Abstract

This work analyses the multivariate Hotelling T^2 control chart for monitoring compositional data (CoDa). A composition is a vector of positive elements which represent parts of a whole and usually add to a constant sum, to which standard multivariate techniques are not appropriate due to their restricted sample space. CoDa have the peculiarity of living in a restricted space. There are many applications where a composition is monitored against time such as in the chemical industry, product composition, impurity profile or gas components analysis.

Firstly, we analyze the proposals found in the literature to control processes in which the quality characteristic is a composition by the use of the T^2 statistic. We have determined that these proposals violate the principle of subcomposicional coherence. This principle states that the inference on a subcomposition (a part of a composition) should be consistent regardless of whether the inference is based on the entire composition or a subcomposition. Furthermore, we have also seen that the classical control regions fall out of the sample space.

Based on these observations, we have proposed a new control chart called compositional T^2 control chart and denoted by T_C^2. It is based on a representation of CoDa onto coordinates in the real space, where the T^2 statistic can be calculated. The coordinates are calculated by the use of a logratio transformation of the components. Conceptually the T_C^2 statistic is the distance from each coordinate to the centre of the coordinates (geometric mean of the composition) by taking into account the correlation among them.

We used the in-control average run length (ARL) as a performance indicator to compare the classical method (T^2) with the compositional method (T_C^2), and we have seen that the T_C^2 has a lower false alarm rate, which remained constant regardless of the centre of the distribution of the data. When the composition is homogeneous, both methods perform well; but the difference between them is more acute when the samples are close to a vertex, which is the case in most real datasets.

In process control, it is not only important to identify the out of control signals, but also to identify the causes of the anomaly in order to carry out corrective actions. This is why we have proposed a graphical method (for 3-part compositions) to interpret the out of control signals in the case of three part compositions based on an appropriate selection of the coordinates.

Finally, we present an algorithm to interpret the causes of the signal in the general case (for more than three components). Both the algorithm and the graphical method are based on finding the log ratio of components that greatly contributes to the overall value of the T_C^2 statistic. The first generalized method is suitable for low dimensional problems and consists on finding the log ratio of components that maximizes the univariate T^2 statistic. The second one is an optimized method for large dimensional problems that simplifies the calculus by transforming the coordinates into the sphere.

Throughout the work we have applied the proposed methods and concepts to simulated examples and a at the end we present a real industrial example from the literature consisting of a dataset of three part compositions.

State of the art

Control charts are a tool of statistical process control (SPC) that ensures the process remains in a state of statistical control, i.e. in the absence of assignable causes and where changes in measures of centre and variability are statistically predictable (common causes).

Most processes require control of multiple variables. The use of multiple univariate control charts does not deliver a useful solution in this situation. The problems are that, the overall probability of signalling a false alarm is not controlled and more seriously the correlation among the variables is ignored (Montgomery (2013)).

The most elementary multivariate control scheme is the Hotelling T^2 control chart (Hotelling (1947)). This graph uses the Mahalanobis distance between the observation (or the average of a group of observations) and the average of the process taking into account the correlation between the variables. More information about using the T^2 can be found in Kenett and Zacks (2014) and Montgomery (2013).

The interpretation of the out-of-control signals of multivariate control charts is important in order to carry out corrective actions to improve the process. This task is particularly difficult in multivariate control schemes due to the use of an overall statistic which takes into account the correlation between variables. A signal in a T^2 value may indicate a change in location or spread of one or more variables as well as a change in the relationship among multiple variables or even a combination of both effects.

For the specific case of the T^2 control chart, various methods have been developed to interpret the signals out of control. One usual method is the one proposed by Mason et al. (1997) based on the decomposition of the T^2 statistic into orthogonal parts directly interpretable called MYT decomposition. The MYT method is described in detail and with examples in the book Mason and Young (2002).

This works focuses on process control when the quality characteristic is a composition. We only found two articles in the literature attempting to implement a control scheme to CoDa where its distinctiveness is mentioned.

A first attempt to implement a control chart for compositional processes is made by Boyles (1997). He develops a chi-square control chart to monitor multinomial and Dirichlet data. The Dirichlet distribution has some very restrictive properties, such as complete subcompositional independence, which makes 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.

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 acontrol charteptance regions. The first one is by performing multiple univariate control charts which is not optimal when a multivariate quality control is desired (Montgomery (2013)). The second method is based on an additive approach (not log-ratio) thus not consistent with CoDa nature.

Main contributions

Up to our knowledge, there is a gap in literature regarding statistical process control methods for processes in which the quality characteristic is a compositional vector. In the analysis of the state of the art, have not found references in which the methodology proposed by Aitchison (1986) is used in a control scheme. This methodology has subsequently led to the development of a series of techniques that have proved adequate to deal with data in restricted spaces.

The main contribution of this work is on adapting the oldest multivariate control scheme, the Hotelling T^2 control chart, and on providing a generalized method for interpreting the compositional control chart signals. We are aware that in recent years there has been huge progress in process control techniques, which now tend to be more complex and adapted to industry requirements. Such an example are the non-parametric methods, which do not require any knowledge or assumption about the initial distribution of data.

The proposed scheme does not provide an innovation regarding the T^2 control chart. It actually inherits the weaknesses of the classical scheme, which are dependence on the normality assumption of the data, the need for a large amount of data for estimating the distribution parameters and the low ability to detect small changes in the mean of the process, among others.

However, we see this work as a first step that will introduce the concepts of compositional methodology in the field of statistical process control which, from our point of view, there is not much awareness about the peculiarity of these data. We observe a certain interest in spreading these concepts, which is reflected by the publication of the two article in peer reviewed journals.

Journal publications

The work submitted in the Ramiro Melendreras award is part of the research work developped during the doctoral studies of Marina Vives Mestres and has partially been published in two peer-reviewed journals indexed in the Science Citation Index. There is also a third article which is under the second review process after taking into account the comments of the referees. The three papers are listed below:

- Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J.A. (2014).

"Out-of-Control signals in three-part compositional T2 control chart". *Quality and Reliability Engineering International*, 30(3), pp. 337-346.

Impact Index in SCI: 0.994 (2013). Engineering Multidisciplinary Q2; Operations Research & Management Science Q3; Engineering Industrial Q3.

- Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J.A. (2014). "Individual T2 Control Chart for Compositional Data". *Journal of Quality Technology*, 46(2), pp. 127-139.

  Impact Index in SCI: 1.271 (2013). Statistics & Probability Q2; Operations Research & Management Science Q2; Engineering Industrial Q2.

- Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J.A. "Signal Interpretation in Hotelling's T2 Control Chart for Compositional Data". Revised in *IIE Transactions.*

  Impact Index in SCI: 1.064 (2013). Operations Research & Management Science Q2; Engineering Industrial Q3.

References

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

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

Hotelling, H. (1947). "Multivariate Quality Control–Illustrated by the Air Testing of Bombsights." In C. Eisenhart, M.W. Hastay and W.A. Willis, eds., Techniques of Statistical Analysis. McGraw-Hill, New York.

Kenett, R., Zacks, S., and Amberti D. (2014). *Modern Industrial Statistics: with applications in R, MINITAB and JMP.* 2nd Edition. Statistics in Practice. Wiley. 592p.

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., and Young, J. C. (2002). *Multivariate statistical process control with industrial applications, 1st ed.* American Statistical Association and Society for Industrial and Applied Mathematics.

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

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.

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.