Published in Journal of Multiple-Valued Logic and Soft Computing (MVLSC), 2024, vol. 43, p. 73-100. Available at https://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-issue-contents/mvlsc-volume-43-number-1-2-2024/

A New Methodological Proposal for Classifying Firms According to the Similarity of Their Financial Structures Based on Combining Compositional Data with Fuzzy Clustering

Xavier Molas-Colomer[1,*], Salvador Linares-Mustarós[1,*], Maria Àngels Farreras-Noguer[1,*] and Joan Carles Ferrer-Comalat[1,*]

[1]*Department of Business Administration, Faculty of Economics and Management - Campus Montilivi, University of Girona, C/ Universitat de Girona, 10 - 17003, Girona (Spain).*

Accepted: January 27, 2024.

The main aim of this paper is to show that the methodology of classifying firms of the same sector according to the similarity of their classical financial ratios results in serious problems of incoherence and, consequently, it is necessary to find a viable alternative. With this objective in mind, simple examples are used to show that the appearance of multiple unsolvable problems make it impossible to accept the validity of grouping firms according to homogeneous groups based on these ratios using the usual clustering techniques. Once this fundamental fact has been verified, the proposal of classifying firms based on the ratios created in compositional data methodology and fuzzy clustering is shown to offer a viable alternative for this purpose. The paper ends with a complete analysis using real data.

Keywords: Compositional data analysis, CoDa, financial ratio, cluster analysis, fuzzy cluster; classifying firms

* Corresponding author: xavier.molas@udg.edu, salvador.linares@udg.edu, angels.farreras@udg.edu; joancarles.ferrer@udg.edu

1 INTRODUCTION

Due to the growing interest in finding valid methods with which to compare the financial statements of different firms in the same sector, financial ratios, that is, ratios that compare accounting figures in these financial statements, have become a point of attraction for researchers and professionals in economic and financial analysis, since they are more focused on relative rather than absolute account magnitudes. Through this comparative analysis, it is possible to infer diagnoses of the financial health of firms in the sector, as well as make more correct decisions with respect to production, sales and export processes, thus enabling the different companies to improve their competitiveness in the sector.

The use of financial ratios has become widespread in many lines of research, including sectorial comparisons [1], determining firm's value [2], firm survival analysis [3], credit scoring [4], assessing the impact of International Financial Reporting Standards [5], prediction of donations to charities [6], accounting restatements [7], and earnings management [8].

However, this paper will focus on another method of statistical data analysis in which the use of financial ratios has become more frequent: cluster analysis, which consists of grouping a set of data points in such a way that points in the same group share more similarities with each other than with those in other groups (see, e.g., [9], [10] and [11]). The purpose of this type of analysis is to be able to classify firms according to the similarity of the structure of their financial statements, searching for different profiles of financial structure, performance or distress.

Unfortunately, the results of these analyses may not be valid, since the literature has reported a number of practical drawbacks in their use as a consequence of their own mathematical construction when employed as variables in classical statistical analyses of the financial health of firms in a given sector.

The first of these drawbacks is related to the fact that most ratios are distributed between zero and infinity, which does not allow for perfectly symmetric data distributions; likewise, ratios tend to have asymmetric distributions because a decrease in denominator produces a greater change in the ratio value than an increase would do [12]. These situations lead to the problem of precluding the use of symmetric probability distributions such as the normal distribution [13], [14], [15] and [16]. Although loss of symmetry in financial ratios is a problem that has been identified for some time [15], [17], [18], and [19], it has not received enough attention in the accounting field to be considered a significant problem. Such asymmetry is also connected to the emergence of spurious outliers in data distributions [14], [17], [18], [20], [21] and [22]; and it can even be the case that such outliers are the main or only source of positive asymmetry in the distributions. However, these outliers do not always reflect atypical management practices, but may arise from having a

small value in the denominator of their corresponding ratios. In the particular case of cluster analysis, asymmetric distributions lead to some clusters being very small [16], [23],and [24], while it is also well known that the presence of such outliers distorts the results of many clustering algorithms and sometimes leads to the appearance of single-member clusters [25], [26] and [27].

The second most important drawback lies in the non-preservation of Euclidean distances among firms in the analysis due to different reasons: on the one hand, choosing a different set of ratios may result in different distances among firms, even if the ratios are computed from exactly the same set of accounting figures; and on the other hand, Euclidean distance is not an appropriate dissimilarity measure for ratios, since even the permutation of accounts between numerator and denominator of the same ratio matters when it comes to Euclidean distance [12], [16] and [22].

In recent years, a line of research investigating a solution to these problems has been developed through the use of a compositional data approach [28], [29]. The validity of this methodology's results has already been widely tested in various fields [30], [31], [32], [33] and [34] and it has been shown to be effective in minimizing or eliminating such problems in the accounting domain [10], [16], [19], [22], [35] and [36]. Although this line of research based on the study of accounting data has already begun to work on cluster analysis (see [37], [38], [39], and [40]), it has never until now sought to converge towards cluster studies with compositional data that have appeared in the main line of fuzzy cluster analysis with compositional data (see [41], [42], [43], and [44]). The present work is the first work that addresses this convergence.

Consequently, after exposing the inadequacy of the use of classical financial ratios in clustering studies on economic sectors, this paper aims to show the possible use of financial ratios based on compositional data analysis (CoDa) methodology, with the ultimate goal of highlighting the possibilities this methodology offers in soft clustering studies. Moreover, different hypothetical and real cluster analyses will also be presented to prove the possibilities of the R-Studio software package for both classical and fuzzy classification, allowing appropriate conclusions to be drawn regarding their comparison.

The paper is organized as follows: Section 2 reviews the basics of the c-means fuzzy clustering by means of a simple example of data points plotted with Excel to demonstrate how R-Studio returns the expected outcome clusters. Section 3 presents another set of data points to focus on the serious problems arising from the use of standard ratios in clustering, but this time one which has been purpose-built to show the emergence of loss of symmetry and spurious outliers in data distribution by clusters. In Section 4, financial ratios based on the CoDa methodology are presented as possible candidates to replace standard financial ratios, showing how this solves the exposed drawbacks when applied to the example given in Section 3. Section 5 shows

a real example to demonstrate the full potential of the new methodology. Finally, Section 6 presents the conclusions.

2 WHY TO USE FUZZY CLUSTERING

Clustering is the task of grouping a set of data points in such a way that points in the same group share more similarities with each other than with those in other groups. It is the most common technique in statistical data analysis and is used in many research fields. Fuzzy-clustering (also referred to as Soft Clustering) is a family of clustering techniques in which each element has a fuzzy degree of group membership, i.e. belongs to two or more clusters. Clusters are identified using similarity measures, which include those relating to connectivity, intensity and distance.

These fuzzy-clustering algorithms arise from the need to solve a deficiency in exclusive clustering (also known as Hard Clustering), which considers that each element can be unambiguously grouped with the elements of its cluster and therefore does not resemble the other elements in the set. A solution to this problem emerged with the introduction of fuzzy logic [45], determining the degree of similarity or membership for each element to each one of the clusters formed. This is achieved by representing the similarity between an element and a group through a function, called the membership function, which takes values between zero and one. Values close to one indicate higher similarity, while those close to zero indicate lower similarity. For example, an apple can be red or green (Hard Clustering), but an apple can also be partially red and partially green (Soft Clustering). Thus, the apple can be both red and green to a certain degree. Instead of the apple belonging to green (green = 1) and not red (red = 0), the apple can belong to green (green = 0.5) and red (red = 0.5). These values are normalized between 0 and 1; however, they are not necessarily values of a probability measure, so the two values do not always need to add 1. Therefore, the problem of fuzzy clustering boils down to finding just such an optimal characterization for the set data points.

One of the most widely used fuzzy-clustering algorithms is the Fuzzy C-means clustering algorithm [46, 47]. With this data clustering technique, the set of data points are grouped into a few number clusters where every data point belongs to each cluster with a certain degree of group membership. More specifically, those data points that lie close to the center of a cluster will have awarded with a high degree of membership to that cluster, and other data points that lie far away from the center will have a low degree of membership to it.

To find an optimal distribution in the clusters, the algorithm follows the steps outlined below:

Step 1: Initialize the data points into desired number of clusters, with the initial randomly assigned value to all fuzzy membership coefficients of each data point.

Step 2: Determine the centroid of each cluster.

Step 3: Determine the distance of each point from each one of the centroids using Euclidean distance.

Step 4: Update the membership values.

Step 5: Repeat Steps 2-4, unless centroids do not change.

Any point *xi* has a set of fuzzy coefficients assigned to it, giving the degree of being in the *j*-th cluster w_{ij}. With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster, or:

$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^{p} \cdot x_i}{\sum_{i=1}^{n} w_{ij}^{p}} \tag{1}$$

where *p* is the parameter that controls how fuzzy the cluster will be (generally taken as 2). The higher it is, the fuzzier the cluster will ultimately be. Then, for Step 4, the membership values can be updated using the following formula,

$$w_{ij} = \left(\sum_{s=1}^{K} \left[\frac{d_{ij}^2}{d_{is}^2} \right]^{\frac{1}{p-1}} \right)^{-1} \tag{2}$$

where *K* is the desired number of clusters and *d*'s are the distances calculated in Step 3.

A brief example of how this algorithm works is provided below, grouping the fictitious set of data points given in Table 1 and represented in Figure 1 into two clusters.

Point	$\mathbf{x_1}$	$\mathbf{x_2}$
1	0	2
2	1	1
3	1	3
4	5	2
5	9	1
6	9	3
7	10	2

TABLE 1
First example of data points

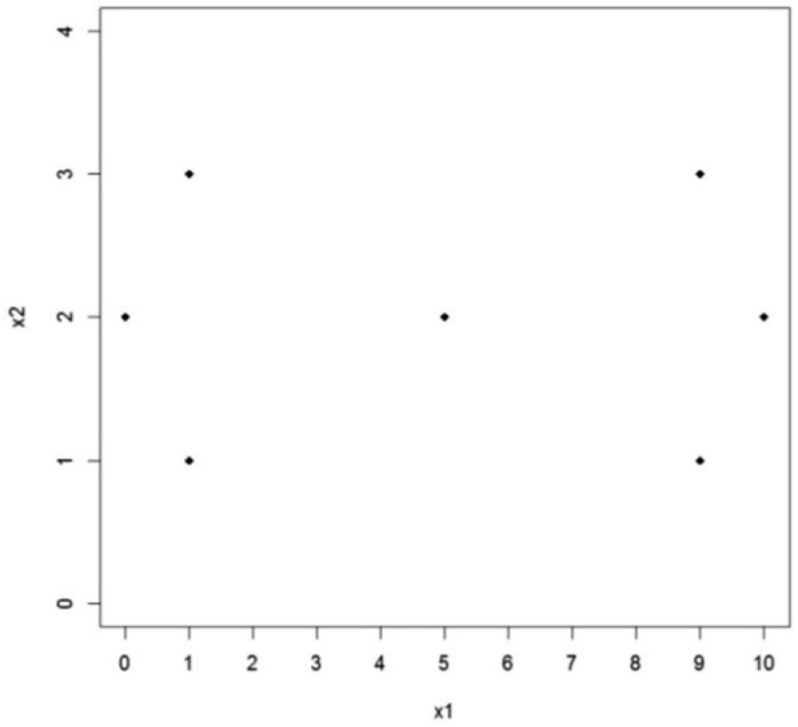

FIGURE 1
Screen plot of data points in Table 1.

The code lines in Box 1 can be used to obtain said fuzzy grouping into two clusters, calculated using the R software and the graphical display of the points.

Values offered by the program for C2 and FC2 can be found in Box 2 and Box 3. These values allow us to verify that the degrees of assignment to the hard clusters in the fuzzy approximation are coincident with the assignment using the classical method.

Figure 2 shows classical clustering into two groups according to object classification C2. Note the loss of symmetry and the arbitrariness of grouping Point 4 in one of the two clusters. In contrast, Figure 3 shows fuzzy grouping into two clusters according to object classification FC2. Here, it is evident that Point 4 belongs equally to both clusters. In Box 3, it is interesting to note that Object 4 belongs with a degree 0.5 to each of two clusters, while in the hard cluster C2 approximation it belongs exclusively to the second cluster. Let us note, then, that the deliberate symmetry of the data results in a perfect symmetry when creating the two clusters by means of the fuzzy logic assumption. This symmetry is perfectly consistent with the membership values of the two fuzzy clusters at the symmetric points with respect to Object 4.

```
# Enter vector x1
>x1=c(0,1,1,5,9,9,10)
# Enter vector x2
> x2=c(2,1,3,2,1,3,2)
# Represent the seven points that make up the elements to be grouped on a
coordinate axis
> plot(x1,x2,ylim=c(0,4),pch=18,xaxp=c(0,10,10))
# A data frame object named "DT" is created, whose rows are the points to
be grouped
> DT <- data.frame(x=x1,y=x2)
# Perform k-means clustering function on a data matrix named “C2”,
which creates the two classic clusters
> C2<-kmeans(DT, 2, iter.max=200, nstart=100)
# Ask for information about the C2 object we have created
> C2
# The library containing the k-means type fuzzy cluster creation function is
installed
> install.packages("fclust")
# Load the library into the program
> library(fclust)
# Perform k-means fuzzy clustering on a data matrix, named “FC2”, to
create the two fuzzy clusters
> FC2=FKM(DT,k=2, RS=25, seed=123)
# Ask for information about the FC2 object we have created
> FC2
```

BOX 1
First Code.

```
Cluster means:
        x       y
1 8.2500000 2
2 0.6666667 2

Clustering vector:
[1] 1 1 1 2 2 2 2
```

BOX 2
Results for C2.

Thus, the use of fuzzy clustering is justified, as it allows a more coherent grouping in a way that is more consistent with human reasoning, because the classical classification methods misses that sometimes there are elements

Number of objects: 7
Number of clusters: 2

Closest hard-clustering partition:

Obj1	Obj 2	Obj 3	Obj 4	Obj 5	Obj 6	Obj 7
1	1	1	2	2	2	2

Membership degree matrix (rounded):

	Clus 1	Clus 2
Obj 1	0.99	0.01
Obj 2	0.98	0.02
Obj 3	0.98	0.02
Obj 4	0.50	0.50
Obj 5	0.02	0.98
Obj 6	0.02	0.98
Obj 7	0.01	0.99

BOX 3
Results for FC2

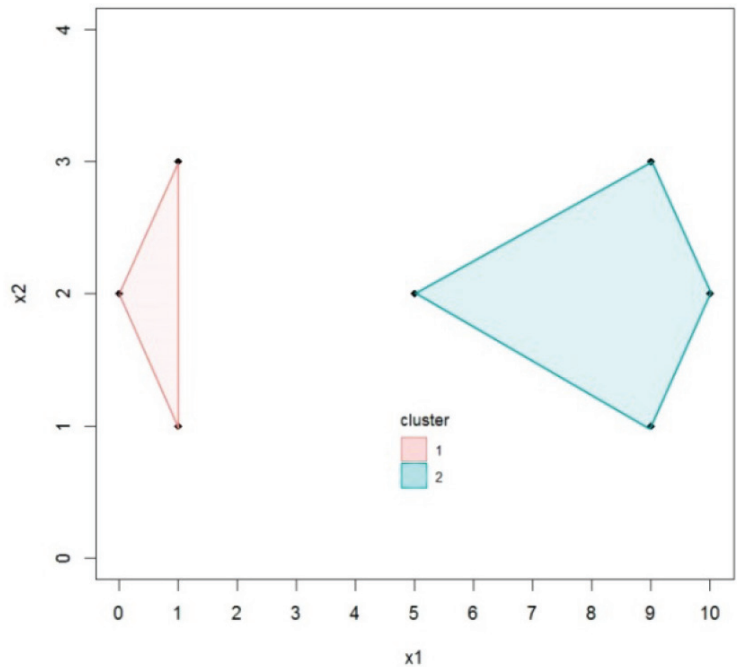

FIGURE 2
Screen plot of data points in two classical clusters

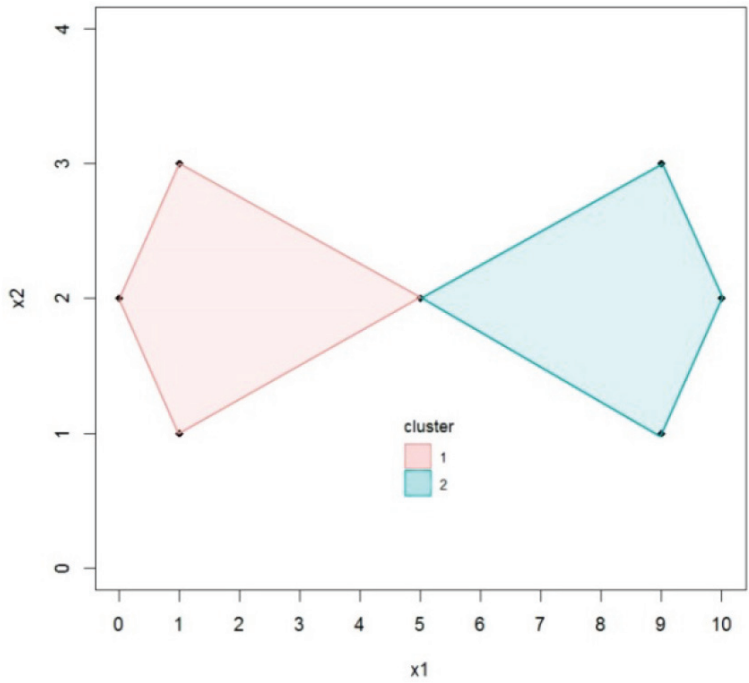

FIGURE 3
Screen plot of data points in two fuzzy clusters.

belonging simultaneously to two groups that are unjustifiable assigned to one group or the other by a classical computational algorithm.

3 WHY NOT TO USE FUZZY CLUSTERING WITH CLASSICAL RATIOS

This section focuses on highlighting the reason why making classifications using classical financial ratios as variables in sectorial statistical analysis should be questioned and perhaps not be considered reliable. In order to understand the problems that arise, we propose as an example the new fictitious set of data points provided in Table 2. Values for these points are not chosen randomly, but purposely, so that they show perfect symmetry when plotted, as reflected in Figure 4.

As can be seen from their arrangement in the graph, the points are expected to be clearly grouped into 3 clusters, with a high degree of membership to their own cluster and a low degree of membership to the others.

If we enter the new data into R-Studio for a three-cluster analysis, as Box 4 shows, for the first three points the returned membership coefficients

Point	x_1	x_2	x_1/x_2	x_2/x_1
1	0,3	8,7	0,03448276	29
2	1,3	9,7	0,13402062	7,46153846
3	1,3	8,7	0,14942529	6,69230769
4	4,5	5,5	0,81818182	1,22222222
5	5	5	1	1
6	5,5	4,5	1,22222222	0,81818182
7	8,7	1,3	6,69230769	0,14942529
8	9,7	1,3	7,46153846	0,13402062
9	8,7	0,3	29	0,03448276

TABLE 2
Second example of data points

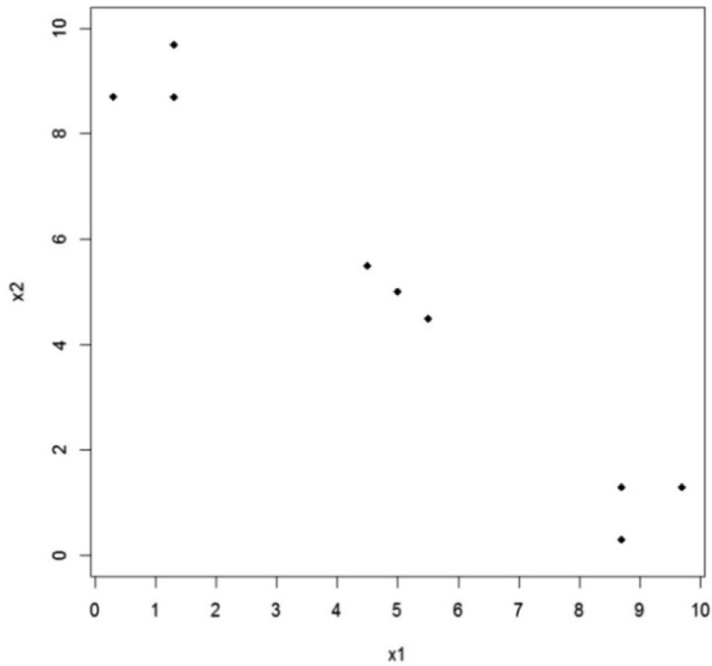

FIGURE 4
Screen plot of data points in Table 2.

display higher values closer to one in the first column, and lower values closer to zero in the other two columns; for the three next points, values closer to one appear in the second column and values closer to zero are

Fuzzy-clustering object of class 'fclust'
Number of objects: 9
Number of clusters: 3
Clustering index values: SIL.F k=3 0.9610499
Closest hard clustering partition:

Obj 1	Obj 2	Obj 3	Obj 4	Obj 5	Obj 6	Obj 7	Obj 8	Obj 9
1	1	1	2	2	2	3	3	3

Membership degree matrix (rounded):

	Clus 1	Clus 2	Clus 3
Obj 1	0.98	0.02	0.00
Obj 2	0.98	0.02	0.00
Obj 3	0.99	0.01	0.00
Obj 4	0.02	0.97	0.01
Obj 5	0.00	1.00	0.00
Obj 6	0.01	0.97	0.02
Obj 7	0.00	0.01	0.99
Obj 8	0.00	0.02	0.98
Obj 9	0.00	0.02	0.98

Note: The fuzzy clusters are been created from x1 and x2 data using the R instructions:

```
> x1=c(0.3,1.3,1.3,4.5,5,5.5,8.7,9.7,8.7)
> x2=c(8.7,9.7,8.7,5.5,5,4.5,1.3,1.3,0.3)
> DT <- data.frame(x=x1,y=x2)
> fC3a=FKM(DT,k=3)
> fC3a
```

BOX 4
Fuzzy clusters created from x1 and x2 data

obtained in the first and third columns, respectively; and finally, for the last three points, values closer to one are obtained in the third column and values closer to zero in the other two columns. This set of results leads, as expected, to the idea that the first three points would form a first cluster, the next three would belong to the second cluster, and the last three would be grouped in the third cluster, as shown in Figure 5.

Let us now return to the main aim of this work and observe what happens when we calculate the ratio x/y while working with ratios in cluster analysis. Entering the new values in R-Studio for a three-cluster analysis as seen in Figure 5, although the returned membership coefficients do not clarify as much as the previous ones, taking a closer look at the subtle differences between the values in Box 5 we see how the first six points show higher

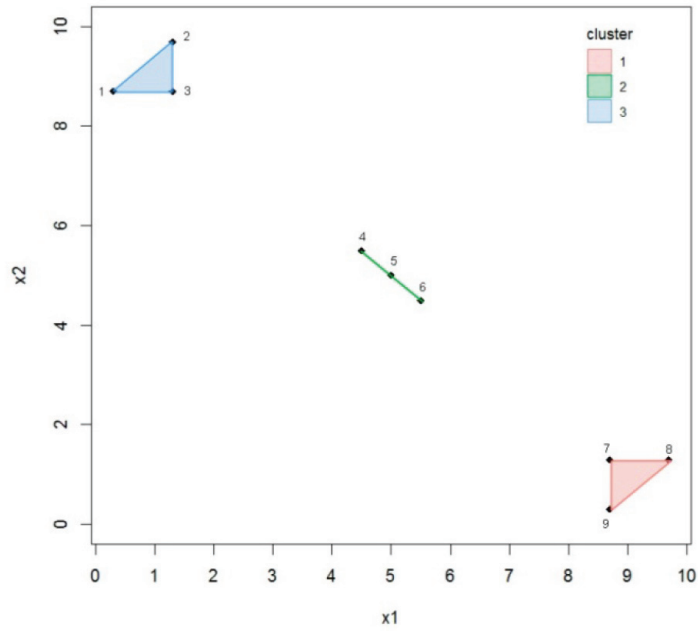

FIGURE 5
Interpretation of a three-cluster group from membership values obtained with R-Studio software.

similar values in the first column, and lower values much closer to zero in the other two columns; for the seventh and eighth points, higher similar values appear in the second column and values much closer to zero are obtained in the first and third columns, respectively; and finally, for the last point, the most astonishing results were obtained with only a value so much closer to one than any other in the third column, and values so much closer to zero than any other in the other two columns.

This set of results may give rise to the idea that the first six points would form a first cluster, the seventh and eighth points would belong to a second cluster, but closer to the first one than expected, and the last point constitutes a third cluster in itself, as shown in the interpretation given in Figure 6.

However, this classification has nothing to do with the three-cluster group expected for that set of data points, as Figure 5 shows. In this case, an asymmetric distribution is presented for this cluster analysis, and the main source of this asymmetry is the emergence of the last value as a possible outlier of the distribution. If we were to rely on the certainty of this interpretation of results, we might perhaps end up eliminating the last point in order to obtain a more accurate classification and repeat the analysis without it. But this makes no sense, since Figure 5 shows that the ninth

```
Fuzzy-clustering object of class 'fclust'
Number of objects: 9
Number of clusters: 3
Clustering index values: SIL.F k=3 0.7919329
Closest hard clustering partition:
Obj 1  Obj 2  Obj 3  Obj 4  Obj 5  Obj 6
  1      1      1      1      1      1
Obj 7  Obj 8  Obj 9
  2      2      3
 Membership degree matrix (rounded):
          Clus1  Clus2  Clus3
Obj 1      0.93   0.06   0.01
Obj 2      0.97   0.03   0.00
Obj 3      1.00   0.00   0.00
Obj 4      0.99   0.01   0.00
Obj 5      0.93   0.06   0.00
Obj 6      0.82   0.17   0.01
Obj 7      0.01   0.99   0.00
Obj 8      0.01   0.99   0.00
Obj 9      0.00   0.00   1.00

Note: The fuzzy clusters are been created from x1/x2 values using the R
instructions:
> r=x1/x2
> Point=c(1,2,3,4,5,6,7,8,9)
> DT <- data.frame(x=Point,y=r)
> fC3b=FKM(DT,k=3)
> fC3b
```

BOX 5
Fuzzy clusters created from x1/x2

point has to be as close to the seventh and eighth points as the first point is to the second and third in order to maintain the symmetric behavior displayed in this example.

Finally, we will demonstrate that it makes no sense to believe that using classical ratios for analysis allows firms to be classified according to the similarity of their financial structures. Note that if we had performed the analysis using the ratio x_2/x_1 instead of the ratio x_1/x_2, then even if the firms had shared the same financial structure, we would have obtained a completely different classification of three groups, with a single outlier different from the one

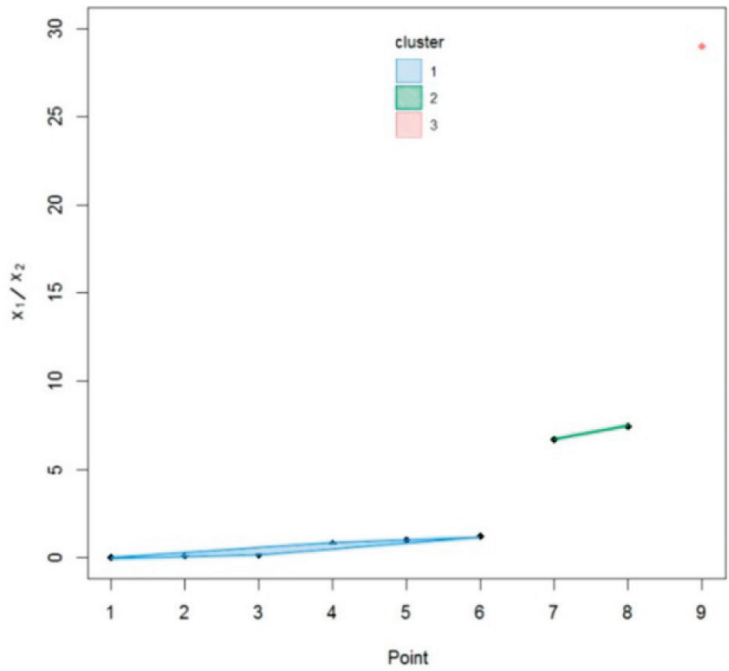

FIGURE 6
First interpretation of a three-cluster group from membership values.

obtained previously. This is confirmed by the data in Box 6 and its clustering representation in Figure 7.

It makes no sense that by simply permuting two values of the firms' financial structure, groupings into clusters would differ completely. For this reason it is therefore recommended that the log-ratios from Compositional Data (CoDa) methodology be used to calculate the financial ratios in order to avoid the asymmetry issues mentioned above. The following section will show that these new ratios avoid or minimize the asymmetry problems presented here.

4 WHY USING FUZZY CLUSTERING WITH LOG-RATIOS FROM COMPOSITIONAL DATA ANALYSIS METHODOLOGY PROVIDES A BETTER SOLUTION

For illustrative purposes, we present a simplified balance sheet (Table 3), which is sufficient to compute the most common *debt and liquidity ratios* [16].

These magnitudes have been selected because, together with operating costs and sales (x_7, x_8), they constitute the basis for a wide array of common financial and management ratios frequently used in accounting. Table 4

Fuzzy-clustering object of class 'fclust'
Number of objects: 9
Number of clusters: 3
Clustering index values: SIL.F k=3 0.7919329
Closest hard clustering partition:

Obj 1	Obj 2	Obj 3	Obj 4	Obj 5	Obj 6	Obj 7	Obj 8	Obj 9
1	2	2	3	3	3	3	3	3

Membership degree matrix (rounded):

	Clus 1	Clus 2	Clus 3
Obj 1	1.00	0.00	0.00
Obj 2	0.00	0.99	0.01
Obj 3	0.00	0.99	0.01
Obj 4	0.01	0.17	0.82
Obj 5	0.00	0.06	0.93
Obj 6	0.00	0.01	0.99
Obj 7	0.00	0.00	1.00
Obj 8	0.00	0.03	0.97
Obj 9	0.01	0.06	0.93

Note: The fuzzy clusters are been created from x2/x1 values using the R instructions:

```
> r=x2/x1
> Point=c(1,2,3,4,5,6,7,8,9)
> DT <- data.frame(x=Point,y=r)
> fC3c=FKM(DT,k=3)
> fC3c
```

BOX 6
Fuzzy clusters created from x2/x1

shows a range of examples of relevant financial ratios that might be computed from x_1 to x_8. This list is not exhaustive, since it could incorporate any possible ratio computed from x_1 to x_8. Thus, it is provided merely as an example in order to show which type of information is carried by x_1 to x_8.

Logarithms of ratios, or simply log-ratios, are the standard transformation commonly used in Compositional Data [32]. Several choices are possible to compute log-ratios. In all cases, they involve a logarithm of a ratio among simple components or among geometric means of components. A scaling constant multiplying the log-ratio may or may not be included. Reformulating financial ratios as log-ratios solves some problems of asymmetry, and

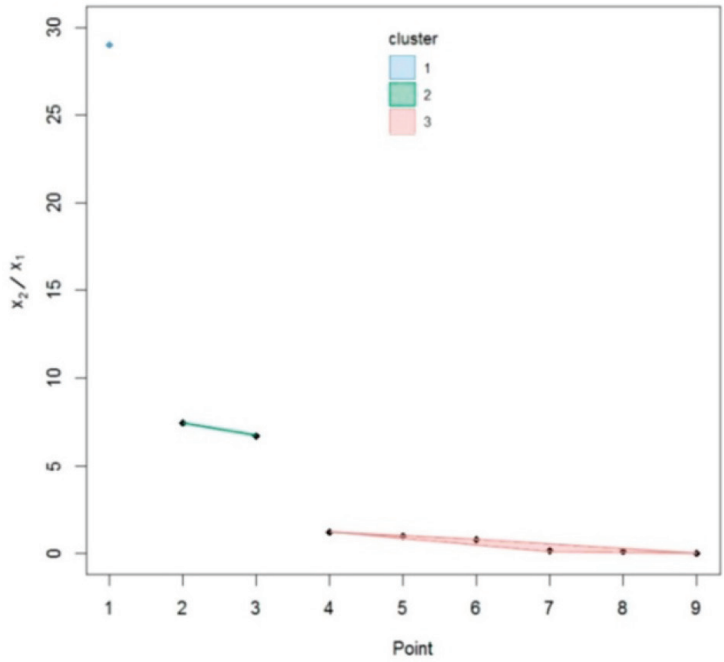

FIGURE 7
Second interpretation of a three-cluster group from membership values.

Assets	Liabilities and equity
x_1 = Fixed assets	x_4 = Equity
x_2 = Inventory	x_5 = Long term debt
x_3 = Quick assets	x_6 = Short term debt

TABLE 3
Simplified balance sheet

also tends to solve problems related to outliers. For ratios involving only two components, like the *acid test ratio*, this can be done in the following straightforward manner:

$$acid\ test \log\_ratio = log\left(\frac{x_3}{x_6}\right) \tag{3}$$

Positive values mean that quick assets are higher than short term debt. Negative values mean the opposite. A zero log-ratio implies equality in both magnitudes,

Name of the ratio	Formal expression
acid test ratio	x_3/ x_6
working capital ratio	$(x_2 + x_3)/x_6$
liability to asset ratio	$(x_5 + x_6)/(x_4 + x_5 + x_6)$
equity to debt ratio	$(x_4)/(x_5 + x_6)$
equity to long term debt ratio	x_4/x_5
current liability to asset ratio	$x_6/(x_4 + x_5 + x_6)$
fixed asset to equity ratio	x_1/ x_4
long term debt to asset ratio	$x_5/(x_4 + x_5 + x_6)$
turnover	$x_8/(x_1+x_2+x_3)$
margin	$(x_8-x_7)/x_8$

TABLE 4
A sample financial ratio list that might be computed from x_1 to x_8

exactly in the same way as unit standard ratios, although in classical ratios the comparison value between the denominator and the numerator is one instead of zero. In this context, symmetry means both that the log-ratio range is from minus infinity to plus infinity, and that permuting the numerator and denominator components leads to the same distance from zero, with only a change in the sign:

$$log\left(\frac{x_3}{x_6}\right) = -log\left(\frac{x_6}{x_3}\right) \tag{4}$$

Furthermore, if one of the components being compared is close to zero, it may lead to an outlying standard ratio when placed in the denominator and to a typical ratio when placed in the numerator. Placement makes no difference over permutation for log-ratios.

We are now going to use this type of transformation in the example we introduced in Section 3, which only requires taking log values of ratios x_1/x_2. A glance at the graph in Figure 8 allows us to observe as the points are repositioned if we change the numerator with de denominator in the log-ratio taking now positive and also negative values in a symmetric way attending the distance from zero.

Entering the new data in R-Studio for a three-cluster analysis, Box 7 shows that the returned membership coefficients for the first three points display higher values closer to one in the first column and lower values closer to zero in the other two columns; for the three next points, values closer to one appear in the second

Fuzzy-clustering object of class 'fclust'
Number of objects: 9
Number of clusters: 3
Clustering index values: SIL.F k=3 0.7891853
Closest hard clustering partition:

Obj 1	Obj 2	Obj 3	Obj 4	Obj 5	Obj 6	Obj 7	Obj 8	Obj 9
1	1	1	2	2	2	3	3	3

Membership degree matrix (rounded):

	Clus 1	Clus 2	Clus 3
Obj 1	0.92	0.06	0.02
Obj 2	0.98	0.01	0.01
Obj 3	0.82	0.15	0.03
Obj 4	0.10	0.86	0.04
Obj 5	0.00	1.00	0.00
Obj 6	0.04	0.86	0.10
Obj 7	0.03	0.15	0.82
Obj 8	0.01	0.01	0.98
Obj 9	0.02	0.06	0.92

Note: The fuzzy clusters are been created from log(x1/x2) values using the R instructions:

```
> r=log(x1/x2)
> Point=c(1,2,3,4,5,6,7,8,9)
> DT <- data.frame(x=Point,y=r)
> fC3c=FKM(DT,k=3)
> fC3c
```

BOX 7
Fuzzy clusters created from log(x1/x2)

column and values closer to zero are obtained in the first and third columns, respectively; and finally, for the last three points, values closer to one are obtained in the third column and values closer to zero in the other two columns.

This set of results gives rise to the conclusion that the first three points would form a first cluster, the next three would belong to the second cluster and the last three would be grouped in the third cluster, as shown in Figure 8. Thus, the initial grouping into three clusters obtained for this set of points at the beginning of the example in Section 3 is recovered. Furthermore, for this same set of data points, there is no longer any spurious outlier, as was obtained when analyzing only the ratio. Similarly, as shown by the data in Box 8, which have been interpreted in Figure 9, the fuzzy-clustering analysis in R of points obtained from the log (x_2/x_1) values of Section 3 gives us the same groupings that were produced in the cluster analysis of the log (x_1/x_2) values. As expected, the sym-

Fuzzy-clustering object of class 'fclust'
Number of objects: 9
Number of clusters: 3
Clustering index values: SIL.F k=3 0.7891853
Closest hard clustering partition:

Obj 1	Obj 2	Obj 3	Obj 4	Obj 5	Obj 6	Obj 7	Obj 8	Obj 9
1	1	1	2	2	2	3	3	3

Membership degree matrix (rounded):

	Clus 1	Clus 2	Clus 3
Obj 1	0.92	0.06	0.02
Obj 2	0.98	0.01	0.01
Obj 3	0.82	0.15	0.03
Obj 4	0.10	0.86	0.04
Obj 5	0.00	1.00	0.00
Obj 6	0.04	0.86	0.10
Obj 7	0.03	0.15	0.82
Obj 8	0.01	0.01	0.98
Obj 9	0.02	0.06	0.92

Note: The fuzzy clusters are been created from log(x2/x1) values using the R instructions:

```
> r=log(x2/x1)
> Point=c(1,2,3,4,5,6,7,8,9)
> DT <- data.frame(x=Point,y=r)
> fC3c=FKM(DT,k=3)
> fC3c
```

BOX 8
Fuzzy clusters created from log(x2/x1)

metry with respect to the mass center of the initial data structure is maintained when the numerator and denominator values are exchanged.

5 A COMPLETE FUZZY-CLUSTER STUDY USING CODA METHODOLOGY WITH REAL DATA

In this section, we will show the construction of a complete fuzzy-cluster study using CoDa methodology. The study analyzes the same data used by Linares-Mustarós, Coenders and Vives-Mestres [16], while using the same CoDa components for the study. This will allow us to perform a comparative study based on the results obtained from applying the classical clustering methodology and the fuzzy clustering methodology developed in the present work.

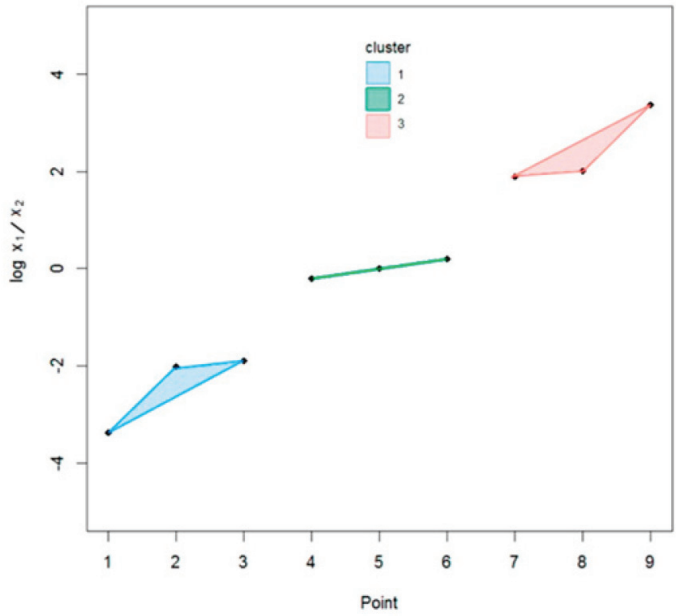

FIGURE 8
Interpretation of a three-cluster group from membership values taking a log-ratio $\log(x_1/x_2)$.

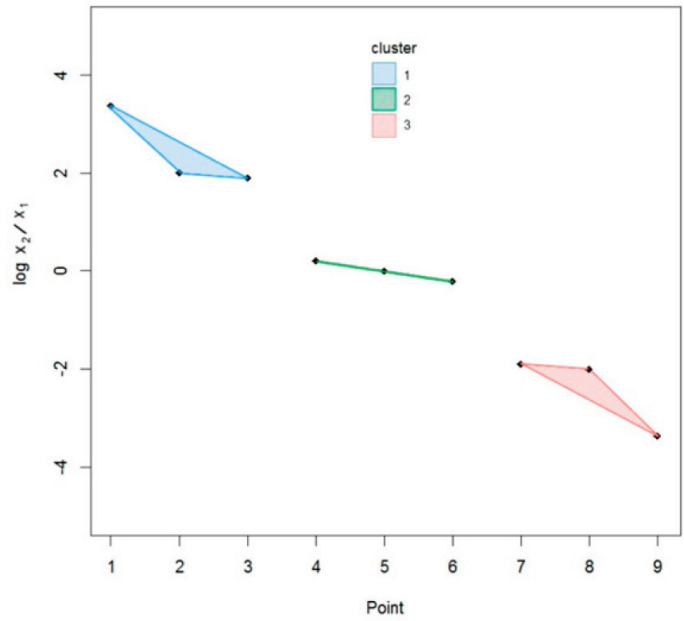

FIGURE 9
Interpretation of a three-cluster group from membership values taking a log-ratio $\log(x_2/x_1)$.

The starting log-ratios constructed as isometric log-ratio coordinates (ilr coordinates) are the log-ratios presented in Table 5, which are interpreted as asset structure (y_1, y_2), liability and equity structure (y_3, y_4), asset, liability and equity structure (y_5), turnover (y_6) and margin (y_7). See [16] for details.

In order to obtain the different classical clusters grouping the data for the variables y_i, the R code lines are written in Box 9.

And the R code lines for obtaining the respective degrees of membership to four fuzzy clusters starting from the DM and Y matrices are in Box 10.

Table 6 presents the results obtained for the first elements of the initial data matrix.

With the aim of demonstrating the potential of working with fuzzy clusters, we will illustrate the presence of elements that exist between these clusters. For instance, Table 7 showcases some objects with relatively high membership in at least two of the clusters. This indicates that a soft classification for these companies is more meaningful than a hard classification.

Finally, it is worth mentioning that, as R provides the means to obtain cluster centers, one can recover the corresponding xi values from the pro-

Name of the log-ratio	**Formal expression**
asset structure 1	$y_1 = \sqrt{\frac{2}{3}} log\left(\frac{\left(x_1 \cdot x_2\right)^{1/2}}{x_3}\right)$
asset structure 2	$y_2 = \sqrt{\frac{1}{2}} log\left(\frac{x_1}{x_2}\right)$
liability and equity structure 1	$y_3 = \sqrt{\frac{2}{3}} log\left(\frac{\left(x_4 \cdot x_5\right)^{1/2}}{x_6}\right)$
liability and equity structure 2	$y_4 = \sqrt{\frac{1}{2}} log\left(\frac{x_4}{x_5}\right)$
asset, liability and equity structure	$y_5 = \sqrt{\frac{9}{6}} log\left(\frac{\left(x_1 \cdot x_2 \cdot x_3\right)^{1/3}}{\left(x_4 \cdot x_5 \cdot x_6\right)^{1/3}}\right)$
turnover	$y_6 = \sqrt{\frac{12}{8}} log\left(\frac{\left(x_7 \cdot x_8\right)^{1/2}}{\left(x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot x_6\right)^{1/6}}\right)$
Margin	$y_7 = \sqrt{\frac{1}{2}} log\left(\frac{x_8}{x_7}\right)$

TABLE 5
Log-ratios computed from x_1 to x_8

```
# Tell the program the path to the directory where the data file to be analyzed is located
setwd("C:/Users/USUARIFCEE/Desktop/cluster/")
# A data matrix named "DM" is created to store the data of the file in columns
DM <-read.table("Sabi_Export_net.csv", header=TRUE, sep=";", dec=",")
# Command by which any column of the matrix will be accessible by giving only the column name
attach(DM)
# The different columns of ilr coordinates are added to the data matrix
DM$y1=sqrt(2/3)*log(((x1*x2)**(1/2))/x3)
DM$y2=sqrt(1/2)*log(x1/x2)
DM$y3=sqrt(2/3)*log(((x4*x5)**(1/2))/x6)
DM$y4=sqrt(1/2)*log(x4/x5)
DM$y5=sqrt(9/6)*log(((x1*x2*x3)**(1/3))/((x4*x5*x6)**(1/3)))
DM$y6=sqrt(12/8)*log(((x8*x7)**(1/2))/((x1*x2*x3*x4*x5*x6)**(1/6)))
DM$y7=sqrt(1/2)*log(x8/x7)
# The columns added to the matrix will be accessible by using its name, as above
attach(DM)
# A new data matrix named "Y" is created with only the data we are interested in to create the clusters
Y= DM[,c("y1","y2","y3","y4","y5","y6","y7")]
# Perform k-means clustering on a data matrix named "C4", which groups the Y data into 4 clusters
C4<-kmeans(Y, 4, iter.max = 200, nstart=100)
# Saving the cluster data in a .CSV file
write.csv2(C4$cluster, file = "Hard_Cluster.csv", row.names = TRUE)
```

BOX 9
Second Code.

vided yi values. To achieve this, perform an inverse transformation of the y1-y7 values to obtain x1-x8, and then normalize them so that x1+x2+x3=100.

If the reader wishes to execute this task, the results for this specific case are provided in Table 8.

6 CONCLUSIONS

This work analyzed the fact that while it represents a valid means of studying the financial reality of single companies, the methodology of standard financial ratios presents serious drawbacks when the ratios are used as variables in

```
# The library containing the k-means type fuzzy cluster creation function is installed
install.packages("fclust")
# Loading the library into the program
library(fclust)
# Perform k-means fuzzy clustering on a data matrix named "FC4", which groups the Y data into 4 clusters
FC4=FKM(Y,k=4, RS=25, seed=123)
# Saving the cluster membership values in two .CSV files. In the first, we save the membership values to each cluster, and in the second, the cluster numbers of maximum membership
write.csv2(FC4$U, file = " Soft_Cluster.csv", row.names = TRUE)
write.csv2(FC4$clus, file = " Soft_Cluster2.csv", row.names = TRUE)
```

BOX 10
Thirst Code.

	Clus 1	**Clus 2**	**Clus 3**	**Clus 4**
Obj 1	0,19	0,27	0,30	0,24
Obj 2	0,43	0.16	0.19	0.22
Obj 3	0,23	0,22	0,25	0,30
Obj 4	0,20	0,29	0,26	0,25
Obj 5	0,19	0,23	0,26	0,32
Obj 6	0,13	0,30	0,31	0,26
Obj 7	0,11	0,25	0,31	0,33
Obj 8	0,22	0,21	0,25	0,32
Obj 9	0,67	0,08	0,11	0,14
Obj 10	0,45	0,17	0,19	0,19

TABLE 6
First ten results in Soft_Cluster.csv

	Clus 1	**Clus 2**	**Clus 3**	**Clus 4**
Obj 98	0.13	0.37	0.03	0.47
Obj 231	0.42	0.19	0.06	0.33
Obj 247	0.50	0.10	0.01	0.39
Obj 311	0.11	0.46	0.05	0.37
Obj 698	0.42	0.13	0.05	0.40
Obj 709	0.44	0.18	0.05	0.33
Obj 713	0.43	0.15	0.08	0.34
Obj 751	0.47	0.13	0.03	0.37
Obj 769	0.47	0.14	0.03	0.36
Obj 791	0.42	0.10	0.03	0.45

TABLE 7
Objects with relatively high membership in at least two of the clusters

			Assets			*Equity*	*Liabilities*		*Exp.*	*Sales*
	Label	**Size**	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
4-fuzzy cluster solution	1	207,5	44,0%	29,6%	26,4%	33,4%	26,8%	39,9%	96,4%	97,1%
	2	219,7	16,6%	43,5%	39,9%	30,8%	16,0%	53,3%	150,1%	147,3%
	3	165,7	17,6%	26,4%	56,0%	49,1%	1,7%	49,2%	166,7%	162,2%
	4	216,0	32,0%	31,8%	36,3%	32,5%	20,4%	47,1%	129,9%	128,6%

TABLE 8
Cluster labels, size and component geometric means scaled to total assets=100

sectorial statistical analyses (in particular clustering analysis), and that this can, for example, generate average values contaminated by spurious outliers or data that must necessarily be positive which can never follow a normal statistical distribution.

The study showed that the cause of the aforementioned problems is the loss of symmetry produced when calculating standard financial ratios, since dividing two numerical values causes a distortion of the symmetry per se. A simple mathematical example was enough to evidence this fact.

Therefore, this work can be classified within the trend of questioning classical ratios that began some years ago. Although the use of logarithms corrected certain problems in the classical methodology, it has been the recent development of CoDa analysis that has allowed us to present an alternative methodology that also solves certain unresolved implicit problems such as the overloading of ratios.

CoDa methods would seem to provide the best solution to address these problems introduced in the research community. A key feature of CoDa is a particular type of logarithmic transformation of ratios, which leads to symmetric distributions that have few or no outliers, with less redundancy of information so that no data reduction is necessary. This type of transformation also ensures that the distances among clustered points are meaningful and only depend on the set of financial accounts considered for the analysis and not on the ratio chosen, minimizing the problem of numerator and denominator permutation in the construction of a ratio.

Furthermore, the choice with regard to placement in the numerator or denominator of the compositional ratio does not modify any other property of the log-ratio coordinate but the sign. This means that it makes no sense to trade in the account categories inside the log-ratio coordinates for cluster analysis purposes, since it results in the same groups comprising an identical number of clusters with the two configurations. Points should not change their membership degree to one or another cluster with only this type of permutation.

In order to highlight the new lines of research arising from this paper, the authors wish to emphasize that the proposed method of using fuzzy clusters

only extends the method described in the previous CoDa and accounting paper published by Linares-Mustarós et al. [16]. Consequently, this work confirms that the method used in this cluster analysis is consistent with the extension of fuzzy methodology. This can be observed in the fact that similar cluster centers were obtained despite using two different classification methods, hard clustering and soft clustering. Clearly, more works like the present one are needed to be able to say with any certainty that CoDa methodology constitutes the most efficient approach. At no time has this been implied here. We have simply aimed to show that the classical methodology presents serious problems and that CoDa methodology would appear to be appropriate for resolving some of these. Thus, from this principal limitation of the original work, a new path is opened up that could be developed by means of MonteCarlo simulations to show the advantages that the proposed new method can have.

On the other hand, the fuzzy method offers cluster membership values that are not necessarily 0 or 1, a major advantage over a non-fuzzy method. In addition, the fuzzy method presents a further great advantage, namely that the membership values for the same point with respect to the various clusters always add up to 1, so the obtained data are in compositional form, which makes it possible to use the results as input data for multivariate data prediction analyses.

Finally, it is worth mentioning that both fuzzy set theory and compositional data analysis methodology are new approaches that have undergone enormous growth in recent years [48-50]. The multiple options that fuzzy set theory offers for working with non-exact values [51-53] and the multiple advantages offered by compositional data analysis [19,36] justify this exponential growth. The present work adheres to these research lines and opens up a new path related to fuzzy cluster analysis with compositional accounting data.

FUNDING

This research was supported by the Spanish Ministry of Science and Innovation/AEI/10.13039/501100011033 and by ERDF A way of making Europe (grant PID2021-123833OB-I00)

REFERENCES

[1] Lucas, A., & Ramires, A. (2022). Directions for management in small and medium hotels and restaurants companies. Geo Journal of Tourism and Geosites, 40(1), 210-217.

[2] Kadim, A., Sunardi, N., & Husain, T. (2020). The modeling firm's value based on financial ratios, intellectual capital and dividend policy. Accounting, 6(5), 859-870.

[3] Kalak, I., Hudson, R. (2016). The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model. International Review of Financial Analysis, 43, 135-145.

[4] Amat, O., Manini, R. & Antón Renart, M. (2017). Credit concession through credit scoring: Analysis and application proposal. Intangible Capital, 13(1), 51-70.
[5] Lueg, R., Punda, P., & Burket, M. (2014). Does transition to IFRS substantially affect key financial ratios in shareholder-oriented common law regimes? Evidence from the UK. Advances in Accounting, 30(1), 241-250.
[6] Trussel, J.M., Parsons, L. M. (2007). Financial reporting factors affecting donations to charitable organizations. Advances in Accounting, 23, 263-285.
[7] Jiang, H., Habib, A., & Zhou, D. (2015). Accounting restatements and audit quality in China. Advances in Accounting, 31(1), 125-135.
[8] Campa, D. (2015). The impact of SME's pre-bankruptcy financial distress on earnings management tools. International Review of Financial Analysis, 42, 222-234.
[9] Cassú, C., Ferrer, J.C., & Bonet, J. (2001). *Classification of several business sectors according to uncertain characteristics*. Handbook of Management under uncertainty, 117-164. Springer.
[10] Rodrigues, L., & Rodrigues, L. (2018). Economic-financial performance of the Brazilian sugarcane energy industry: An empirical evaluation using financial ratio, cluster and discriminant analysis. *Biomass and bioenergy*, 108, 289-296.
[11] Reyes-Ruiz, G., & Hernández-Hernández, M. (2021). Fuzzy clustering as a new grouping technique to define the business size of SMEs through their financial information. Journal of Intelligent & Fuzzy Systems, 40(2), 1773-1782
[12] Frecka, T. J., Hopwood, W.S. (1983). The effects of outliers on the cross-sectional distributional properties of financial ratios. Accounting Review, 58(1), 115-128.
[13] Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. The Accounting Review 51(1), 90-96.
[14] Ezzamel, M., Mar-Molinero, C. (1990). The distributional properties of financial ratios in UK manufacturing companies. Journal of Bussiness Fiunance & Accounting, 17(1), 1-29.
[15] Mcleay, S., Omar, A. (2000). The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios. The British Accounting Review, 32(2), 213-230.
[16] Linares-Mustarós, S., Coenders, G., & Vives-Mestres, M. (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. Advances in Accounting, 40, 1-10.
[17] Lev, B., Sunder, S. (1979). Methodological issues in the use of financial ratios. Journal of Accounting and Economics, 1(3), 187-210.
[18] Cowen, S. S., Hoffer, J. A. (1982). Usefulness of financial ratios in a single industry. Journal of Business Research, 10(1), 103-118.
[19] Carreras-Simó, M., & Coenders, G. (2021). The relationship between asset and capital structure: a compositional approach with panel vector autoregressive models. Quantitative Finance and Economics, 2021, vol. 5, núm. 4, p. 571-590.
[20] So, J. C. (1987). Some empirical evidence on the outliers and the non-normal distribution of financial ratios. Journal of Business Finance & Accounting, 14(4), 483–496.
[21] Watson, C.J. (1990). Multivariate distributional properties, outliers, and transformation of financial ratios. The Accounting Review, 65(3), 682-695.
[22] Creixans-Tenas, J., Coenders, G., & Arimany-Serrat, N. (2019). Corporate social responsibility and financial profile of Spanish private hospitals. Heliyon 5 (10), e02623.
[23] Santis, P., Albuquerque, A., & Lizarelli, F. (2016). Do sustainable companies have a better financial performance? A study on Brazilian public companies. Journal of Cleaner Production, 133, 735-745.
[24] Yoshino, N., & Taghizadeh-Hesary, F. (2015). Analysis of credit ratings for small and medium-sized enterprises: Evidence from Asia. Asian Development Review, 32(2), 18–37.
[25] Yap, B. C. F., Mohamed, Z., & Chong, K. R. (2014). The effects of the financial crisis on the financial performance of malaysian companies. Asian Journal of Finance & Accounting, 6(1), 236–248.
[26] Feranecová, A., Krigovská, A. (2016). Measuring the performance of universities through cluster analysis and the use of financial ratio indexes. Economics & Sociology, 9(4), 259-271.

[27] Sharma, S., Shebalkov, M., & Yukhanaev, A. (2016). Evaluating banks performance using key financial indicators–a quantitative modeling of Russian banks. The Journal of Developing Areas, 50(1), 425–453.
[28] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139-160.
[29] Aitchison, J. (1986). The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman and Hall, London.
[30] Pawlowsky-Glahn, V., Buccianti, A. (2011). Compositional data analysis. Theory and applications. Wiley, New York.
[31] Van den Boogaart, K.G., Tolosana-Delgado, R. (2013). Analyzing compositional data with R. Springer. Berlin.
[32] Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado R. (2015). Modeling and analysis of compositional data. Wiley, Chichester.
[33] Filzmoser, P., Hron, K., & Templ, M. (2018). Applied compositional data analysis with worked examples in R, Springer, New York.
[34] Greenacre, M. (2018). Compositional data analysis in practice. Chapman and Hall/CRC press, New York.
[35] Carreras-Simó, M., Coenders, G. (2020). Principal component analysis of financial statements. A compositional approach. Revista de Métodos Cuantitativos para la Economía y la Empresa, 29, 18-37.
[36] Arimany-Serrat, N., Farreras-Noguer, M., & Coenders, G. (2022). New developments in financial statement analysis. Liquidity in the winery sector. Accounting, 8(3), 355-366.
[37] Saus-Sala, E., Farreras-Noguer, A., Arimany-Serrat, N., Coenders, G. (2021): Compositional DuPont analysis. A visual tool for strategic financial performance assessment. In Filzmoser P., Hron K., Martín-Fernández J.A., Palarea-Albaladejo J. (Eds.) Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn. Springer Nature, Cham: 189-206.
[38] Jofre-Campuzano, P., Coenders, G. (2022): Compositional classification of financial statement profiles: The weighted case. Journal of Risk and Financial Management, 15, 12: 546.
[39] Saus-Sala, E., Farreras-Noguer, M. À., Arimany-Serrat, N., Coenders, G. (2023): Análisis de las empresas de turismo rural en Cataluña y Galicia: rentabilidad económica y solvencia 2014 – 2018. Cuadernos del CIMBAGE, 25, 1: 33-54.
[40] Coenders, G. (in press): Application aux ratios financiers. In Gégout-Petit, A., Bertrand, F., Thomas-Agnan, C. (Éds.) Données de Composition. Société Française de Statistique/Éditions TECHNIP, Paris.
[41] Butler, B. M., Palarea-Albaladejo, J., Shepherd, K. D., Nyambura, K. M., Towett, E. K., Sila, A. M., & Hillier, S. (2020). Mineral–nutrient relationships in African soils assessed using cluster analysis of X-ray powder diffraction patterns and compositional methods. *Geoderma*, 375, 114474.
[42] Zhou, S., Zhou, K., Wang, J., Yang, G., & Wang, S. (2018). Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. *Frontiers of Earth Science*, 12, 491-505.
[43] Hron, K., & Filzmoser, P. (2013). Robust diagnostics of fuzzy clustering results using the compositional approach. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* (pp. 245-253). Springer Berlin Heidelberg.
[44] Palarea-Albaladejo, J., Martín-Fernández, J. A., & Soto, J. A. (2012). Dealing with distances and transformations for fuzzy C-means clustering of compositional data. *Journal of classification*, 29, 144-169
[45] Zadeh L.A. (1965). Fuzzy sets. Information and Control 8, 338-353.
[46] Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57.
[47] Bezdek, J.C. (1981), "Pattern Recognition with Fuzzy Objective Function Algorithm", Plenum, NY.
[48] Merigó, J. M., Gil-Lafuente, A. M., & Yager, R. R. (2015). An overview of fuzzy research with bibliometric indicators. Applied Soft Computing, 27, 420-433.

[49] Coenders, G., & Ferrer-Rosell, B. (2020). Compositional data analysis in tourism: review and future directions. Tourism Analysis, 25(1), 153-168.

[50] Navarro-Lopez, C., González Morcillo, C., Mulet-Forteza, C., & Linares-Mustarós, S. (2021). A Bibliometric Analysis of the 35th anniversary of the paper" The Statistical Analysis of Compositional Data" by John Aitchison (1982). Austrian Journal of Statistics, 2021, vol. 50, núm. 2, p. 38-55.

[51] Linares-Mustarós, S., Ferrer-Comalat, J. C., Corominas-Coll, D., & Merigó, J. M. (2019). The ordered weighted average in the theory of expertons. International Journal of Intelligent Systems, 34(3), 345-365.

[52] Linares-Mustarós, S., Ferrer-Comalat, J. C., Corominas-Coll, D., & Merigo, J. M. (2021). The weighted average multiexperton. Information Sciences, 557, 355-372.

[53] Ferrer-Comalat, J. C., Corominas-Coll, D., & Linares-Mustarós, S. (2020). Fuzzy logic in economic models. Journal of Intelligent & Fuzzy Systems, 38(5), 5333-5342.