Received: 28 November 2020 | Revised: 7 March 2021 | Accepted: 1 April 2021

DOI: 10.1002/qre.2910

SPECIAL ISSUE ARTICLE

WILEY

Measurement, selection, and visualization of association rules: A compositional data perspective

A Compositional Data perspective on Association Rules

Marina Vives-Mestres[1,2] | **Ron S. Kenett**[3] | **Santiago Thió-Henestrosa**[1] | **Josep Antoni Martín-Fernández**[1]

[1] Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, POLITÈCNICA IV, Campus Montilivi, Girona 17003, Spain

[2] Clinical Statistics, Curelator Inc., 210 Broadway #201, Cambridge, MA 02139, USA

[3] KPA Group and Samuel Neaman Institute, Technion, Raanana, Israel

Correspondence
Marina Vives-Mestres, Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain
Email: marina.vives@imae.udg.edu

Abstract

Association rule mining is a powerful data analytic technique used for extracting information from transaction databases with a collection of itemsets. The aim is to indicate what item goes with what item (ie, an association rule) in a set of collected transactions. It is extensively used in text analytics of text records or social media. Here we use Compositional Data analysis (CoDa) techniques to generate new visualizations and insights from association rule mining. These CoDa methods show the relationship between itemsets, their strength, and direction of dependency. Moreover, after expressing each association rule as a contingency table, we discuss two statistical tests to guide identification of the relevant rules by analyzing the relative importance of the elements of the table. As an example, we use these visualizations and statistical tests for investigating the association of negative mood emotions to various types of headache/migraine events. Data for those analysis comes from N1-Headache™, a digital platform where individual users record attacks and symptoms as well as their daily exposure to a list of potential factors.

KEYWORDS

Aitchison geometry, association rule, independence test, measure of interestingness, odds ratio test, simplex representation

1 | INTRODUCTION

A semantic database is formed by attributes and transactions. The attributes are binary variables $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_n\}$ called items; and the transactions are the row vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. For example, in web clickstream analysis, the web pages visited are items in a web session (a transaction). In market basket analysis, a transaction is a single visit of a customer to the supermarket and the attributes are the list of products or items bought. Given $\mathbf{A}, \mathbf{B} \subseteq \mathbf{I}$ two itemsets (sets of items) with $\mathbf{A} \cap \mathbf{B} = \emptyset$, an association rule (AR) is an implication of the form $\{\mathbf{A} \Rightarrow \mathbf{B}\}$. Here, the itemsets $\mathbf{A}$ and $\mathbf{B}$ are, respectively, called the antecedent or left-hand-side (LHS) itemset and consequent or right-hand-side (RHS) itemset. This expresses a relationship

TABLE 1 *AR* contingency table (**T**) for the *AR* {A ⇒ B}

	B	**^{c}B**
A	x_1	x_2
^{c}A	x_3	x_4

of the type IF THEN and does not imply a timed sequence or a causality link. The AR {onions, potatoes}⇒{burger} is a popular example in market basket analysis. The typical AR analysis deals with binary variables however continuous rules can be also defined: {age > 25} ⇒ {total purchase > 50€}. Using AR mining one can detect and extract useful information from for example, unstructured semantic data commonly organized in large media, operational and customer relations databases.[1] Applications of AR mining are found in a wide range of fields such as improving the quality of production processes,[2] defect detection,[3] or health surveillance.[4,5] The general purpose of *AR* mining is to discover the associations between items for predicting future transactions. Our approach assumes independence between transactions.[6] When independence cannot be assumed, other techniques such as sequential pattern analysis may be applied.[7,8]

In identifying ARs worth acting on, one applies measures of association, also called "measures of interestingness," that provide prioritized sorted lists of ARs. Let {**A** ⇒ **B**} be the AR of interest. Let x_1 be the support (relative frequency of occurrence) of both **A** and **B**; x_2 the support of only **A**; x_3 the support of only **B**; and x_4 the relative frequency of transactions where neither **A** nor **B** occur. In other words, let n_k be the number of transactions which satisfy the conditions in x_k, $k = 1, \ldots, 4$, where the total number of transactions is $\Sigma n_k = m$, and $x_k = n_k/m$. Table 1 shows that x_k, $k = 1, \ldots, 4$, respectively, estimates P(**A** ∩ **B**), P(**A** ∩ ^{c}**B**), P(^{c}**A** ∩ **B**), P(^{c}**A** ∩ ^{c}**B**).

We present below six measures of interestingness, implemented in the "arules" R package[9]:

- support{**A** ⇒ **B**} $= n_1/m = x_1$, where n_1 is the number of transactions verifying the rule, informs of the proportion of transactions that verify the AR and it is an unbiased and consistent estimator.[6]
- confidence{**A** ⇒ **B**} = support{**A** ⇒ **B**}/ support{**A**} $= x_1/(x_1 + x_2)$, where support{**A**} is the relative frequency of transactions containing the antecedent. It can be interpreted as an asymptotically unbiased and consistent estimator of a conditional probability.[6]
- lift{**A** ⇒ **B**} = confidence{**A** ⇒ **B**}/support{**B**} $= x_1/[(x_1 + x_2)\cdot(x_1 + x_3)]$. It can be interpreted as a deviation under independence of the itemsets.[6,10] When lift is smaller (greater) than 1, the knowledge that **A** holds causes a negative (positive) effect on the probability of **B**. For lift = 1, there is no effect, that is, there is no association between the itemsets.
- *RLD*{**A** ⇒ **B**}, the Relative Linkage Disequilibrium[11] captures the level of dependence of the *AR* by measuring the relative Euclidean distance of the *AR* from its linear projection on a surface with lift = 1.
- $OR(AR)$ = odds(**B**/**A**)/odds(**B**/^{c}**A**) $= (x_1x_4)/(x_2x_3)$, (odds ratio) described as a measure of interestingness.[12] The value $OR(AR) = 1$ indicates independence of itemsets, $OR(AR) > 1$ a positive effect, and $OR(AR) < 1$ a negative effect. It is an unnormalized measure that ranges between 0 and $+\infty$.
- Yule′s Q $(AR) = \frac{x_1x_4 - x_2x_3}{x_1x_4 + x_2x_3} = OR^*\ (AR)$ is a normalized version of the *OR* through the transformation function $(OR-1)/(OR+1)$ that ranges between −1 and +1.

Lift, *RLD*, *OR* and Yule's measures of interestingness can be classified as measures where the "interest" is expressed by dependence,[6] that is, one is measuring the association between the antecedent (LHS) and consequent (RHS). In general, an interestingness measure M should satisfy three key properties[12]:

- P1: $M = 0$ if A and B are statistically independent;
- P2: M monotonically increases with $P(A \cap B)$ when $P(A)$ and $P(B)$ remain the same;
- P3: M monotonically decreases with $P(A)$ (or $P(B)$) when the rest of the parameters ($P(A \cap B)$ and $P(B)$ or $P(A)$) remain unchanged.

For example, the Yule's Q measure possesses these three properties and the $OR(AR)$ verifies properties P2 and P3. An analogous discussion can be developed with other measures included in "arules" package in terms of the survey presented in Geng and Hamilton.[13]

Visualization is a key aspect to understand and retain knowledge after *AR* mining. Most common visualizations represent rules by their measures of interestingness and by the items they are made of (**A** and/or **B**). The package "arulesViz"[14] in R includes very good visualization tools including interactive features.

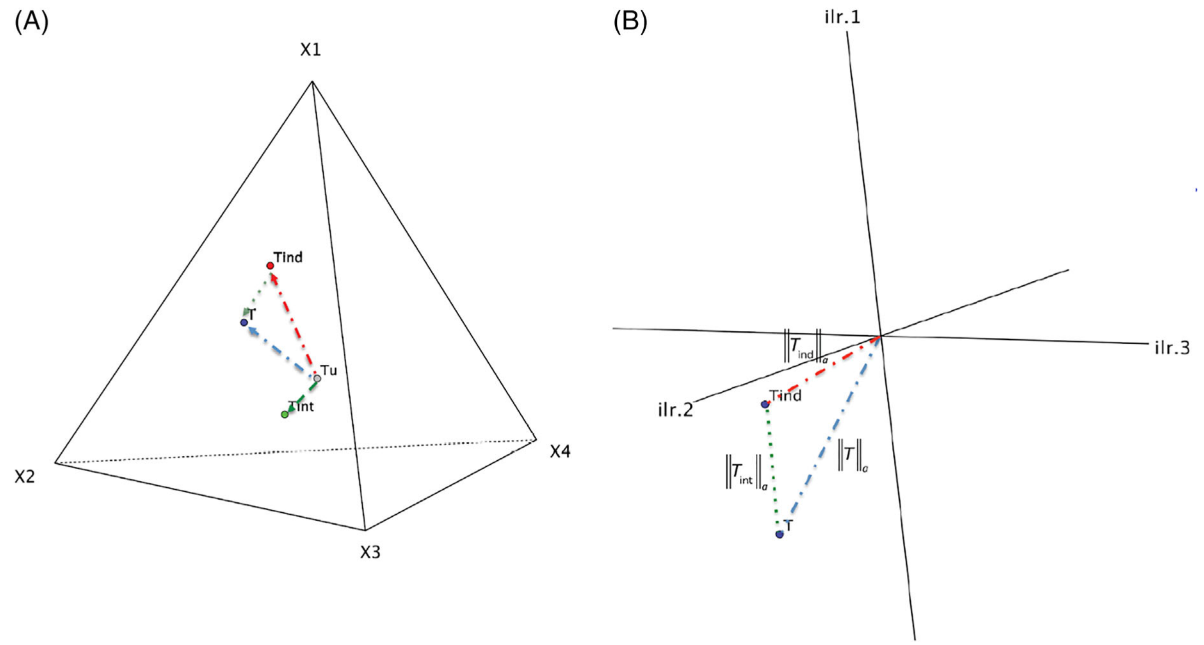

FIGURE 1 Table decomposition: (A) in the simplex $\mathbf{T} = \mathbf{T}_{\text{ind}} \oplus \mathbf{T}_{\text{int}}$, $\mathbf{T}_u$ indicates the center of S^4; (B) in the *ilr*-coordinates space $ilr(\mathbf{T}) = ilr(\mathbf{T}_{\text{ind}}) + ilr(\mathbf{T}_{\text{int}})$. The table $\mathbf{T}$ (blue) is orthogonally projected to table $\mathbf{T}_{\text{ind}}$ (red) in the plane $< ilr_2, ilr_3 >$. The dotted green line represents the norm of the table $\mathbf{T}_{\text{int}}$

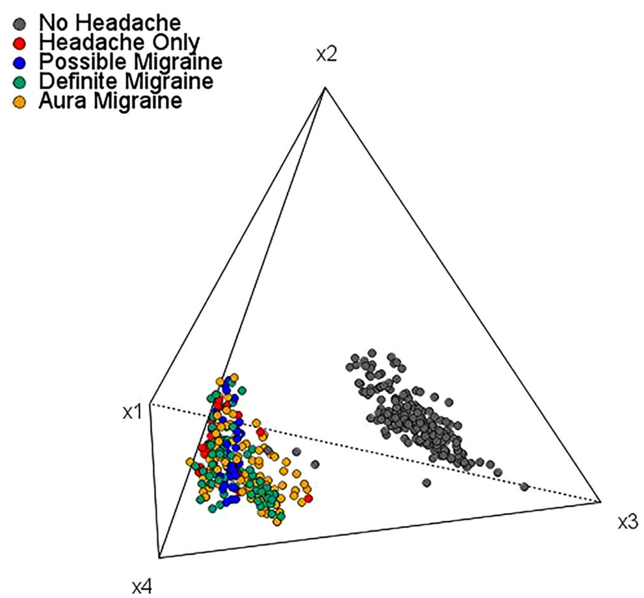

FIGURE 2 Quaternary diagram for significant *AR*s. Observations are colored according the consequent of the *AR*s

This paper is about the application of Compositional Data (CoDa) analysis methods[15] to text or unstructured semantic data. For more on CoDa see http://www.compositionaldata.com or the introductory textbooks.[16,17] The compositional methods used in this work for analyzing *AR* are introduced in Section 2. In Section 3, two tools for *AR* visualization are presented as well as several measures of interestingness from a CoDa perspective and finally an adaptation of two test for significance. A simple example is presented in Section 4.1, where all the concepts introduced are applied and the results are interpreted and in Section 4.2 the analysis of a real large database illustrates its real world application. Finally, in Section 5, some concluding remarks are presented. The programming of the techniques discussed in this article was carried out using the CoDaPack package[18] and the "arules" R package.[9] The artwork was created using both CoDaPack (Figures 1–3) and R[19] (Figure 4).

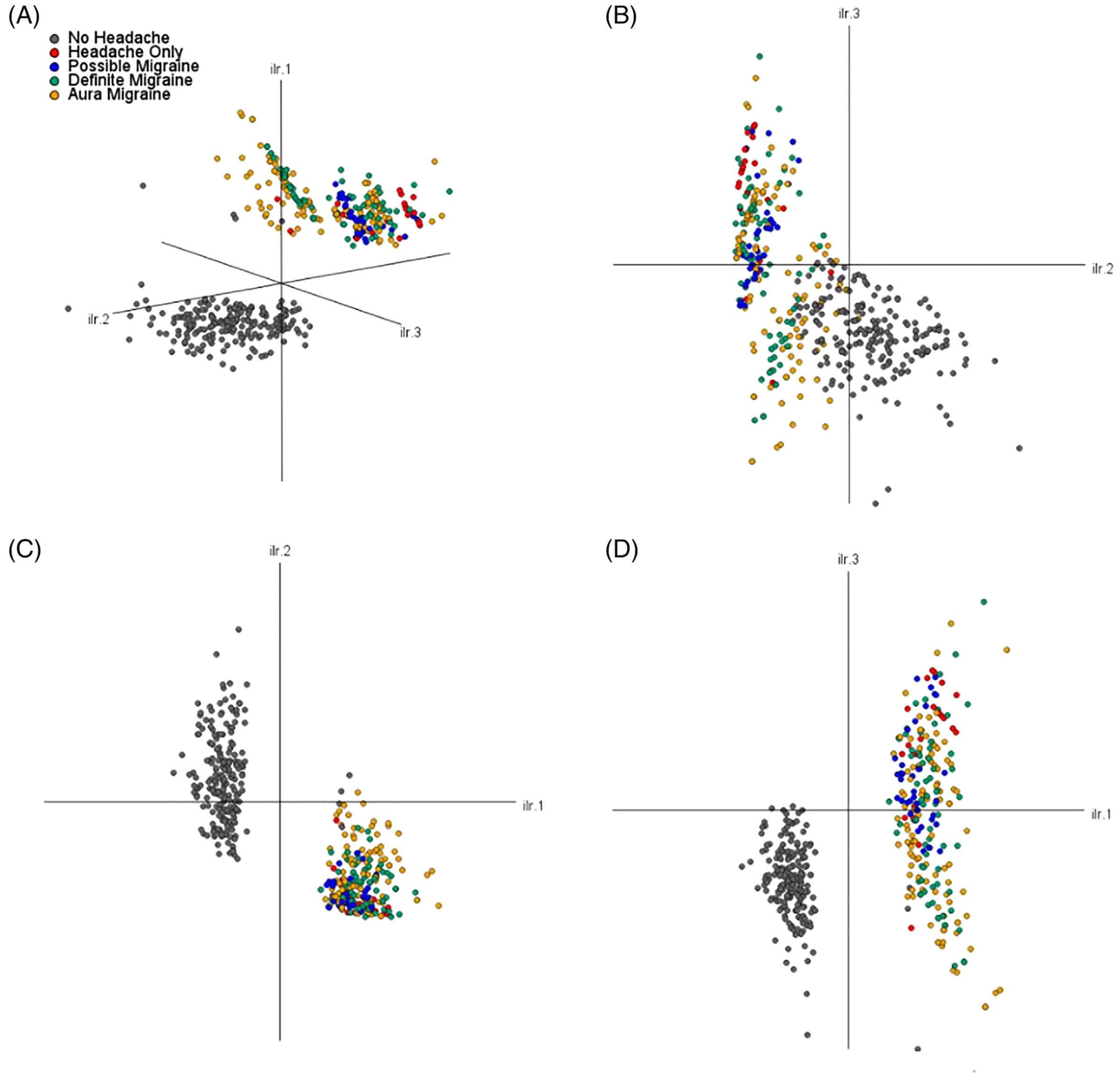

FIGURE 3 3D *ilr* plot for significant *AR*s and three projections. Observations are colored according the consequent of the *AR*s

2 | CODA AND CONTINGENCY TABLES

Each association rule can be expressed as a contingency table **T** (Table 1) and can be represented on the unit simplex.[11] The unit simplex is defined as $S^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D) \in R^D / x_k > 0, \mathrm{k} = 1, \ldots, D \text{ and } \Sigma x_k = 1\}$. Consequently, the 2 × 2 contingency table **T** (Table 1) from an *AR* can be considered as a composition[15] of S^4.

The simplex S^D has its own geometry, different from the unconstrained classical Euclidean geometry.[20] The three basic operations of this particular geometry are: perturbation, powering, and inner product. These basic operations provide a Euclidean structure of dimension D-1 to the simplex space.[20] It allows to analyze CoDa, such as a contingency table, with standard multivariate methods applied on transformed coordinates. The important initial step in implementing standard multivariate techniques to CoDa is to construct orthonormal bases for getting the orthonormal log-ratio (olr) coordinates[21] olr(**x**) of a composition **x**. When one uses a sequential binary partition[22] to construct these olr bases one can express any table **T** in terms of three coordinates, called isometric log-ratio coordinates: $\mathbf{ilr}(\mathrm{T}) = (ilr_1, ilr_2, ilr_3)$. For example, the composition represented by Table 1 (**T**) can be expressed in terms of the *ilr*-coordinates[22,23]:

$$ilr(\mathbf{T}) = \left(\frac{1}{2} \ln \left(\frac{x_1 x_4}{x_2 x_3} \right), \frac{\sqrt{2}}{2} \ln \left(\frac{x_1}{x_4} \right), \frac{\sqrt{2}}{2} \ln \left(\frac{x_2}{x_3} \right) \right). \tag{1}$$

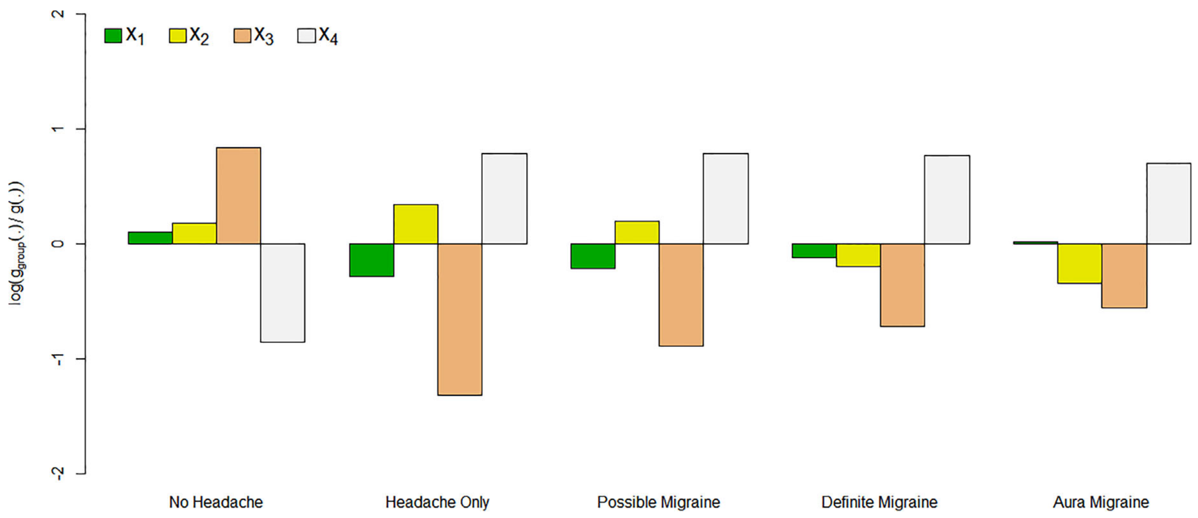

FIGURE 4 Geometric mean barplot representing the logratio of the geometric mean of table **T** for each group ($g_{group}(.)$) divided by the overall geometric mean ($g(.)$)

TABLE 2 Table of independence $\mathbf{T_{ind}}$ of *AR* $\{A \Rightarrow B\}$ (without closure for simplicity)

	B	^{c} **B**
A	$x_1\sqrt{x_2x_3}$	$x_2\sqrt{x_1x_4}$
^{c} **A**	$x_3\sqrt{x_1x_4}$	$x_4\sqrt{x_2x_3}$

One important benefit of such a representation is the ease of interpretation of the *ilr*-coordinates: the three terms indicate the level of dependence in the table and therefore provide measures of dependence. The first coordinate is related to the *OR* measure in that $ilr_1(\mathbf{T}) = 1/2{\cdot}\ln(OR(AR))$ and $OR(AR) = e^{2ilr_1(\mathrm{T})}$. This monotonic functional relation indicates that both values have the same ranking. The second coordinate is about the relationship between the estimates of the probabilities $\mathrm{P}(\mathbf{A} \cap \mathbf{B})$ and $\mathrm{P}(^c\mathbf{A} \cap {}^c\mathbf{B})$. Whereas the third coordinate represents the relationship between $\mathrm{P}(\mathbf{A} \cap {}^c\mathbf{B})$ and $\mathrm{P}(^c\mathbf{A} \cap \mathbf{B})$.

Table **T** can be decomposed[24] into the table of independence ($\mathbf{T}_{\mathrm{ind}}$, shown on Table 2) and the table of interaction ($\mathbf{T}_{\mathrm{int}}$, shown on Table 3) that have the property: $ilr(\mathbf{T}) = ilr(\mathbf{T}_{\mathrm{ind}}) + ilr(\mathbf{T}_{\mathrm{int}})$. To construct those tables we need to define the multiplicative column and row marginal vectors that are $Gc = \mathrm{C}(\sqrt{x_1x_3}, \sqrt{x_2x_4})$ and $Gr = \mathrm{C}(\sqrt{x_1x_2}, \sqrt{x_3x_4})$, respectively, where C means the closure operation $\mathrm{C}\,(\mathbf{x}) = (\frac{x_1}{\Sigma x_k}, \ldots, \frac{x_D}{\Sigma x_k})$. Note that table $\mathbf{T}_{\mathrm{ind}}$ corresponds to independence because $\mathbf{T}_{\mathrm{ind}} = (\mathbf{T}_{\mathrm{ind}})_{\mathrm{ind}}$. The table of interaction ($\mathbf{T}_{\mathrm{int}}$) is derived by applying the perturbation operation to subtract table $\mathbf{T}_{\mathrm{ind}}$ from **T** ($\mathbf{T}_{\mathrm{int}} = \mathbf{T} \ominus \mathbf{T}_{\mathrm{ind}}$). For more on tables decomposition, see Egozcue et al.[24]

Table 4 shows the *ilr*-coordinates of tables **T**, $\mathbf{T}_{\mathrm{ind}}$, and $\mathbf{T}_{\mathrm{int}}$. Let $\|\mathbf{T}\|_a = \|\mathrm{olr}(\mathbf{x})\|$ be the Aitchison norm of a table **T**.[20] Then, $\|\mathrm{T}\|_a^2 = \|\mathrm{T_{ind}}\|_a^2 + \|\mathrm{T_{int}}\|_a^2$, that is, one has a decomposition of the squared Aitchison norm of table **T**, that is invariant under a change of orthonormal basis.

From Equation (1) it can be deduced that zero values in table **T** are not allowed. It is responsibility of the analyst to decide if those zeros are assumed as "true" structural zeros or, instead, they are produced by the sampling design. When the analyst assumes that the zeros are true values, the common decision is to analyze them separately. On the other hand, when the analyst assumes that zeros in a table **T** are a consequence of the sampling design then the zeros can be replaced

TABLE 3 Table of interaction $\mathbf{T}_{\mathrm{int}}$ of *AR* $\{A \Rightarrow B\}$ (without closure for simplicity)

	B	^{c} **B**
A	$1/\sqrt{x_2x_3}$	$1/\sqrt{x_1x_4}$
^{c} **A**	$1/\sqrt{x_1x_4}$	$1/\sqrt{x_2x_3}$

TABLE 4 *ilr*-coordinates of tables **T**, $\mathbf{T_{ind}}$, and $\mathbf{T}_{int}$ of *AR* {$A \Rightarrow B$} using the basis defined in Equation (1)

ilr-coordinates	ilr_1	ilr_2	ilr_3
T	$\frac{1}{2}\ln(\frac{x_1x_4}{x_2x_3})$	$\frac{\sqrt{2}}{2}\ln(\frac{x_1}{x_4})$	$\frac{\sqrt{2}}{2}\ln(\frac{x_2}{x_3})$
$\mathbf{T}_{ind}$	0	$\frac{\sqrt{2}}{2}\ln(\frac{x_1}{x_4})$	$\frac{\sqrt{2}}{2}\ln(\frac{x_2}{x_3})$
$\mathbf{T}_{int}$	$\frac{1}{2}\ln(\frac{x_1x_4}{x_2x_3})$	0	0

by a small value using a Bayesian-multiplicative replacement.[25] Consequently, hereafter, we assume that all values in a table **T** are nonzero.

3 | COMPOSITIONAL DATA AND ASSOCIATION RULES

3.1 | CoDa measures for assessing independence in a table

The simplicial deviance (*SD*) is a measure of independence in a generic table,[24] which, for a table **T** (Table 1), is defined as

$$SD(\mathrm{T}) = \|\mathrm{T}_{int}\|_a^2 = \frac{1}{4}\ln^2\left(\frac{x_1x_4}{x_2x_3}\right) = ilr_1^2\ (\mathrm{T}), \tag{2}$$

where $ilr_1(\mathbf{T})$ is the first *ilr*-coordinate of **T**. We can interpret the strength of the *AR* by the value of the ilr_1 coordinate. In other words, the closer ilr_1 gets to zero, the more independence between itemsets **A** and **B**. More precisely:

- $ilr_1(\mathbf{T}) < 0$: negative repelling effect between itemsets (**A** true, **B** less likely true)
- $ilr_1(\mathbf{T}) = 0$: independence
- $ilr_1(\mathbf{T}) > 0$: positive attractive effect (**A** true, **B** more likely true)

Note that under the standard concept of independence (and being **x** normalized to 1) $x_1 = (x_1 + x_2)(x_1 + x_3)$, or $x_1x_4 - x_2\,x_3 = 0$, or $x_1\,x_4 = x_2\,x_3$, which can be formulated as

$$\frac{x_1x_4}{x_2x_3} = 1 \Leftrightarrow \ln\left(\frac{x_1x_4}{x_2x_3}\right) = 0 \Leftrightarrow ilr_1\ (\mathrm{T}) = 0 \Leftrightarrow SD = 0.$$

However, the decomposition of $\|\mathrm{T}\|_a^2$ suggests that a same *SD* value may be obtained with different sizes of the norm of **T**. Due to that fact, the relative simplicial deviance (*RSD*) was introduced,[24] which normalizes *SD*

$$RSD(\mathrm{T}) = \frac{SD}{\|\mathrm{T}\|_a^2} = \frac{ilr_1^2(\mathrm{T})}{ilr(\mathrm{T})^2}. \tag{3}$$

RSD takes values in an interval [0, 1], with $RSD = 0$ for the independence and $RSD = 1$ for the maximum association; that is, $\mathbf{T} = \mathbf{T}_{int}$, which corresponds to that **T** is purely interaction and the independent part is uniform.

We can combine the benefits of interpretation and fulfillment of the three properties of a measure *M* described in Section 1 by defining the unnormalized compositional measure of association

$$C\ (AR) = ilr_1\ (\mathrm{T}). \tag{4}$$

It is more difficult to interpret the strength of the association because the measure $C(AR)$ takes values in $(-\infty, \infty)$. On the other hand, when $C(AR) = 0$, or when it takes values not significantly different from zero, indicates that **A** and **B** are statistically independent (property P1 or equivalently $\mathbf{T} = \mathbf{T}_{ind}$). Among the number of possibilities to normalize a measure that ranges between $-\infty$ and $+\infty$ as $C(AR)$ (Eq. 4), one can select the hyperbolic tangent function[26] $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$ with the property that:

$$C^*\ (AR) = \tanh(C(AR)) = OR^*\ (AR) = \text{Yule's}\ Q\,(AR).$$

This demonstrates the properties of the Yule's Q measure. Finally note that by its definition, $ilr_1(\mathbf{T})$ verifies the property P1 described in Section 1. Because the unnormalized version of measure $OR(AR)$ verifies properties P2 and P3, then $ilr_1(\mathbf{T})$ also verifies these two properties.[12] On the other hand, by its definition, the measure $SD(AR)$ does not possess these two properties.

3.2 | CoDa-*AR* visualization

A composition in S^4 is commonly visualized in a quaternary diagram: a tetrahedron in which each point, that is, table $\mathbf{T} = (x_1, x_2, x_3, x_4)$, is plotted at a distance x_1 to the face opposite to vertex x_1, a distance x_2 to the face opposite to vertex x_2, a distance x_3 to the face opposite to vertex x_3 and a distance x_4 to the face opposite to vertex x_4. As an example, table $\mathbf{T} = (0.4, 0.35, 0.2, 0.05)$ is represented in Figure 1A. The closer an *AR* lies to a vertex of the tetrahedron the higher the value of that component is in table $\mathbf{T}$. If the *AR* lies near an edge, it means that the two components represented in the edge are the prevailing ones in the table. While if the *AR* is in the center of the tetrahedron, it means that all components of the table are represented in alike proportions. The decomposition of table $\mathbf{T}$ can also be visualized in the quaternary diagram. Figure 1A shows such a decomposition. Importantly, each table $\mathbf{T}$ can also be represented in R^3 by means of their *ilr* coordinates. The *ilr* coordinates of the table represented in Figure 1A are $ilr(\mathbf{T}) = (-0.63, 1.47, 0.40)$ and are shown in Figure 1B. Again we can visualize the decomposition of $ilr(\mathbf{T})$ in Figure 1B into the vector $ilr(\mathbf{T}_{\text{int}})$ (green) and its orthogonal projection to the plane $< ilr_2, ilr_3 >$, the $ilr(\mathbf{T}_{\text{ind}})$ (red).

The two plots on Figure 1 play a much important role when it comes to represent multiple *AR* as will be later seen in Section 4.2. Representing rules in a quaternary diagram allows visualizing the raw data from which the rules are made of and identify trends, patterns, and similarities. Dots inside the quaternary diagram can be colored according to other measures of interestingness or the antecedent/consequent of the rules. The graphical representation of the *ilr* coordinates of a set of rules has the same advantages than the quaternary diagram representation, but has an extra advantage in that the independence plane ($ilr1 = 0$) can easily be identified; the further the dot lies form the independence plane, the stronger the dependence between the consequent and the antecedent.

Other tools developed for visualizing CoDa can also be used for *AR* visualization. As an example the geometric mean barplot[27] (shown in the example on Figure 4) is an option for describing differences between groups, for example, according to the consequent. It shows the log-ratio geometric mean of the group and the whole geometric mean. Large bars in the plot indicate large differences in the means on a specific component in that group with respect to the overall mean.

3.3 | Lift and relative linkage disequilibrium versus $ilr_1(\mathbf{T})$

Given a specific *AR*, lift can be interpreted by a comparingits value with "1": a lift higher than 1 indicates "stickiness" of the precedent and antecedent while a lift lower than 1 indicate "repulsion." Technically, lift measures how similar is the value x_1 to the product of corresponding additive column and row marginal vectors $(x_1 + x_2)(x_1 + x_3)$. Note that table $\mathbf{T}$ components are closed. On the other hand, RLD measures the similarity between the value x_1 and the product $(x_1 + x_2)(x_1 + x_3)$ via the subtraction $\mathrm{D(AR)} = \mathrm{x_1} - (\mathrm{x_1} + \mathrm{x_2})(\mathrm{x_1} + \mathrm{x_3}) = \mathrm{x_1 x_4} - \mathrm{x_2 x_3}$, which measures disequilibrium (D).[11] With no disequilibrium, or independence $\mathrm{D(AR)} = \mathrm{x_1 x_4} - \mathrm{x_2\, x_3} = 0$. Importantly, $D(AR)$ takes values in [–1, 1] and it can be shown that

$$\text{lift (AR)} = 1 + \frac{\mathrm{D(AR)}}{(\mathrm{x_1} + \mathrm{x_2})(\mathrm{x_1} + \mathrm{x_3})}.$$

A value $D(AR) < 0$ indicates a negative repelling effect; $D(AR) = 0$ corresponds to independence; and a positive attraction effect corresponds to $D(AR) > 0$. The definition of $D(AR)$ produces some difficulties and in Kenett and Salini[10] (page 153) it is pointed out that: "… points closer to the edges of the simplex will have intrinsically smaller values of D."

To solve this difficulty, the measure $RLD = D(AR)/D_M$ is proposed,[10] where D_M is the Euclidean distance between the projection on the simplex of table $\mathbf{T}$ and the surface $D(AR) = 0$. *RLD* thus normalizes the location effect of a table, within the simplex space. The *RLD* takes values in an interval [0, 1], with $RLD = 0$ for the independence and $RLD = 1$ for the extreme association detected by the measure $D(AR)$. For examples of *RLD* applications and a simple algorithm for computing *RLD*, see Refs. (10), (11), (28), (29). However, since the Euclidean distance is not coherent with the simplicial geometry,[30] one could also use the Aitchison distance between two compositions **x** and **y**: $d_a(\mathrm{x}, \mathrm{y}) = \|olr(\mathrm{x}) - -olr(\mathrm{y})\|$, which is invariant under a change of basis. This distance evaluates relative changes in the data components.[31]

From the definition of *RLD*, one can easily deduce that tables **T**, where one or more values in the vector **x** are equal to zero, are not interesting to analyze. Indeed, if only one value in the vector **x** is equal to zero then $RLD = 1$, that is, the point **x** takes the maximal distance to the surface $D = 0$. One has the same situation when the pair $\{x_1, x_4\}$ or the pair $\{x_2, x_3\}$ are equal to zero. On the other hand, when the other possible pairs are zero or three values are zero, then $D = 0$, that is, one has independence. However, for the case of three values equal to zero the index *RLD* can be misleading because it suggests that the itemsets **A** and **B** are associated. For example, when **T** in Table 1 is equal to $\mathbf{x} = (0, 1, 0, 0)$ the estimate for $\mathrm{P}(\mathbf{A} \cap {}^c\mathbf{B})$ is 1, suggesting that the antecedent **A** is never followed by the consequent **B**. Hereafter, we assume that all values in a table **T** are nonzero.

3.4 | CoDa measures of significance

Asymptotic confidence intervals for the "support" and "confidence" measures of interestingness can be determined.[6] In this section, we discuss how to determine if an interestingness measure expressed by "dependence,"[6] is statistically different from random noise using a compositional approach. The library "arules" from package R provides a function to find rules in which the antecedent and the consequent significantly depend on each other (ie, are dependent). The function uses the classical chi-squared test and Fisher's exact test for contingency tables. In this work, we consider two simplicial approaches derived from two different sources: first, the adaptation of the classical chi-squared test for contingency tables[24]; and second, a new adaptation of Haldane's test for odds-ratios[32] to evaluate the measure of interestingness significance.

To evaluate the significance of both *SD* and *RSD* measures, Egozcue et al[24] introduced a bootstrap algorithm. For a large database, this procedure is computationally time consuming and still an approximative method. Analyzing the independence in a table **T** is equivalent to testing the significance of the hypothesis H_0: $ilr_1(\mathbf{T}) = 0$ which is equivalent to H_0: $\mathbf{T} = \mathbf{T}_{\mathrm{ind}}$, where $\mathbf{T}_{\mathrm{ind}}$ takes the form given in Table 2.[24] The formula for the chi-squared statistic is

$$\chi^2 = \mathrm{m} \cdot \sum_{\mathrm{k}=1}^{4} \frac{(x_k - g_k)^2}{\mathrm{g_k}}, \tag{5}$$

where m is the number of transactions, (x_1, x_2, x_3, x_4) are the proportions in a table **T** and (g_1, g_2, g_3, g_4) the values in $\mathbf{T}_{\mathrm{ind}}$ (Table 2). The statistic of Equation (5) follows a chi-squared distribution with one degree of freedom ($\chi^2_{0.05,1} = 3.8415$). ARs where the statistic takes values greater than the chi-squared 95% quantile are labeled as significant.

Assuming normality, a 95% confidence interval ($z_{0.025} = 1.96$) for an odds ratio is[31]

$$\left(\exp\left(\ln(OR) - 1.96\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}\right), \exp\left(\ln(OR) + 1.96\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}\right)\right).$$

Here, using the connection between *OR* and $ilr_1(T)$, we propose to adapt this formula to define the corresponding test ($\alpha = 0.05$) for $C(AR)$. With this approach, *AR*s where

$$\left|\frac{2 \cdot ilr_1(T)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}}\right| > 1.96 \tag{6}$$

are considered statistical significant and relevant for the study. In practice, we suggest applying both criteria, in Equations (5) and (6), to discard the ARs that are nonsignificant.

TABLE 5 Tables T, $\mathbf{T}_{\mathrm{ind}}$, and $\mathbf{T}_{\mathrm{int}}$: (a) $\mathbf{T}$ in counts (proportions); (b) $\mathbf{T}_{\mathrm{ind}}$; (c) $\mathbf{T}_{\mathrm{int}}$

(a)		
T	**Cereal**	**Not cereal**
Basketball	2000 (0.4)	1750 (0.35)
Not basketball	1000 (0.2)	250 (0.05)
(b)		
$\mathbf{T}_{\mathrm{ind}}$	**Cereal**	**Not cereal**
Basketball	0.54	0.25
Not basketball	0.14	0.07
(c)		
$\mathbf{T}_{\mathrm{int}}$	**Cereal**	**Not cereal**
Basketball	0.17	0.33
Not basketball	0.33	0.17

4 | EXAMPLES OF APPLICATION

4.1 | A simple example: basketball and cereals

As an example, consider a questionnaire where young people are asked if they like basketball and if they eat cereals for breakfast ($m = 5000$). Table 5 shows the three *AR* tables corresponding to the full data ($\mathbf{T}$), independence ($\mathbf{T}_{\mathrm{ind}}$) and interaction ($\mathbf{T}_{\mathrm{int}}$), and they are plotted in S^4 in Figure 1A and according to their *ilr* coordinates in Figure 1B.

It can be verified that $\mathbf{T} = (0.4, 0.35, 0.2, 0.05) = \mathbf{T}_{\mathrm{ind}} \oplus \mathbf{T}_{\mathrm{int}} = (0.54, 0.25, 0.17, 0.07) \oplus (0.17, 0.33, 0.33, 0.17)$, where "$\oplus$" is the perturbation operation.[15] The vector of *ilr*-coordinates is $ilr(\mathbf{T}) = (-0.63, 1.47, 0.40)$ so $C(AR) = -0.63$ and $C^*(AR) = -0.56$. The negative values of the compositional measures of association correspond to a negative effect, that is, given that a young person likes basketball, it is less likely that he/she eats cereal for breakfast. The positive sign of $ilr_2(\mathbf{x}) = 1.47$ indicates that it is more likely that a young person likes both products than none. Moreover, because $ilr_3(\mathbf{x}) = 0.40$ is positive, we can assume that people that only like one of them, prefer basketball.

The simplicial deviance is equal to $SD = 0.39$ that normalizes to $RSD = 0.14$. When the testing procedure for independence is applied,[24] we obtain both P-values below 0.5×10^{-4}, indicating a significant interaction. Moreover the chi-squared statistic in Equation (5) is 501.6 also indicating a significant dependence and the value from Equation (6) is |–16.1| clearly greater than the threshold values of 1.96 thus again indicating dependence.

4.2 | Application: CoDa-*AR* measure applied to N1-Headache™ data

Migraine is a common disabling disease, affecting approximately 1 billion people worldwide or 11.79% of the population.[33] Migraine is not a "bad headache™"; it can cause severe pain for hours to days and is often accompanied by nausea/vomiting, sensitivity to light, sound, and odors. Aura (bright spots, flashes or wavy, zigzag vision) may occur before or after migraine. N1-Headache™ is an app that enables daily self-monitoring of headache risk factors as well as symptoms, medication, and quality of life (https://n1-headache.com/).

We are interested on understanding how negative mood factors are related to the type of headache events classified according to the International Classification of Headache Disorders 3rd edition (ICHD-3). Each event (day) is classified (from less to more severe) as nonheadache, headache-only, possible migraine, definite migraine or aura migraine. Note that "headache" refers to any type of head pain, including headache-only and all types of migraine. The classification is clinically relevant for both migraine diagnosis and treatment. Detecting mood associations with headache is important because it might help develop early interventions that might help improve patient condition.

In this study, a transaction is a daily questionnaire answered by a user and the attributes are eight negative mood factors (stress, anxiety, irritability, lack of happiness, sadness, angriness, boredom, lack of relaxedness) each answered on a 0-10 scale that have been categorized each into low/high at an individual level according to the individual pattern of response, for example, stress = 5 can be a high value for one individual but it can be a low for another.

There were 462 individuals that answered ninety or more daily questionnaires each and represented 65 929 transactions. For each individual, the rules having as a consequent the type of headache day and as an antecedent the low/high

TABLE 6 Geometric mean of table **T** components (x_i) grouped by consequent

	x_1	x_2	x_3	x_4
No headache	0.18	0.15	0.51	0.12
Headache only	0.12	0.18	0.06	0.60
Possible migraine	0.13	0.16	0.09	0.60
Definite migraine	0.14	0.11	0.11	0.59
Aura migraine	0.16	0.09	0.13	0.55

TABLE 7 Mean (standard deviation) of *ilr* coordinates *SD*, *RSD*, *C*(*AR*) *C**(*AR*) for significant ARs grouped by consequent

	$\text{ilr}_1 = C(AR)$	ilr_2	ilr_3	*SD*	*RSD*	*C**(*AR*)
No headache	−0.67 (0.26)	0.29 (0.52)	−0.85 (0.44)	0.52 (0.24)	0.39 (0.24)	−0.57 (0.20)
Headache only	0.96 (0.21)	−1.14 (0.25)	0.79 (0.78)	0.96 (0.41)	0.30 (0.15)	0.73 (0.09)
Possible migraine	0.85 (0.17)	−1.10 (0.18)	0.39 (0.60)	0.74 (0.29)	0.31 (0.10)	0.68 (0.09)
Definite migraine	1.00 (0.22)	−1.02 (0.25)	−0.02 (0.94)	1.05 (0.46)	0.37 (0.14)	0.75 (0.09)
Aura migraine	1.02 (0.22)	−0.87 (0.37)	−0.23 (0.86)	1.12 (0.63)	0.41 (0.19)	0.75 (0.10)

value of each of the eight negative mood factors were computed. Then redundant rules were removed and a Bonferroni correction was applied to account for multiple testing. Rules being significant using any of the criteria presented in this paper (Equations 5 and 6) were retained.

From 5646 rules identified, 1297 were found to be significant by both methods and 746 only by the chi-square one. After the Bonferroni correction, 439 rules were left significant by any of the two methods and those are the ones we further analyze. Note that the Fisher's exact test for contingency tables detected 199 significant rules, from which 160 are common with the selected ones, and the classical chi-squared test detected 336, from which 306 are common with the selected ones. Both the Fisher's exact test and the chi-squared test were corrected for multiple testing (Bonferroni).

The 439 rules were found on 158 individuals each contributing on average with 2.7 rules ($SD = 3.7$). Retained rules are plotted on the quaternary diagram in Figure 2 and their *ilr*-coordinates in Figure 3 (snapshot of the 3D plot and three projections).

Table 6 shows the geometric mean of table **T** components (x_i) by consequent and Figure 4 shows the geometric mean bar plot that describes the differences between groups. Table 7 shows the *ilr* coordinates for each group (consequent) as well as the summary (mean and standard deviation) of the CoDa measures of interestingness presented in this paper.

Figure 2 shows that nonheadache days (gray) have overall greater values of x_3 compared to the rest of headache days; that is, they tend to occur associated with low negative mood factors. Moreover, most of them (98%) have negative values of ilr_1 (and $C(AR)$) which describes a negative effect, that is, given that a person has high negative mood emotions, it is less likely that he/she has a headache-free day. Nonheadache days also have, on average, a negative ilr_3 value as can be seen on Figure 3, which means that it is more likely to have nonheadache days with low negative mood emotions than headache days with high negative mood emotions. This because individuals in the sample have low migraine frequency; on average 35% of days are headache days.

The normalized compositional measure of association ($C^*(AR)$) has a negative average when the consequent refers to nonheadache days, again indicating that they are more likely found with low negative mood emotions. This measure is positive when there are headache days on the consequent and its value is closer to |1| and have lower variability indicating that all types of headache days are strongly associated with negative mood emotions.

Association rules including headache only and possible migraine on the consequent have very similar tables (see Table 6) and CoDa association measures (see Table 7); that is, the association between those events and negative mood emotions is very similar. Moreover, we observe that as the severity of the headache event increases (from headache only, to possible, then definite and finally aura) the proportion of x_1 and x_3 increase, x_2 decrease, and x_4 remains approximately the same (Figure 4). This leads to a reduction on the negative average ilr_2 (moving toward zero) as the headache severity increases (Table 7) meaning that the more severe is the event the more likely it is to find it associated with high negative mood. Moreover, among the rules in which the consequent is a headache there is a decrease on the ilr_3 value with the increase of the headache severity. This happens because, the proportion allocated into x_2 and x_3 changes its weigh toward

x_3 with the increase of severity. This means that the more severe is the migraine event on the consequent, the more likely is that given a high negative mood there is a headache event.

5 | CONCLUSIONS

An *AR* is associated to a two by two contingency table, which can be analyzed as a composition. The CoDa geometry provides interesting visualization techniques that are needed when a large number of rules are analyzed. Compositions of 2×2 tables are naturally represented in the quaternary diagram (S^4), which allows visualizing a set of *AR* and their overall behavior in a 3D plot. Moreover, compositions can be represented by means of their *ilr* coordinates and we have presented such a transformation that enhances the interpretability of the components. The visualization of the *AR* in terms of their *ilr* coordinates has the advantage that the independence plane can easily be identified. The *ilr* plots are unique features of the CoDa analysis.

We propose here a new compositional measure of interestingness $C(AR)$ and its normalized version $C^*(AR)$. These measures have properties derived from *OR* and Yule's *Q*, respectively. Moreover, two tests are provided to confirm the significance of a compositional measure of interestingness. Significant *AR*s exhibit either repelling or attractive relations between antecedent and consequent. We have also reviewed two compositional measures of independence, *SD* and *RSD*. All them are coherent with the simplicial geometry of the simplex and the sample space of contingency tables corresponding to *AR*. In addition, the relation between these CoDa *AR* measures and other common measures of *AR* facilitates the interpretation of negative and positive effects between itemsets. The principles of coherence and scalability, that are fundamental to CoDa, are especially relevant to *AR* mining. This paper demonstrates how this can be implemented and interpreted.

The N1-Headache$^{\text{TM}}$ application shows the value of a compositional data analysis of association rules. We have used it to extract relevant information from a large dataset by using both effective visualizations of ARs and the use of the statistical tests for identifying ARs different than random. Moreover, we have used the CoDa measures of interestingness to understand the size and sign of the effect of headache events related to level of negative mood emotions.

ACKNOWLEDGMENTS

We would like to acknowledge Curelator Inc. for providing anonymized data. This research has been supported by the Spanish Ministry of Economy, Industry and Competitiveness under the project CODAMET (Ref: RTI2018-095518-B-C21)

ORCID

Marina Vives-Mestres https://orcid.org/0000-0001-8304-257X
Ron S. Kenett https://orcid.org/0000-0003-2315-0477
Santiago Thió-Henestrosa https://orcid.org/0000-0003-2555-6489
Josep Antoni Martín-Fernández https://orcid.org/0000-0003-2366-1592

REFERENCES

1. Agrawal R, Imielienski T, Swami A. Mining association rules between sets of items in large databases. *Proceedings of the Conference on Management of Data*. New York: ACM Press; 1993:207-216.
2. Kamsu-Foguem B, Rigal F, Mauget F. Mining association rules for the quality improvement of the production process. *Expert Syst Appl*. 2013;40(4):1034-1045.
3. Wei-Chou C, Shian-Shyong T, Ching-Yao W. A novel manufacturing defect detection method using association rule mining techniques. *Expert Syst Appl*. 2005;29(4):807-815.
4. Ma L, Tsui FC, Hogan WR, Wagner MM, Ma H. A framework for infection control surveillance using association rules. *AMIA Annual Symposium Proceedings*; 2003: 410-414.
5. Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: a survey. *Artific Intell Rev*. 2017;47:313-340.
6. Weiß CH. Statistical mining of interesting association rules. *Statistics Comput*. 2008;18:185-194.
7. Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episodes in event sequences. *Data Mining Knowl Discov*. 1997;1:259-289.
8. Agrawal R, Srikant R, Mining sequential patterns. In: *11th International Conference on Data Engineering (ICDE '95)*, Taipeh, Taiwan, 1995: 3-14. https://doi.org/10.1109/ICDE.1995.380415
9. Hahsler M, Grun B, Hornik K. arules—a computational environment for mining association rules and frequent item sets. *J Stat Softw*. 2005;14(15):1-25.

10. Kenett RS, Salini S. Measures of association applied to operational risks (Chapter 9). In: Kenett RS, Raanan Y, eds. *Operational Risk Management*. Chichester, UK: John Wiley & Sons, Ltd; 2011.
11. Kenett RS. On an exploratory analysis of contingency tables. *J R Stat Soc Ser D*. 1983;32(4):395-403.
12. Tan PN, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. *Inform Syst*. 2004;29(4):293-313.
13. Geng L, Hamilton HJ. Interestingness measures for data mining: a survey. *ACM Comput Surveys*. 2006;38(3):9.
14. Hahsler M. arulesViz: visualizing association rules with R. *R Journal*. 2017;9(2):163-175.
15. Aitchison J. The statistical analysis of compositional data. *Monographs on Statistics and Applied Probability*. London, UK: Chapman and Hall Ltd; 1986. Reprinted 2003 with additional material by The Blackburn Press, London, UK.
16. Pawlowsky V, Buccianti A, eds. *Compositional Data Analysis: Theory and Applications*. Chichester, UK: Wiley; 2011.
17. van den Boogaart KG, Tolosana-Delgado R. *Analyzing Compositional Data with R*. Berlin: Springer; 2013.
18. Comas-Cufí M, Thió-Henestrosa S, CoDaPack 2.0: a stand-alone, multi-platform compositional software. In: *Proceedings of the 4th International Workshop on Compositional Data Analysis*, Egozcue JJ, Tolosana-Delgado R, Ortego MI, eds. Spain: Sant Feliu de Guíxols; 2011. http://www.compositionaldata.com/codapack.php. (Accessed 26 February 2020)
19. Core Team R. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018. https://www.R-project.org/.
20. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Modeling and Analysis of Compositional Data*. Chichester, UK: John Wiley & Sons; 2015.
21. Martín-Fernández JA. Comments on: Compositional data: the sample space and its structure by Egozcue, J.J. and Pawlowsky-Glahn, V. TEST, 2019; 28(3):653-657.
22. Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. *Mathemat Geol*. 2005;37:795-828.
23. Facevikova K, Hron K, Statistical analysis of compositional 2 × 2 tables. In: *Proceedings of the 5th International Workshop on Compositional Data Analysis*, Hron K, Filzmoser P, Templ M, eds. Austria: Vorau; 2013. ISBN: 978-3-200-03103-6
24. Egozcue JJ, Pawlowsky-Glahn V, Templ M, Hron K. Independence in contingency tables using simplicial geometry. *Commun Stat—Theor M*. 2013;44:3978-3996.
25. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometr Intell Lab Syst*. 2015;143:85-96.
26. Prados F, Boada I, Prats A, et al. Analysis of new diffusion tensor imaging anisotropy measures in the 3P-plot. *J Magn Reson Imag*. 2010;31(6):1435-1444.
27. Martín-Fernandez JA, Daunis-i-Estadella J, Mateu-Figueras G. On the interpretation of differences between groups for compositional data. *SORT*. 2015;2:231-252.
28. Kenett RS, Salini S. Relative linkage disequilibrium in tracking web search patterns. In *Proceedings of CLADAG*, Florence, Italy; 2010.
29. Kenett RS, Salini S. Relative linkage disequilibrium applications to aircraft accidents and operational risks. *T Mach Learn Data Min*. 2008;1(2):83-96.
30. Palarea-Albaladejo J, Martín-Fernández JA, Soto J. Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *J Classif*. 2012;29:144-169.
31. Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V. Logratio analysis and compositional distance. *Math Geol*. 2000;32(3):271-275.
32. Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. *Ann Hum Genet*. 1956;20(4):309-311.
33. Lipton RB, Bigal ME, Diamond M, Freitag F, Reed ML, Stewart WF. Migraine prevalence, disease burden, and the need for preventive therapy. *Neurology*. 2007;68(5):343-349.

AUTHOR BIOGRAPHIES

Dr. Vives-Mestres has published 12 papers in indexed peer-reviewed journals in a wide range of areas (ie, statistical quality control, economy, materials, and headache) and has participated in more than 20 international conferences, some as an invited speaker. She is a visiting professor at the Universitat de Girona (Spain) and analytics lead at Curelator Inc. (USA). Marina Vives-Mestres studied industrial engineering at the Universitat Politècnica de Catalunya (UPC) in Barcelona (Spain) within the framework of a double diploma with the Institut des Sciences Appliquées de Lyon (France). She earned her master's degree on statistics and operational research (UPC and Universitat de Barcelona); became certified as a Six Sigma Black Belt; and completed a PhD on statistics applied to process monitoring at the Universitat de Girona. She is the recipient of two awards from the European Network for Business and Industrial Engineering for the best student presentation (2015) and for the best presentation made by a nonacademic (2018).

Professor Ron Kenett is chairman of the KPA Group, Israel, Chairman of the Data Science Society at AEAI, senior research fellow at the Samuel Neaman Institute, Technion, Haifa, Israel, and research professor at the University of Turin, Italy. He is an applied statistician combining expertise in academic, consulting, and business domains. Ron is

member of the Public Advisory Council for Statistics Israel, member of the of the executive academic council, Wingate academic college for sports education, member of the INFORMS QSR advisory board, member of the advisory board of DSRC, the University of Haifa Data Science Research Center and member of the board of directors in several start-up companies. He is past president of the Israel Statistical Association (ISA) and of the European Network for Business and Industrial *Statistics* (ENBIS), authored and coauthored over 250 papers and 14 books on topics such as data science, industrial statistics, biostatistics, health care, customer surveys, multivariate quality control, risk management, system and software testing, and information quality. The KPA Group he founded in 1994 is a leading Israeli firm focused on generating insights through analytics. He was awarded the 2013 Greenfield Medal by the Royal Statistical Society and, in 2018, the Box Medal by the European Network for Business and Industrial Statistics. BSc in mathematics (with first class honors) from Imperial College, London University and PhD in mathematics from the Weizmann Institute of Science, Rehovot, Israel.

Dr Santiago Thió-Henestrosa has a degree in computer science (UPC) and a PhD in computer science (UPC) with a work framed in local descriptive factor analyses. He is a full professor in the area of statistics and operations research at the Department of Computer Science, Applied Mathematics and Statistics at the University of Girona. His research focuses on the statistical analysis of compositional data; he has published theoretical contributions to the development of a specific methodology for compositional data and applications in different fields. He is also a coauthor of the CoDaPack software. He has published in different specialized journals: *Computers & Geosciences, Mathematical Geology*, *Computational Statistics*, *Applied Stochastic Models in Business and Industry*. Among other positions, he has been deputy director of the Superior Polytechnic School and he is secretary of the Computer Science Applied Mathematics and Statistics Department of the University of Girona.

Dr Martín-Fernández holds a degree in mathematics (Universitat Autònoma de Barcelona) and a PhD in statistics by the Universitat Politècnica de Catalunya, which was qualified as the award with special distinction for the 2000-2001 academic course. Since 1990, he is working in the Computer Science, Applied Mathematics and Statistics Department of the University of Girona (UdG). His area of interest is the statistical analysis of compositional data. Currently he is the principal investigator of the CoDa-research group at the UdG. Since June 2015 he is the treasurer of the CoDa Association, which aims to bring together scientists interested in developing methods for compositional data modeling and their application in different fields such as geology, biology, economics, or social sciences, to name a few.

How to cite this article: Vives-Mestres M, Kenett RS, Thió-Henestrosa S, Martín-Fernández JA. Measurement, selection and visualization of association rules: A compositional data perspective. *Qual Reliab Eng Int*. 2021;1-13. https://doi.org/10.1002/qre.2910