Analecta
REVIEW OF FACULTY OF ENGINEERING
Analecta Technica Szegedinensia
ISSN 2064-7964

Vol. 11, No. 2 2017

# INVESTIGATION OF MULTIVARIATE STATISTICAL PROCESS CONTROL IN R ENVIRONMENT

***[1]MIHALKÓ, József; [2]RAJKÓ, Róbert***
[1]Institute of Process Engineering, Faculty of Engineering, University of Szeged,
5-7. Moszkvai krt., Szeged, Hungary, H-6725, Szeged, Hungary,
e-mail: jozsefmihalko@gmail.com
[2]Institute of Process Engineering, Faculty of Engineering, University of Szeged,
5-7. Moszkvai krt., Szeged, Hungary, H-6725, Szeged, Hungary,
e-mail: rajko@mk.u-szeged.hu

**ABSTRACT**

At the first stage of our work, the theoretical knowledge needed to use the multivariate statistical process control (MSPC) was explored. Last year, we clarified the sometimes confused concepts, equations, and formulas [1]. At the second stage, R project simulation studies and some food industrial practical model investigations are carried out for confirming the MSPC advantages compared with the univariate ones. Furthermore, we analyse, using principal component analysis (PCA), what could cause the outlying values. Moreover, we will demonstrate how to use the MYT-decomposition.

Keywords: Multivariate statistical process control, MYT-decomposition, principal component analysis

## 1. INTRODUCTION

During the statistical process control (Statistical Process Control, SPC) interventions happen in a particular process of product manufacturing if we know a cause, which can affect the value of the quality characteristics (e.g., mass) in a wrong way. Among the main tools of the SPC, the control charts [2] can be mentioned.

## 2. METHODS

Within the statistical process control, univariate and multivariate process controlling methods can be distinguished. The main difference between these two methods is that in the case of univariate methods (Univariate SPC, USPC) a known variable is being interpreted with one or more – not artificial – variables. While in the case of multivariate methods (Multivariate SPC, MSPC) several known variables are being interpreted with fewer artificially constructed variables [3]. The Fig. 1 shows the difference between these two methods, e.g. the oblique position of the ellipse indicates the co-changing (correlation) of the variables.

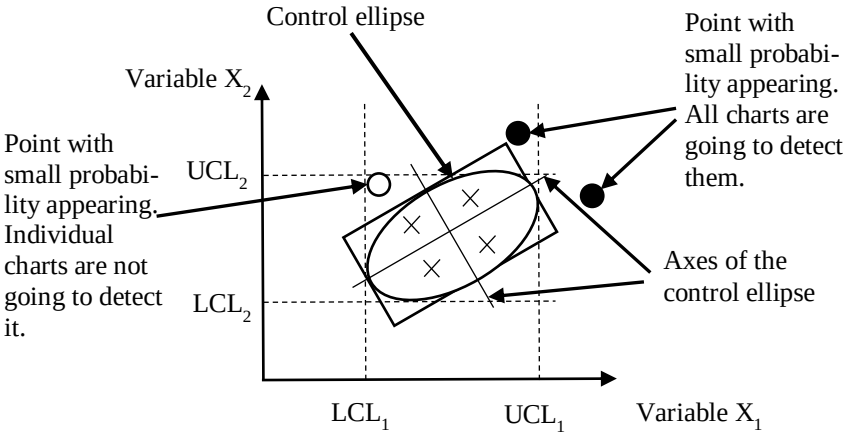


***Figure 1. The comparison of univariate and multivariate methods [2]***

ISSN 2064-7964

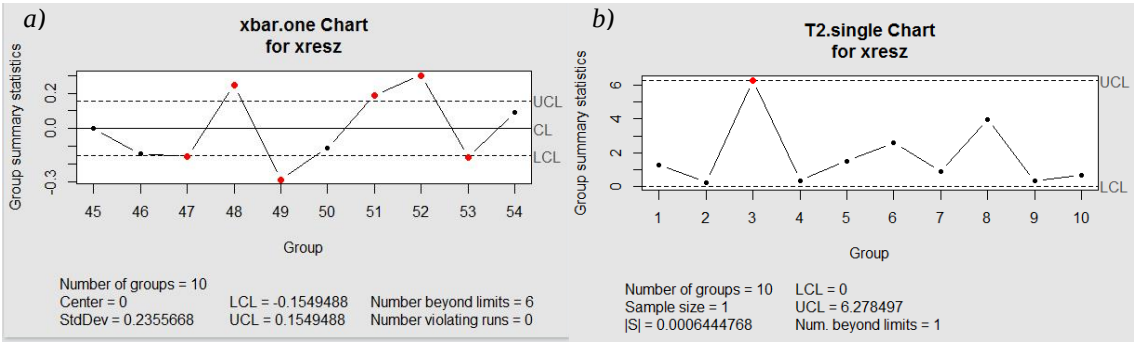


***Figure 2. a) Mean-chart and b) $T^2$-chart (with a confidence level of 95%), obtained by generating random numbers***

The exact distribution of the $T^2$-statistics ($T^2 = n\,(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu)$) depends on two aspects [2]:

- on the one hand, it matters whether we work with individual or sub-grouped data;
- on the other hand, whether we carry out a backdated analysis for stabilisation (Phase I.) and supervise the current process for monitoring (Phase II.).

In Phase II., however, it is hard to determine what could have caused the signal's alteration from the acceptance range. Possibly it was one of the quality characteristics, or the co-changing of one or more variables, or the change of the covariance. Several methods have been developed for this problem, for example, the principal component analysis and MYT-decomposition [4]. Ref. [4] describes some examples of a special calculation scheme of MYT-decomposition.

## 3. RESULTS AND DISCUSSION

Last year, at the conference called 22nd International Symposium on Analytical and Environmental Problems [1], we presented the necessary theoretical knowledge for using the MSPC in a poster presentation. We made clear the types of the applicable knowledge and their roles, and we collected the main advantages and disadvantages of the application of the MSPC compared to the USPC [1].

In the second stage of our work, firstly we carried out some simulation examinations executed in the R-environment [5], which we are going to detail from now on. In the case of one of the algorithms, we made an X-bar chart (mean-chart) (Fig. 2*a)*), applicable as the instrument of univariate process controlling methods, by using data series obtained by generating 100 uniform distributed random numbers (with the help of the "runif"-command). The CL stands for the centre line, UCL stands for upper control limit of the acceptance range, while LCL means its lower control limit. In the Fig. 2*b)*, we can see the $T^2$-chart, applicable to MSPC, where the "rmvnorm"-command was used for generating random numbers.

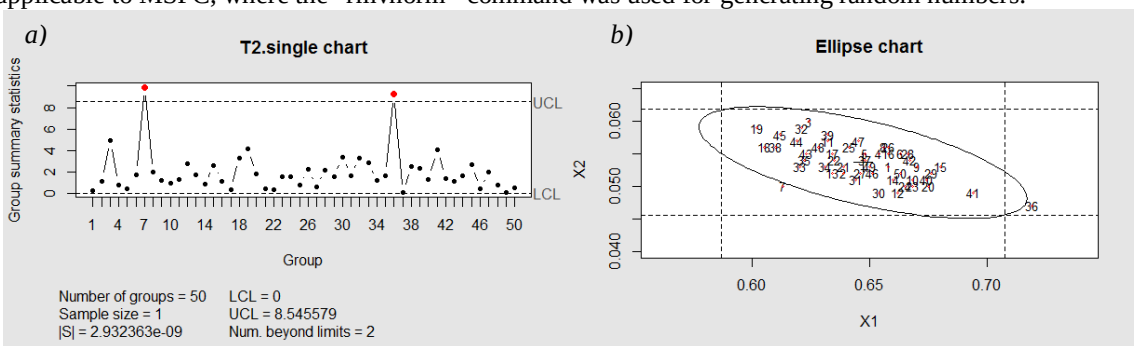


***Figure 3. a) $T^2$-chart and b) the control chart involving a control ellipse about the salt- and water-content of marinated ham, based on the example of Ref. [6] (with a confidence level of 99%)***

ISSN 2064-7964

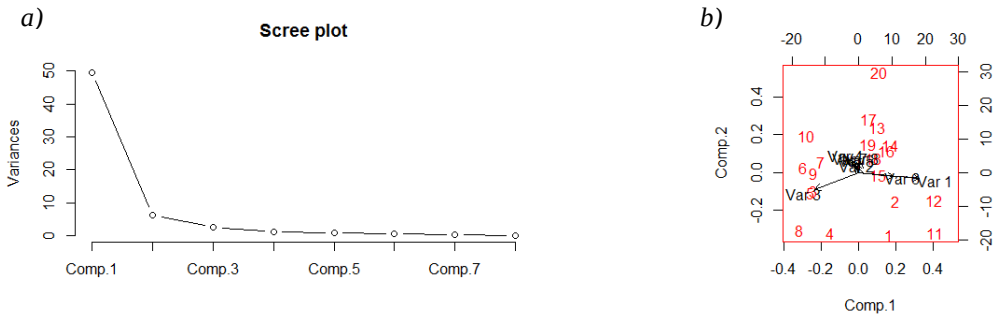


*Figure 4. a) Scree plot and b) biplot in the viewpoint of the variables of the maturing of the Parma ham*

After the implementation of the algorithms, the next step was the control (done in an R-project) of the measurement data published in Ref. [6] with the aim of reconstructing the Hotelling's $T^2$-chart and the control chart containing the control ellipse (Fig. 3*a)*).

We could reconstruct the $T^2$-chart, but we did not get back the control chart containing the control ellipse like published in the referred literature [6], due to an error message: („group sizes must be larger than one"). With fixing up the source code of the command „ellipseChart", we could finally display the control chart (Fig. 3*b)*).

After that, we generated illustrations of practical application to prove the advantages of the MSPC compared to the USPC. One of the practical application illustrations is the process control, presented with the use of the measurement data of the maturing process of Parma ham published by the University of Copenhagen [7]. Considering other experiences as well, it can be established that it is not necessary to mature the hams for 15-18 months from the date of the production, roughly it is enough to mature them for a year (11-12 months), at least with taking into consideration only the analytical measurement data.

After this, we examined (in R-project, with the use of the method of the principal component analysis) that from among the 8 measured variables (i.e, water-, salt- and protein-content, two organoleptic judgements, and the L, a and b parameters belonging to the colour measurement), which will have prominent impact to the process. In the scree plot (Fig. 4*a)*) we can see that there are three main components. With the help of the biplot (Fig. 4*b)*), it can be claimed that the 3rd variable (protein-content, it is an impact factor as well), the 1st and the 6th factors (water-content and the L-parameters) differ significantly from the other six variables.

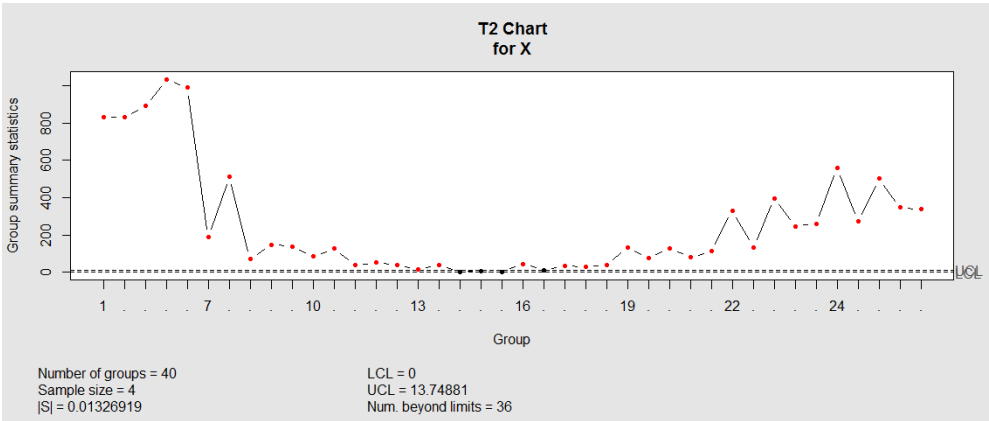


*Figure 5. Process tracking in the viewpoint of the variables of the maturing of salami (with a confidence level of 95%)*

ISSN 2064-7964

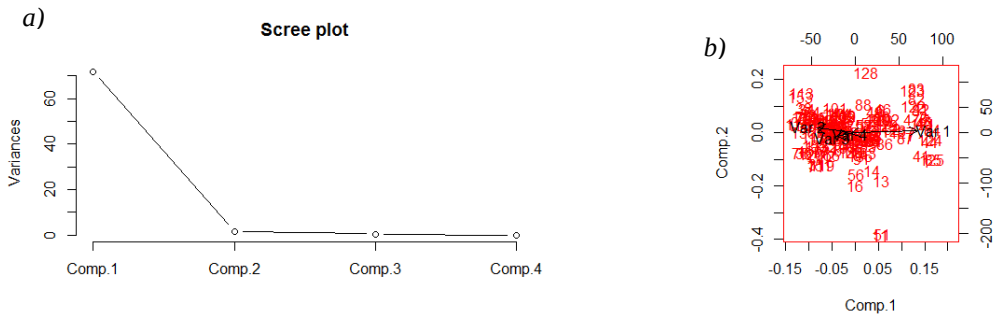


***Figure 6. a) Scree plot and b) biplot in the viewpoint of the variables of the maturing of salami***

The other practical application illustration was the process tracking of the curing of a private label salami, whose measurement data we thank Csaba Rostás and his supervisor, Ferenc Eszes.
In the Fig. 5, the parts of the curing of the salami can be easily distinguished:

- re-moistening and softening of the crust happened in the first days after the smoking,
- the drying of the product started with the 7$^{th}$ day,
- from the 10$^{th}$ day to the 19$^{th}$ the curing of the salami stabilised
- then dehydration happened again with ending the stable process.

Furthermore, in the scree plot (Fig. 6*a)*), we can see that there are two principal components. The biplot (Fig. 6*b)*) indicates that the first variable and the second variable (the variables were declared confidential by the private company) differ strongly from each other and the other quality characteristics.
Lastly, we are going to present an example of the application of MYT-decomposition, or in other words, we are going to find the answer to the question that what could cause the signal's alteration from the acceptance range. For this we used the $T^2$-statistics obtained from the measurement data of Ref. [6] (the $T^2$-chart in Fig. 3*a)*). In Fig. 3*a)*, we can see that there are two outliers, these are the 7$^{th}$ and the 36$^{th}$ samples. We present the detailed calculation through the 7$^{th}$ sample first. Since this process control is bivariate, we can decompose the $T^2$-value (obtained from the 7$^{th}$ sample) in two ways [4]:

$$T^2 = T_1^2 + T_{2.1}^2 \tag{1}$$

$$T^2 = T_2^2 + T_{1.2}^2 \tag{2}$$

The value of the 7$^{th}$ sample is 9,888, and the upper control limit (UCL) is 8,546, so it can be seen that the $T^2$-value exceeds the UCL. Thus, it is necessary to calculate the unconditional $T^2$-values ($T_1^2$ and $T_2^2$) with a new upper control limit. We summarise these values in Table 1.
The values of Table 1 show that the unconditional $T^2$-values fall within the acceptance range (their value is lower than the value of UCL), therefore we have not found the cause of the fault indication yet.

***Table 1. Unconditional $T^2$-values and the UCL value belonging to them***

| $T_1^2$ | $T_2^2$ | $UCL$ |
|---|---|---|
| 2,038 | 1,544 | 7,326 |

*Table 2. Conditional $T^2$-values and the UCL-value belonging to them*

| $T^2_{2.1}$ | $T^2_{1.2}$ | $UCL$ |
|---|---|---|
| 7,850 | 8,344 | 7,491 |

Thus, we have to calculate the conditional $T^2$-values (with the help of the equation (1) and equation (2)), and the upper control limit, which we summarise in Table 2.
Since the conditional $T^2$-values exceed the upper control limit, the signal's alteration from the acceptance range in the case of the 7th sample happens, because the correlation of the water- and salt-content does not follow the one observable for the other measurement points.

## 4. CONCLUSIONS

In this article, we presented the simulation examinations developed in an R-project, and the practical application methods, with which we illustrated the advantages of the MSPC compared to the USPC. However, it is important to notice that these measurement data are not designed for process control (both in the case of the process tracking of Parma ham and privately labeled salami). We also investigated on published data what caused the signal's alteration from the acceptance range in the Phase II., for which we can use principal component analysis or the so-called MYT-decomposition.
Our aim, in the future, is to perform the control of a specific technological process in practice, with the help of an industrial partner, for which we have to develop an optimal experiment plan as well.

## 5. ACKNOWLEDGEMENTS

Supported by the UNKP-17-2 New National Excellence Program of The Ministry of Human Capacities.

## REFERENCES

[1] J. Mihalkó, R. Rajkó (2016): Advantages and drawbacks of using multivariate statistical process control in food industry, Proceedings of the 22nd International Symposium on Analytical and Environmental Problems, Szeged, Hungary, October 10, 2016, 385-389.
[2] M. Rogalewicz, Some Notes on Multivariate Statistical Process Control, Management and Production Engineering Review, 3 (4) (2012), 80-86.
[3] O. R. Mohana Rao, K. V. Subbaiah, K. N. Rao, T. S. Rao, Application of multivariate control chart for improvement in quality of hotmetal – a case study, International Journal for Quality Research, 7 (4) (2013), 623-640.
[4] R. L. Mason, N. D. Tracy, J. C. Young, A practical approach for interpreting multivariate $T^2$ control chart signals, Journal of Quality Technology, 29 (4) (1997), 396-406.
[5] https://www.rstudio.com/
[6] A. Ittzés, E. Zukál, Többváltozós folyamatszabályozás alkalmazási lehetőségei a húsiparban (Application of multivariate process control in the meat industry), A Hús, 9 (3) (1999), 179-183.
[7] http://www.models.life.ku.dk/ParmaHam_Fluor