International Journal on
Advanced Science
Engineering
Information Technology

Vol.11 (2021) No. 5
ISSN: 2088-5334

The Monitoring of Dirichlet Compositional Data

Reham W. Elshaer [a,*], Aya A. Aly [a]

[a] *Statistics Department, Cairo University, 1 Gamaa Street, Giza; 12613, Egypt*
Corresponding author: [*]*rehamelshaer@feps.edu.eg*

***Abstract*— Compositional data are used in many applications such as Cement, Asphalt, and many other Chemical industries. Such data represent random variables whose values must sum up to a certain constant. Quality engineers and technicians require monitoring compositional data and detecting the source of the irregularity in the process as soon as it happens. Throughout the literature, complicated methods were introduced to monitor compositional data. Such methods are computationally complex and can lead to difficulties in interpreting the results. The Dirichlet distribution is commonly used in the literature to model compositional data. In this study, we propose three simple methods to monitor the mean vector of the Dirichlet distribution. The first method is based on a MEWMA control chart. The second method is based on transforming the Dirichlet random variables into beta random variables and then monitoring them using multiple EWMA control charts, while the third method uses multiple EWMA control charts for transformed independent random variables. Using a simulation technique, the performance of the three methods is investigated, and the three methods performed very well under different sample sizes, many random variables, and values of the distribution parameters. When the process is out-of-control, the source of the out-of-control signal can be detected using Method 2 and Method 3. Method 2 maintained its good performance with a probability 0.99 of correctly detecting the source of the signal. Method 3 performed well except for the case of Dirichlet parameter values less than one. However, it maintained almost a probability of correct detection of at least 90% in most cases. The three proposed methods are simple, do not need complicated calculations, and can easily be applied and used by practitioners.**

***Keywords*— Compositional data; Dirichlet distribution; phase II monitoring; quality control; MEWMA chart.**

Manuscript received 8 Oct. 2020; revised 17 Mar. 2021; accepted 23 Apr. 2021. Date of publication 31 Oct. 2021.

I. INTRODUCTION

It is well-known that quality plays an important role in the success of many industries and services. Low-quality products will never survive in the competitive market. Quality has many definitions, one of which is that "quality is inversely proportional to variability" [1]. The set of statistical tools and techniques used to measure, maintain and improve the quality of products and services are referred to as statistical quality control (SQC). Statistical process control (SPC) is a subarea of SQC in which the quality of a product or a process is maintained by monitoring one or more quality characteristics.

In some cases, these quality characteristics are highly correlated, and their sum must be equal to one, and the data are called "compositional data." Compositional data (CoDa) analysis is applied in many industrial, economic, and psychological applications. Foley [2] is divided the individual's activities in a usual day into six components, treated them as compositional data, and compared their effect on active and non-active individuals. Quinn [3] treated the DNA counts as compositional data and used the appropriate statistical tools to analyze them. Compositional data is defined in the simplex space and not the real space to account for the sum constraint. Since the first research on CoDa by Pearson [4], many researchers have studied different ways to overcome the positivity and the sum constraints. Pawlowsky-Glahn and Buccianti [5] discussed and compared many of the work done on CoDa. It was mentioned that the Dirichlet distribution is the only known distribution so far to be defined in the simplex space with independent structure. This distribution is constructed from independent Gamma variables with the same scale parameters [6].

In literature, most of the work done to monitor CoDa, in general, used the log-ratio transformations. Aitchison [6] showed different log-ratio transformations to compositional data. These transformations were used to transform the data from the simplex space to the real space, which is mathematically easier to deal with, and thus, standard unconstrained multivariate techniques can be used. Praus [7] used the log transformations mentioned in Egozcue [8] to study the compositional data of treated wastewaters.

Pawlowsky-Glahn [9] made a comprehensive literature review of compositional data techniques proposed in the literature and their implementation in R.

Boyles [10] suggested a modified Chi-square chart to monitor compositional data and compared it to the T^2 control charts based on the log-ratio transformations. He argued that for practitioners and technicians -who are usually non-mathematicians- it is difficult and less practical to deal with these log-transformations, which require more complicated calculations. The author concluded that although the method used was less sensitive in extreme cases, it worked well to detect assignable causes.

Yang *et al.*[11] used the T^2 control chart to monitor compositional data and suggested multiple univariate control charts to detect the source of the out-of-control signal without using the log transformations of the compositional data. They used the standard normal approximation to compute the control limits. They concluded that the T^2 control chart accounted for the correlated nature of the data and resulted in lower out-of-control ARL. However, it failed to detect which component is the source of the out-of-control signal. Vives-Mestres *et al.* [12] suggested using a Hotelling T_c^2 control chart after transforming the compositional data using the isometric log-ratio (ilr) introduced [13]. They stated that one of the difficulties of using this transformation is the absence of a unique orthonormal basis in the real space to transform the compositional components from the simplex S^p to the real space R^{p-1}, besides the components cannot take zero values as this method depends on dividing the components by their geometric mean. They assumed that the compositional data follow a log-normal distribution which will not be the case in this paper. Vives-Mestres *et al* [12] compared their method to the traditional T^2 control chart after deleting one of the components. They concluded that both methods perform well in the homogeneous cases where the data are concentrated in the center, while their method performs better in the extreme cases where the data are concentrated in the vertex. They used fixed control limits in both methods depending on the normal distribution, therefore in extreme cases the traditional method failed to maintain the assumed in-control ARL while their method did. Vives-Mestres *et al* [14], [15] introduced some methods to decompose the T^2 control chart based on the ilr transformations to detect which component is the source of the out-of-control signal. Vives-Mestres *et al.* [14] found the percentage of correct detection of their method does not exceed 50%. Vives-Mestres *et al.* [15] continued their previous work and proposed two methods that can be used for cases where the number of compositional parts exceeds three. One of the proposed methods performed worse as the number of variables increases, and the other required using all possible log-ratios at every stage where a signal occurs; thus, the variance-covariance matrix changes.

Tran *et al.* [16] proposed a MEWMA control chart based on the ilr transformations to monitor CoDa. They used a Markov chain to assess the performance of their proposed method. They concluded that their method outperformed the T_c^2 control chart proposed by Vives-Mestres [12].

Up to our knowledge, no research was introduced to monitor the Dirichlet random variables. Such variables have different nature and a well-known distribution, unlike the compositional data studied in the previous papers. In addition, although Multivariate Exponentially Weighted Moving Average (MEWMA) control charts are usually preferred in monitoring multivariate quality characteristics in Phase II than the Hotelling T^2 control chart, they were used only in Tran [16] to monitor compositional data. The MEWMA charts are preferred because they use information from previous samples and not only information from the current sample being monitored, which makes them quicker in detecting small to moderate shifts. The remaining sections of this paper are arranged as follows. The Dirichlet probability distribution is presented in section 2. In section 3, the MEWMA control chart is proposed to monitor the Dirichlet random variables. Additionally, another two methods are proposed in the same section to detect quickly the source of the out-of-control signal using the special characteristics of the Dirichlet probability distribution. The proposed methods do not require complicated calculations and are easily applied. Moreover, the proposed methods have proved their efficiency for various sample sizes and various dimensions for the data. Simulations are done to assess the performance of the three proposed methods, and their results are presented in section 4. Final conclusions and future work are presented in section 5.

II. Material and Method

A. Dirichlet Probability Distribution

As for the Beta distribution, the Dirichlet random variables are related to a set of independent gamma variables, as shown below.

Let a set of p independent Gamma random variables be:

$$Z_i \sim \boldsymbol{G}\text{amma}(a_i, 1), i = 1,2, \ldots\ldots. p \qquad (1)$$

where $a_1, \ldots\ldots, a_p > 0$, and they are the shape parameters of the Gamma distribution. The random variables Y_i it is defined as.

$$Y_i = \frac{Z_i}{Z_1 + Z_2 + \ldots. Z_p}, i = 1,2, \ldots\ldots p \qquad (2)$$

The vector $Y = (Y_1, Y_2, \ldots\ldots, Y_p)$ is said to follow the *Dirichlet* $(a_1, \ldots\ldots, a_p)$ distribution , where $a_1, \ldots\ldots, a_p > 0$ and $Y_1 + Y_2 + \ldots. Y_p = 1$. The density of the subvector $(\mathrm{Y}_1, \ldots., \mathrm{Y}_{\mathrm{p}\text{-}1})$ is given by:

$$f(y_1, \ldots., y_{p-1}) = \begin{cases} \frac{\Gamma(a_1 + \ldots.. + a_p)}{\Gamma(a_1)\ldots..\Gamma(a_p)} y_1^{a_1 - 1} \ldots. y_p^{a_p - 1} & (y_1, \ldots.. y_{p-1}) \in M, \\ 0 & (y_1, \ldots.. y_{p-1}) \notin M, \end{cases} \qquad (3)$$

where $y_p = 1 - y_1 - \ldots. - y_{p-1}$ and M are defined as

$$M = \{(y_1, \ldots., y_{p-1}): y_1 > 0, \ldots., y_{p-1} > 0, y_1 + \ldots + y_{p-1} < 1\}$$

The marginal distribution of each variable from the random vector Y is distributed as Beta distribution with parameters $(a_j, \varphi - a_j)$, where $\varphi = a_1 + \ldots. + a_p$, j=1,...,p and is called the concentration parameter. Therefore, the Dirichlet distribution is considered as the multivariate extension of the Beta distribution with mean, variance, and covariance as follows:

$$E(Y_j) = \frac{a_j}{\varphi}, j = 1, \ldots\ldots\ldots, p \qquad (4)$$

$$Var(Y_j) = \frac{a_j(\varphi - a_j)}{\varphi^2(\varphi+1)}, j = 1,2,\ldots..p \qquad (5)$$

$$Cov(Y_j, Y_s) = -\frac{a_j a_s}{\varphi^2(\varphi+1)}, j \neq s; j,s = 1,..,p \qquad (6)$$

The concentration parameter φ (also known as the scale parameter) determines the shape of the Dirichlet distribution, as shown in Figure.1. For values of $a_i < 1$, the distribution concentrates in the corners and along the boundaries of the simplex. For values of $a_i > 1$, the distribution tends toward the center of the simplex. For values of a_i=1, the concentration parameter is equal to p, and the Dirichlet distribution becomes a uniform distribution in the p-1 simplex.

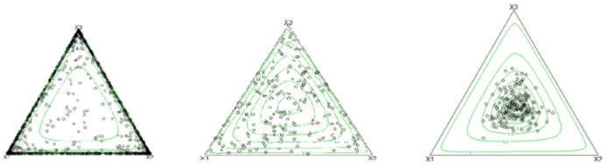

Fig. 1 The distribution of 3 Dirichlet random variables with $a_1 = a_2 = a_3 = 0.1, a_1 = a_2 = a_3 = 1$ and $a_1 = a_2 = a_3 = 10$ respectively from left to right

B. The Multivariate EWMA Control Chart (MEWMA-Method

The means of the Dirichlet random variables are functions of the Dirichlet parameters as shown in equation (4), and thus any shift in the parameters will be reflected in the mean. Therefore, the first proposed method is a multivariate EWMA control chart to monitor the means of the dirichlet random variables. An amendment was made to encounter the singularity of the variance-covariance matrix, as the p^{th} dirichlet random variable is a linear combination of the other (p-1) random variables. The pth random variable was dropped to overcome this, and any shifts in its parameter will be reflected in the chart statistic. Vives-Mestres *et al.* [12] criticized using the compositional data without transforming them to log-ratio variables due to having a higher probability of false alarms when using fixed control limits based on the normal distribution. To overcome this drawback, we will use simulations to find control limits that would sustain the in-control ARL assumed.

The EWMA chart statistics are defined for the i^{th} sample as follows:

$$E_i = \lambda\bar{Y}_i + (1-\lambda)E_{(i-1)}, i = 1,2,\ldots., \qquad (7)$$

where E_0 is the target mean vector of the (p-1) variables, λ is a constant that determines the weight given to current observations compared to the previous ones, $0 < \lambda \leq 1$ and the Hotelling T^2statistic is given by:

$$\mathrm{T}_{\mathrm{i}}^2 = \mathrm{n}(\mathrm{E}_{\mathrm{i}} - \mathrm{E(Y)})'\Sigma_{\mathrm{E}}^{-1}(\mathrm{E}_{\mathrm{i}} - \mathrm{E(Y)}) \qquad (8)$$

where $\Sigma_{\mathrm{E}} = \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2\mathrm{i}})\Sigma$, and Σ is the variance-covariance matrix of the Dirichlet random variables $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots\ldots, \mathrm{Y}_{\mathrm{p}-1})$.

C. EWMA Control Charts (Method 2)

The main challenge in the multivariate control charts is detecting which variable is the source of the out-of-control signal. Many approaches were introduced in the literature, "using univariate control charts with Bonferroni control limits" is one of them. It was proposed by Alt and Jain [17] as a method to detect the source of out-of-control signal in Phase I analysis. Alt and Jain [17] adjusted the control limits of the univariate Shewhart control charts to give the required overall false alarm probability. Although this method ignores the correlation between the variables, it can be used to detect which variable is the source of the signal. In the case of Dirichlet distribution, the correlation matrix depends only on the parameters of the Dirichlet distribution. Therefore, using the univariate control charts will not ignore the correlation between the variables, as they are monitoring the Dirichlet parameters. Therefore, no detected signal means in-control parameters and an unchanged correlation matrix. The covariance of the Dirichlet random variables is proportional to the product of their means, as shown in equation (6). As mentioned before, the Dirichlet distribution is a multivariate extension of the beta distribution. Therefore, a transformation was made from the p Dirichlet random variables to (p-*1*) Beta random variables that are only correlated through the p^{th} Dirichlet random variable.

Let

$$X_j = \frac{Y_j}{Y_j+Y_p} = \frac{Z_j}{Z_j+Z_p} j = 1,2,\ldots..,p-1, X_j \sim \text{Beta}(a_j, a_p) \qquad (9)$$

where $a_{.j}, a_p > 0$ and are the shape parameters of the Beta distribution.

Afterward, (p-*1*) EWMA control charts are introduced to monitor the means of the (p-*1*) Beta random variables, and their control limits are chosen to give the desired overall out-of-control average run length (ARL). The EWMA chart statistics are defined for the i^{th} sample and the j^{th} variable as follows:

$$E_{ij} = \lambda\bar{X}_j + (1-\lambda)E_{(i-1)j}, i = 1,2,\ldots., j = 1,2,\ldots.p-1 \qquad (10)$$

where E_{0j} is the target mean of the jth Beta random variable.

Assuming no shifts occurring in the distribution of the p^{th} Dirichlet random variable, shifts occurring in the distribution of the j^{th} Dirichlet random variable will only be detected by the j^{th}EWMA control chart. As shown in equation 9, each variable of the newly introduced Beta random variables is a function of the corresponding Dirichlet random variable and the p^{th} Dirichlet random variable. Therefore, shifts occurring in the pth Dirichlet random variable distribution will be detected by some or all of the EWMA charts. This will be assessed later using simulations.

D. Independent EWMA control charts (Method 3)

Ongaro and Migliorati [18] stated that partitioning the Dirichlet random variables into subsets and dividing each element in the subset by their sum will make these subsets independent from each other. Using this proposition, a method is introduced to transform the p Dirichlet random variables into p-1 independent random variables, and thus separate EWMA control charts can be used to monitor the means of these independent variables. A proof for this transformation is found in the Appendix.

Define the following new p-1 random variables:

$$Y_{j.1} = \frac{Y_j}{1-Y_1}, j = 2,3,\ldots\ldots,p \qquad (11)$$

Therefore, $Y_{j.1} = (Y_{2.1}, \ldots\ldots, Y_{p.1})$ will be a random vector distributed as *Dirichlet* $(a_2, \ldots\ldots, a_p)$ where $a_2, \ldots\ldots, a_p >$ 0and $Y_{2.1} + \ldots. Y_{p.1} = 1$

Now Y_1 is independent of the new set of the p-1 Dirichlet random variables $Y_{j.1}$.

Define new p-2 random variables:

$$Y_{j.1,2} = \frac{Y_{j.1}}{1-Y_{2.1}}, j = 3, \ldots\ldots, p \qquad (12)$$

Therefore $Y_{j.1,2} = (Y_{3.1,2}, \ldots\ldots, Y_{p.1,2})$ will be a random vector distributed as *Dirichlet* $(a_3, \ldots\ldots, a_p)$ where $a_3, \ldots\ldots, a_p > 0$ and $Y_{3.1,2} + \ldots . Y_{p.1,2} = 1$.

Now $Y_{2.1}$ is independent of the (p-2) Dirichlet random variables $Y_{j.1,2}$. Also Y_1 is independent of these variables.

This approach will be continued until the (p-1)th transformation is done as follows:

$$Y_{j.1,2,\ldots,(p-2)} = \frac{Y_{j.1,\ldots,(p-3)}}{1-Y_{(p-2).1,\ldots,(p-3)}}, j = (p-1), p \qquad (13)$$

Now $Y_{(p-1).1,2,\ldots,(p-2)}$&$Y_{p.1,2,\ldots,(p-2)}$ are Beta distributed with parameters (a_{p-1}, a_p), and they are independent of the variables $Y_1, Y_{2.1}, \ldots\ldots, Y_{(p-2).1,2,\ldots(p-3)}$.

In this section, *(p-1)* EWMA control charts are introduced to monitor the means of the (p-1) independent random variables. The control limits of the EWMA charts are chosen to give the desired overall out-of-control average run length (ARL). The chart statistics for the p-1 EWMA charts are defined for the *i*th sample and the *j*th transformed variable as follows:

$$E_{ij.K} = \lambda \bar{Y}_{j.K} + (1-\lambda) E_{(i-1)j.K}, i = 1,2, \ldots., j = 1,2, \ldots . p-1, \quad (14)$$

where *K* is a vector of the variables removed to attain the independence and $E_{0j.K}$ is the target mean of the *j*th transformed random variable.

The following flow chart summarizes the methodology and the implementation process:

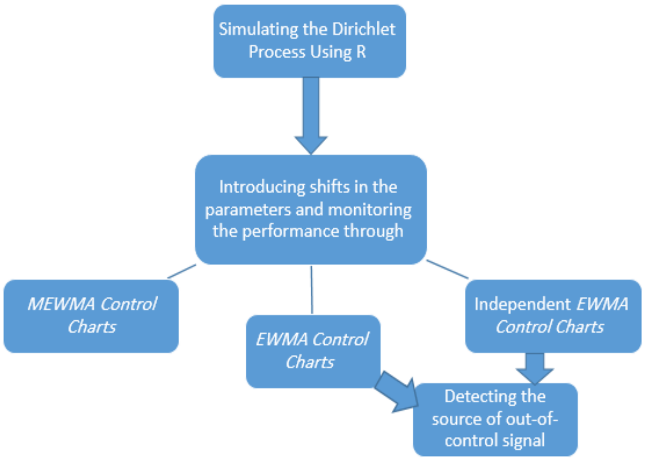
Simulating the Dirichlet Process Using R
Introducing shifts in the parameters and monitoring the performance through
MEWMA Control Charts
EWMA Control Charts
Independent EWMA Control Charts
Detecting the source of out-of-control signal

Fig. 2 Flow chart summarizing the methodology

III. Result and Discussion

Simulations of 100,000 runs were carried out to compare the performance of the three proposed methods. Shifts in the parameters were introduced in a multiplicative way, i.e., the out-of-control parameter $a_i^* = \delta a_i$. The shifts values used are $\delta = 0.2$, 0.5,0.8,1.2, 1.5, and 2. Models with a different number of variables were examined: three, five, and eight variables. Different simulation scenarios were considered based on changing the number of variables p (p=3,5,8), changing the sample size (n=5,10,15), and changing the shape of the Dirichlet distribution (a_is less than one, equal to one, and greater than one). The upper control limit *h* was chosen to ensure that the three methods have the same in-control ARL of 370. The three competing methods are compared based on the out-of-control ARL performance in the first subsection, and afterward, the probability of correctly detecting the source of the out-of-control signal is compared for methods 2 and 3 in the following subsection. The probability of correct detection was computed by dividing the number of runs when the chart -monitoring the variable with the shifted parameter gives an out-of-control signal solely without any signal from the remaining charts- by the total number of runs. For example, referring to equations 9 and 10 and using method 2, if a shift occurred for Dirichlet random variable Y_j, then the probability of correct detection = $\frac{no.\ of\ runs|E_j|>h\ and\ all|E_{i\neq j}|<h}{total\ number\ of\ runs}$.

A. *Comparing the out-of-control ARL performance of the three proposed methods*

In the first part of this section, the simulations of the three proposed methods at p=3 with different sample sizes and different values of the Dirichlet parameters are presented in detail. Tables I, II, and III show the out-of-control ARL values for the competing methods at different shift sizes.

TABLE I
COMPARISON OF THE OUT-OF-CONTROL ARL FOR a_1=10, a_2=12, a_3=20 FOR THE 3 METHODS WITH SAMPLE SIZES 5,10 &15

δ	Method 1			Method 2			Method 3		
	n=5, h=54.5			n=5, h=1.03			n=5, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	1.1	1	1	1	1	1	1.1	1	1
0.5	2.2	2	1.7	2.1	2	1.6	2.1	2	1.7
0.8	7.4	6.6	5.4	7.6	6.9	5.2	7.2	6.8	5.7
1.2	8.9	8.1	7.4	9.6	8.8	7.8	8.4	8.7	8.8
1.5	2.8	2.7	2.6	2.9	2.8	2.8	2.7	2.8	3.1
2	1.7	1.6	1.7	1.8	1.8	1.9	1.5	1.8	1.9
	n=10, h=109.5			n=10, h=1.03			n=10, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	1	1	1	1	1	1	1	1	1
0.5	1.7	1.5	1.1	1.5	1.4	1.1	1.5	1.4	1.1
0.8	4.5	4.1	3.5	4.5	4.1	3.4	4.3	4.2	3.7
1.2	5.2	4.9	4.4	5.5	5.1	4.7	4.9	5.1	5.2
1.5	2.1	2	1.8	2	2	2.1	1.9	2	2.2
2	1.1	1	1	1.1	1.1	1.2	1	1.1	1.3
	n=15, h=164.5			n=15, h=1.03			n=15, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	1	1	1	1	1	1	1	1	1
0.5	1.2	1.1	1	1.1	1.1	1	1.1	1.1	1
0.8	3.4	3.2	2.8	3.5	3.2	2.7	3.3	3.2	2.9
1.2	3.9	3.7	3.5	4.2	3.9	3.6	3.8	3.9	3.9
1	1.7	1.7	1.6	1.8	1.7	1.9	1.6	1	1.9
2	1	1	1	1	1	1	1	1	1

TABLE II

COMPARISON OF THE OUT-OF-CONTROL ARL FOR a_1=1, a_2=1, a_3=1 FOR THE 3 METHODS WITH SAMPLE SIZES 5,10 &15

δ	Method 1			Method 2			Method 3		
	n=5, h=54.5			n=5, h=1.01			n=5, h=1.02		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	3	3	3.3	2.8	2.8	2.6	3.2	2.8	2.8
0.5	7.5	7.3	7.5	7	7	5.8	7.4	6.7	6.6
0.8	55	55	55	56	56	38	66	46	46
1.2	83	83	83	102	102	66	68	100	100
1.5	15	15	15	19	19	13	13	18	18
2	5.8	5.8	5.8	7.2	7.2	5.9	5.4	6.9	7
	n=10, h=109.5			n=10, h=1.02			n=10, h=1.02		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	2.2	2.3	2.2	2.1	2	2	2.2	2.1	2.1
0.5	4.6	4.6	4.5	4.3	4.3	3.7	4.5	4.2	4.2
0.8	28	28	27	29	29	20	28	25	24
1.2	40	40	40	51	51	33	35	45	45
1.5	8.1	8	7.9	9.7	9.7	7.5	7.4	9.2	9.1
2	3.6	3.6	3.6	4.3	4.3	3.7	3.4	4.2	4.3
	n=15, h=164.5			n=15, h=1.02			n=15, h=1.02		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	2	2	2	1.8	1.8	1.7	1.9	1.8	1.8
0.5	3.6	3.6	3.6	3.4	3.4	3	3.5	3.3	3.2
0.8	18	18	18	19	19	14	18	17	16
1.2	26	26	26	32	32	21	24	29	29
1.5	5.8	5.8	5.8	6.9	6.9	5.6	5.5	6.6	6.5
2	2.9	2.9	2.9	3.4	3.4	2.9	2.7	3.3	3.1

TABLE III

COMPARISON OF THE OUT-OF-CONTROL ARL FOR a_1=0.2, a_2=0.3, a_3=0.4 FOR THE 3 METHODS WITH SAMPLE SIZES 5,10 &15

δ	Method 1			Method 2			Method 3		
	n=5, h=54.5			n=5, h=1.03			n=5, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	8	5.3	4.1	5.9	4.5	2.9	7.9	4.4	3.5
0.5	26	14	11.1	18.6	12.9	7.9	30.4	12.7	9.8
0.8	206	113	80.4	186	117	55.1	340	100	70.4
1.2	132	151	199	127	139	177	114	166	256
1.5	33	33.1	383	32.3	33.2	35.3	29.8	35.8	50
2	11	11	11.8	11	11	12	10.2	11.5	14.8
	n=10, h=109.5			n=10, h=1.03			n=10, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	4.6	3.4	2.8	3.7	2.9	2.1	4.5	2.9	2.4
0.5	11.7	8	6.5	9.2	7.2	3.3	11.6	6.9	5.9
0.8	102	61.7	46.3	87.1	60.7	17.1	131	53.2	41.4
1.2	84.7	84.4	95.3	81.5	81.3	28.9	75.6	86.6	109
1.5	17.2	15.8	16.5	16.8	15.9	6.6	16	16.2	18.9
2	6.3	6	6.3	6.3	6.2	3.3	6	6.2	7.2
	n=15, h=164.5			n=15, h=1.03			n=15, h=1.03		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
0.2	3.6	2.7	2.3	2.9	2.3	1.9	3.5	2.3	2.1
0.5	8	5.8	4.9	6.4	5.2	3.9	7.9	5.2	4.5
0.8	64.5	40.9	31.9	52.3	38.7	22.5	73.8	35.6	28.8
1.2	60.5	56.4	59.9	57.4	55.9	47.4	55.4	56.9	65.9
1.5	11.9	10.7	10.8	11.3	10.8	10.2	11.2	10.8	12
2	4.7	4.5	4.7	4.7	4.6	4.8	4.5	4.7	5.2

As shown in the tables, the performance of the competing methods is nearly the same when the parameters' values are greater than one. Method 1 maintains the same good performance when monitoring any of the parameters compared to the other two methods, as shown in Table II. However, they all perform worse when the parameters' values are less than one. This can be explained as the shifts are represented as multiple of the parameters. Increasing the sample size enhances the performance of the three methods, as shown in Tables I, II, and III. However, in this case, Method 2 is slightly better than the other two methods, as shown in Table III. Simulations showed that the three methods show slightly higher out-of-control ARL when increasing the number of the variables monitored for different values of the Dirichlet Parameters. However, increasing the sample size improves the performance of the three methods.

B. Detecting the Source of Out-of-control Signal

In this section, an assessment of the ability of Methods 2 and 3 to correctly detect the source of the out-of-control signal is presented under different scenarios. Method 2 has a higher power of detecting the source of the out-of-control signal than Method 3, for monitoring three Dirichlet random variables, as shown in Tables IV, V and VI. This can be explained as, in method 2 each variable has a beta distribution with its parameter a_j and the p^{th} variable's parameter a_p only. However, for Method 3, although the variables are independent, the first variable has the parameters of the latter variables defining its distribution. Therefore, shifts in the parameters of the latter variables may appear as well in the charts of the previous variables as an out-of-control signal. The percentage of correct detection for Method 2 is always 99% in all cases except for the case where the parameters take values equal or less than one for shifts δ =0.8 and 1.2.

This can be justified as follows: since shifts are defined as multiples of the parameters, then a shift of size 0.8 where a_i= 0.1, results in a shifted parameter of size 0.08, which is very close to the in-control parameter. However, this could be resolved by increasing the sample size. As shown in Table IV, increasing the sample size from n= 5 to n=10 then to n=15 increased the probability of detecting a shift of size 0.8 in a_3=0.1 from 0.76 to 0.88 then to 0.93**.**

Moreover, increasing the sample size enhances the performance of detection of both methods under any values of the Dirichlet parameters. If small shifts need to be correctly detected, a larger sample size is recommended. The same conclusions apply when increasing the numbers of variables p to 5 or 8; tables can be sent upon request.

TABLE IV
COMPARISON OF PROBABILITY OF CORRECT DETECTION FOR a_1=10, a_2=12, a_3=20 FOR METHODS 2 AND 3 WITH SAMPLE SIZES 5,10 &15

Method 2						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	1	1	1	1	1	1
0.5	1	1	0.99	1	1	1
0.8	0.99	0.99	0.99	0.99	0.99	0.99
1.2	0.99	0.99	0.99	0.99	0.99	0.99
1.5	0.99	0.99	1	0.99	1	1
2	0.99	0.99	1	1	1	1
Method 3						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	1	0.98	1	1	0.98	1
0.5	1	0.96	1	1	0.96	1
0.8	0.99	0.94	0.99	0.99	0.94	0.99
1.2	0.99	0.95	0.99	0.99	0.95	0.99
1.5	0.99	0.98	1	0.99	0.98	1
2	1	0.98	1	1	0.98	1

TABLE V
COMPARISON OF PROBABILITY OF CORRECT DETECTION FOR a_1=1, a_2=1, a_3=1 FOR METHODS 2 AND 3 WITH SAMPLE SIZES 5,10 &15

Method 2						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	0.99	0.99	1	0.99	1	1
0.5	0.99	0.99	0.99	0.99	0.99	0.99
0.8	0.93	0.93	0.96	0.97	0.98	0.98
1.2	0.87	0.87	0.94	0.94	0.96	0.96
1.5	0.98	0.98	0.99	0.99	0.99	0.99
2	0.99	0.99	0.99	0.99	0.99	0.99
Method 3						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	0.99	0.91	1	0.9	1	0.88
0.5	0.99	0.85	0.99	0.87	0.99	0.88
0.8	0.92	0.72	0.97	0.8	0.98	0.82
1.2	0.92	0.81	0.96	0.82	0.97	0.83
1.5	0.98	0.87	0.99	0.87	0.99	0.87
2	0.99	0.88	0.99	0.87	0.99	0.87

TABLE VI
COMPARISON OF PROBABILITY OF CORRECT DETECTION FOR a_1=0.2, a_2=0.3, a_3=0.4 FOR METHODS 2 AND 3 WITH SAMPLE SIZES 5,10 &15

Method 2						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	0.99	0.99	0.99	0.99	0.99	1
0.5	0.98	0.98	0.99	0.99	0.99	0.99
0.8	0.76	0.83	0.88	0.92	0.93	0.95
1.2	0.85	0.89	0.9	0.9	0.93	0.94
1.5	0.97	0.97	0.98	0.99	0.99	0.99
2	0.99	0.99	0.99	0.99	0.99	0.99
Method 3						
	n=5		n=10		n=15	
Shift	a_1	a_2	a_1	a_2	a_1	a_2
0.2	0.99	0.93	0.99	0.95	0.99	0.96
0.5	0.97	0.86	0.99	0.92	0.99	0.93
0.8	0.6	0.61	0.84	0.76	0.91	0.82
1.2	0.87	0.89	0.91	0.92	0.93	0.92
1.5	0.97	0.98	0.99	0.97	0.99	0.97
2	0.99	0.99	0.99	0.98	0.99	0.98

The effect of increasing the number of random values monitored is shown in Table VII. Method 2 performs better than method 3 in the three cases for *p*=3,5, and 8. The performance of Method 2 remains the same with increasing the number of variables except for shifts δ=0.8 and 1.2, as its performance gets worse. Method 3 performs nearly the same at shifts δ=0.2, 1.5, and 2 when increasing the number of variables being monitored. However, its performance gets worse with the other shifts. The same conclusion is reached when the dirichlet parameters take values less and greater than one. Tables can be available upon request.

TABLE VII
COMPARISON OF PROBABILITY OF CORRECT DETECTION FOR N=10, BETWEEN METHODS 2 AND 3 FOR $a_1 = a_2 = \ldots = a_8 = 1$ AND P=3,5 & 8

Method 2																
	p=3			*p*=5					*p*=8							
	a_1	a_2	a_3	a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
0.2	1	0.99	-	1	1	1	1	-	0.99	0.99	0.99	0.99	0.99	0.99	0.99	-
0.5	0.99	0.99	-	0.99	0.99	0.99	0.99	-	0.98	0.98	0.99	0.95	0.99	0.99	0.98	-
0.8	0.96	0.97	-	0.94	0.94	0.94	0.94	-	0.65	0.76	0.81	0.4	0.84	0.86	0.76	-
1.2	0.94	0.94	-	0.88	0.88	0.88	0.88	-	0.71	0.70	0.66	0.72	0.64	0.61	0.70	-
1.5	0.99	0.99	-	0.98	0.98	0.98	0.98	-	0.96	0.96	0.97	0.95	0.96	0.96	0.96	-
2	0.99	0.99		0.99	0.99	0.99	0.99	-	0.99	0.99	0.99	0.99	0.99	0.99	0.99	-
Method 3																
	p=3			*p*=5					*p*=8							

	a_1	a_2	a_3	a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
0.2	1	0.90	-	0.99	0.99	0.97	0.91	-	0.99	0.98	0.98	0.62	0.97	0.94	0.96	-
0.5	0.99	0.9	-	0.96	0.97	0.94	0.85	-	0.85	0.94	0.95	0.4	0.92	0.87	0.89	-
0.8	0.97	0.8	-	0.92	0.86	0.78	0.67	-	0.4	0.42	0.43	0.35	0.48	0.52	0.46	-
1.2	0.96	0.92	-	0.94	0.93	0.92	0.77	-	0.73	0.75	0.79	0.73	0.83	0.76	0.6	-
1.5	0.99	0.97	-	0.99	0.99	0.98	0.85	-	0.95	0.97	0.98	0.95	0.99	0.97	0.89	-
2	0.99	0.97	-	0.99	0.99	0.99	0.88	-	0.99	0.99	0.99	0.99	0.99	0.97	0.91	-

IV. CONCLUSION

In this paper, three easily applicable methods were proposed to monitor Dirichlet data. Two of which are used to detect the source of the out-of-control signal. The first method is based on a MEWMA control chart. The second method is based on transforming the Dirichlet random variables to beta random variables and then monitoring them using multiple EWMA control charts. The third method uses multiple EWMA control charts for transformed independent random variables. To assess the performance of the three methods, different simulation scenarios were used to represent various cases.

The three methods performed well and were nearly the same except for values of the Dirichlet parameters less than one. However, increasing the sample size enhanced the performance of the suggested methods. Method 2 performed better than the other two methods for a Dirichlet distribution with parameters less than one. Increasing the number of variables to be monitored increases the out-of-control ARL for the three methods, which can be overcome by increasing the sample size.

When the process is out-of-control, the source of the out-of-control signal can be detected using Method 2 and Method 3. Method 2 maintained its good performance with a probability 0.99 of correctly detecting the source of the signal. Method 3 performed well except for the case of parameter values less than one. However, it maintained almost a probability of correct detection of at least 90% in most cases, unlike Vive, which [14] approach had a percentage of correct detection not exceeding 50%. For future work, a comparison might need to be held to compare the performance of the proposed methods to the ilr Hotelling T^2 control chart used by Vives-Mestres et al [14], [15]

REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control, Sixth Edition*. 2009.

[2] L. Foley, D. Dumuid, A. J. Atkin, T. Olds, and D. Ogilvie, “Patterns of health behaviour associated with active travel: A compositional data analysis,” *Int. J. Behav. Nutr. Phys. Act.*, 2018, doi: 10.1186/s12966-018-0662-8.

[3] T. P. Quinn, I. Erb, G. Gloor, C. Notredame, M. F. Richardson, and T. M. Crowley, “A field guide for the compositional analysis of any-omics data,” *Gigascience*, 2019, doi: 10.1093/gigascience/giz107.

[4] K. Pearson, “Mathematical contributions to the theory of evolution. — On a form of spurious correlation which may arise when indices are used in the measurement of organs,” *Proc. R. Soc. London*, vol. 60, no. 359–367, pp. 489–498, Dec. 1897, doi: 10.1098/rspl.1896.0076.

[5] V. Pawlowsky-Glahn and A. Buccianti, *Compositional Data Analysis: Theory and Applications*. 2011.

[6] J. Aitchison, *The Statistical Analysis of Compositional Data*. 1986.

[7] P. Praus, “Robust multivariate analysis of compositional data of treated wastewaters,” *Environ. Earth Sci.*, 2019, doi: 10.1007/s12665-019-8248-6.

[8] J. J. Egozcue, V. Pawlowsky-Glahn, and G. B. Gloor, “Linear association in compositional data analysis,” *Austrian J. Stat.*, 2018, doi: 10.17713/ajs.v47i1.689.

[9] V. Pawlowsky-Glahn, “Peter Filzmoser, Karel Hron, Matthias Templ: Applied compositional data analysis, with worked examples in R,” *Stat. Pap.*, vol. 61, no. 2, pp. 921–922, 2020, doi: 10.1007/s00362-020-01163-7.

[10] R. A. Boyles, “Using the chi-square statistic to monitor compositional process data,” *J. Appl. Stat.*, 1997, doi: 10.1080/02664769723567.

[11] G. Yang, D. B. H. Cline, R. L. Lytton, and D. N. Little, “Ternary and Multivariate Quality Control Charts of Aggregate Gradation for Hot Mix Asphalt,” *J. Mater. Civ. Eng.*, 2004, doi: 10.1061/(asce)0899-1561(2004)16:1(28).

[12] M. Vives-Mestres, J. Daunis-I-Estadella, and J. A. Martín-Fernández, “Individual T2 control chart for compositional data,” *J. Qual. Technol.*, 2014, doi: 10.1080/00224065.2014.11917958.

[13] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, “Isometric Logratio Transformations for Compositional Data Analysis,” *Math. Geol.*, 2003, doi: 10.1023/A:1023818214614.

[14] M. Vives-Mestres, J. Daunis-I-Estadella, and J. A. Martín-Fernández, “Out-of-control signals in three-part compositional T2 control chart,” 2014, doi: 10.1002/qre.1583.

[15] M. Vives-Mestres, J. Daunis-i-Estadella, and J. A. Martín-Fernández, “Signal interpretation in Hotelling’s T2 control chart for compositional data,” *IIE Trans. (Institute Ind. Eng.*, 2016, doi: 10.1080/0740817X.2015.1125042.

[16] K. P. Tran, P. Castagliola, G. Celano, and M. B. C. Khoo, “Monitoring compositional data using multivariate exponentially weighted moving average scheme,” *Qual. Reliab. Eng. Int.*, 2018, doi: 10.1002/qre.2260.

[17] F. Alt and K. Jain, “Multivariate quality controlMultivariate quality control,” in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and C. M. Harris, Eds. New York, NY: Springer US, 2001, pp. 544–550.

[18] A. Ongaro and S. Migliorati, “A generalization of the dirichlet distribution,” *J. Multivar. Anal.*, 2013, doi: 10.1016/j.jmva.2012.07.007.

APPENDIX

p INDEPENDENT GAMMA RANDOM VARIABLE

Let a set of p independent Gamma random variables be:

$$Z_i \sim \boldsymbol{G}\text{amma}(a_i, 1) i = 1,2, \ldots\ldots . p$$

The *(p-1)* random variables $Y_1, Y_{2.1}, \ldots . . Y_{(p-1).1,2,\ldots(p-2)}$ can be written as functions of the gamma random variables as shown later.

The joint distribution of the p independent Gamma random variables is given by:

$$f(z_1, z_2, \ldots . , z_p) = \frac{1}{\Gamma a_1 \ldots . . \Gamma a_p} z_1^{a_1-1} \ldots \ldots . z_p^{a_p-1} e^{-(z_1+\ldots+z_p)}, z_i \geq 0 \forall i = 1, \ldots . p$$

A transformation of variables will be done, so a p^{th} variable need to be introduced

$$S = z_1 + \ldots . + z_p$$

For the sake of simplifying the equations, Y_j^* will be used to represent the j^{th} transformed Dirichlet random variable $y_{j.1,\ldots,(j-1)}$.

$$\begin{aligned}
z_1 &= y_1^* s \\
z_2 &= y_2^*(s - z_1) = y_2^*(s - y_1^* s) = y_2^* s(1 - y_1^*) \\
z_3 &= y_3^*(s - z_1 - z_2) = y_3^*(s - y_1^* s - y_2^* s(1 - y_1^*)) \\
&= y_3^* s(1 - y_1^*)(1 - y_2^*) \\
&\vdots \\
&\vdots \\
z_j &= y_j^* s(1 - y_1^*) \ldots \ldots (1 - y_{j-1}^*) \\
&\vdots \\
&\vdots \\
z_p &= s(1 - y_1^*) \ldots \ldots (1 - y_{p-1}^*)
\end{aligned}$$

Define the Jacobian as

$$|J| = \begin{vmatrix} \partial z_1 / \partial y_1^* & \cdots & \partial z_1 / \partial y_{p-1}^* & \partial z_1 / \partial s \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \partial z_p / \partial y_1^* & \cdots & \partial z_n / \partial y_{p-1}^* & \partial z_p / \partial s \end{vmatrix}$$

$$= \begin{vmatrix} s & 0 & 0 & \cdots & \cdots & \cdots & y_1^* \\ -y_2^* s & (1-y_1^*)s & 0 & \cdots & \cdots & \cdots & y_2^*(1-y_1^*) \\ -y_3^*(1-y_2^*)s & -y_3^*(1-y_1^*)s & (1-y_1^*)(1-y_2^*)s & 0 & \cdots & \cdots & y_3^*(1-y_1^*)(1-y_2^*) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ -y_j^*(1-y_2^*)\ldots(1-y_{j-1}^*)s & \cdots & \cdots & \cdots & \cdots & 0 & y_j^*(1-y_1^*)\ldots(1-y_{j-1}^*) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -(1-y_2^*)\ldots(1-y_{p-1}^*)s & \cdots & \cdots & \cdots & \cdots & \cdots & (1-y_1^*)\ldots(1-y_{p-1}^*) \end{vmatrix}$$

Now *S* will be taken as common factor from *p-1* columns, $(1 - y_1^*)$ from *p-2* columns,. $(1 - y_j^*)$ from *(p-1-j)* columns, and the jacobian will be

$$|J| = \mathrm{s}^{\mathrm{p\text{-}1}}(1 - y_1^*)^{p-2} \ldots . . (1 - y_j^*)^{p-1-j} \ldots . . (1 - y_{p-2}^*)$$

$$\begin{vmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & y_1^* \\ -y_2^* & 1 & 0 & \cdots & \cdots & \cdots & y_2^*(1-y_1^*) \\ -y_3^*(1-y_2^*) & -y_3^* & 1 & 0 & \cdots & \cdots & y_3^*(1-y_1^*)(1-y_2^*) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ -y_j^*(1-y_2^*)\ldots(1-y_{j-1}^*) & \cdots & \cdots & \cdots & \cdots & 0 & y_j^*(1-y_1^*)\ldots(1-y_{j-1}^*) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -(1-y_2^*)\ldots(1-y_{p-1}^*) & \cdots & \cdots & \cdots & \cdots & \cdots & (1-y_1^*)\ldots(1-y_{p-1}^*) \end{vmatrix}$$

After doing some operations on the determinant, as multiplying the first column by $(1 - y_1^*)$ and adding it to the last column, resulting in a column whose first element is 1 and the rest is zeros. Then multiplying column j by $-(1 - y_j^*)$ and adding the *jth* column to the *(j-1)th* column resulting into a column starting with 1, then -1 and then zeros for all *j*=2,.....,*(p-1)*

$$|J| = \mathrm{s}^{\mathrm{p\text{-}1}}(1 - y_1^*)^{p-2} \ldots . . (1 - y_j^*)^{p-1-j} \ldots . . (1 - y_{p-2}^*)$$

$$\begin{vmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & 1 \\ -1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & -1 & 0 \end{vmatrix}$$

Solving this determinant will give a solution of 1, so the joint distribution of the transformed variables is given by

$$\begin{aligned}
f(y_1^*, y_2^*, \ldots . , y_{p-1}^*, s) &= f(y_1^* s, y_2^*(1 - y_1^*)s, \ldots\ldots , y_{p-1}^*(1 - y_1^*) \ldots (1 - y_{p-2}^*)s, (1 - y_1^*) \ldots (1 - y_{p-1}^*)s) * |J|
\end{aligned}$$

$$= \frac{1}{\Gamma a_1 \ldots . . \Gamma a_p} [y_1^* s]^{a_1-1} \ldots\ldots . \left[(1 - y_1^*) \ldots . . (1 - y_{p-1}^*)s\right]^{a_p-1} e^{-s} s^{\mathrm{p\text{-}1}} (1 - y_1^*)^{p-2} \ldots . . (1 - y_j^*)^{p-1-j} \ldots . . (1 - y_{p-2}^*)$$

$$= \frac{1}{\Gamma a_1 + \ldots + a_p} s^{a_1+\ldots+a_p-1} e^{-s} \frac{\Gamma a_1 + \ldots + a_p}{\Gamma a_1 . \Gamma a_2 + \ldots + a_p} y_1^{*a_1-1} (1 - y_1^*)^{a_2+\ldots+a_p-1} \ldots . \frac{\Gamma a_j + a_{j+1} + \ldots + a_p}{\Gamma a_j . \Gamma a_{j+1} + \ldots + a_p} y_j^{*a_j-1} (1 - y_j^*)^{a_{j+1}+\ldots+a_p-1}$$

$$\frac{\Gamma a_{p-1} + a_p}{\Gamma a_{p-1} . \Gamma a_p} y_{p-1}^{*a_{p-1}-1} (1 - y_{p-1}^*)^{a_p-1}, s \geq 0, 0 \leq y_j^* \leq 1 \forall j = 1, \ldots . p - 1$$

Since the P.d.fs of the transformed variables can be separated, then they are independent random variables. *S* follows Gamma distribution with parameters $(a_1 + \ldots + a_p, 1)$, and y_j^* follows Beta distribution with parameters $(a_j, a_{j+1} + \ldots + a_p)$.