B. (2014). Minkowski Generalizations of Wardâs Method in Hierarchical Clustering. *Journal of Classification*, **31**(2), 194–218.

Maleki, S. and Bingham, C. (2019). Robust hierarchical clustering for novelty identification in sensor networks: With applications to industrial systems. *Appl. Soft Comput.*, **85**(C).

Mongi, C. E., Langi, Y. A. R., Montolalu, C. E. J. C., and Nainggolan, N. (2019). Comparison of hierarchical clustering methods (case study: data on poverty influence in north sulawesi). *IOP Conference Series: Materials Science and Engineering*, **567**(1), 012048.

Ogasawara, Y. and Kon, M. (2021). Two clustering methods based on the ward's method and dendrograms with interval-valued dissimilarities for interval-valued data. *International Journal of Approximate Reasoning*, **129**, 103–121.

Pandey, A. and Malviya, K. (2018). Enhancing test case reduction by k-means algorithm and elbow method. *International Journal of Computer Sciences and Engineering*, **6**, 299–303.

Patel, S., Sihmar, S., and Jatain, A. (2015). A study of hierarchical clustering algorithms. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 537–541.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ros, F. and Guillaume, S. (2019). A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications*, **128**, 96–108.

Senthilnath, J., pathre balakrishna, S., Rajendra, R., Suresh, S., Kulkarni, S., and Benediktsson, J. (2019). Hierarchical clustering approaches for flood assessment using multi-sensor satellite images. *International Journal of Image and Data Fusion*, **10**, 28–44.

Shi, P., Zhao, Z., Zhong, H., Shen, H., and Ding, L. (2020). An improved agglomerative hierarchical clustering anomaly detection method for scientific data. *Concurrency and Computation: Practice and Experience*, **33**.

Sokal, R., Michener, C., and of Kansas, U. (1958). *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas.

Strauss, T. and von Maltitz, M. (2017). Generalising Ward's method for use with Manhattan distances. *PLoS ONE*, **12**.

Szekely, G. J. and Rizzo, M. L. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification*, **22**(2), 151–183.

Unglert, K., Radić, V., and Jellinek, M. (2016). Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra. *Journal of Volcanology and Geothermal Research*, **320**, 58–74.

Vijaya, Sharma, S., and Batra, N. (2019). Comparative study of single linkage, complete linkage, and Ward method of agglomerative clustering. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 568–573.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Min. Knowl. Discov.*, **10**, 141–168.