Advances in Economics and Business 5(9): 500-509, 2017 http://www.hrpub.org
DOI: 10.13189/aeb.2017.050903

Correspondence Analysis in Ternary Diagrams – A Graphical Approach to Improving the Validity of the Interpretation of Point Clouds

Jan Frederik Graff

School of Business and Economics, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Germany

Abstract Correspondence Analysis (CA) is a statistical procedure which is frequently used in market research. However, the neglecting of the third axis in the graphical representation of CA point clouds entails the risk of the faulty clustering of points and the misinterpretation of data. To meet this problem this paper proposes ternary diagrams as an approach to achieving a valid representation and interpretation of three-dimensional point clouds when plotted two-dimensionally. As CA coordinates do not necessarily meet the mathematical prerequisites ternary diagrams require, a procedure will be presented which transforms coordinates in order to meet the prerequisites.

Keywords Correspondence Analysis, Ternary Diagrams, Point Clouds, Clustering, Transformation

1. Introduction

A few years ago, the author worked as a research consultant in methods & statistics for a market research company. In our projects we frequently used Correspondence Analysis (CA) to detect multivariate relationships on a nominal scale level (e.g., "Does feature x apply to product y? Yes/No"). This method was very popular among our customers, because its output is a coordinate system in which variables with a tight association with each other are located close to each other. Thus, the graphical representation of complex multivariate relationships enables an intuitive and quite convenient interpretation of data and clustering of points. This requires the reduction of the potentially *n*-dimensional point cloud to two dimensions, because the representation in classical coordinate systems is per definition two-dimensional. However, this causes problems, as distances between points can easily be under- or overestimated when one dimension is missing. Of course, it is possible to relinquish the graphical interpretation and look at the pure coordinates and/or use a classical cluster analysis to cluster points in meaningful sub-clouds, but the market researcher tends to stick to the graphical visualization because it is intuitive, convenient, and impressive in presentations to customers. Thus, it has to be looked for a way to visualize CA point clouds without neglecting an axis and hereby avoid under- or overestimation of point distances, faulty clustering of points, and faulty interpretation of the data.

Correspondence Analysis (CA) is a descriptive, dimensionality-reducing procedure of multivariable statistics which allows a vivid graphical representation of complex correlations of two (or more) categorical variables as a point cloud in a (theoretically) three- (or more) dimensional space. It is widely used in social science, psychology, medicine, and in the area of market research [1]. Similarities between variables are converted into distances on three or more dimensions, and the positions of the variables, represented by points, are converted into coordinates. It is assumed that a dependency between rows and columns exists which can be explained by latent, unobserved variables. Later on, the variables will be presented as a three-dimensional point cloud. The three axes can then intuitively be interpreted as these latent variables. Sometimes, the points can be clustered manually to identify subgroups of variables closely associated to each other. Thus, the complete graphic representation of the results in a so called mapping [2] requires a coordinate system with three axes. This entails problems, as the graphic representation can, ultimately, only be two-dimensional. Software solutions therefore usually offer the option of printing three mappings, each representing two of the three axes. This produces a xy-, an xz-, and an yz-coordinate system. However, variance explained by a third axis is often neglected. If this share of variance is of a significant level, the neglection of one axis can lead to massive under- or overestimations of distances and thereby to faulty clustering and faulty interpretation of the data. Greenacre [2], who gives an excellent introduction

in theory and application of correspondence analysis, states: "If a large percentage of the total inertia lies along other principal axes [than the first two principal axes] then it means that some points are not being well represented with respect to the first two principal axes." [2].

Very little research has yet been done on alternative ways of plotting the point cloud of a CA. Backhaus et al. [1] point out that if more than two axes are needed to achieve a satisfactory explained variance, the researcher has to decide which two are the most important. These should be plotted. However, a two-axes-plotting is never possible without a loss of information [1]. Whitlark and Smith [3] state that: "Relying on a two-dimensional map may be risky. In our experience, it is rare to see a two-dimensional map tell a complete or even an accurate story" [3]. And Greenacre [2] adds: "Since the actual 2-dimensional display shows the projection onto the plane and does not show which points lie close to the plane and which are further off, we need to consider additional information if we are to interpret the display correctly." [2].

The aim of this paper is not to introduce a procedure which provides better clustering than a classical multivariate cluster-analysis but to develop graphical presentations of the data in order to facilitate a valid manual clustering compared to a two-dimensional mapping. To meet this goal, this paper proposes the use of ternary diagrams as a way to represent CA point clouds three-dimensionally. In ternary diagrams the sides of a triangle represent the three axes of a coordinate system, with the result that information from the third axis also remains in the two-dimensional representation. However, to be plottable in a ternary diagram, points need to meet some mathematical conditions – something that points in a CA point cloud usually do not.

After a short introduction to the key notions of the Correspondence Analysis, the basic concept of ternary diagrams will be explained. Then, a novel mathematical transformation is developed which transfers the coordinates of a CA point cloud in a way which makes them plottable in ternary diagrams. The following section gives an evaluation of the transformation using a data example taken from the field of market research, which includes a comparison of a manual clustering based on ternary diagrams where the clustering is realized by the Quick-Cluster procedure in IBM SPSS 22. A mathematical problem – the so called 1/3-problem – and possible solutions will be discussed in the last section before the conclusion.

2. Key Notions of Correspondence Analysis

The database of a CA consists of frequencies n_{ij} in a contingency table K with I rows R_i and J columns C_j (Table 1).

Table 1. General Dataset of a Correspondence Analysis

	C_1	C_2	…	C_J	$\sum$
R_1	n_{11}	n_{12}	…	n_{1J}	r_1
R_2	n_{21}	n_{22}	…	n_{2J}	r_2
⋮			…		⋮
R_I	n_{I1}	n_{I2}	…	n_{IJ}	r_I
$\sum$	c_1	c_2	…	c_J	1

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1J} \\ \vdots & \ddots & \vdots \\ p_{I1} & \cdots & p_{IJ} \end{pmatrix} \quad (1)$$

is the matrix of relative frequencies with

$$N = \sum_i \sum_j n_{ij} \quad (2)$$

and

$$p_{ij} = \frac{n_{ij}}{N} \quad (3)$$

We define

$$r_i = \sum_{j=1}^{J} p_{ij} \quad (4)$$

and

$$c_j = \sum_{i=1}^{I} p_{ij} \quad (5)$$

as the masses of the rows and columns [2]. The frequencies in the cells are normalized with r_i and c_j, respectively. As a result, we get two matrices

$$R = \begin{pmatrix} \frac{p_{11}}{r_1} & \cdots & \frac{p_{1J}}{r_1} \\ \vdots & \ddots & \vdots \\ \frac{p_{I1}}{r_I} & \cdots & \frac{p_{IJ}}{r_I} \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} \frac{p_{11}}{c_1} & \cdots & \frac{p_{1J}}{c_J} \\ \vdots & \ddots & \vdots \\ \frac{p_{I1}}{c_1} & \cdots & \frac{p_{IJ}}{c_J} \end{pmatrix}.$$

Each row of R is called the i_{th} row profile of K. Each column of C is called the j_{th} column profile of K [2]. The columns, interpreted as J axes, put up a J-dimensional space C_J, which the row profile should be plotted in, analogously the rows. Obviously, the J-/I-dimensionality cause problems because plotting more than three-dimensional spaces makes an understanding of the mapping more difficult. Due to this, the number of dimensions of C_J and R_I is reduced to three (C_3, R_3), and both are integrated into one coordinate system. Thereby, three conditions have to hold:

- The more similar the profiles are, the closer the points representing the respective row/column should lie to each other in the coordinate system. The distance is measured using the Euclidean distance.

- The reduction of the overall variance of K over all p_{ij}, the so-called *Total Inertia T*, should be minimal, while T is given as

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - r_i \times c_j)}{r_i \times c_j} \quad (6)$$

- The additive decomposition of T should be maximal [2].

After further calculations we get XYZ-coordinates for each row and column, so points can easily be plotted (Figure 1). The axes of the integrated space can be interpreted as latent variables which explain the variance of the conditional frequencies. Thus, for a complete plotting of results, a coordinate system with three axes is required. This poses some difficulties, as a valid graphical representation may ultimately be two-dimensional. Plotting software solutions (e.g., SPSS, mapwise, XGobi) therefore usually offer the option of plotting three different mappings, where always one of the three axes is omitted. Hereby, an XY-, an XZ- as well as a YZ-coordinate system are produced [4, 5]. This may cause problems for the interpretation of results, because only two axes can be considered simultaneously. This is why practitioners often choose the two axes with the highest explained variance and ignore the third one, thereby neglecting important distance information. Figure 1 shows an example of how a missing dimension can cause invalidities in the graphical representation. In Table 2 the mean Euclidean distances are given. It becomes obvious that the distance between point C and point 4 is massively underestimated if only the axes X and Y are taken into account. The mean Euclidean distance increases from 10.2 to 23.5 when axis Z is included into the calculation. The same holds for the distance between the points 11 and 23. Here, the mean Euclidean distance increases from 6.3 to 8.9. Analogously, the distance between point D and point 16 is heavily overestimated in the two-dimensional mapping. The mean Euclidean distance is reduced from 20.9 to 15.7 when the Z-coordinate is included into the calculation. This invalid representation of distances in the two-dimensional mapping can easily lead to a faulty clustering of data. This is dangerous, because the aim of CA is to show graphically which row points (traits) are associated with which column point (trait carrier) in the perception of potential customers. Thus, faulty clustering can lead to wrong interpretations of these associations and foster invalid conclusions in the manner that, for example, a point is included into one cluster but truly belongs to another one. In the worst case such errors can let a public relation campaign fail because traits of one product are stressed although they are – according to the customer's perception – not in the least associated with this product. To sum up, the main problems with non-three-dimensional plottings of a CA point cloud are: *overestimation of distances*, *underestimation of distances*, *faulty clustering*, *and faulty interpretation.*

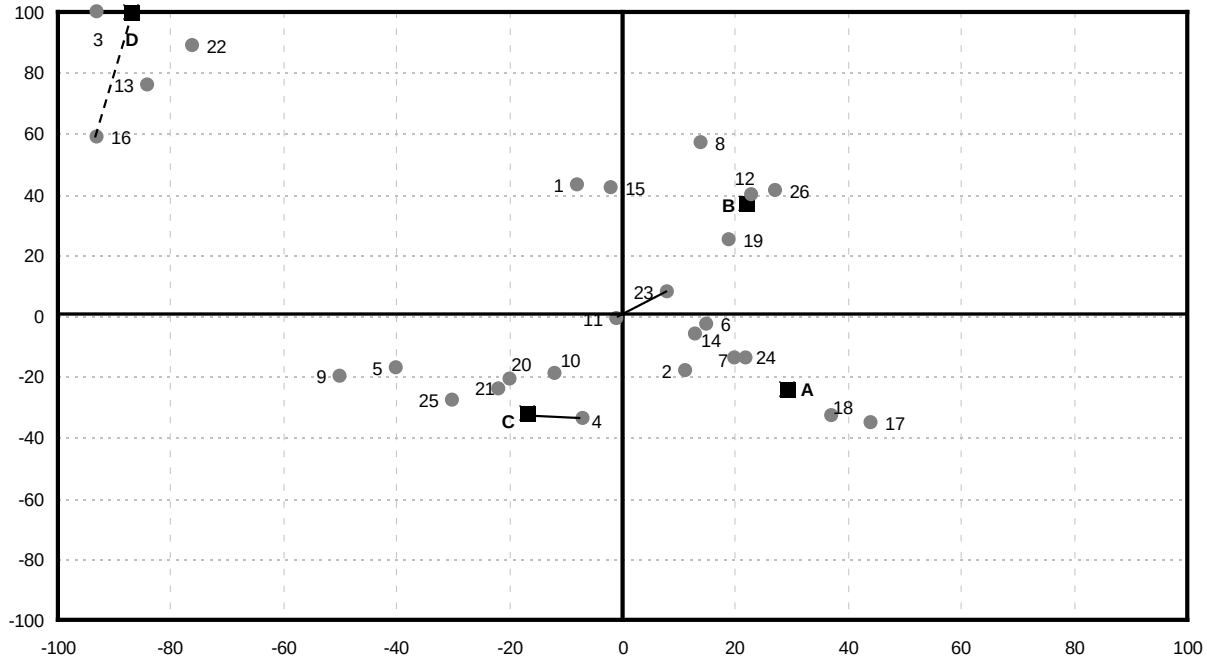

Figure 1. Underestimation (–) and overestimation (- - -) of point distances in a two-dimensional mapping. Z-axis not plotted

Table 2. Mean Euclidean distances of example points in Figure 1

	Mean Euclidean Distance		
Points	XY	XYZ	
C, 4	10.2	23.5	Underestimation
11, 23	6.3	8.9	Underestimation
D, 16	20.9	15.7	Overestimation

3. A Simple Example

One reason why CA is so widely used in social science and market research lies in the relatively convenient way, which complex correlations between categorical variables can be intuitively interpreted using mappings (Figure 1). With a mapping, the researcher interprets the relative distances of the points from each other and their relative positions on the axes. Additionally, the market researcher uses the mapping to cluster points which lie close to each other and sufficiently far apart from others. Thereby, sub-clouds of points emerge which are assumed to be substantially associated with each other. Most often, one column point (■ in Figure 1) makes up more or less the center of a sub-cloud consisting of several row points (● in Figure 1). In market research, for example, the column points usually represent variations of a product, brands, or firms, and the row points represent features, ratings, or (potential) purchasers [1]. Thus, a column point represents some kind of an object or a person that certain characteristics are associated with. In other words, the object carries some traits. Therefore, row points will henceforth be referred to as *traits*, and column points as *trait carriers*.

As an example, we assume that a market researcher is interested in how potential buyers perceive four different car brands (A, B, C, D) in regard to certain characteristics. Therefore, she compiles up a questionnaire containing an item matrix in which the surveyed participants are to assess whether a given statement applies to brand A, brand B, brand C, and/or brand D. Multiple answers are possible. Possible statements for the example in Figure 1 are given in Table 3. The point cloud in the mapping in Figure 2 reveals that the points 3, 13, 16, and 22 form a subcloud around point D. Thus, 3, 13, 16, and 22 are traits which the surveyed participants perceived as applying to the trait carrier D. The car brand D is considered to be luxury, to be a status symbol, to be sportive, and to be providing driving pleasure. Analogously, the brands A, B, and C are associated with the traits located close to the trait carrier points in the mapping. In contrast to the cluster around D, the clusters around A, B, and C are more difficult to separate, because the clusters lay relatively close to each other. Especially in regard to the points in the center between the three clusters, 11 and 23, it is almost impossible to decide to which cluster they belong just by visual inspection. Furthermore, it is possible that points now located seemingly clearly in one cluster actually belong to a different cluster, because of the invalid representation of distances. Thus, what is needed is either a classical multivariate cluster analysis or a graphical representation which avoids the under- or overestimation of distances in order to get a clearer distinction of the clusters. Of course, a cluster analysis provides a statistically correct clustering of points. However, because of its convenience and vividness we want to maintain the option of a graphical manual clustering and therefore look for a way to make it more valid.

Table 3. Item matrix as dataset for a correspondence analysis of the association of four car brands with 26 car-related statements

	Brand	A	B	C	D
1	Is reliable	□	□	□	□
2	Provides good service	□	□	□	□
3	Is luxury	□	□	□	□
4	Has a good price-performance ratio	□	□	□	□
5	Is technically of the highest standard	□	□	□	□
6	Has a modern design	□	□	□	□
7	Is economical	□	□	□	□
8	Is rather something for men	□	□	□	□
9	Is rather something for women	□	□	□	□
10	Is rather a brand for younger people	□	□	□	□
11	Is rather a brand for older people	□	□	□	□
12	Has a low fuel consumption	□	□	□	□
13	Is a status symbol	□	□	□	□
14	Is comfortable	□	□	□	□
15	Is practical	□	□	□	□
16	Is sportive	□	□	□	□
17	Is a modern brand	□	□	□	□
18	Maintenance expense is low	□	□	□	□
19	Has good emission data	□	□	□	□
20	Posseses a high safety standard	□	□	□	□
21	Is an innovative brand	□	□	□	□
22	Provides driving pleasure	□	□	□	□
23	Is old-fashioned	□	□	□	□
24	Is always a reasonable decision	□	□	□	□
25	Is good for transport	□	□	□	□
26	Is useful for families	□	□	□	□

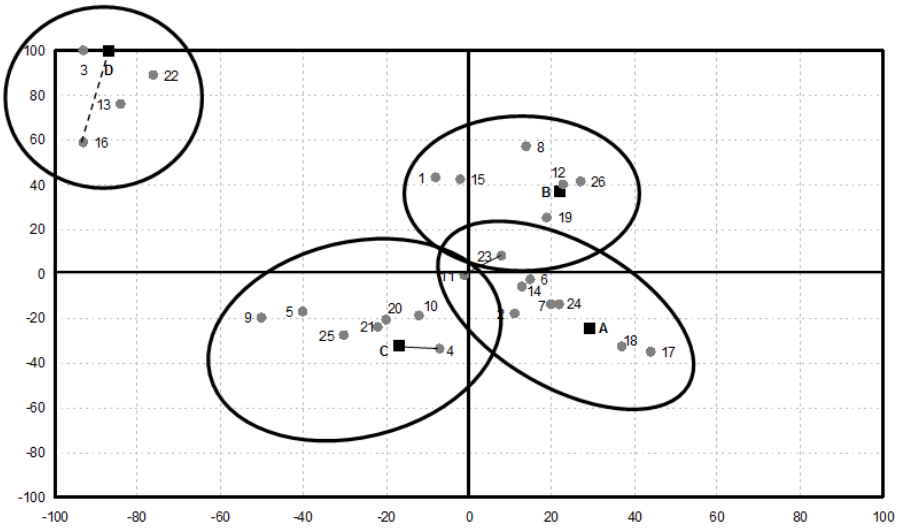

Figure 2. CA points cloud representation in a two-axis coordinate system with a simple two-dimensional mapping (layout similar to SPSS). Points are manually clustered with obvious fuzziness around the points 11 and 2.

We assume that the clustering of points would be more precise if there were a way to visualize the third axis in order to get better distance information. For example, from the previous section and Figure 1 we know that the distance between point C and point 4 is underestimated in the two dimensional mapping. Thus, it is imaginable that under a valid representation of the point distances, C and 2 would not be positioned in the same cluster, but that 2 could be located in the cluster around A, which would lead to different characterization of brands C and A. Analogously, it could be demonstrated that the distance between the points 11 and 23 is underestimated. Given a valid representation which does not neglect the third axis, 11 and 23 would probably lie further apart and thereby make the clusters around C and B easier to separate.

4. Ternary Diagrams

The concept of ternary (triangular) diagrams was first published in 1958 by the Austin, Texas geologist Edmund Sneed together with Robert Folk from the Ohio Oil Company. These diagrams are therefore sometimes referred to as Sneed & Folk diagrams. Sneed and Folk applied ternary diagrams for the first time in their study on particle morphogenesis in which samples of pebbles from the Lower Colorado River were described by their characteristic mineralogical composition from three rock types – quartz, chert, and limestone [6]. Since then, ternary diagrams have primarily been used in geology and related earth sciences, but also in chemistry, economics, and social sciences. Due to the fact that ternary diagrams can only depict the relative shares of three components of a whole, the typical application of ternary diagrams is the representation of attributes that can be completely divided into three such components. For example: chemistry (characterizing the mixture ratio of a chemical put together from three different components), geology (characterizing soil samples composed of sand, clay, and silt), agriculture sciences (characterizing the structure of agricultural areas composed of three types of land, e.g., farm land, meadow land, and rangelands), or social sciences (characterizing the age structure of groups covering three groups of age). Different kinds of objects (e.g., categories of chemicals or soils) can be categorized by their location in the diagram, which represents a certain combination of attributes. Sometimes, it can be reasonable to divide the ternary diagram into segments in order to distinguish groups of objects which have a similar characteristic combination of the three components. In ternary diagrams the third axis, which is missing in a Cartesian coordinate system, is considered by mapping the coordinates into a coordinate system that has the form of an equilateral triangle. Each side represents one dimension and is scaled accordingly (Figure 3).

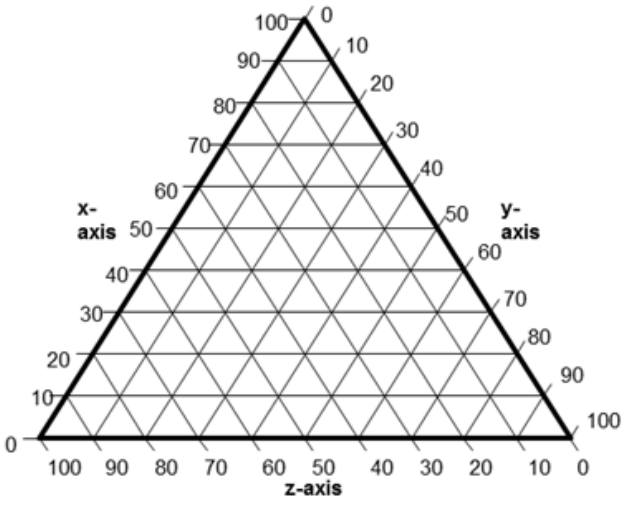

Figure 3. Ternary Diagram

The scale lines of an axis are not placed orthogonally on it, but run as a family of parallels toward the counterclockwise next side. That way, each scale line subtends the next side at the value w at the position $w' = 100 - w$. The base axis has its point of origin in the bottom right corner of the triangle. The scale lines are orientated as parallel skeins at the right axis of the triangle. The origin of the right axis is at the top of the triangle whereas the value 100 is in the bottom right corner and the scale parallels are orientated at the left axis. The origin of the left axis is located in the left corner, the value 100 in the triangle's apex. The scaling lines run parallel to the baseline of the triangle. For better understanding, the scale lines are extended over and above the axis. Hereby, it is possible to see which parallel skein belongs to which axis. As the representation is two-dimensional, in spite of having three axes there exist only two degrees of freedom. Therefore, only complementary attributes – attributes whose values sum up to a constant sum (i.e. 1 or 100) – can be represented using conventional ternary diagrams. Furthermore in ternary diagrams coordinates smaller than zero cannot be mapped, so that for any point P_i with the coordinates (x_i, y_i, z_i) it must hold that:

$$x_i + y_i + z_i = 100 \tag{1}$$

and

$$(x_i, y_i, z_i) \in [0, 100] \tag{2}$$

We name (1) the constant sum condition and (2) the non-negativity condition. Because of the limited applicability of ternary diagrams due to the constant sum condition, Sneed and Folk [6] developed a modified diagram in the context of the systematization of glacial stone. In this diagram, the coordinates are calculated by the proportions of the characteristic attributes: For an object P_i with the attributes A, B, and C and the characteristics a, b, and c the Sneed-Folk-coordinates (x_i, y_i, z_i) are given by

$$x_i = \frac{c}{a} \tag{3}$$

$$y_i = \frac{b}{a} \tag{4}$$

$$z_i = \frac{a-b}{a-c} \tag{5}$$

If $a, b, c \in \mathbb{R} \geq 0$, at least one of the following conditions must hold in order to avoid the coordinates x', y', $z' < 0$, which are not possible per scale:

$$a \geq b \wedge a > c \tag{6}$$

$$b \geq a \wedge c > a \tag{7}$$

However, the condition does not hold in the case of correspondence analysis. Thus, the Sneed & Folk algorithm cannot be used to transform CA-coordinates according to the constant sum condition and the non-negativity condition. Only conventional ternary diagrams are suitable for the presentation of results of a correspondence analysis.

5. The Transformation

Cartesian coordinates defining points on three axes are the result of a classical CA. To allow plotting these points in a ternary diagram, it is necessary to transform the coordinates according to the constant sum condition and the non-negativity condition (see previous section). In this section such a transformation will be proposed.

The constant sum condition defines a level in a three-dimensional space which cuts coordinate axes in (100, 0, 0), (0, 100, 0) and (0, 0, 100). If the examined points are already part of this level or even build a triangle with the vertices (100, 0, 0), (0, 100, 0) and (0, 0, 100), there is no transformation necessary and the triangle can be visualized[1]. If the points are located in the above defined level but not yet in the triangle, it is possible to perform a centric elongation to bring them in [7]. In general, not all points will be on this level, so that at first coordinates ascertained for an orthogonal three-axis system are transformed onto a range of values in such a way that, after the transformation, it holds that

$$x_i' + y_i' + z_i' = 100 \tag{8}$$

and

$$x_i', y_i', z_i' \in [0,100] \tag{9}$$

This is ensured through the following transformation: An exponential function of the transformed values reduces the codomain of the coordinates to $\mathbb{R} \geq 0$. Another option would be a simple adjustment by adding the vector

$$\vec{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

which brings all points into the 1st octant. Compared to the adjustment method, the use of an exponential function to standardize the values of the coordinates has the advantage that small values stay small and great values become much greater. This is important to still ensure a spread of the values over the whole codomain after the standardization. Previously, every coordinate is divided by 100 to bring the coordinates into a codomain which generates values of manageable size during the application of the exponential function. The scaling factor s influences the strength of the exponential function and thereby the spread of the point cloud in the direction of the diagram corners in the graphical representation. In practical use, a scaling factor between 2.5 and 3 proved itself. Without a scaling factor there will be a clustering in the middle of the diagram (see proof in section 6). Afterwards, for every coordinate, the percentage of the sum of the standardized coordinates is evaluated:

1 Unless a barycentric transformation is wanted. A description of the transformation into barycentric coordinates is, inter alia, given in Coxeter (1980).

$$x_i' = \frac{\exp\left[\left(\frac{x_i}{100}\right)*s\right]}{\exp\left[\left(\frac{x_i}{100}\right)*s\right]+\exp\left[\left(\frac{y_i}{100}\right)*s\right]+\exp\left[\left(\frac{z_i}{100}\right)*s\right]}*100 \quad (10)$$

$$y_i' = \frac{\exp\left[\left(\frac{y_i}{100}\right)*s\right]}{\exp\left[\left(\frac{x_i}{100}\right)*s\right]+\exp\left[\left(\frac{y_i}{100}\right)*s\right]+\exp\left[\left(\frac{z_i}{100}\right)*s\right]}*100 \quad (11)$$

And from (1) follows that

$$z_i' = 100 - x_i' - y_i' \quad (12)$$

with $x_i, y_i, z_i \in [0,100]$ and $s > 0$.

The coordinates x_i', y_i', z_i' generated by this means can be drawn in a ternary diagram either by hand or using a specialized software like Tri-plot[2] [8].

6. Proof

Does the transformation without scaling factor lead to a clustering of points in the center of the triangle? A proof is provided in two steps. Step 1: It is to be demonstrated, that after the transformation (with or without scaling factor) no point can be in the corners of the triangle (extremum). Step 2: It is to be demonstrated that the expansion of the point cloud towards the corners is achieved by the scaling factor.

Step 1: After the transformation, extreme values such as P' = (100,0,0) in the corners of the triangle are not possible anymore. To illustrate this, we search for the coordinates (x_p, y_p, z_p) of the point P that, after the transformation without scaling factor, is located as extremum P' = (100,0,0) in a corner of the diagram.

Thus, we search (x_p, y_p, z_p), for which holds

$$x'_P = \frac{\exp\left(\frac{x_P}{100}\right)}{\exp\left(\frac{x_P}{100}\right)+\exp\left(\frac{y_P}{100}\right)+\exp\left(\frac{z_P}{100}\right)} = 1 \quad (13)$$

x'_P can only be 1 if numerator equals denominator.

As $\exp\left(\frac{x_P}{100}\right)$ appears both in the numerator and in the denominator, numerator and denominator can only be equal if

$$\exp\left(\frac{y_P}{100}\right)+\exp\left(\frac{z_P}{100}\right) = 0 \quad (14)$$

However, this is not possible, because all terms of the sum are positive. Thus, it always holds that

$$\exp\left(\frac{k}{100}\right) > 1, \quad \forall \ k \ \text{ with } \ k \in \{x, y, z\} \quad (15)$$

Step 2: It has to be shown that the implementation of a scaling factor s results in a shift of the transformed points towards the corners of the triangle, so that

$$k_s' \geq k' \quad (16)$$

Whereby, let

$$k' = \frac{\exp\left(\frac{x}{100}\right)}{\exp\left(\frac{x}{100}\right)+\exp\left(\frac{y}{100}\right)+\exp\left(\frac{z}{100}\right)} \quad (17)$$

be a coordinate from a transformation without scaling factor and

$$k_s' = \frac{\exp\left[\left(\frac{x}{100}\right)*s\right]}{\exp\left[\left(\frac{x}{100}\right)*s\right]+\exp\left[\left(\frac{y}{100}\right)*s\right]+\exp\left[\left(\frac{z}{100}\right)*s\right]} \quad (18)$$

be a coordinate after a transformation with scaling factor. That (16) holds, can be shown as follows.

The transformation of the extremum $k = (x_{\max},0,0)$ is considered, first without scaling factor:

$$\begin{aligned} k'(x_{\max}) &= \frac{\exp\left(\frac{x_{\max}}{100}\right)}{\exp\left(\frac{x_{\max}}{100}\right)+2\exp\left(\frac{0}{100}\right)} \\ &= \frac{\exp\left(\frac{x_{\max}}{100}\right)}{\exp\left(\frac{x_{\max}}{100}\right)+2} < \exp\left(\frac{x_{\max}}{100}\right) \end{aligned} \quad (19)$$

Hence, a scaling factor s is implemented into the transformation formula and we prove whether $k_s' \geq k_s$ holds for the resulting coordinates k_s'. With scaling factor follows

$$k_s'(x_{\max}) = \frac{\exp\left[s\left(\frac{x_{\max}}{100}\right)\right]}{\exp\left[s\left(\frac{x_{\max}}{100}\right)\right]+2} < \exp\left[s\left(\frac{x_{\max}}{100}\right)\right] \quad (20)$$

As the exponential function is continuous and monotonically increasing, it always holds that

$$\exp\left[s\left(\frac{x_{\max}}{100}\right)\right] > \exp\left(\frac{x_{\max}}{100}\right), \quad \forall \ x, s \quad (21)$$

However, this is only the case as long as $x = y = z$ does not hold. Otherwise, a special case occurs, which we call the *1/3-problem*.

2 Tri-plot was developed by Graham & Midgley (2000) and is available as a free download including detailed description on http://www.lboro.ac.uk/research/phys-geog/tri-plot/index.html

7. Evaluation of the Procedure

Comparisons with classical two-dimensional mappings show that the distances between points in the ternary diagram are depicted more validly than in the two-dimensional mapping. Actually neighboring points remain neighboring in the ternary diagram (Figure 4), while points with neighboring x- and y-coordinates but very different z-coordinates, which were not realized in the two-dimensional mapping, are strictly separated in the ternary diagram. In a two-dimensional mapping, the points can also lie further apart than how the coordinates show over all three axes. In this case, the points move closer together in the ternary diagram. The larger the variance explanation by the z-axis, the more obvious these effects will appear. Figure 4 shows the mapping of the same point cloud as in Figure 2, now with transformed points and plotted into a ternary diagram. Looking at the sample points mentioned above, it becomes obvious that the transformation meets its goals. The points whose distance had been underestimated in the two-dimensional mapping (C-4, 11-23) are now located clearly apart while the points whose distance had been overestimated (D-16) are now located very close to each other. As a consequence, the identification of subclouds becomes more convenient. The two-dimensional mapping, for example, was fuzzy about which cluster point 11 and point 23 belong to. In the ternary diagram they are clearly positioned in different clusters. The cluster solution in Figure 4 was drawn by hand and verified using the Quick-Cluster procedure in SPSS 22. Thus, we can assume that the representation of the point cloud in ternary diagram instead of a two-dimensional mapping provides a more convenient and more valid way of visualization. The free-hand clustering of points is – at least in this example – obvious and in line with the statistical clustering via a cluster analysis.

8. The 1/3-Problem

The proof given in section 6 demonstrates that the implementation of a scaling factor *s* leads to a shift of the transformed points towards the triangle's corners. In this section we will demonstrate that this is not the case if it holds that $x = y = z$. Under this condition, a problem occurs which we call the *1/3-problem*. Two ways to solve the 1/3-problem will be presented later on in this section.

Given a three-dimensional space R^3 with the axes *x*, *y*, and *z*, every axis proceeds from the origin $U = (0,0,0)$ towards $(x_{max},0,0)$, $(0,y_{max},0)$, or $(0,0,z_{max})$, while $x_{max} = y_{max} = z_{max} < \infty$ and $x_{max} = y_{max} = z_{max} > 0$. We consider the transformation of the extreme values $E_{min} = (x_{min}, y_{min}, z_{min}) = U = (0,0,0)$ and $E_{max} = (x_{max}, y_{max}, z_{max})$. $k_i(E_{min})$ is the i^{th} coordinate of the point E_{min} with $i \in \{1,2,3\}$. Then:

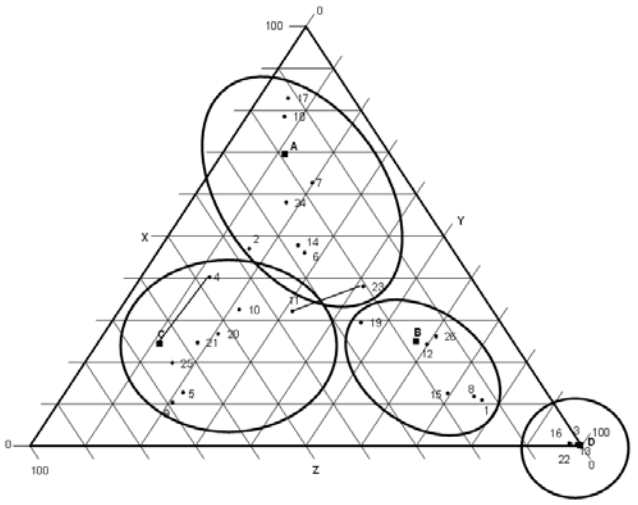

Figure 4. Representation of the mapping from Figure 1 in a ternary diagram. Clusters calculated using the Quick-Cluster procedure in SPSS 22. Note the formerly underestimated point distances between the points C and 4, and 11 and 23 as well as the formerly overestimated point distances between points D and 16.

$$k_i'\ (E_{\min}) = \frac{\exp\left(\frac{k_i(E_{\min})}{100}\right)}{\sum_{i=1}^{3}\exp\left(\frac{k_i(E_{\min})}{100}\right)} = \frac{1}{3}, \forall\, i \qquad (22)$$

and E_{min} becomes E'_{min} with $x'_{min} = y'_{min} = z'_{min} = 1/3$.

Analogously it holds for the transformation of E_{max} that:

$$k_i'\ (E_{\max}) = \frac{\exp\left(\frac{k_i(E_{\max})}{100}\right)}{\sum_{i=1}^{3}\exp\left(\frac{k_i(E_{\max})}{100}\right)} = \frac{1}{3}, \forall\, i \qquad (23)$$

and E_{max} becomes E'_{max} with $x'_{max} = y'_{max} = z'_{max} = 1/3$. Obviously, after the transformation, extreme values are not possible anymore if $x = y = z$ applies, because all points fall together in (1/3,1/3,1/3), the centroid of the triangle.

Solution I – Independent Scaling Factors

Every coordinate gets its individual scaling factor $s_i > 1$ within the transformation formula. The scaling factors are independent of one another. For the transformation it applies thereafter that

$$x'_P = \frac{\exp\left(\frac{x_P}{100}s_x\right)}{\exp\left(\frac{x_P}{100}s_x\right) + \exp\left(\frac{y_P}{100}s_y\right) + \exp\left(\frac{z_P}{100}s_z\right)} \qquad (24)$$

$$z'^{'}_P = \frac{\exp\left(\frac{z_P}{100}s_z\right)}{\exp\left(\frac{x_P}{100}s_x\right) + \exp\left(\frac{y_P}{100}s_y\right) + \exp\left(\frac{z_P}{100}s_z\right)} \qquad (25)$$

$$y'_P = \frac{\exp\left(\frac{y_P}{100} s_y\right)}{\exp\left(\frac{x_P}{100} s_x\right) + \exp\left(\frac{y_P}{100} s_y\right) + \exp\left(\frac{z_P}{100} s_z\right)} \quad (26)$$

with $s_x \neq s_y \neq s_z$.

Solution II – Interdependent scaling factors

The scaling factors are constantly interdependent, which means that every coordinate has its individual scaling factor, which is defined via another one. In that case, (24), (25), and (26) still hold, but from now on with: $s_x = s$, $s_y = s_x + a$, and $s_z = s_x - a$. Thus, for the transformation it now holds that

$$x'_P = \frac{\exp\left(\frac{x_P}{100} s\right)}{\exp\left(\frac{x_P}{100} s\right) + \exp\left(\frac{y_P}{100}(s+a)\right) + \exp\left(\frac{z_P}{100}(s-a)\right)} \quad (27)$$

$$y'_P = \frac{\exp\left(\frac{y_P}{100}(s+a)\right)}{\exp\left(\frac{x_P}{100} s\right) + \exp\left(\frac{y_P}{100}(s+a)\right) + \exp\left(\frac{z_P}{100}(s-a)\right)} \quad (28)$$

$$z'_P = \frac{\exp\left(\frac{z_P}{100}(s-a)\right)}{\exp\left(\frac{x_P}{100} s\right) + \exp\left(\frac{y_P}{100}(s+a)\right) + \exp\left(\frac{z_P}{100}(s-a)\right)} \quad (29)$$

The optimal dependence (the optimal a) is still to be defined.

9. Conclusions

Mapping the point clouds of correspondence analyses in two-dimensional coordinate systems is a popular instrument to vividly visualize the results of a correspondence analysis. It allows the presenting and interpreting of associations between points or clusters of point graphically and intuitively. However, the neglect of a third axis and ignorance of variance information accounting for this axis entails the risk of under- and overestimating point distances. In the worst case, this invalid graphical representation can lead to faulty clustering and misleading interpretations. Of course, faulty clustering can be avoided by applying a classical cluster analysis, but we want to maintain the possibility of manual graphical clustering as a convenient method to present and interpret multivariate statistics in the context of market research. Thus, the contribution of this paper is not to develop an optimized cluster algorithm, but to find a way to visualize three-dimensional point clouds in a two-dimensional point cloud in order to enable valid manual clustering.

In this paper we introduced the concept of ternary diagrams as a simple and cheap option to visualize three-dimensional point clouds in a two-dimensional diagram without losing variance information. The mathematical necessity to transform the coordinates (x_i, y_i, z_i) into (x'_i, y'_i, z'_i) to meet certain conditions was explained. We introduced two conditions which have to be met by the transformed points, the *non-negativity condition* (x_i, y_i, $z_i \in$ [0,100]) and the *constant sum condition* ($x_i' + y_i' + z_i' = 100$). Both conditions are met by the transformation of the coordinates using (10), (11), and (12). By employing a scaling factor s, the transformation also optimizes the distribution of the points over the area with the aim of a clearly arranged representation and more simply clustering and interpretation. A challenge is the so called *1/3-problem*, which means that all points $P = (x_p, y_p, z_p)$ for which holds $x_p = y_p = z_p$ accumulate in the point (1/3, 1/3, 1/3), the centroid of the triangle, after the transformation. This problem can be solved by applying different scaling factors for the three axes. We presented two possible solutions in this study. Using a simple example taken from the context of market research and marketing, we managed to illustrate the challenges of plotting CA point clouds two-dimensionally. The example could also show that transforming the points and plotting them in a ternary diagram leads to reduce under- and overestimations of point distances and to a simplified clustering. The comparison of the manual clusters with the cluster solution realized by a classical cluster analysis showed that the manual clusters in our example match the calculated ones. We take this as evidence of the validness of our procedure.

To sum up, ternary diagrams are an easy and cost-efficient alternative to classical *xy*-diagrams. Complicated coordinate-comparisons are no longer required. The interpretation of the results is also more valid, because there is no loss of information and all axes are considered simultaneously. The usage of ternary diagrams is especially useful for correspondence analysis with large share of explained variance on the *z*-axis (from approximately 10 %).

10. Appendix

Table 4. Exact coordinates of the mapping example in Figure 1 and Figure 2

Point	X	Y	Z
A	28.9	-23.6	-21.5
B	22.1	37.4	15.5
C	-16.8	-31.9	23.3
D	-87.1	100.0	-100.0
1	-7.9	42.8	-4.7
2	11.1	-18.1	4.3
3	-92.6	100.0	-100.0
4	-6.9	-34.4	-2.1
5	-40.2	-17.1	21.8
6	14.6	-2.7	-4.3
7	20.4	-13.5	-33.3
8	14.4	56.9	28.5
9	-50.0	-20.1	20.1
10	-12.1	-19.2	-0.8
11	-1.1	-1.3	4.5
12	22.9	39.8	13.2
13	-84.3	75.6	-92.7
14	13.5	-6.4	-7.3
15	-1.5	41.7	15.7
16	-93.1	58.7	-78.1
17	44.1	-35.4	-31.3
18	36.8	-33.3	-28.8
19	18.6	24.8	20.5
20	-20.3	-21.4	3.3
21	-22.2	-24.3	9.0
22	-76.3	89.2	-81.6
23	7.6	7.9	-19.2
24	21.6	-14.0	-10.4
25	-30.0	-28.0	14.1
26	26.9	41.2	7.2

REFERENCES

[1] K. Backhaus, B. Erichson, W. Plinke and R. Weiber, Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Springer, Berlin Heidelberg, 2006.

[2] M. J. Greenacre, Theory and applications of correspondence analysis. Academic Press, Orlando, Fla, 1984.

[3] D. B. Whitlark, S. M. Smith, Using Correspondence Analysis to Map Relationships, Marketing Research, Vol. 13, No. 3, 22–27, 2001.

[4] D. F. Swayne, D. Cook, A. Buja, XGobi: Interactive Dynamic Data Visualization in the X Window System, Journal of Computational and Graphical Statistics, Vol. 7, 113–130, 1998.

[5] O. Nenadic, M. Greenacre, Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package, Journal of Statistical Software, Vol. 20, No. 3, 1–13, 2007.

[6] E. D. Sneed, R. L. Folk, Pebbles in the Lower Colorado River, Texas. A study of particle morphogenesis, Journal of Geology, Vol. 66, No. 2, 114–150, 1958.

[7] H. Schupp, Elementargeometrie. Uni-Taschenbücher, Vol. 669. Schöningh, Paderborn, 1977.

[8] D. J. Graham, N. G. Midgley, Technical Communication. Representation of particle shape using triangular diagrams: An Excel spreadsheet method, Earth Surface Processes and Landforms, Vol. 25, No. 13, 1473–1477, 2000.