Individual T^2 Control Chart for Compositional Data

MARINA VIVES-MESTRES, JOSEP DAUNIS-I-ESTADELLA
and JOSEP-ANTONI MARTÍN-FERNÁNDEZ
Universitat de Girona, PIV-Campus Montilivi, 17071 Spain

The usual Hotelling T^2 control chart is not appropriate for monitoring processes where the quality characteristic is a mixture. The composition of mixtures are vectors of positive elements that represent parts of a whole, to which standard multivariate techniques are not appropriate due to their restricted sample space. There are many applications where a mixture is monitored against time, such as in the chemical industry, product composition, impurity profile, or gas components analysis. In this paper, a multivariate control chart for individual compositional observations based on the T^2 statistic is proposed and compared with the typical one in terms of ARL. We show how results are more consistent with compositional data nature and illustrate implementation in a real-world example.

Key Words: ARL; Hotelling's T^2 Statistic; Log Ratio; Mixture data; Multivariate Control Chart; Simplex; Statistical Process Control.

Introduction

In statistical process control, to monitor simultaneously multiple quality characteristics taking into account the correlation among the variables, a Hotelling's T^2 control chart (CC) is commonly used. Explanations on the use of the T^2 statistic can be found in Tracy et al. (1992), Kenett and Zacks (1998) and Montgomery (2009).

Ms. Vives-Mestres is a Ph.D. candidate in the Department of Computer Science, Applied Mathematics and Statistics. Her email address is marina.vives@udg.edu.

Dr. Daunis-i-Estadella is a titular Professor in the Department of Computer Science, Applied Mathematics and Statistics. His email address is pepus@imae.udg.edu.

Dr. Martín-Fernández is a titular Professor in the Department of Computer Science, Applied Mathematics and Statistics. His email address is josepantoni.martin@udg.edu.

In a T^2 CC the following statistic is calculated for each individual observation $\mathbf{x}$ in $\mathbb{R}^p$:

$$T_t^2 = (\mathbf{x}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}) \qquad (1)$$

Where $\mathbf{x}_t$, $t = 1, \ldots, m$, are p-variate observations assumed to be mutually independent and multivariate normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

In a first stage, called Phase I, the process is brought into a state of statistical control and the process parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from a sample of size m by the sample arithmetical mean $\bar{\mathbf{x}}$ and the sample covariance matrix $\mathbf{S}$, respectively. In a second stage (Phase II), the control scheme is developed and the estimates are used to compute the upper control limit (UCL). At each arrival of a new observation, the T^2 is compared with UCL to verify if the in-control state has changed. A discussion on how to compute control limits in both phases can

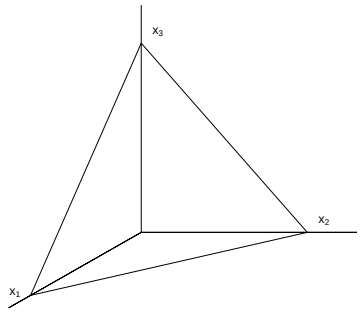
be found in Tracy et al. (1992).

We consider the case in which the quality characteristic being monitored is a mixture or a compositional vector $\mathbf{x} = (x_1, \ldots, x_p)$ with non-negative elements that add to a constant κ (for simplicity, often taken to be 1). Classical data units are weight or volume percent, *ppm* or molar proportions. Due to the constant sum, compositional data (CoDa) live in a restricted sample space of dimension $p-1$. It has already been demonstrated (e.g., Aitchison and Egozcue (2005) or Pawlowsky-Glahn and Buccianti (2011) and references therein) that standard multivariate techniques assuming that the sample space is $\mathbb{R}^p$ are not appropriate for restricted spaces. Note that, in this article, compositional data or composition refers to the composition or components of a mixture.

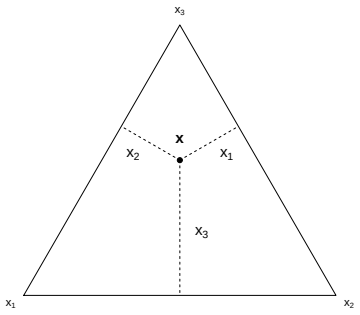
The sample space of CoDa is the Simplex $\mathcal{S}^p$, where p represents the number of variables in the composition. When $p = 3$, the composition lies in an equilateral triangle in $\mathbb{R}^3$ (Figure 1a), although it is more common to represent the data in the ternary diagram (Figure 1b), which is an equivalent representation.

If a sample $\mathbf{x} = (x_1, x_2, x_3)$ lies near the center of the triangle (close to $(1/3, 1/3, 1/3)$ when $\kappa = 1$), the sample is homogeneous, as all components are present in a similar proportion. On the contrary, if the sample is close to an edge or vertex, there are one or two components, respectively, that are present in minor quantity in the composition.

Although the real space is a linear vector space with Euclidean metric structure, the typical geometry used therein (sum, multiplication, orthogonality...) is not appropriate for CoDa (Bacon-Shone (2011)). To illustrate this assertion, consider the pairs of percentages 1–2% and 39–40%. The absolute difference (Euclidean distance) in both cases is $1 = 2-1 = 40-39$. But a more adequate measure to describe the way in which



(a) Simplex in $\mathbb{R}^3$



(b) Ternary diagram

Figure 1: Two different but equivalent representations of the Simplex with $p = 3$ in $\mathbb{R}^3$ (a) and in the ternary diagram (b).

they are unlike each other is to use the relative difference; thus, in the first case, the relative increase is 100% and in the second case is less than 3%.

It is possible to equip the Simplex with the structure of a metric vector space with specific operations that allow one to solve compositional problems with its specific algebraic-geometric structure. However, there is another approach based on log-ratio transformations that enables a one-to-one representation in an unconstrained space where standard multivariate techniques can be applied (Egozcue et al. (2003)). The transformed variables are called *coordinates*. We use the second approach here.

When the variable $\mathbf{x}$ is a composition, the covariance matrix in Equation (1) is singular and thus cannot be inverted to compute the T^2 statistic. This is due to multicollinearity problem, always encountered in CoDa, caused by the restricted sum of the components. We have found in the litera-

ture three different scenarios when performing a T^2 control scheme to compositional variables:

1. Mason and Young (2001) suggest eliminating one of the variables involved in the collinearity and then computing the T^2. Another suggestion is to rebuild the covariance matrix by eliminating the eigenvectors corresponding to the near-zero eigenvalues thus to compute the inverse of the covariance matrix with the largest ones.

2. Measurement errors make the covariance matrix near singular. Although collinearity exists, it can remain undetected. In that case, the T^2 statistic is severely distorted: signalling observations are no longer credible and the control procedure does not make sense (Mason and Young (2001)).

3. When only some parts of the whole composition (subcomposition) are included in the analysis, no collinearity problem exists because no constant sum is defined. This is the case in articles such as Mason et al. (1997), Mason et al. (2001), Mason and Young (2001), Ortiz-Estarelles (2001), Gonzalez-de la Parra (2003), among others. In that case, the T^2 statistic can be computed without difficulties but the results may not be coherent with the original data.

None of the above strategies take into account the peculiarity of CoDa, which is the case in most of the literature reviewed. We only found two articles where the distinctiveness of CoDa is mentioned.

A first attempt to implement a CC for compositional processes is made by Boyles (1997). He develops a chi-square CC to monitor multinomial and Dirichlet data. The Dirichlet distribution has some very restrictive properties, such as complete subcompositional independence, which makes it impossible to model any reasonable dependence structure for CoDa. Boyles (1997) uses simple descriptive graphs to compare the χ^2 chart with a T^2 chart based on a log-ratio transformation using as a divisor the last component of the composition (known as additive log-ratio transformation - alr) with the main drawback that is a non-isometric transformation. It is found that the T^2 chart based on log-ratios is more sensitive than the χ^2, but the author states that "the computational complexity of the optimal approach [...] makes it impractical in many shopfloor situations". We consider that the advantages of using the correct "optimal approach" go beyond the "computational complexity", considering the recent advances in automated manufacturing (Stoumbos et al. (2000)).

Another proposal for monitoring compositional data is made by Yang et al. (2004), where they control the quantity of different sizes of aggregates for the asphalt industry. They propose two ways of defining acceptance regions. The first one is by performing multiple univariate control charts, which is not optimal when a multivariate quality control is desired (Montgomery (2009)). The second method is based on an additive approach (not log-ratio), thus not consistent with CoDa nature.

In this paper, we demonstrate that applying a typical T^2 CC to CoDa in any of the above-mentioned situations is not useful, and propose an alternative methodology based on a log-ratio approach. In the following section, theory on CoDa is reviewed. In Section the inconsistencies of typical solutions are exposed through simple examples and a new CC is proposed in Section . It is compared in terms of ARL with the typical T^2 CC in Section and finally an example from the industry is used to demonstrate its applicability.

CoDa theory

The Simplex $\mathcal{S}^p$ is defined mathematically as $\mathcal{S}^p = \{\mathbf{x} \in \mathbb{R}^p | x_i > 0, \Sigma_{i=1}^p x_i = \kappa\}$. Compositions provide information about relative values of components; its total sum is not informative. Therefore, every statement about a composition can be stated in terms of ratios of components (Aitchison (1986)).

Aitchison observed that log-ratios are more easily handled than standard ratios and proposed a new methodology based on the former. Those projections enable representation of CoDa in the real space (coordinates), where standard unconstrained multivariate statistics can be applied. Inference therein is translatable back into compositional statements.

Handling data with ratios enables working with different constant sums κ or scaling the composition to a given value (usually 1). This operation is called *closure* and does not affect the ratios between components:

$$\mathcal{C}(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^p x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^p x_i}, \ldots, \frac{\kappa \cdot x_p}{\sum_{i=1}^p x_i}\right]$$

A *subcomposition* $\mathbf{x_s}$ of a composition $\mathbf{x}$ is obtained applying a closure operation to the subvector $[x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$ of $\mathbf{x}$. Subindexes $i_1, i_2, \ldots, i_s$ tell which parts are selected in the subcomposition.

There are three main transformations, but the one that has better properties is the ilr (isometric log-ratio) first presented in Egozcue et al. (2003). The ilr transformation switches from compositions in the Simplex $\mathcal{S}^p$ to an orthonormal basis in the real space $\mathbb{R}^{p-1}$. Unfortunately, there is no unique basis in the real space. One method for determining a basis is by using a sequential binary partition of the components (Egozcue and Pawlowsky-Glahn (2005)), known as balances. An explicit transformation formula for one such basis is:

$$\begin{aligned} \mathbf{x} &\xrightarrow{\text{ilr}} \mathbf{z} = (z_1, \ldots, z_{p-1})' \\ z_i &= \sqrt{\frac{i}{i+1}} \log \frac{x_{i+1}}{\sqrt[i]{\prod_{j=1}^i x_j}} \end{aligned} \tag{2}$$

for $i = 1, \ldots, p-1$. The inverse transformation, which recovers the composition from its coordinates, is called ilr^{-1}. Note that zeros are not allowed in Equation (2). Those elements must be previously replaced with specific techniques that can be found in Martín-Fernández et al. (2011) and references therein.

There are two main conditions that should be fulfillled by any statistical analysis applied to compositions: *scale invariance* and *subcompositional coherence* (Aitchison (1986)).

The scale invariance principle emphasizes the idea that a composition provides information only about relative values, so ratios of components are the relevant entities to study. In that case, the value of the constant sum κ is not relevant because the ratio remains unchanged. In practical situations, this means that analysis should be the same whether the data set is in proportions, percentages or ppm.

The subcompositional coherence principle demands that, whenever working with the full composition or with a subcomposition, inference about relationships within the common parts should be the same. Working with ratios or, equivalently, log-ratios, involves not only scale invariance but automatically subcompositional coherence because ratios within a subcomposition are equal to the corresponding ratios within the full composition.

Inconsistences of typical T^2 applied to CoDa

In this section we show that the typical solutions for the three scenarios described in

the introductory section are not consistent with compositional nature because they fail to fulfill the condition of subcompositional coherence. We will not cover the principle of scale invariance because it is not violated by the typical T^2 statistic from Equation 1. It can be proved with simple algebraic operations that the T^2 value would be the same whether the data units are, for example, in proportions or percentages.

When one variable is deleted in order to avoid collinearity, the resulting confidence interval is an hyper-ellipse, which has to be drawn in the Simplex. Hyper-elliptical shapes have to be avoided in restricted spaces because limits can easily drop out of the sample space. We illustrate this with two examples in $\mathcal{S}^3$.

We simulate 79 samples of a dataset that we call "arch shaped" because of its contour and another dataset called "vertex data" of 48 samples near the vertex x_1. Both are drawn in Figure 2. As the variance-covariance matrix is singular, we use only the first two components to calculate the T^2 statistic. Whatever variable is removed from the composition, the value of the T^2 statistic is going to be the same (Barceló-Vidal et al. (1999)).

We set a control limit for Phase I with unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows

$$UCL = \frac{(m-1)^2}{m} B_{(\frac{\alpha}{2};\frac{p}{2};\frac{m-p-1}{2})} \qquad (3)$$

Where $B_{(\frac{\alpha}{2};\frac{p}{2};\frac{m-p-1}{2})}$ is the $1-\alpha$ percentile of the beta distribution and the values m and p are the sample size and the number of variables, respectively (Tracy et al. (1992)). For $\alpha = 0.03$ the control limits are $UCL = 6.788$ and $UCL = 6.641$ for the "arch shaped" and the "vertex data", respectively. It can be seen in Figure 2 that the resulting contour ellipses fall out of the sample space and do not follow the distribution of the samples.

Many real datasets have this "arch shaped" structure, such as the volcanic gas

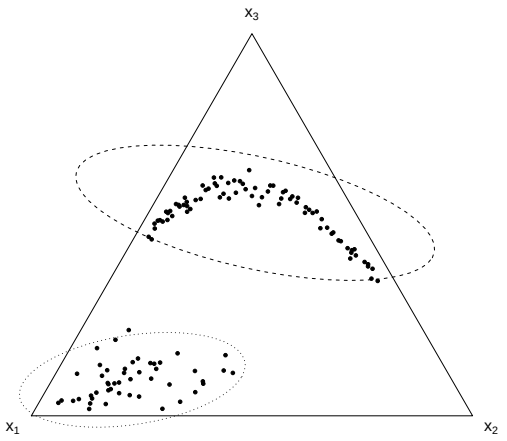


Figure 2: Contour ellipses for the "Arch Shaped" (dashed lines) and "Vertex data" (dotted lines)

chemistry of a volcano system or the reaction at equilibrium of hydrochloric acid (Buccianti (2011)) or sediment samples at different water depths and many others from industrial and scientific applications that are given in Aitchison (1986). Many other datasets with non homogeneus compositions like "vertex data" can be found in this same reference.

No satisfactory solution is obtained by deleting from the covariance matrix the eigenvector corresponding to the smallest eigenvalue. In the "arch shaped" example, data lives in $\mathbb{R}^3$ in a plane perpendicular to the vector $(1, 1, 1)$. Performing a principal component analysis retaining only the first two components would be equivalent to select a plane intersecting the previous one. A plane will never fit a data set with this particular shape. A detailed study of this effect is given in Aitchison (1986).

When the covariance matrix is near-singular and collinearity is not detected, the effect of computing the T^2 statistic with the corresponding covariance matrix is disastrous and the CC is no longer credible (Mason and Young (2001)). If prior knowledge about the nature of the data is available, this situation can be settled with a simple closure operation.

The third situation described in the introductory section, which consists on working with subcompositions, is analysed through a simple simulated example. We simulate 50 observations $\mathbf{x} = (x_1, x_2, x_3)$ and add an extra observation (square (■) in Figure 3), which is an outlier. It can be seen from Figure 3 that the outlier does not have an extreme value of x_1 (the smallest x_1 is limited by the dashed line) and neither does x_2 (the smallest x_2 is limited by the dotted line). However, the outlier has a large value of x_3, as it is far from the solid line that limits the highest x_3 value from the dataset.

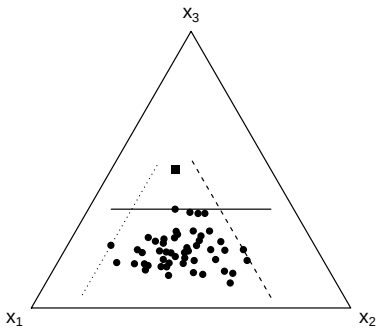
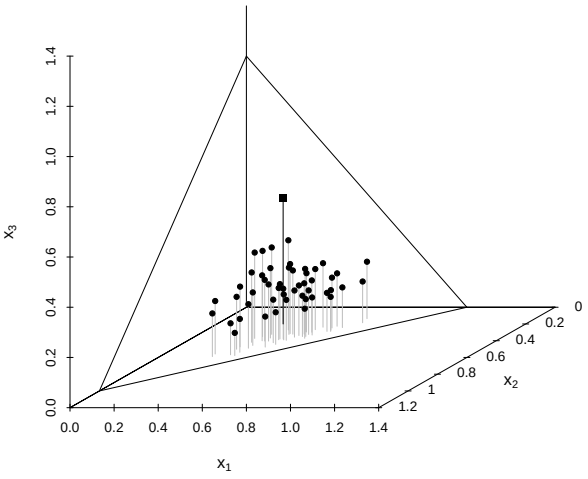
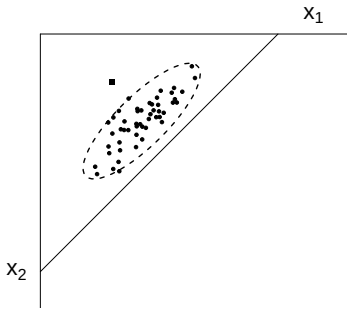


Figure 3: Simulated dataset of 50 observations and an extra outlier (■) with limiting lines for the smallest x_1 (dotted line), the smallest x_2 (dashed line) and the highest x_3 (solid line).

If, instead of working with the full composition the practitioner only collects data from the two first variables, then the dataset is the result of the projection of the previous Simplex into the plane $x_3 = 0$ in $\mathbb{R}^2$, as shown in Figure 4a. When a contour region is settled in this plane (Figure 4b), it can be seen that the outlier is also found to be out-of-control. However, when attempting to identify the cause of the anomaly, it can be interpreted that, given the value of x_1, the value of x_2 is not where it should be (or vice versa). This conclusion is not coherent with original data, as we have seen that the problem with this observation was in the value of x_3. So this approach is not subcompositionally coherent.



(a) Subcomposition (x_1, x_2)



(b) Projection into $x_3 = 0$

Figure 4: Working with the subcomposition (x_1, x_2) means projecting the dataset into the plane $x_3 = 0$ (a). The contour region of the projection leads to a wrong conclusion on the causes of the anomaly in the outlier ■ (b).

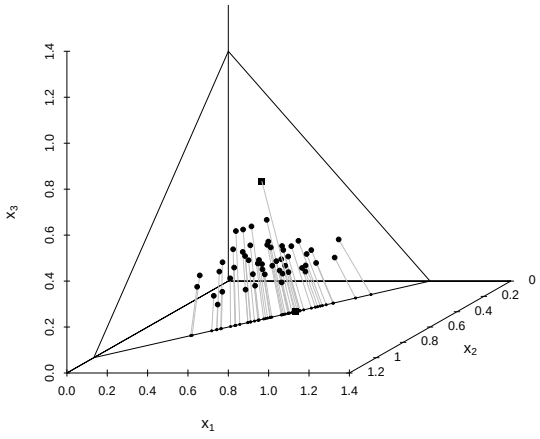


Figure 5: The closed subcomposition is the result of the projection into the edge (x_1, x_2). Note how in this subcomposition, sample ■ is not an outlier anymore.

If aware of the nature of the data, using a subcomposition would mean a closure operation of the subvector of selected components. In the example, the $\mathcal{C}(x_1, x_2)$ is equivalent to project the dataset into the edge $x_1 + x_2 = 1$ (Figure 5). In that case the outlier is no longer atypical – it lies in the middle of the dataset – which is consistent with the whole original composition.

When analysing log-ratios between components, the results must be the same whether the whole composition or a subcomposition is used (subcompositional coherence principle) or even if the subcomposition is closed or not (scale invariance principle).

CoDa T^2 Control Chart

Based on the theory above, we now describe the proposed method for calculating the T^2 statistic for compositional variables. This method is consistent with the CoDa nature because it fulfills the conditions of scale invariance and subcompositional coherence.

Given $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, a p-part composition and $\mathbf{z} = (z_1, \ldots, z_{p-1})$, its ilr coordinates as defined in Equation 2, the CoDa T^2 statistic (T^2_C) is defined as

$$(T^2_C)_t = (\mathbf{z}_t - \boldsymbol{\mu}_z)' \boldsymbol{\Sigma}_z^{-1} (\mathbf{z}_t - \boldsymbol{\mu}_z) \quad (4)$$

Where $\mathbf{z}_t$ is the coordinate of the observed composition at time t and $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are the mean vector and the variance covariance matrix of the log-ratio coordinates. In practice, it is necessary to estimate both values in Phase I, as is done with typical standard methods.

It can be easily demonstrated that the T^2_C statistic is not affected by the basis used for calculating the ilr coordinates. In practice, the user would select a basis that is convenient for easy interpretation (Egozcue and Pawlowsky-Glahn (2005)).

We assume that the in-control observation vectors $\mathbf{z}_t$, $i = 1, \ldots, m$ are i.i.d. multivariate normal random vectors ($\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$) with common mean vector and covariance matrix. In that case, the vector $\mathbf{x}_t$ is said to follow a normal distribution on the Simplex: $\mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ (Mateu-Figueras et al. (2013) and Mateu-Figueras and Pawlowsky-Glahn (2008)). This is a natural assumption because, as stated in Buccianti (2011), "Whenever there is a change in a composition by an independent process able to produce random variation, log-rationed data tend to become normally distributed".

Conceptually, we are comparing each observation with the geometric mean because the $\boldsymbol{\mu}_z$ from Equation 4 is the coordinate of the geometric mean of the raw composition. This is a better measure of center than the arithmetic mean because usually the univariate distributions of compositions do not follow normal distributions (for which it is convenient to use T^2) but log-normal distributions do (Aitchison (1986) and Buccianti (2011)).

The control limit (UCL) of the T^2_C control chart is calculated in the same way as that in standard multivariate control charts for individual observations (Tracy et al. (1992)) but is applied to the ilr coordinates.

We apply the T_C^2 CC to the examples described in the previous section: "arch shaped" and "vertex data". The control regions are drawn in the Simplex and in the coordinate space ($\mathbb{R}^2$) for both examples in Figure 6. It can be seen that the well-known elliptical contour is only found in the coordinate space. The same contour is deformed when transformed back to $\mathcal{S}^3$ due to the special geometry in the Simplex. The coordinates follows a multivariate normal distribution in both examples.

Comparative study

We compare the in-control performance of T_C^2 CC with the one obtained after deleting one variable and computing the typical T^2 when parameters are known. The run length (RL) distribution and its average (ARL) and percentiles are used as performance indicators. A simulation program similar to that used in Champ et al. (2005) for unknown parameters is used.

The data in which the performance of both methods is going to be tested is considered normal data in the Simplex $\mathcal{N}_{\mathcal{S}^3}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ with known parameters

$$\boldsymbol{\mu}_z = (0, 0) \quad \boldsymbol{\Sigma}_z = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$$

Where $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are the parameters of the multivariate normal distribution in the coordinate space in $\mathbb{R}^2$.

Without loss of generality we consider a log-ratio uncorrelated (diagonal) covariance matrix and a mean vector which is located in the centre of the ternary diagram. We also consider seven more mean vectors (covariance matrix remains unchanged) going from the center of the ternary diagram to the vertex x_3 (Table 1). Thus, comparison of both methods is going to be done across these eight scenarios considering homogeneous and heterogeneous compositions.

For better understanding of the cases considered, simulated samples of size 30 have been drawn in the ternary diagram with mean vectors numbered 0, 3 and 7 from Table 1.

Table 1: Values of the mean vector considered in the simulation

	$\boldsymbol{\mu}_x$		
Scenario	x_1	x_2	x_3
0	0.33	0.33	0.33
1	0.29	0.29	0.42
2	0.25	0.25	0.50
3	0.21	0.21	0.58
4	0.17	0.17	0.67
5	0.12	0.12	0.75
6	0.08	0.08	0.83
7	0.04	0.04	0.92

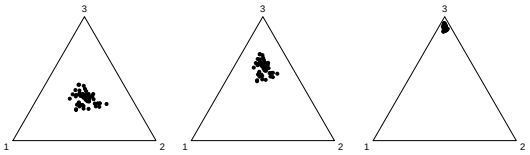


Figure 7: Samples from a $\mathcal{N}_{\mathcal{S}^3}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ with $\boldsymbol{\mu}_z$ equal to the coordinates of scenarios 0, 3 and 7.

The corresponding parameters of the composition ($\mathbf{x}$) in $\mathcal{S}^3$ are $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_x$. The mean of $\mathbf{x}$ is calculated by $\boldsymbol{\mu}_x = \text{ilr}^{-1}(\boldsymbol{\mu}_z)$ but there is no exact formula for obtaining $\boldsymbol{\Sigma}_x$ from $\boldsymbol{\Sigma}_z$. We estimate it from the ilr^{-1} of one million samples from a $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ for each of the mean vectors of Table 1. This matrix is considered as the known covariance matrix of $\mathbf{x}$.

The control limit for known parameters follows a χ_p^2 distribution with p degrees of freedom. In both CC (T_C^2 and T^2), the control limit is set at $UCL = 10.597$ with $\alpha = 0.005$.

The simulation program is outlined as follows:

1. Generate a random vector $\mathbf{z}_t$ from a $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ to represent the new process information observed at time t and calculate $\mathbf{x}_t = \text{ilr}^{-1}(\mathbf{z}_t)$. Compute T_C^2

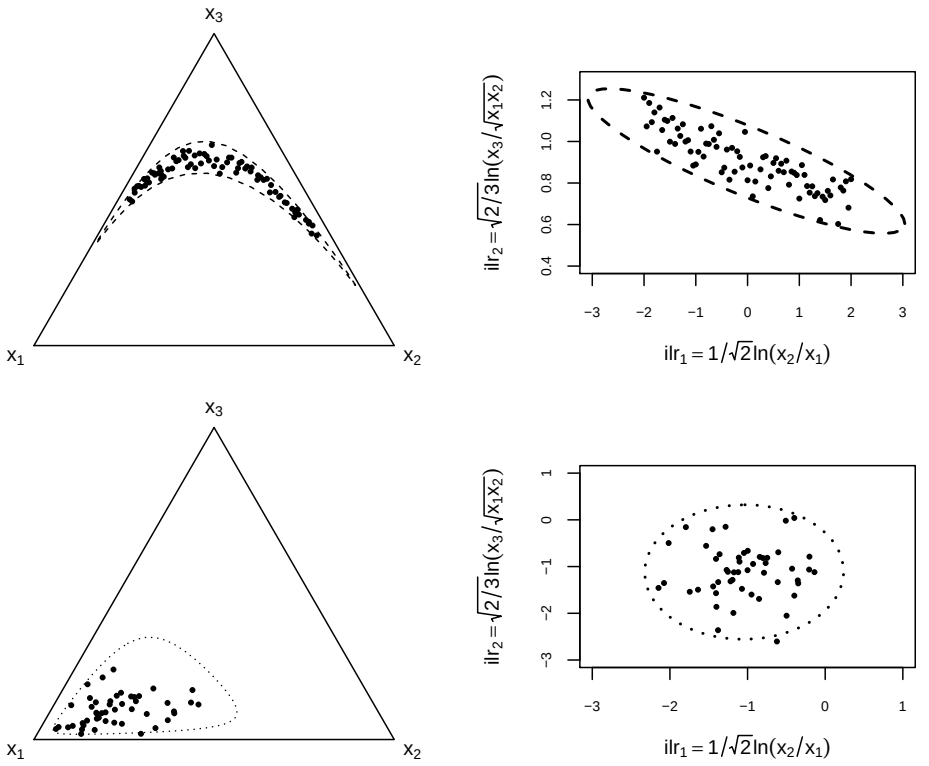


Figure 6: "Arch Shaped" (above) and "Vertex data" (below) with contour limits using T_C^2 and the corresponding ilr coordinates with the same contour region.

from Equation 4 and T^2 from Equation 1 after deleting one component (i.e., the last one).

2. Compare both values (T_C^2 and T^2) with the control limit. If no signal is observed, then go to step 1. If a signal is observed in one of them, retain the run length value and go to step 1 until a signal is observed in the other one. Once there is a run length value for both processes, go to step 3.

3. Record both run lengths in a separate vector.

4. Repeat steps 1-3 until the desired number of repetitions has been completed (100,000).

5. Repeat steps 1 to 4 for each of the 8 mean vectors $\boldsymbol{\mu}_z$.

As a result, we have $100,000$ RL values for each of the eight values of $\boldsymbol{\mu}_z$. The RL follows a geometric distribution with mean ARL$= 1/\alpha$ when the chart statistics are independent and identically distributed and the control limits are constant. For $\alpha = 0.005$ the mean is ARL$= 200$.

This is always true for the T_C^2 chart, as can be seen in Figure 8, but the ARL of the T^2 decreases as the distribution of the composition moves to the vertex.

From Table 2, it can be seen that the ARL in the CoDa approach is near the theoretical value of 200 with some errors due to the simulation. On the contrary, the mean and the quantiles of the RL of the typical approach are decreasing as we move through the different scenarios, which means that the chart will signal more often (false alarms) when samples are near the vertex.

The increased false alarms are due to the fact that contour ellipses of the typical approach contain extreme observations or even values that are not in the sample space as in the "vertex data" of Figure 2.

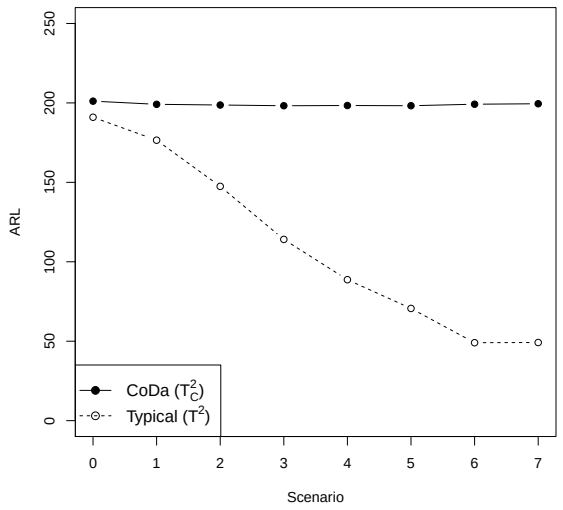


Figure 8: ARL of the T^2_C compared to the typical T^2 for the eight simulated scenarios

Table 2: The in-control performance of the T^2_C (CoDa) and the typical T^2 CC (Typical) with known parameters.

Scenario		ARL	SDRL	Q10	Q50	Q90
0	CoDa	201.12	201.91	21	139	463
	Typical	190.99	190.12	20	132	442
1	CoDa	199.12	198.36	21	138	459
	Typical	176.54	176.57	18	122	408
2	CoDa	198.68	199.28	21	137	459
	Typical	147.48	148.26	15	102	340
3	CoDa	198.22	198.4	21	137	456
	Typical	114.1	115.17	12	79	264
4	CoDa	198.37	200.37	20	137	458
	Typical	88.72	89.18	9	61	206
5	CoDa	198.24	198.5	20	137	457
	Typical	70.64	70.96	7	49	163
6	CoDa	199.18	200.62	21	137	458
	Typical	49.06	49.6	5	34	114
7	CoDa	199.47	200.07	21	138	459
	Typical	49.19	49.63	5	34	113

For the example in scenario 7 with $\boldsymbol{\mu}_x = (0.04, 0.04, 0.92)$ an extreme value such as $\mathbf{x}_t = (0.01, 0.01, 0.98)$ is not likely to occur considering the natural distribution of the composition. Such a change would mean that components 1 and 2 had reduced by one fourth. With a log-ratio approach, this observation would be found as an out of control observation (T^2_C CC). Despite the extremeness of this observation, in the typical T^2 CC it would be found to be in control.

When the composition is homogeneous, both methods perform well. But the difference between both methods is more acute when the samples are close to the vertex. We show an application of a real vertex dataset in next section.

Example

Here we supply an example of an industrial application using the data of Holmes and Mergen (1993), which is also used in Sullivan and Woodall (1996) and reproduced in Montgomery (2009). The data describes the particle size distribution (percentage by weight) for a plant in Europe. There are 56 observations with three components L, M, and S, denoting the percentages classified as large, medium and small, respectively. The dataset is in Table 3.

As the sum of each row is constant, in the three previous articles, authors decided to suppress one component: "Only the first two columns are used in the analysis since the total of the percentages is always 100 and the variance-covariance matrix will not invert under these conditions " (Holmes and Mergen (1993)) or "Since a dependency exists, only the first two components of each observation were used " (Sullivan and Woodall (1996)). The removed component in the three cases is the percentage of S.

A typical T^2 CC was set up to ensure that the particle size distribution was being manufactured in a consistent manner. The arithmetic mean vector is $\bar{\mathbf{x}}' = (5.682, 88.220, 6.098)$ and the sample covariance matrix is $\mathbf{S}_1$. Note that $\mathbf{S}_1$ is degenerate.

$$\mathbf{S}_1 = \begin{pmatrix} 3.770 & -5.495 & 1.725 \\ -5.495 & 13.529 & -8.033 \\ 1.725 & -8.033 & 6.308 \end{pmatrix}$$

Sullivan and Woodall (1996) proposed another estimator of the covariance matrix called the *sample covariance matrix of successive observations*, denoted by $\mathbf{S}_5$, also known as $\mathbf{S}_D$ in other references (Williams et al. (2006)).

This estimator does not perform very well for detecting outliers but has good properties for detecting sustained step shifts in the mean vector. In Williams et al. (2006), the distribution of the $T^2_{\mathbf{S}_5}$, which is the Hotelling statistic using $\mathbf{S}_5$, as an estimator of the covariance matrix, is studied. Recommendations on the calculations of the UCL for given historical data set size (m) and variables (p) are given.

For a $p < 10$, when $m > p^2 + 3p$ the UCL has to be calculated using the χ^2 distribution. With a false alarm probability of $\alpha = 0.003$, the limit is set at UCL= 11.618. Instead, Sullivan and Woodall (1996) obtained a limit of 11.35 from simulation that corresponds to a false-alarm probability of 0.003 for each of the 56 independent observations. The authors also suggest a limit of 10.55 for the CC using $\mathbf{S}_1$. Both limits suggested by Sullivan and Woodall (1996) will be considered in this example.

The sample covariance matrix of successive differences $\mathbf{S}_5$ for this example is

$$\mathbf{S}_5 = \begin{pmatrix} 1.562 & -2.093 & 0.531 \\ -2.093 & 6.721 & -4.628 \\ 0.531 & -4.628 & 4.097 \end{pmatrix}$$

which is also a singular matrix.

With only considering the first two components of each observation (i.e., the first two columns and rows of the variance-covariance estimators), the typical T^2 statistic is calculated using $\mathbf{S}_1$ and $\mathbf{S}_5$ and CC of Figure 9 are obtained. A control chart with $T^2_{\mathbf{S}_1}$ do not detect any outlier, while observations 26, 45 and 52 are found out of control by $T^2_{\mathbf{S}_5}$ CC.

In the ternary diagram (Figure 10), samples are located near the vertex corresponding with 100% of medium (M) particle sizes, as it is the predominant component.

Control regions defined by $T^2_{\mathbf{S}_1}$ and $T^2_{\mathbf{S}_5}$ are drawn in Figure 10. The well-known elliptic profile of the Hotelling statistic is obtained in the ternary diagram.

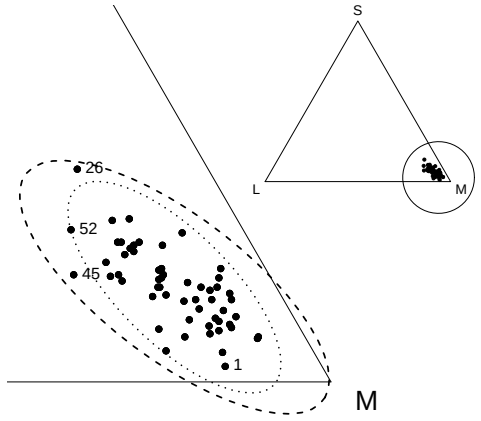


Figure 10: Zoom of the right-hand vertex of the ternary diagram with control regions using $\mathbf{S}_1$ (dashed) and $\mathbf{S}_5$ (dotted) as sample covariance estimators

It is easy to see how both elliptical profiles admit in the in-control region observations that are not in the sample space. At the same time, if a shift occurs in the direction of increasing the percentage of M, it is never going to be detected by the CC computed using $\mathbf{S}_1$. The same occurs if $\mathbf{S}_5$ is used and a shift occurs in the direction of observation 1, which has the smallest values of S.

As the data handled in this example is compositional, it is more convenient to apply the T^2_C from Equation 4. For that particular example, a basis has been selected so that the projected coordinates $\mathbf{z} = (z_1, z_2)$ are positive.

$$z_1 = \frac{1}{\sqrt{2}} \log \frac{M}{S}$$
$$z_2 = \sqrt{\frac{2}{3}} \log \frac{\sqrt{MS}}{L}$$

The results of this projection can be found in Table 3, as well as the values of

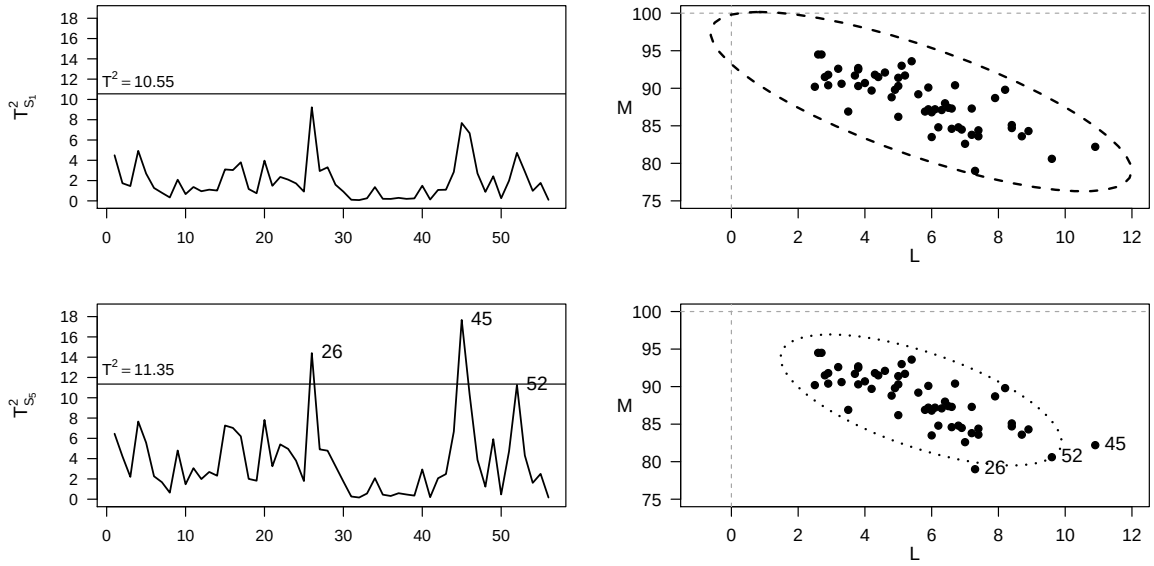


Figure 9: Typical T^2 CC using $\mathbf{S}_1$ and $\mathbf{S}_5$ as sample covariance estimators and its contour regions in the $L - M$ plane.

the T_C^2 statistic computed using Equation 4 with sample mean and covariance estimators

$$\bar{\mathbf{z}} = (1.959, 1.118)'$$
$$\mathbf{S}_z = \begin{pmatrix} 0.134 & 0.133 \\ 0.133 & 0.153 \end{pmatrix}$$

and with upper control limit $UCL = 10.472$ from Equation 3 with $\alpha = 0.003$. Note that $\bar{\mathbf{z}}$ corresponds to the coordinate of the sample geometric mean of the compositional dataset: $\mathbf{G}_x = (5.40, 89.03, 5.58)$

If a T_C^2 CC is computed, then the elliptical control region is obtained in the real space $\mathbb{R}^2$ of the coordinates obtained after applying an ilr transformation (Figure 11 left). The control region in the Simplex is shown in Figure 11 (right).

It can be seen that observations detected as outliers with the typical approach are no longer considered as atypical under the CoDa approach. Observation 26 has a large absolute value of S and small M; thus, the log-ratio between M and S (z_1) is small but not that much different from the other ratios. If compared with the geometric mean, we see that a measure of centre of the ratio M vs S is 16 and for observation 26 is not that different: 5.7.

Observation 45 has large L if compared directly with other values of large particle size, although the value of z_2 – which ratio has L as a denominator – is not that small compared with other ratios. And finally, observation 52 signals in the typical T^2 CC because of its small M, although the relative proportions between components are perfectly met.

On the other hand, observation 1, which has the lowest value of S and the third highest value of M, is now detected as atypical due to the hight log-ratio between M and S (z_1). Again, if compared with the geometric mean we obtain a ratio M vs S of 93, which is almost 6 times more than the same ratio of the geometric mean. The resulting T_C^2 CC is shown in Figure 12.

Conclusions

In this paper, we proposed a multivariate Hotelling T^2 control chart (T_C^2) suitable for monitoring individual composition of a mix-

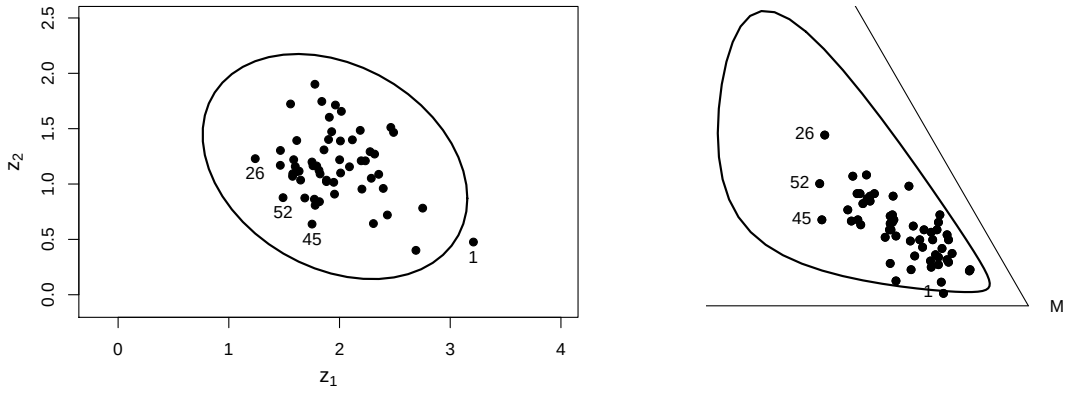


Figure 11: Contour region in the ilr plane (left) and in the Simplex (right). Only observation 1 is out-of-control.

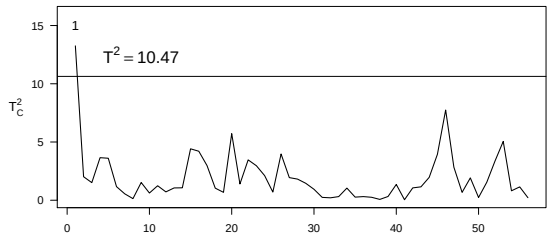


Figure 12: T_C^2 control chart.

ture. The proposed control chart is based on an ilr transformation of the data that moves the data from a restricted space into a non restricted space where the standard T^2 control chart can be applied. Conclusions are then translatable back to the restricted space.

We showed that the typical approach of applying a T^2 control chart after deleting one variable of the composition is not consistent with the CoDa nature. When the dataset lies near the vertex, the control region allows for samples out of the sample space. Also, when the dataset has specific shapes, e.g., like an arch, the typical method does not provide a reasonable model of the data.

Our simulation study showed that, assuming that compositions of mixtures follow a normal distribution on the Simplex, the performance of the T_C^2 is better than the typical T^2 in terms of in-control ARL, specially when samples are close to a vertex. When samples are homogeneous, both methods perform well.

A final promising research topic includes studying other estimators of the covariance matrix of the transformed coordinates, such as the covariance matrix from the vector differences between successive observations. This is a more robust estimate and the resulting control chart has been demonstrated to be more effective in detecting step or ramp shifts in the mean vector. The new estimate applied to the coordinates is likely to detect step and ramp shifts in terms of ratios of components.

Acknowledgment

This work has been partially financed by the Agència de Gestió d'Ajuts Universitaris i de Recerca (Ref: 2009SGR424) and by the Ministerio de Ciencia e Innovación (Ref:MTM2012-33236 and Ref: MTM2009-13272). The authors wish to thank the Universitat de Girona for the grant given to Marina Vives-Mestres during her doctoral studies. The authors appreciate the helpful comments by the Editor and the referee.

Table 3: Example from Holmes and Mergen (1993) showing the data and the ilr coordinates and the corresponding T_C^2 statistic

i	L	M	S	z_1	z_2	T_C^2	i	L	M	S	z_1	z_2	T_C^2
1	**5.40**	**93.60**	**1.00**	**3.21**	**0.48**	**13.26**	29	7.40	83.60	9.00	1.58	1.07	1.45
2	3.20	92.60	4.20	2.19	1.48	2.02	30	6.80	84.80	8.40	1.63	1.12	0.94
3	5.20	91.70	3.10	2.40	0.96	1.51	31	6.30	87.10	6.60	1.82	1.09	0.24
4	3.50	86.90	9.60	1.56	1.72	3.65	32	6.10	87.20	6.70	1.81	1.12	0.21
5	2.90	90.40	6.70	1.84	1.75	3.60	33	6.60	87.30	6.10	1.88	1.02	0.31
6	4.60	92.10	3.30	2.35	1.09	1.17	34	6.20	84.80	9.00	1.59	1.22	1.04
7	4.40	91.50	4.10	2.20	1.21	0.55	35	6.50	87.40	6.10	1.88	1.03	0.27
8	5.00	90.30	4.70	2.09	1.16	0.14	36	6.00	86.80	7.20	1.76	1.17	0.31
9	8.40	85.10	6.50	1.82	0.84	1.52	37	4.80	88.80	6.40	1.86	1.31	0.25
10	4.20	89.70	6.10	1.90	1.40	0.62	38	4.90	89.80	5.30	2.00	1.22	0.07
11	3.80	92.50	3.70	2.28	1.29	1.24	39	5.80	86.90	7.30	1.75	1.20	0.32
12	4.30	91.80	3.90	2.23	1.21	0.72	40	7.20	83.80	9.00	1.58	1.09	1.35
13	3.70	91.70	4.60	2.12	1.40	1.06	41	5.60	89.20	5.20	2.01	1.10	0.04
14	3.80	90.30	5.90	1.93	1.47	1.07	42	6.90	84.50	8.60	1.62	1.11	1.06
15	2.60	94.50	2.90	2.46	1.51	4.41	43	7.40	84.40	8.20	1.65	1.04	1.15
16	2.70	94.50	2.80	2.49	1.47	4.21	44	8.90	84.30	6.80	1.78	0.81	1.96
17	7.90	88.70	3.40	2.31	0.64	2.98	**45**	**10.90**	**82.20**	**6.90**	**1.75**	**0.64**	**3.94**
18	6.60	84.60	8.80	1.60	1.16	1.04	46	8.20	89.80	2.00	2.69	0.40	7.75
19	4.00	90.70	5.30	2.01	1.39	0.68	47	6.70	90.40	2.90	2.43	0.72	2.84
20	2.50	90.20	7.30	1.78	1.90	5.72	48	5.90	90.10	4.00	2.20	0.95	0.68
21	3.80	92.70	3.50	2.32	1.27	1.39	49	8.70	83.60	7.70	1.69	0.87	1.92
22	2.80	91.50	5.70	1.96	1.71	3.46	50	6.40	88.00	5.60	1.95	1.02	0.24
23	2.90	91.80	5.30	2.02	1.66	2.96	51	8.40	84.70	6.90	1.77	0.86	1.54
24	3.30	90.60	6.10	1.91	1.60	2.11	**52**	**9.60**	**80.60**	**9.80**	**1.49**	**0.88**	**3.37**
25	7.20	87.30	5.50	1.95	0.91	0.70	53	5.10	93.00	1.90	2.75	0.78	5.05
26	**7.30**	**79.00**	**13.70**	**1.24**	**1.23**	**3.97**	54	5.00	91.40	3.60	2.29	1.05	0.81
27	7.00	82.60	10.40	1.47	1.17	1.94	55	5.00	86.20	8.80	1.61	1.39	1.15
28	6.00	83.50	10.50	1.47	1.30	1.81	56	5.90	87.20	6.90	1.79	1.16	0.22

References

Aitchison, J., (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London (UK). 416 p.

Aitchison, J., and Egozcue, J. J. (2005). "Compositional Data Analysis: Where Are We and Where Should We Be Heading?". *Mathematical Geology*, 37, 7, pp. 829-850.

Bacon-Shone, J. (2011). "A short history of compositional data analysis". In: Pawlowsky-Glahn, V., and Buccianti, A. *Compositional Data Analysis: Theory and Applications*. John Wiley, p. 400. Chichester (UK).

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (1999). Letter to the editor on "Singularity and Nonnormality in the Classification of Compositional Data" by G. C. Bohling, J. C. Davis, R. A. Olea, and J. Harff. *Mathematical Geology*, 31, 5, pp. 581-586.

Boyles, R. (1997). "Using the chi-square statistic to monitor compositional process data". *Journal of Applied Statistics*, 24, 5, pp. 589-602.

Buccianti, A. (2011). "Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach". In: Pawlowsky-Glahn, V., and Buccianti, A. *Compositional Data Analysis: Theory and Applications*. John Wiley, p. 400. Chichester (UK).

Champ, C. W., Jones-Farmer, L. A., and Rigdon, S. E. (2005). "Properties of the T^2 control chart when parameters are estimated". *Technometrics*, 47, 4, pp. 437-445.

Egozcue, J. J., and V. Pawlowsky-Glahn (2005). "Groups of parts and their balances in compositional data analysis". *Mathematical Geology*, 37, 7, pp. 795–828.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). "Isometric log-ratio transformations for compositional data analysis". *Mathematical Geology*, 35, 3, pp. 279–300.

Gonzalez-de la Parra, M., and Rodriguez-Loaiza, P. (2003). "Application of the Multivariate T^2 Control Chart and the MasonTracyYoung Decomposition Procedure to the Study of the Consistency of Impurity Profiles of Drug Substances". *Quality Engineering*, 16, 1, pp. 127-142.

Holmes, D. S., and Mergen, A. E. (1993). "Improving the performance of the T^2 control chart". *Quality Engineering*, 5, 4, pp. 619-625.

Kenett, R., and Zacks, S. (1998). *Modern industrial statistics: design and control of quality and reliability.* Duxbury Press Pacific Grove. Wadsworth Publishing.

Martín-Fernández, J.A., Palarea-Albaladejo, J., and Olea, R. A. (2011). "Dealing with zeros". In: Pawlowsky-Glahn, V., and Buccianti, A. *Compositional Data Analysis: Theory and Applications.* John Wiley, p. 400. Chichester (UK).

Mason, R. L., Tracy, N. D., and Young, J. C. (1997). "A practical approach for interpreting multivariate T^2 control chart signals". *Journal of Quality Technology*, 29, 4, pp. 396-406.

Mason, R. L., Chou, Y.-M., and Young, J. C. (2001). "Applying hotellings T^2 statistic to batch processes". *Journal of Quality Technology*, 33, 4, pp. 466-479.

Mason, R. L., and Young, J. C. (2001). *Multivariate statistical process control with industrial applications, 1st ed.* American Statistical Association and Society for Industrial and Applied Mathematics.

Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2008). "A critical approach to probability laws in geochemistry". *Mathematical Geosciences*, 40, 5, pp. 489–502.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J.J. (2013). "The normal distribution in some constrained sample spaces". *Statistics and Operations Research Transactions*, 37, 1, pp.29–56

Montgomery, D. C. (2013). *Statistical quality control: a modern introduction, 7th ed.*. Asia: John Wiley & Sons.

Ortiz-Estarelles, O., Martín-Biosca, Y., Medina-Hernández, M. J., Sagrado, S., and Bonet-Domingo, E. (2001). "Multivariate data analysis of quality parameters in drinking water". *The Analyst*, 126, 1, pp. 91-96.

Pawlowsky-Glahn, V., and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications.* (V. Pawlowsky-Glahn and A. Buccianti, Eds.) John Wiley., p. 400. Chichester (UK).

Stoumbos, Z. G., Reynolds, M. R., Ryan, T. P., and Woodall, W. H. (2000). "The State of Statistical Process Control as We Proceed into the 21st Century". *Journal of the American Statistical Association*, 95, 451, pp. 992–998.

Sullivan, J. H., and Woodall, W. H. (1996). "A Comparison of Multivariate Control Charts for Individual Observations". *Journal of Quality Technology*, 28, 4, pp. 398-408.

Tracy, N. D., Young, J. C., and Mason, R. L. (1992). "Multivariate control charts for individual observations". *Journal of Quality Technology*, 24, 2, pp. 88-95.

Williams, J. D., Woodall, W. H., Birch, J., and Sullivan, J. H. (2006). "Distribution of Hotellings T^2 Statistic Based on the Successive Differences Estimator". *Journal of Quality Technology*, 38, 3, pp. 217-229.

Yang, G., Cline, D., Lytton, R., and Little, D. (2004). "Ternary and multivariate quality control charts of aggregate gradation for hot mix asphalt". *Journal of materials in civil engineering*, 10, pp. 28-34.