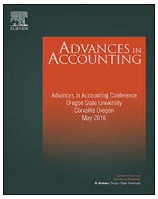
Contents lists available at ScienceDirect

Advances in Accounting

journal homepage: www.elsevier.com/locate/adiac



Financial performance and distress profiles. From classification according to financial ratios to compositional classification

Salvador Linares-Mustarós[a], Germà Coenders[b,*], Marina Vives-Mestres[c]

[a] Department of Business, University of Girona, Campus Montilivi, Faculty building of Economics and Business, C. Universitat 10, 17003 Girona, Spain
[b] Department of Economics, University of Girona, Campus Montilivi, Faculty Building of Economics and Business, C. Universitat 10, 17003 Girona, Spain
[c] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, Building P4. C. Maria Aurèlia Capmany, 61, 17003 Girona, Spain

ARTICLE INFO

Keywords:
Compositional data analysis (CoDa)
Financial ratio
Cluster analysis
Ilr coordinate
Aitchison's distance

JEL classification:
C49
G30
M41

ABSTRACT

Financial ratios are often used in cluster analysis to classify firms according to the similarity of their financial structures. Besides the dependence of distances on ratio choice, ratios themselves have a number of serious problems when subject to a cluster analysis such as skewed distributions, outliers, and redundancy. Some solutions to overcome those drawbacks have been proposed in the literature, but have proven problematic. In this work we put forward an alternative financial statement analysis method for classifying firms which aims at solving the above mentioned shortcomings and draws from compositional data analysis. The method is based on the use of existent clustering methods with standard software on transformed data by means of the so-called isometric logarithms of ratios. The method saves analysis steps (outlier treatment and data reduction) while defining distances among firms in a meaningful way which does not depend on the particular ratios selected. We show examples of application to two different industries and compare the results with those obtained from standard ratios.

1. Introduction

Financial ratios, i.e., ratios comparing the magnitudes of accounts in financial statements, constitute a case of researchers' and professionals' interest in relative rather than absolute account magnitudes. From the classical work on bankruptcy prediction by Altman (1968), the use of financial ratios has spread along and across many research lines (Willer do Prado et al., 2016), such as stock market returns (e.g., Dimitropoulos, Asteriou, & Koumanakos, 2010), firm survival analysis (e.g., Kalak & Hudson, 2016), credit scoring (e.g., Amat, Manini, & Antón Renart, 2017), assessing the impact of International financial reporting standards (e.g., Lueg, Punda, & Burkert, 2014), predicting donations to charitable organizations (e.g., Trussel & Parsons, 2007), accounting restatements (e.g., Jiang, Habib, & Zhou, 2015), and earnings manipulation (e.g., Campa, 2015). This article focuses on another frequent use of financial ratios: to classify firms according to similarity of the structure of their financial statements, searching for different profiles of financial structure, performance or distress. Since the seminal works of Cowen and Hoffer (1982), and Gupta and Huefner (1972), through the relevant contributions by Dahlstedt, Salmi, Luoma, and Laakkonen (1994); Ganesalingam and Kumar (2001); Mar Molinero, Apellaniz Gomez, and Serrano Cinca (1996); Serrano Cinca (1998); and Voulgaris, Doumpos, and Zopounidis (2000), the interest in clustering firms according to their financial ratios remains current (Feranecová & Krigovská, 2016; Lukason & Laitinen, 2016; Luptak, Boda, & Szucs, 2016; Martín-Oliver, Ruano, & Salas-Fumás, 2017; Momeni, Mohseni, & Soofi, 2015; Santis, Albuquerque, & Lizarelli, 2016; Sharma, Shebalkov, & Yukhanaev, 2016; Yoshino & Taghizadeh-Hesary, 2015; Yoshino, Taghizadeh-Hesary, Charoensivakorn, & Niraula, 2016).

Despite the popularity of financial ratios, the financial and statistical literature has long reported a number of serious practical drawbacks of their use. The first of them has to do with the fact that most ratios are distributed between zero and infinity and thus make fully symmetric distributions impossible to achieve. Ratios also tend to have asymmetric distributions because decreases in the denominator produce larger changes in the ratio value than increases do (Frecka & Hopwood, 1983). Both phenomena tend to produce distributions with positive skewness and preclude using symmetric probability distributions such as the normal (e.g., Deakin, 1976; Ezzamel & Mar-Molinero, 1990; Kane, Richardson, & Meade, 1998; Martikainen, Perttunen, Yli-Olli, & Gunasekaran, 1995; Mcleay & Omar, 2000; So, 1987). Asymmetry is also connected to the commonly reported outliers (e.g., Cowen & Hoffer, 1982; Ezzamel & Mar-Molinero,

[*] Corresponding author.
E-mail addresses: salvador.linares@udg.edu (S. Linares-Mustarós), germa.coenders@udg.edu (G. Coenders), marina.vives@udg.edu (M. Vives-Mestres).

http://dx.doi.org/10.1016/j.adiac.2017.10.003
Received 13 September 2017; Received in revised form 16 October 2017; Accepted 25 October 2017

1990; Lev & Sunder, 1979; So, 1987; Watson, 1990). It can even be the case that outliers are the main or only source of positive asymmetry in the distributions (Frecka & Hopwood, 1983). These outliers do not always reflect atypical management practices but can also result from a small value of the denominator of the ratio (e.g., Ezzamel & Mar-Molinero, 1990; Kane et al., 1998). In the particular case of cluster analysis, asymmetric distributions lead to some clusters being very small (e.g., Feranecová & Krigovská, 2016; Santis et al., 2016; Sharma et al., 2016; Yoshino et al., 2016; Yoshino & Taghizadeh-Hesary, 2015). It is also well known that the presence of outliers distorts the results of many clustering algorithms, and oftentimes it even leads to one-member clusters (e.g., Feranecová & Krigovská, 2016; Sharma et al., 2016; Yap, Mohamed, & Chong, 2014).

The second major drawback has to do with redundancy of the over 100 ratios currently in use (Chen & Shimerda, 1981; Pindado & Rodrigues, 2004; Pohlman & Hollinger, 1981). Oftentimes, redundancy occurs to such an extent that “there is no absolute test for the importance of variables” (Barnes, 1987, 455) and “to identify those ratios which contain complete information about a firm while minimising duplication cannot be achieved purely by logic” (Barnes, 1987, 456). In extreme cases there is an exact dependency between ratios. For instance, the inverse of the liability to asset ratio is the equity to debt ratio plus one. In cluster analysis such redundancy has often led to groups not capturing proper distinct profiles. Solutions with overall ordered groups labelled as “healthy, in between, less healthy”, “highly distressed, mildly distressed, not distressed”, “dynamic, medium, weak”, or “good performers, average performers, poor performers” abound (e.g., Ganesalingam & Kumar, 2001; Momeni et al., 2015; Voulgaris et al., 2000; Yap et al., 2014; Yoshino et al., 2016; Yoshino & Taghizadeh-Hesary, 2015). In cluster analysis, redundancy has one further consequence: it increases distances among firms along the added redundant information, which is tantamount to inadvertently giving this redundant information greater weight in the results (e.g., Aldenderfer & Blashfield, 1984).

The third major drawback has to do with arbitrariness of Euclidean distance among firms. On the one hand, a different set of ratios leads to different distances among firms, even if ratios are computed from exactly the same set of financial accounts. On the other hand, Euclidean distance is not an appropriate dissimilarity measure for ratios. Even placement of accounts in the numerator or in the denominator of the same ratio matters to Euclidean distance. This is because increases in the numerator and in the denominator are not treated in the same way (Frecka & Hopwood, 1983). Let us consider the simplest possible case in which only two financial accounts x_1 and x_2 are of interest. Only two ratios are possible: $r_1 = x_1/x_2$ and $r_2 = x_2/x_1$. Let us consider three firms A, B, and C, such that $x_{1A} = 1$, $x_{2A} = 1$, $x_{1B} = 1$, $x_{2B} = 2$, $x_{1C} = 2$, $x_{2C} = 1$. The ratio values are $r_{1A} = r_{2A} = 1$, $r_{1B} = 0.5$, $r_{2B} = 2$, $r_{1C} = 2$, $r_{2C} = 0.5$. Intuitively, the ratios r_1 and r_2 should contain the same information about firms. However, Euclidean distances computed from r_1 are d(A,B) = 0.5, d(A,C) = 1, and d(B,C) = 1.5, while Euclidean distances computed from r_2 are d(A,B) = 1, d(A,C) = 0.5, and d(B,C) = 1.5. In other words, when using r_1 firms A and B would tend to cluster together and when using r_2 firms A and C would tend to cluster together. Unclear distances which depend on arbitrary decisions and even on a permutation of numerator and denominator can only threaten the results of cluster analysis (Martín, 1998).

As regards the problem related to asymmetry and outliers, some form of transformation and/or outlier trimming has often been applied. These include transformations such as Box-Cox (e.g., Ezzamel & Mar-Molinero, 1990; Mcleay & Omar, 2000; Watson, 1990), logs (e.g., Cowen & Hoffer, 1982; Deakin, 1976; Sudarsanam & Taffler, 1995), ranks (e.g., Kane et al., 1998; Lueg et al., 2014), square roots (e.g., Deakin, 1976; Frecka & Hopwood, 1983; Martikainen et al., 1995), weight of evidence (e.g., Nikolic, Zarkic-Joksimovic, Stojanovski, & Joksimovic, 2013); outlier trimming (e.g., Ezzamel & Mar-Molinero, 1990; Frecka & Hopwood, 1983; Lev & Sunder, 1979; Martikainen et al., 1995; So, 1987; Watson, 1990); and outlier winsorization (e.g., Lev & Sunder, 1979).

Both transformation and outlier treatment have proved problematic. Not only is there uncertainty about which transformation to apply or which outliers to remove. There is also uncertainty regarding whether one should first remove outliers and then transform to account for the remaining non-normality or first transform and then remove the remaining outliers (e.g., Ezzamel & Mar-Molinero, 1990). The log transformation is especially appealing, given its wide understanding and ease of interpretation as relative change in the economic and financial fields. It is also theoretically justified when the numerator and the denominator follow a log-normal distribution. Empirically it is also often reported to yield acceptable results (Sudarsanam & Taffler, 1995). However, as shown above, there is no consensus on the transformation issue, and in some cases more than one transformation has been shown to yield approximately normal ratios (Buijink & Jegers, 1986).

As regards the redundancy problem, many clustering studies use data reduction methods prior to the analysis, either to compute a few aggregated functions of ratios or to select a few relevant and distinct ratios. These strategies include principal component analysis (e.g., Cowen & Hoffer, 1982; Dimitropoulos et al., 2010; Martín-Oliver et al., 2017; Sharma et al., 2016; Yoshino et al., 2016; Yoshino & Taghizadeh-Hesary, 2015), grey relation analysis (e.g., Ho & Wu, 2006), factor analysis (e.g., Feranecová & Krigovská, 2016; Lukason & Laitinen, 2016; Yap et al., 2014), self-organising feature maps (e.g., Serrano Cinca, 1998), multidimensional scaling (e.g., Mar Molinero et al., 1996), or cluster analysis on the transposed data matrix, to define groups of ratios instead of groups of firms (e.g., Nikolic et al., 2013; Serrano Cinca, 1998). While this is generally sound practice, it adds an extra step to the analysis, and it is often not clear which data reduction method should be preferred for a particular problem.

To the best of our knowledge, the distance issue has not been solved in the financial literature, but it has been solved in other scientific fields, from which we draw below.

The aim of this article is to put forward an alternative financial statement analysis method for classifying firms from the structure of their financial statements, which aims at solving the above mentioned shortcomings and draws from the compositional data analysis (CoDa) literature. CoDa is the standard methodological toolbox to analyse the relative importance of magnitudes in fields such as biology, chemistry and geology. A key feature of CoDa is a particular type of log transformation of ratios, which tends to lead to symmetric distributions with few or no outliers, and to less redundancy, thus making data reduction less necessary. This transformation also ensures that the distances among clustered cases are meaningful and that they only depend on the set of financial accounts which is considered for the analysis and not on ratio choice. Once this transformation has been carried out, standard clustering methods and software can be used, which is an attractive possibility for applied researchers.

The article is organized as follows. Section 2 reviews the basics of CoDa. Section 3 deals with the proposal to use an alternative financial statement analysis method based on CoDa. Section 4 presents two numerical real-data examples of cluster analysis in high tech and low tech manufacturing industries. Results are compared to those obtained when using standard financial ratios. Section 5 summarizes the main results and makes suggestions for further research.

2. Compositional data analysis

2.1. Compositional data

Compositional Data are positive vector variables carrying information about the relative size of their *D* components to one another (Aitchison, 1986; Barceló-Vidal & Martín-Fernández, 2016):

$$\mathbf{x} = (x_1, x_2, \cdots, x_D) \text{ with } x_j > 0, \quad j = 1, 2, \cdots, D$$

In our case, the components represented by the x_j variables are different financial accounts, such as inventories, sales, operating expenses, equity, or accounts receivable.

The standard methodological framework for dealing with compositional data (known as Compositional data analysis - CoDa) was born in the fields of geology and chemistry. These disciplines typically focus interest on the relative importance of the parts of the whole rock or substance which is analysed, while the size of the rock or chemical sample is deemed irrelevant. After the seminal works of Aitchison (1982, 1986) thirty five years of development have led to a well-established standard CoDa toolbox which is covered in text books (van den Van den Boogaart & Tolosana-Delgado, 2013; Pawlowsky-Glahn & Buccianti, 2011; Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015). Recently, CoDa has also been applied in finance and in management to answer research questions concerning relative magnitudes. Examples include crowdfunding (Davis, Hmieleski, Webb, & Coombs, 2017), financial markets (Ortells, Egozcue, Ortego, & Garola, 2016), market segmentation (Ferrer-Rosell and Coenders, in press; Ferrer-Rosell, Coenders, & Martínez-Garcia, 2016), market share (Morais, Thomas-Agnan, & Simioni, 2017), shopping basket mining (Kenett, Martín-Fernandez, & Vives-Mestres, 2017), patents (Hingley, 2017), consumer research (Ferrer-Rosell, Coenders, & Martínez-Garcia, 2015; Ferrer-Rosell, Coenders, Mateu-Figueras, & Pawlowsky-Glahn, 2016; Vives-Mestres, Martín-Fernández, & Kenett, 2016), quality management (Vives-Mestres, Daunis-i-Estadella, & Martín-Fernández, 2014, 2016), and management education (Batista-Foguet, Ferrer-Rosell, Serlavós, Coenders, & Boyatzis, 2015; Mateu-Figueras et al., 2016).

From a historical perspective, CoDa was born as a method to analyse parts of a whole (Aitchison, 1986). Recently, the emphasis in CoDa has shifted to the sheer interest in relative size of any set of positive magnitudes (Barceló-Vidal & Martín-Fernández, 2016). Interesting applications of CoDa to data which do not represent parts of any whole are in Ortells et al. (2016) and Azevedo Rodrigues, Daunis-i-Estadella, Mateu-Flgueras, and Thió-Henestrosa (2011). This is the case in financial statement analysis, in which, for instance, sales and assets are not parts of any whole and the asset turnover ratio compares the magnitudes of both, in relative terms.

CoDa can involve using specialised methods and software on the raw data (e.g., Palarea-Albaladejo & Martín-Fernández, 2015; Thió-Henestrosa & Martín-Fernández, 2005; Van den Boogaart & Tolosana-Delgado, 2013). However, using existent standard statistical methods with standard software on transformed data (the so-called *coordinates*, see Mateu-Figueras, Pawlowsky-Glahn, & Egozcue, 2011) tends to be a more attractive possibility for applied researchers. The latter approach is fully feasible in the cluster-analysis case (e.g., Ferrer-Rosell and Coenders, in press) and we embrace it in this article. In cluster analysis, transformations and distances are, of course, interrelated, as shown below.

2.2. Compositional transformations

Logarithms of ratios are the standard transformation in CoDa (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015). Several choices are possible to compute log-ratios. In all cases they involve a logarithm of a ratio among single components or among geometric means of components. A scaling constant multiplying the log-ratio may or may not be included. A log-ratio involving only two components might be computed as:

$$\log \frac{x_1}{x_2}$$

where log stands for the natural logarithm. Positive values mean that x_1 is larger than x_2. Negative values show the opposite. A zero log-ratio implies equality of both magnitudes, exactly in the same way as a unit standard ratio.

A log-ratio is symmetric in the sense that its range is from minus infinity to plus infinity. Besides, permuting the numerator and denominator components leads to the same distance from zero:

$$\log \frac{x_1}{x_2} = -\log \frac{x_2}{x_1}$$

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. For log-ratios placement makes no difference.

Another argument for log-ratios is that they tend to be approximately normally distributed (Aitchison, 1986). Normality after a log transformation is justified by the additive log-normal distribution (Aitchison, 1982) and by the compositional equivalent to the central limit theorem (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015).

Even if log-ratios solve the asymmetry and outlier problems, reformulating all financial ratios in the literature as log-ratios would leave us with the same redundancy problems encountered in standard financial ratio analysis. It can be show that just $D-1$ *log-ratios* contain all information about the relative importance of D components (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015). For cluster analysis purposes it is also important that log-ratios are coherent with a meaningful notion of distance. Both aims are achieved by the so-called transformation into *isometric log-ratio (ilr) coordinates* (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, & Barceló-Vidal, 2003).

Ilr coordinates can be easily formed from a *sequential binary partition* (SBP) of components. To create the first ilr coordinate, the complete composition x = $(x_1, x_2, \ldots, x_D)$ is split into two groups of components: one for the numerator and the other for the denominator of the log-ratio. In the following step, one of the two groups is further split to create the second ilr coordinate. A SBP always implies $D-1$ ilr coordinates.

In any step of the SBP, when the y_j ilr coordinate is created, a group containing $r+s$ components is split into two: r components ($x_{n1,} \ldots, x_{nr}$) are placed in the numerator, and s components ($x_{d1,} \ldots, x_{ds}$) in the denominator. The ilr coordinate is a scaled log-ratio of the geometric means of each group of components (Egozcue et al., 2003):

$$y_j = \sqrt{\frac{rs}{r+s}} \log \frac{(x_{n1} \cdots x_{nr})^{1/r}}{(x_{d1} \cdots x_{ds})^{1/s}}$$

The choice about placement in the numerator or the denominator will not modify any property of the log-ratio but the sign. A positive sign of the coordinate implies greater importance of the components in the numerator as compared to those in the denominator. The scaling constant $\sqrt{\frac{r\ s}{r+s}}$ plays a role in the distances defined in the next section.

2.3. Compositional distances

As stated in the introduction, when applied to financial ratios, Euclidean distance depends on the particular set of ratios computed from the financial accounts of interest, and even yields different results when permuting which part is in the numerator and which in the denominator.

The commonest distance measure used in CoDa, which is called *Aitchison's distance* (Aitchison, 1983; Aitchison, Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000) solves these drawbacks. Aitchison's distance between compositions x and x* is computed from the differences in the logarithms of the ratios of all pairwise comparisons of components:

$$\mathrm{d}(\mathbf{x}, \mathbf{x}^*) = \sqrt{\frac{1}{2D} \sum_{i \neq j} \left(\log \frac{x_i}{x_j} - \log \frac{x_i^*}{x_j^*} \right)^2} \text{ with } i = 1, 2, \cdots, D; j = 1, 2, \cdots, D$$

Under Aitchison's distance all parts appear both in the numerator and in the denominator and are compared to all other parts. Thus, distance depends only on the selected financial accounts which are included in x, and not on the particular ratios or log-ratios selected. Other attractive properties are described in Aitchison (1992) and Martín (1998).

In connection with Aitchison's distances, ilr coordinates have the attractive property that Euclidean distances computed using ilr coordinates as data:

$$\mathrm{d}(\mathbf{x},\mathbf{x}^{*}) = \sqrt{\sum_{j} (y_j - y_j^{*})^2} \text{with } j = 1, 2, \cdots, D-1$$

equal Aitchison's distances computed from the raw composition. This holds for any choice of SBP (Egozcue et al., 2003). On the basis of this property, the method suggested in Section 3 involves first computing ilr coordinates and then applying standard cluster analysis methods with standard software using ilr coordinates as data.

3. Step-by-step CoDa method for the analysis of financial statements

3.1. Account selection

Without loss of generality, the method is described on the basis of a simplified balance sheet (Fig. 1), which is enough for computing the commonest *debt and liquidity ratios* (ratios indicating whether assets can pay for the whole debt, and quick assets and other current assets can pay for short term debt; e.g., Linares-Mustarós, Farreras-Noguer, Ferrer-i-Comalat, & Rabaseda-Tarrés, 2012).

Common liquidity ratios which can be computed from Fig. 1 include the *acid test ratio* = x_3/x_6, also known as *quick ratio*, and the *current ratio* = $(x_2 + x_3)/x_6$. Typical debt ratios include the *liability to asset ratio* = $(x_5 + x_6)/(x_4 + x_5 + x_6)$, the *equity to debt ratio* = $(x_4)/(x_5 + x_6)$, the *equity to long term debt ratio* = x_4/x_5, the *current liability to asset ratio* = $(x_6)/(x_4 + x_5 + x_6)$, and the *long term debt to asset ratio* = $(x_5)/(x_4 + x_5 + x_6)$.

By just adding x_7 = operating expenses and x_8 = sales from the *profit and loss statement* (also known as income statement), liquidity and debt information is complemented with that carried by *profitability and activity ratios*. Examples are *operating margin* = $(x_8\text{-}x_7)/x_7$ and *asset turnover* = $x_8/(x_1 + x_2 + x_3)$.

The method we present does not imply using those particular ratios. We only want to raise the point that we aim to classify firms based on comparable information to that provided by such ratios. Of course, greater detail can be included in the balance sheet at will, for instance by distinguishing several types of inventories and quick assets, and any other account from the profit and loss statement can be included at wish.

3.2. Ilr coordinate computation

While any SBP will do the job, an attractive and interpretable possibility to compute ilr coordinates is to divide assets on the one hand and liabilities and equity on the other. Then one can continue to define partitions and ilr coordinates within assets within liabilities and equity.

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

Fig. 1. Simplified balance sheet.

A very simple procedure is to order the balance sheet from less to more liquid assets and from longer to shorter term liabilities. For assets we could have the following ilr coordinates:

$$y_1 = \sqrt{\frac{2}{3}}\log\left(\frac{(x_1x_2)^{1/2}}{x_3}\right),\ y_2 = \sqrt{\frac{1}{2}}\log\left(\frac{x_1}{x_2}\right)$$

and for liabilities:

$$y_3 = \sqrt{\frac{2}{3}}\log\left(\frac{(x_4x_5)^{1/2}}{x_6}\right),\ y_4 = \sqrt{\frac{1}{2}}\log\left(\frac{x_4}{x_5}\right)$$

For instance, y_1 compares quick assets (x_3) with the geometric average of less liquid assets (fixed assets –x_1– and inventories –x_2). A larger y_1 coordinate is interpreted as a larger weight of fixed assets and inventories compared to quick assets.

The ilr coordinate comparing assets with liabilities and equity is:

$$y_5 = \sqrt{\frac{9}{6}}\log\left(\frac{(x_1x_2x_3)^{1/3}}{(x_4x_5x_6)^{1/3}}\right)$$

An additional coordinate compares the set of profit and loss accounts to the set of balance sheet accounts:

$$y_6 = \sqrt{\frac{12}{8}}\log\left(\frac{(x_7x_8)^{1/2}}{(x_1x_2x_3x_4x_5x_6)^{1/6}}\right)$$

Finally, the set of profit and loss accounts can be partitioned in any order:

$$y_7 = \sqrt{\frac{1}{2}}\log\left(\frac{x_8}{x_7}\right)$$

y_7 is readily interpreted as a substitute for operating margin. The reader must be reminded that CoDa compares positive magnitudes. Note that we avoid constructing log-ratios using profits (which may be negative) with no loss of information, since any information on profitability is included by considering expenses and revenues separately, which cannot be negative.

SBPs can be represented in an intuitive way by means of a tree diagram. The implied SBP by y_1 to y_7 is in Fig. 2.

Even if components are compared in a particular way in the SBP, all information comparing any accounts is included in the existent ilr coordinates. This is so irrespective of the chosen SBP. For instance:

- a linear combination of y_3 and y_4 compares equity (x_4) with both debt accounts (x_5 and x_6) and thus conveys similar information as the equity to debt ratio. It must be taken into account that log ratios substitute geometric means for sums:

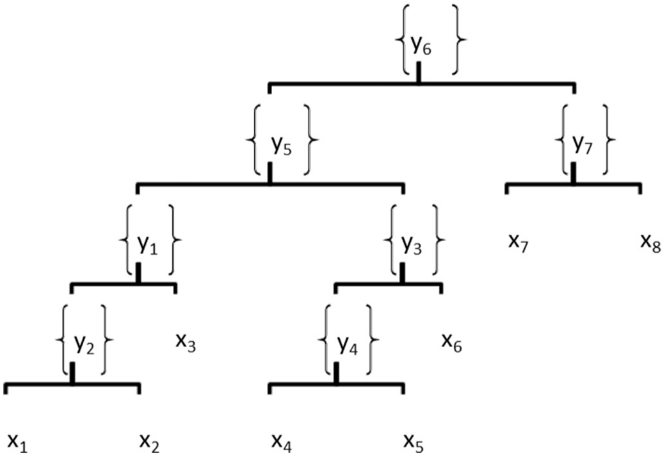


Fig. 2. Tree diagram showing the SBP. Coordinates (y_1 to y_7) and components which they compare (x_1 to x_8).

$$\frac{1}{2}\sqrt{\frac{3}{2}}y_3 + \frac{3}{4}\sqrt{\frac{2}{1}}y_4 = \log\left(\frac{(x_4x_5)^{1/4}x_4^{3/4}}{x_6^{1/2}x_5^{3/4}}\right) = \log\left(\frac{x_4}{(x_5x_6)^{1/2}}\right)$$

- a linear combination of y_1 and y_3 compares quick assets and short term debt in a similar way as the acid test ratio. It must be taken into account that quick assets exceed short term debt when long term debt and equity exceed the remaining assets; hence the additional term at the end of the equation:

$$\sqrt{\tfrac{3}{2}}y_3 - \sqrt{\tfrac{3}{2}}y_1 = \log\left(\tfrac{x_3}{x_6}\right) + \log\left(\tfrac{(x_4x_5)^{1/2}}{(x_1x_2)^{1/2}}\right)$$

- a linear combination of y_6 and y_7 is interpreted analogously to asset turnover:

$$\sqrt{\frac{8}{12}}y_6 + \sqrt{\frac{1}{2}}y_7 = \log\left(\frac{x_8}{(x_1x_2x_3x_4x_5x_6)^{1/6}}\right)$$

This does not imply that the researcher or the professional should embark in the tedious calculations of such linear combinations of ilr coordinates. Quite on the contrary, we mean to show that this effort is unnecessary because the information about the relative importance of any sets of components is already contained in the ilr coordinates.

3.3. Cluster analysis

Once ilr coordinates have been computed, any standard cluster analysis method using Euclidean distances (e.g., Wards' method or the *k*-means method which assume Euclidean distances, or other standard hierarchical clustering methods for which Euclidean distances are an option) can be applied with standard software, providing equivalent results to Aitchison's distances. Standardization of ilr coordinates is not desirable because it modifies distances and would thus make Euclidean distances no longer be equivalent to Aitchison's distances.

3.4. Interpretation

Once the cluster analysis has been carried out, clusters are described with geometric means of the original components x_1 to x_8 computed within each cluster. In CoDa, geometric rather than arithmetic means of components are always used as estimates of the centre of a dataset. This is coherent with a focus on relative rather than absolute information. Since different clusters might contain firms of different sizes, cluster geometric means are better interpreted after some sort of normalization, for instance by multiplying all geometric means in a given cluster by the appropriate constant so that *total assets* = $(x_1 + x_2 + x_3) = 100$. The practitioner's favourite standard financial ratios can also be computed from those normalized geometric means as the representative financial ratios in each cluster.

As in any cluster analysis, it is very useful to relate the classification to external non-financial variables to enrich interpretation. Since this differs in no way from standard practice, we do not give it further consideration in this article.

4. Applications

4.1. Clustering method

In the first application we attempt to mirror common cluster analysis practice, both with standard financial ratios and with ilr coordinates. For an overview on cluster analysis see Everitt, Landau, Leese, and Stahl (2011).

The most popular clustering methods in the applied literature on clustering with financial ratios are Ward's method (e.g., Dahlstedt et al., 1994; Ganesalingam & Kumar, 2001; Martín-Oliver et al., 2017; Voulgaris et al., 2000) and the *k*-means method (e.g., Feranecová & Krigovská, 2016; Luptak et al., 2016; Momeni et al., 2015; Yap et al., 2014). Both implicitly use Euclidean distances. Using the *k*-means method with Ward's solution as initial clustering is strongly recommended in the cluster analysis literature, and applications in the financial field start to appear (e.g., Santis et al., 2016). This approach both provides a refinement of Ward's solution and a sensible initial solution for the *k*-means method, which is very sensitive to the initial clustering choice.

It is well known that variables with larger variances have a larger weight on the classification, and standardization of financial ratios constitutes sound practice. As argued above, standardization is not convenient on ilr coordinates.

4.2. Ilr coordinates and ratios

In our application we draw from the unstandardized ilr coordinates (y_1 to y_7) and from the 12 standardized financial ratios used by one of the most cited articles (Voulgaris et al., 2000), which can all be computed from x_1 to x_8.

r_1 = Current assets/Current liabilities = *current ratio* = $(x_2 + x_3)/(x_6)$
r_2 = (Current assets − Inventory)/Current liabilities = *acid test ratio* = $(x_3)/(x_6)$
r_3 = (Long-term debt + equity)/Fixed assets = $(x_5 + x_4)/(x_1)$
r_4 = Long-term debt/Total assets = *long term debt to asset ratio* = $(x_5)/(x_4 + x_5 + x_6)$
r_5 = Total debt/Total assets = *liability to asset ratio* = $(x_5 + x_6)/(x_4 + x_5 + x_6)$
r_6 = Equity/Long term debt = *equity to long term debt ratio* = $(x_4)/(x_5)$
r_7 = Current liabilities/Total assets = *current liability to asset ratio* = $(x_6)/(x_4 + x_5 + x_6)$
r_8 = Inventory × 360/Sales = $(x_2 \times 360)/(x_8)$
r_9 = Sales/Fixed assets = $(x_8)/(x_1)$
r_{10} = Profit/Sales = $(x_8 - x_7)/(x_8)$
r_{11} = Profit/Equity = $(x_8 - x_7)/(x_4)$
r_{12} = Profit/Total assets = $(x_8 - x_7)/(x_4 + x_5 + x_6)$

Another often cited article using a similar number of analogous ratios which are also computable from x_1 to x_8 is that by Cowen and Hoffer (1982).

4.3. Data source

The data come from the SABI (Iberian Balance sheet Analysis System) database, developed by INFORMA D & B in collaboration with Bureau Van Dijk, and contains financial statements of over 2 million Spanish companies and > 500,000 Portuguese ones. The database was last accessed 28/8/2017. Search criteria included:

- NACE (Statistical classification of economic activities in the European Community) code 21.2: Manufacture of pharmaceutical preparations.
- Availability of data for 2015.

n = 168 companies fulfilled the search criteria. Firms which were not operating (zero revenues, n = 15) or bankrupt (n = 14) were excluded. A number of firms had zero values in some components, which

Table 1
Skewness and kurtosis coefficients of ilr coordinates (left) and financial ratios (right).

	Skewness	Kurtosis		Skewness	Kurtosis
y_1	− 0.4	2.9	r_1	1.8	3.3
y_2	0.2	0.2	r_2	2.1	4.8
y_3	− 0.6	0.0	r_3	10.4	109.4
y_4	0.6	0.1	r_4	1.1	− 0.1
y_5	0.5	0.5	r_5	0.3	− 1.1
y_6	− 0.2	0.3	r_6	8.4	78.5
y_7	− 1.1	13.1	r_7	1.2	1.0
			r_8	2.2	8.7
			r_9	10.3	107.6
			r_{10}	− 5.3	44.3
			r_{11}	4.2	26.3
			r_{12}	1.5	7.3

made it impossible to compute either ratios or coordinates. The final usable sample was $n = 110$.

4.4. Descriptive statistics

Table 1 shows standard financial ratios to globally have larger skewness than ilr coordinates, which is generally positive, thus indicating positive asymmetry. Kurtosis serves as an indication of the presence of outliers and also tends to affect ratios to a much larger extent than ilr coordinates. Tables 2 and 3 show Pearson correlations as a measure of redundancy. Some financial ratio correlations are extremely high. No similar phenomenon is observed in ilr coordinates. In order to gather further evidence on redundancy, we submitted the matrices in Tables 2 and 3 to principal component analyses. In the principal component analysis of the ilr coordinate correlation matrix (Table 2), the last component explains 2.4% of the variance. In the principal component analysis of the financial ratio correlation matrix (Table 3), the last four components explain only 0.0%, 0.0%, 0.1% and 1.3% of the variance, respectively, and are thus redundant to a much greater degree. Principal component screeplots are in Fig. 3.

4.5. Cluster analysis

The dendrogram using Ward's method on the unstandardized ilr coordinates y_1 to y_7 (Fig. 4) shows a 3-group solution to be appropriate, which was used as initial solution for the $k -$ means method.

Table 4 shows cluster geometric means for components. For ease of interpretation, the geometric means of x_1 to x_3, x_7 and x_8 have been multiplied row-wise by the appropriate constant so that *total assets* $= x_1 + x_2 + x_3 = 100$. In the same vein, the geometric means of x_4 to x_5, and x_6 have been multiplied by the appropriate constant so that $x_4 + x_5 + x_6 = 100$. These magnitudes can be interpreted on their own as the balance sheet and profit and loss statement of a typical firm in the cluster, with a normalized balance sheet size equal to 100. Standard financial ratios can validly be computed from these geometric means for ease of interpretation of the cluster financial profiles. For instance, the acid test ratio of the representative firm in cluster 1 in the 3-cluster coordinate solution is easily obtained as 26.8/32.9. In this way, traditional ratios can be used to enrich the CoDa cluster interpretation if one wishes to do so.

Table 2
Pearson correlations among ilr coordinates.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_1	1.000	0.072	0.180	− 0.344	− 0.100	− 0.318	− 0.215
y_2	0.072	1.000	0.340	− 0.233	− 0.408	− 0.421	− 0.063
y_3	0.180	0.340	1.000	− 0.470	− 0.549	− 0.629	− 0.047
y_4	− 0.344	− 0.233	− 0.470	1.000	0.768	0.557	0.216
y_5	− 0.100	− 0.408	− 0.549	0.768	1.000	0.534	0.145
y_6	− 0.318	− 0.421	− 0.629	0.557	0.534	1.000	0.249
y_7	− 0.215	− 0.063	− 0.047	0.216	0.145	0.249	1.000

A standard graphical representation of cluster analysis results in CoDa is the geometric means barplot. This plot depicts the log-ratio of the cluster geometric mean of each component over the overall geometric mean of the component. Positive bars show above average components for that particular cluster and negative bars below average components. Fig. 5 shows this plot for the 3-cluster coordinate solution.

At the top of Table 4, Cluster 1 in the 3-cluster solution on ilr coordinates (49 firms) shows a profile of financial distress, with small margin and return on assets (when comparing x_8 to x_7, and their difference to total assets = 100), short term liquidity problems (short term debt –x_6– exceeds quick assets –x_3), very high leverage (low equity –x_4– as compared to liabilities), and very high fixed assets (x_1) as compared to equity. Clusters 2 (49 firms) and 3 (12 firms) show distinct financially healthy profiles. Cluster 2 contains a profile which is characterised by the highest margin and profitability. It also has the best acid test ratio, stemming from an asset composition with a low share of fixed assets and a large share of quick assets (Fig. 5). Cluster 3 shows a profile which reveals the option not to have long term debt (Fig. 5), with a higher turnover and a lower margin than Cluster 2. Both clusters thus reveal different strategic choices rather than a gradation of financial distress.

In the two next blocks of Table 4, we compare results by using the same combination of clustering methods on the standardized ratios r_1 to r_{12}. Since some clusters are built around outliers, in order to find 3 substantial clusters we need to move to the 6-cluster solution. Cluster 1 (47 firms) is reminiscent of cluster 1 in the ilr-coordinate case (38 firms are actually included in this cluster by both methods). Cluster 2 (50 firms) is reminiscent of Cluster 2 in the ilr-coordinate case (33 firms are actually included in this cluster by both methods). Cluster 3 (9 firms) markedly differs from the previous solution and contains an extremely healthy profile, with the highest margin, the highest return on assets, the lowest fixed asset to equity ratio, and the lowest liability to asset ratio. Clusters 1, 2 and 3 thus provide a gradation from more to less financial distress. Clusters 4 to 6 are built around 4 outliers often arising from small values in the denominator of certain ratios (one case has a long term debt and equity to fixed assets ratio $r_3 = 466.7$ and a sales to fixed assets ratio $r_9 = 511.2$; one case has a profit to equity ratio $r_{11} = 3.14$ and a profit to total assets ratio $r_{12} = 0.69$; one case has $r_{11} = 2.65$; and one case has an equity to long term debt ratio $r_6 = 7903.03$).

As expected, when working with financial ratios, either outliers need to be removed from the sample prior to performing a cluster analysis or the number of clusters has to be increased. On the contrary, thanks to the fact that ilr coordinates have much lower kurtosis and are less prone to outliers, there seems to be no need to remove outliers. As expected, working with financial ratios leads to redundancy problems. As repeatedly shown in the literature, given the fact that financial ratios provide redundant and overlapping information on financial distress, clusters often reflect a mere gradation from more to less distress, which is also the case in this application. On the contrary, at least in in this application, working with ilr coordinates has yielded clusters with distinct financial profiles.

4.6. Second application

This application differs from the former in some important respects, including a low tech industry, a much larger sample size and a different clustering method. Having said this, it uses the same data source and selection procedures on firms belonging to NACE code 14.1 (Manufacture of wearing apparel, except fur apparel). The usable sample size is $n = 809$. In the same manner as with the previous application, skewness and kurtosis are much larger for ratios than for coordinates. Six out of twelve ratios have absolute skewness higher

Table 3
Pearson correlations among financial ratios.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
r_1	1.000	0.976	0.074	− 0.326	− 0.657	0.087	− 0.575	− 0.077	0.052	0.207	− 0.083	0.169
r_2	0.976	1.000	0.101	− 0.299	− 0.610	0.110	− 0.538	− 0.220	0.079	0.199	− 0.053	0.181
r_3	0.074	0.101	1.000	− 0.102	− 0.081	0.028	− 0.010	− 0.094	0.997	0.020	− 0.013	0.003
r_4	− 0.326	− 0.299	− 0.102	1.000	0.684	− 0.185	− 0.058	0.036	− 0.116	− 0.231	− 0.103	− 0.281
r_5	− 0.657	− 0.610	− 0.081	0.684	1.000	− 0.219	0.688	0.094	− 0.063	− 0.175	0.173	− 0.170
r_6	0.087	0.110	0.028	− 0.185	− 0.219	1.000	− 0.116	− 0.119	0.032	− 0.016	− 0.060	− 0.078
r_7	− 0.575	− 0.538	− 0.010	− 0.058	0.688	− 0.116	1.000	0.093	0.029	− 0.010	0.339	0.046
r_8	− 0.077	− 0.220	− 0.094	0.036	0.094	− 0.119	0.093	1.000	− 0.102	− 0.137	− 0.237	− 0.235
r_9	0.052	0.079	0.997	− 0.116	− 0.063	0.032	0.029	− 0.102	1.000	0.033	0.031	0.036
r_{10}	0.207	0.199	0.020	− 0.231	− 0.175	− 0.016	− 0.010	− 0.137	0.033	1.000	0.440	0.672
r_{11}	− 0.083	− 0.053	− 0.013	− 0.103	0.173	− 0.060	0.339	− 0.237	0.031	0.440	1.000	0.738
r_{12}	0.169	0.181	0.003	− 0.281	− 0.170	− 0.078	0.046	− 0.235	0.036	0.672	0.738	1.000

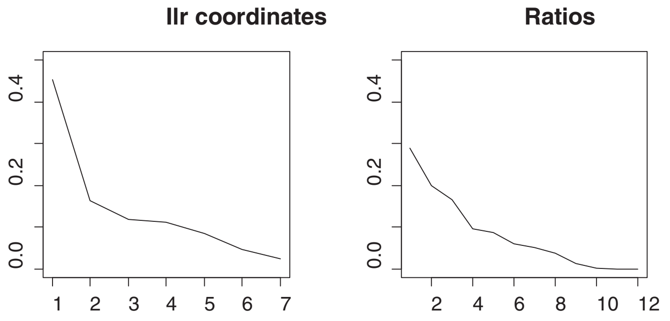


Fig. 3. Principal component analysis screeplots. Proportion of explained variance by each principal component.

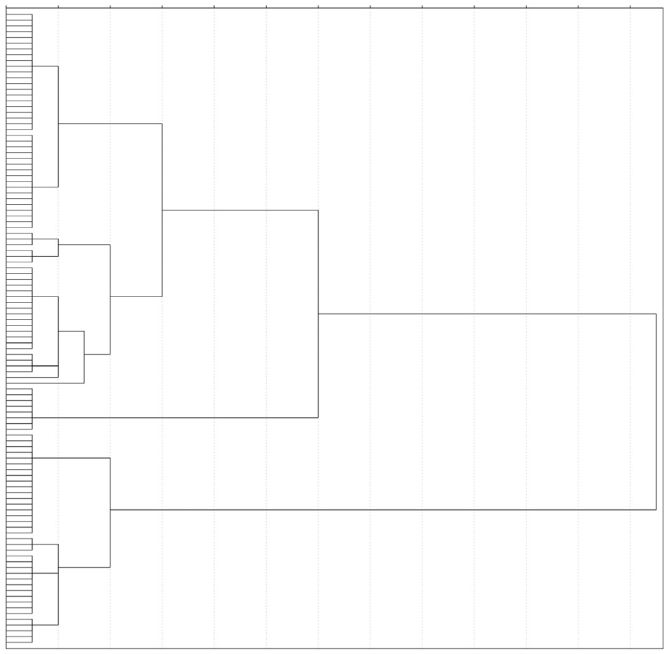


Fig. 4. Ward's method hierarchical clustering dendrogram using ilr coordinates.

than 10 while the highest absolute skewness among coordinates is 3.4. Redundancy also follows the same pattern as in the previous application.

With such a large sample size, hierarchical cluster analysis becomes less practical. We used only the *k*-means method by selecting the best of 100 solutions with random initial clustering. The number of clusters was selected from a scree plot of the within-cluster sums of squares, suggesting a 4-cluster solution with ilr coordinates (Fig. 6). A 7-cluster solution with ratios r_1 to r_{12} provided 4 substantial clusters and three clusters with one or two outliers for comparison (Table 5).

As it is often the case, when clustering highly skewed variables some clusters are very large and some very small (the largest and smallest cluster when using standard financial ratios contain 368 and 10 firms, respectively). Under both solutions, cluster 1 shows a financial distress profile with negative operating profit and debt problems. Under both solution cluster 2 seems to have liquidity problems. Similarity between both solutions ends here. The remaining two clusters in the ilr solution reveal distinct healthy profiles with a strategic orientation to have either low long-term debt or low inventory (Fig. 7). The remaining two clusters in the standard ratio solution are on the one hand a very large cluster with generally healthy firms and a very small cluster with negative operating profit and very low leverage.

5. Discussion, limitations and future research

The proposed financial statement analysis method based on CoDa boils down to computing alternative measures of the relative importance of components in the balance sheet and the profit and loss statement, called ilr coordinates. Ilr coordinates are simply scaled logarithms of ratios of geometric means of components, as arranged in a SBP, which can be represented as a tree diagram. Standard cluster analysis can subsequently be applied with the researchers' favourite clustering method and software. Results can be presented to a statistically non-sophisticated readership just as typical financial statement profiles for each cluster, from which even standard financial ratios can be computed.

Ilr coordinates are constructed in such a way as to reduce their mutual redundancy and are sparing in number. This is so because ilr coordinates define an orthonormal basis (Egozcue et al., 2003). Once *D* strictly positive relevant components of interest in the balance sheet and the profit and loss statement have been selected, $D-1$ coordinates are enough. CoDa can be understood as the manner in which a minimum number of coordinates can be selected which carrys all information about the relative importance of any component to any other. Being a particular case of log-ratios, ilr coordinates tend to be symmetric (Aitchison, 1982, 1986).

Euclidean distance computed using ilr coordinates as data is equivalent to Aitchison's distance, the most commonly used measure of compositional distance. This distance focuses on the relative size of components and is thus coherent with the aim of financial ratio analysis. The selected components in the balance sheet and the profit and loss statement are what matters to distances and not the chosen ilr coordinates. Any SBP constructed from the same set of accounts leads to the same Aitchison's distances. The worries in the literature about the choice of the best ratios, hampered by their mutual redundancy and

Table 4
Cluster labels, size and component geometric means scaled to total assets = 100 (pharmaceutical industry).

			Assets			Equity	Liabilities		Expen.	Sales
	Label	Size	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
3-cluster solution (ilr coordinates)	1	49	63.9	9.3	26.8	34.7	32.4	32.9	72.9	74.4
	2	49	31.7	14.1	54.2	72.4	3.5	24.1	100.7	115.9
	3	12	43.1	14.1	42.8	72.0	0.2	27.8	112.7	119.9
3-cluster solution (ratios)	1	71	54.3	13.2	32.6	43.0	15.3	41.7	95.9	99.0
	2	38	37.2	9.9	52.9	83.2	1.7	15.1	77.8	91.7
	3	1	0.2	3.4	96.5	70.4	0.2	29.4	70.5	77.3
6-cluster solution (ratios)	1	47	58.3	12.0	29.7	33.2	24.7	42.1	85.1	86.0
	2	50	45.0	14.2	40.8	75.3	2.7	22.1	97.3	106.5
	3	9	22.7	5.2	72.2	86.0	3.5	10.4	54.3	72.3
	4	2	16.5	3.7	79.8	12.4	4.7	82.9	108.7	172.8
	5	1	43.5	2.5	54.0	85.6	0.0	14.4	72.2	72.4
	6	1	0.2	3.4	96.5	70.4	0.2	29.4	70.5	77.3

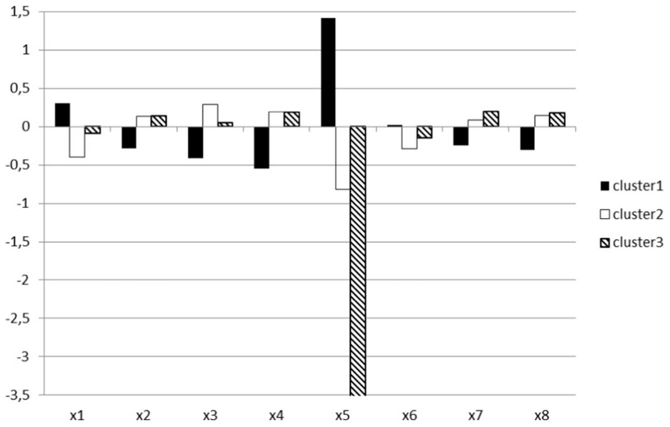


Fig. 5. Geometric means barplot of the 3-cluster solution (ilr coordinates).

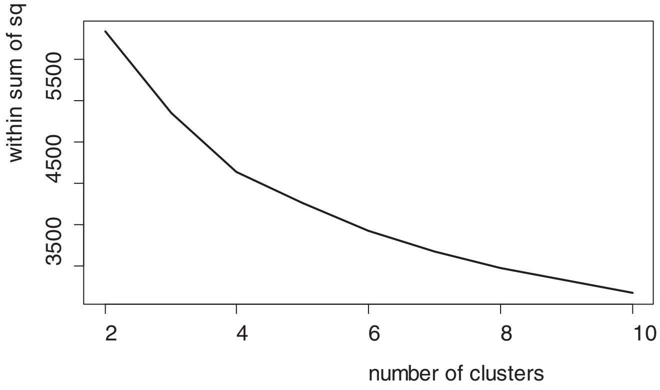


Fig. 6. Scree plot of *k*-means within-cluster sums of squares using ilr coordinates.

thus interchangeability, and by the lack of deep theoretical grounds favouring one ratio over the other, are avoided when using CoDa. Dependence on the choice of numerator and denominator is also solved.

Our applications show our financial statement analysis method to effectively reduce redundancy, outliers, skewness; and to lead to a more interpretable clustering solution in terms of financial profiles rather than a mere gradation of financial distress, to a lower dispersion of cluster sizes, and to no clusters including only a few outliers. The need to perform a careful and partly controversial outlier trimming prior to the analysis seems to have been circumvented, at least in our data sets. Since ilr coordinates reduce redundancy, the need to perform a prior data reduction, for instance, a principal component analysis, is diminished. CoDa can be understood as a manner of keeping analysis steps down to a minimum, as compared to standard ratio analysis.

A commonly mentioned drawback of log-ratio transformations is that data may contain no zero values (e.g. Martín-Fernández, Palarea-Albaladejo, & Olea, 2011). This drawback is largely also present in standard financial ratio analysis, in which zeros are not allowed either, at least for components in the denominator. In the last years CoDa has developed an advanced toolbox for zero treatment and further research can include the application of the standard procedures for the replacement of zeros in CoDa, by extending the methods in Martín-Fernández et al. (2011) and Palarea-Albaladejo and Martín-Fernández (2015) to the financial case. This would make financial statement analysis possible even when some accounts of interest equal zero.

The objective of this article was not to establish the superiority of any clustering algorithm or any criterion for deciding the number of clusters over any other. The approach we suggest concerns the way of treating data and distances prior to clustering and is compatible with any clustering algorithm supporting Euclidean distance. Having said this, further research can include using alternative clustering methods such as mixture models (e.g., Comas-Cufí, Martín-Fernández, & Mateu-Figueras, 2016; Ferrer-Rosell, Coenders, and Martínez-Garcia, 2016), or fuzzy clustering (Palarea-Albaladejo, Martín-Fernández, & Soto, 2012).

Our analysis approach makes it possible to include any other positive magnitude whose size one wishes to compare with financial accounts in relative terms, which also deserves further research. This holds even for non-monetary magnitudes, such as the number of employees, as is often done when using ratios for managerial purposes and strategy or performance assessment. The required magnitudes can simply be added to the list of selected profit and loss accounts, in any order.

Further research can also extend the proposed CoDa approach to predict failure, bankruptcy, distress, survival time, stock market returns, or other variables, as in the research stream started by Altman (1968). Ilr coordinates just need to be included as predictors in the chosen standard statistical or econometric model (Coenders, Martín-Fernández, & Ferrer-Rosell, in press). No serious outlier, non-linearity and collinearity problems are expected.

Last but not least, alternative CoDa methods have been developed in order to take into account not only composition but also firm size (Coenders et al., in press; Ferrer-Rosell, Coenders, Mateu-Figueras et al., 2016; Pawlowsky-Glahn, Egozcue, & Lovell, 2015). Further research in this arena is promising in case researchers believe that proportionality among accounts does not tell the whole story in financial statement analysis.

Table 5
Cluster labels, size and component geometric means scaled to total assets = 100 (wearing apparel industry).

			Assets			Equity	Liabilities		Expen.	Sales
	Label	Size	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
4-cluster solution (ilr coordinates)	1	255	51.0	35.3	13.8	30.0	35.4	34.7	59.6	57.2
	2	310	9.3	51.5	39.2	24.7	16.0	59.3	151.2	156.1
	3	129	18.0	25.4	56.6	57.7	0.4	41.9	144.5	148.1
	4	115	39.3	2.8	57.9	43.1	19.3	37.6	133.0	141.0
7-cluster solution (ratios)	1	152	38.2	34.3	27.4	16.7	59.2	24.2	101.1	94.4
	2	368	20.5	43.6	35.9	19.5	8.5	72.0	159.8	165.5
	3	275	29.1	21.7	49.2	67.7	6.0	26.3	131.1	137.1
	4	10	44.5	9.2	46.3	89.1	6.6	4.3	18.2	14.2
	5	1	0.0	56.0	44.0	10.7	79.4	9.9	119.3	122.3
	6	2	25.3	71.3	3.4	2.3	32.5	65.3	1.4	0.3
	7	1	34.3	42.0	23.8	87.9	0.0	12.1	68.6	63.5

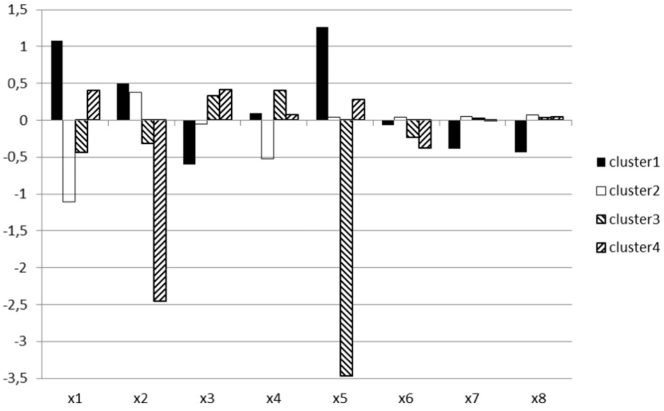


Fig. 7. Geometric means barplot of the 4-cluster solution (ilr coordinates).

Acknowledgements

This work was supported by the Catalan Autonomous Government [grant number 2014SGR551], the Spanish Health Ministry [grant number CB06/02/1002], FEDER/Spanish Economy and Competitiveness Ministry [grant number MTM2015-65016-C2-1-R], and the University of Girona [grant numbers MPCUdG2016/069, MPCUdG2016/098, and MPCUdG2016/107]. The authors want to thank Pepus Daunis-i-Estadella, Oriol Amat-i-Salas, Raffaele Manini, Josep M. Argilés-Bosch, and the members of the research group on statistics and compositional data analysis and of the Girona statistics seminar, for their very helpful comments to earlier versions of this manuscript. Funding sources have had no involvement in the study design, in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. Authors declare not to have used any copyrighted material from other sources (including the Internet) and not to have any competing interests.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B: Methodological, 44*(2), 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika, 70*(1), 57–65. http://dx.doi.org/10.1093/biomet/70.1.57.

Aitchison, J. (1986). The statistical analysis of compositional data. *Monographs on statistics and applied probability*. London: Chapman and Hall.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology, 24*(4), 365–379. http://dx.doi.org/10.1007/BF00891269.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology, 32*(3), 271–275. http://dx.doi.org/10.1023/A:1007529726302.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Beverly Hills, CA: Sage.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Amat, O., Manini, R., & Antón Renart, M. (2017). Credit concession through credit scoring: Analysis and application proposal. *Intangible Capital, 13*(1), 51–70. http://dx.doi.org/10.3926/ic.903.

Azevedo Rodrigues, L., Daunis-i-Estadella, J., Mateu-Flgueras, G., & Thió-Henestrosa, S. (2011). Flying in compositional morphospaces: Evolution of limb proportions in flying vertebrates. In V. Pawlowsky-Glahn, & A. Buccianti (Eds.). *Compositional data analysis. Theory and applications* (pp. 235–254). New York, NY: Wiley.

Barceló-Vidal, C., & Martín-Fernández, J. A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics, 45*(4), 57–71. http://dx.doi.org/10.17713/ajs.v45i4.142.

Barnes, P. (1987). The analysis and use of financial ratios: A review article. *Journal of Business Finance & Accounting, 14*(4), 449–461. http://dx.doi.org/10.1111/j.1468-5957.1987.tb00106.x.

Batista-Foguet, J. M., Ferrer-Rosell, B., Serlavós, R., Coenders, G., & Boyatzis, R. E. (2015). An alternative approach to analyze ipsative data. Revisiting experiential learning theory. *Frontiers in Psychology, 6*(1742), 1–10. http://dx.doi.org/10.3389/fpsyg.2015.01742.

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R.* Berlin: Springer.

Buijink, W., & Jegers, M. (1986). Cross-sectional distributional properties of financial ratios in Belgian manufacturing industries: Aggregation effects and persistence over time. *Journal of Business Finance & Accounting, 13*(3), 337–362. http://dx.doi.org/10.1111/j.1468-5957.1986.tb00501.x.

Campa, D. (2015). The impact of SME's pre-bankruptcy financial distress on earnings management tools. *International Review of Financial Analysis, 42*, 222–234. http://dx.doi.org/10.1016/j.irfa.2015.07.004.

Chen, K. H., & Shimerda, T. A. (1981). An empirical analysis of useful financial ratios. *Financial Management, 10*(1), 51–60.

Coenders, G., Martín-Fernández, J. A., & Ferrer-Rosell, B. (2017). When relative and absolute information matter. Compositional predictor with a total in generalized linear models. *Statistical Modelling*. http://dx.doi.org/10.1177/1471082X17710398 (in press).

Comas-Cufí, M., Martín-Fernández, J. A., & Mateu-Figueras, G. (2016). Log-ratio methods in mixture models for compositional data sets. *SORT-Statistics and Operations Research Transactions, 40*(2), 349–374.

Cowen, S. S., & Hoffer, J. A. (1982). Usefulness of financial ratios in a single industry. *Journal of Business Research, 10*(1), 103–118. http://dx.doi.org/10.1016/0148-2963(82)90020-0.

Dahlstedt, R., Salmi, T., Luoma, M., & Laakkonen, A. (1994). On the usefulness of standard industrial classifications in comparative financial statement analysis. *European Journal of Operational Research, 79*(2), 230–238. http://dx.doi.org/10.1016/0377-2217(94)90354-9.

Davis, B. C., Hmieleski, K. M., Webb, J. W., & Coombs, J. E. (2017). Funders' positive affective reactions to entrepreneurs' crowdfunding pitches: The influence of perceived product creativity and entrepreneurial passion. *Journal of Business Venturing, 32*(1), 90–106. http://dx.doi.org/10.1016/j.jbusvent.2016.10.006.

Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review, 51*(1), 90–96.

Dimitropoulos, P. E., Asteriou, D., & Koumanakos, E. (2010). The relevance of earnings and cash flows in a heavily regulated industry: Evidence from the Greek banking sector. *Advances in Accounting, 26*(2), 290–303. http://dx.doi.org/10.1016/j.adiac.2010.08.005.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35*(3), 279–300. http://dx.doi.org/10.1023/A:1023818214614.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis.* Chischester: Wiley.

Ezzamel, M., & Mar-Molinero, C. (1990). The distributional properties of financial ratios in UK manufacturing companies. *Journal of Business Finance & Accounting, 17*(1), 1–29. http://dx.doi.org/10.1111/j.1468-5957.1990.tb00547.x.

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. http://dx.doi.org/10.14254/2071-789X.2016/9-4/16.

Ferrer-Rosell, B., & Coenders, G. (2017). Destinations and crisis. Profiling tourists' budget share from 2006 to 2012. *Journal of Destination Marketing & Management*. http://dx.doi.org/10.1016/j.jdmm.2016.07.002 (in press).

Ferrer-Rosell, B., Coenders, G., & Martínez-Garcia, E. (2015). Determinants in tourist expenditure composition- the role of airline types. *Tourism Economics, 21*(1), 9–32. http://dx.doi.org/10.5367/te.2014.0434.

Ferrer-Rosell, B., Coenders, G., & Martínez-Garcia, E. (2016). Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. *Tourism Analysis, 21*(6), 589–602. http://dx.doi.org/10.3727/108354216X14713487283075.

Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G., & Pawlowsky-Glahn, V. (2016). Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics, 22*(2), 269–291. http://dx.doi.org/10.5367/te.2016.0548.

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.

Ganesalingam, S., & Kumar, K. (2001). Detection of financial distress via multivariate statistical analysis. *Managerial Finance, 27*(4), 45–55. http://dx.doi.org/10.1108/03074350110767132.

Gupta, M. C., & Huefner, R. J. (1972). A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research, 10*(1), 77–95.

Hingley, P. (2017). Forecasting patent filings at the European Patent Office (EPO) using compositional data analysis techniques. In K. Hron, & R. Tolosana-Delgado (Eds.). *Proceedings of the seventh international workshop on compositional data analysis CoDaWork 2017* (pp. 97–106). CoDa-Association. Downloaded 10-8-2017 from In: *http://www.compositionaldata.com/codawork2017/proceedings/ProceedingsBook2017_May30.pdf*.

Ho, C. T., & Wu, Y. S. (2006). Benchmarking performance indicators for banks. *Benchmarking: An International Journal, 13*(1/2), 147–159. http://dx.doi.org/10.1108/14635770610644646.

Jiang, H., Habib, A., & Zhou, D. (2015). Accounting restatements and audit quality in China. *Advances in Accounting, 31*(1), 125–135. http://dx.doi.org/10.1016/j.adiac.2015.03.014.

El 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. http://dx.doi.org/10.1016/j.irfa.2015.11.009.

Kane, G. D., Richardson, F. M., & Meade, N. L. (1998). Rank transformations and the prediction of corporate failure. *Contemporary Accounting Research, 15*(2), 145.

Kenett, R. S., Martín-Fernandez, J. A., & Vives-Mestres, M. (2017). Association rules and compositional data analysis: Implications to big data. In K. Hron, & R. Tolosana-Delgado (Eds.). *Proceedings of the seventh international workshop on compositional data analysis CoDaWork 2017* (pp. 107–115). CoDa-Association. Downloaded 10-8-2017 from In: *http://www.compositionaldata.com/codawork2017/proceedings/ProceedingsBook2017_May30.pdf*.

Lev, B., & Sunder, S. (1979). Methodological issues in the use of financial ratios. *Journal of Accounting and Economics, 1*(3), 187–210. http://dx.doi.org/10.1016/0165-4101(79)90007-7.

Linares-Mustarós, S., Farreras-Noguer, M. A., Ferrer-i-Comalat, J. C., & Rabaseda-Tarrés, J. (2012). New sectorial financial ratio. The liquidity return ratio. *Cuadernos del CIMBAGE, 15*, 57–72.

Lueg, R., Punda, P., & Burkert, 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. http://dx.doi.org/10.1016/j.adiac.2014.03.002.

Lukason, O., & Laitinen, E. K. (2016). Failure processes of old manufacturing firms in different European countries. *Investment Management and Financial Innovations, 13*(2), 310–321.

Luptak, M., Boda, D., & Szucs, G. (2016). Profitability and capital structure: An empirical study of French and Hungarian wine producers in 2004-2013. *Business Systems Research Journal, 7*(1), 89–103. http://dx.doi.org/10.1515/bsrj-2016-0007.

Mar Molinero, C., Apellaniz Gomez, P., & Serrano Cinca, C. (1996). A multivariate study of Spanish bond ratings. *Omega, 24*(4), 451–462. http://dx.doi.org/10.1016/0305-0483(96)00008-4.

Martikainen, T., Perttunen, J., Yli-Olli, P., & Gunasekaran, A. (1995). Financial ratio distribution irregularities: Implications for ratio classification. *European Journal of Operational Research, 80*(1), 34–44. http://dx.doi.org/10.1016/0377-2217(93)E0134-J.

Martín, M. C. (1998). Performance of eight dissimilarity coefficients to cluster a compositional data set. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, & Y. Baba (Eds.). *Data science, classification, and related methods* (pp. 162–169). Tokyo: Springer.

Martín-Fernández, J. A., Palarea-Albaladejo, J., & Olea, R. A. (2011). Dealing with zeros. In V. Pawlowsky-Glahn, & A. Buccianti (Eds.). *Compositional data analysis. Theory and applications* (pp. 47–62). New York, NY: Wiley.

Martín-Oliver, A., Ruano, S., & Salas-Fumás, V. (2017). The fall of Spanish cajas: Lessons of ownership and governance for banks. *Journal of Financial Stability*. http://dx.doi.org/10.1016/j.jfs.2017.02.004.

Mateu-Figueras, G., Daunis-i-Estadella, J., Coenders, G., Ferrer-Rosell, B., Serlavós, R., & Batista-Foguet, J. M. (2016). Exploring the relationship between two compositions using canonical correlation analysis. *Metodološki Zvezki, 13*(2), 131–150.

Mateu-Figueras, G., Pawlowsky-Glahn, V., & Egozcue, J. J. (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn, & A. Buccianti (Eds.). *Compositional data analysis. Theory and applications* (pp. 31–42). New York, NY: Wiley.

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. http://dx.doi.org/10.1006/bare.1999.0120.

Momeni, M., Mohseni, M., & Soofi, M. (2015). Clustering stock market companies via k-means algorithm. *Kuwait Chapter of the Arabian Journal of Business and Management Review, 4*(5), 1–10.

Morais, J., Thomas-Agnan, C., & Simioni, M. (2017). Using compositional and Dirichlet models for market-share regression. *Toulouse School of Economics Working Papers, 17*(804), 1–21. Downloaded 17-7-2017 from https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2017/wp_tse_804.pdf.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. *Expert Systems with Applications, 40*(15), 5932–5944. http://dx.doi.org/10.1016/j.eswa.2013.05.022.

Ortells, R., Egozcue, J. J., Ortego, M. I., & Garola, A. (2016). Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indices. In J. A. Martín-Fernández, & S. Thió-Henestrosa (Vol. Eds.), *Compositional data analysis. Springer proceedings in mathematics & statistics*. *Vol. 187*. *Compositional data analysis. Springer proceedings in mathematics statistics* (pp. 145–166). Cham, CH: Springer.

Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*. *Vol. 143*. *Chemometrics and Intelligent Laboratory Systems* (pp. 85–96). . http://dx.doi.org/10.1016/j.chemolab.2015.02.019.

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*(2), 144–169. http://dx.doi.org/10.1007/s00357-012-9105-4.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis. Theory and applications.* New York, NY: Wiley.

Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015). Tools for compositional data with a total. *Statistical Modelling, 15*(2), 175–190. http://dx.doi.org/10.1177/1471082X14535526.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data.* Chichester, UK: Wiley.

Pindado, J., & Rodrigues, L. F. (2004). Parsimonious models of financial insolvency in small companies. *Small Business Economics, 22*(1), 51–66. http://dx.doi.org/10.1023/B:SBEJ.0000011572.14143.be.

Pohlman, R. A., & Hollinger, R. D. (1981). Information redundancy in sets of financial ratios. *Journal of Business Finance & Accounting, 8*(4), 511–528. http://dx.doi.org/10.1111/j.1468-5957.1981.tb00832.x.

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. http://dx.doi.org/10.1016/j.jclepro.2016.05.180.

Serrano Cinca, C. (1998). From financial information to strategic groups: A self organising neural network approach. *Journal of Forecasting, 17*(4–5), 415–428. http://dx.doi.org/10.1002/(SICI)1099-131X(1998090)17:5/6<415::AID-FOR705>3.0.CO;2-X.

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. http://dx.doi.org/10.1353/jda.2016.0015.

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. http://dx.doi.org/10.1111/j.1468-5957.1987.tb00108.x.

Sudarsanam, P. S., & Taffler, R. J. (1995). Financial ratio proportionality and intertemporal stability: An empirical analysis. *Journal of Banking & Finance, 19*(1), 45–60. http://dx.doi.org/10.1016/0378-4266(94)00044-4.

Thió-Henestrosa, S., & Martín-Fernández, J. A. (2005). Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology, 37*(7), 773–793. http://dx.doi.org/10.1007/s11004-005-7379-3.

Trussel, J. M., & Parsons, L. M. (2007). Financial reporting factors affecting donations to charitable organizations. *Advances in Accounting, 23*, 263–285. http://dx.doi.org/10.1016/S0882-6110(07)23010-X.

Vives-Mestres, M., Daunis-i-Estadella, J., & Martín-Fernández, J. A. (2014). Out-of-control signals in three-part compositional T^2 control chart. *Quality and Reliability Engineering International, 30*(3), 337–346. http://dx.doi.org/10.1002/qre.1583.

Vives-Mestres, M., Daunis-i-Estadella, J., & Martín-Fernández, J. A. (2016). Signal interpretation in Hotelling's T^2 control chart for compositional data. *IIE Transactions, 48*(7), 661–672. http://dx.doi.org/10.1080/0740817X.2015.1125042.

Vives-Mestres, M., Martín-Fernández, J. A., & Kenett, R. (2016). Compositional data methods in customer survey analysis. *Quality and Reliability Engineering International, 32*(6), 2115–2125. http://dx.doi.org/10.1002/qre.2029.

Voulgaris, F., Doumpos, M., & Zopounidis, C. (2000). On the evaluation of Greek industrial SME's performance via multicriteria analysis of financial ratios. *Small Business Economics, 15*(2), 127–136. http://dx.doi.org/10.1023/A:1008159408904.

Watson, C. J. (1990). Multivariate distributional properties, outliers, and transformation of financial ratios. *The Accounting Review, 65*(3), 682–695.

Willer do Prado, J., de Castro Alcântara, V., de Melo Carvalho, F., Carvalho Vieira, K., Cruz Machado, L. K., & Flávio Tonelli, D. (2016). Multivariate analysis of credit risk and bankruptcy research data: A bibliometric study involving different knowledge fields (1968–2014). *Scientometrics, 106*(3), 1007–1029. http://dx.doi.org/10.1007/s11192-015-1829-6.

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. http://dx.doi.org/10.5296/ajfa.v6i1.5314.

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. http://dx.doi.org/10.1162/ADEV_a_00050.

Yoshino, N., Taghizadeh-Hesary, F., Charoensivakorn, P., & Niraula, B. (2016). Small and medium-sized enterprise (SME) credit risk analysis using bank lending data: An analysis of Thai SMEs. *Journal of Comparative Asian Development, 15*(3), 383–406. http://dx.doi.org/10.1080/15339114.2016.1233821.