Multivariate Statistical Process Control by Individual Observations

V N Klyachkin, I N Karpunina

Abstract— ***Individual observations are often in use for multivariate statistical process control instead of instantaneous sampling. Some questions arise related to a covariance matrix assessment and determining the position of limits in the Hotelling chart and generalized variance. There was developed an algorithm of control by individual observations on the basis of potable water purity and heat sink cone processing parameters completed tests. There were also offered some control efficiency improvement methods through the ways of data normalization, warning limits use, random structures analysis***

Keywords and word combinations – generalized variance, Hotelling chart, model average mid- range, normalization.

I. INTRODUCTION

The main function of engineering process statistical control is detecting the process stability violation. The diagnosis is made on the statistical data , ensuring the detection of non-random reasons for the process unsteadiness and enabling the use of the control affect before the quality controlled data of the product go beyond the tolerance limits.

The most popular tool of such a control is Stewhart chart , specified by the standards. It is alleged that the quality is characterized by one parameter, or some independent parameters. At certain time the instantaneous sampling is taken in size of 3 to 10 units to plot the charts, and the result of their measuring gives the statistical characteristic of the process. As a rule the monitoring of mid-range level of the procedure (mean charts) and its dispersion (standard deviation and range charts) is done.

Frequently such measuring takes a lot of expenses and labor cost. In this case the charts for individual values are used: the scope of instantaneous sampling is equal to one. The practice shows that during multivariate process control some data of quality are independent (and they are monitored with Stewhart chart), but the data of the other part are correlated between each other. In this situation the independent control for some data can lead to significant errors, so it is necessary to use multivariate methods [1-2]. Mid- range level of the process in multivariate control is assessed with the use of Hotelling statistics. The distribution of this statistics in the control procedure by individual monitoring and the peculiarities of covariance matrix assessment were analyzed in articles [3-5]. The control of multivariate dispersion may be performed with the help of generalized variance? i.e. covariance matrix determinant [1].

Some problems appear hereto, related to the efficiency of multivariate control by individual observations i.e stability violation timely detection

II. PROBLEM SETTING UP

By the efficiency of control chart we mean its sensitivity to possible process unsteadiness. The main characteristics of the control efficiency are the model mid- range length: the number of observations done between the process stability violation start and the unsteadiness detection.

Hotelling statistics trends to a multivariate normal distribution, which does not necessarily happen in real engineering procedures. Unsteadiness of the process leads to control sensitivity limit deterioration. There are no general methods of multivariate normal distribution check. It is necessary to check the stability of each data as minimum, then, if necessary, normalization steps are taken.

The important issue, determining the control efficiency, is the knowledge of Hotelling statistics as well: i.e. the width of confidence limit is determined with the inverted distribution. If during the use of instantaneous sampling this problem is solved, then there are different ways acceptable for the individual observations to assess the covariance matrix , and different distributions correspond to these ways.

The monitoring efficiency may be increased by the use of the warning limit, as well as the analysis of non- random models on the chart. [6]. The interpretation of a multivariate chart presupposes the answer to the question, which of multiple controlled parameters failed

Finally, a row of questions appears during the multivariate dispersion process control with the use of generalized variance charts plotting. Strictly speaking, it is impossible to make such a chart based on individual observations, so one must find a proper way out from this situation.

III. THEORY

3.1. Process mid-range level control

Let's assume that in a process we control p data $x = (x_1, \dots, x_p)$, of common normal distribution. Hotelling algorithm admits the computation for each instantaneous sampling t ($t = 1, \dots, m$) of the statistics [1-2]

$$T_t^2 = n(\bar{\mathbf{x}}_t - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_t - \boldsymbol{\mu}_0), \quad (1)$$

Revised Manuscript Received on April 12, 2019

V N Klyachkin, Ulyanovsk State Technical University, Ulyanovsk, RUSSIA

I N Karpunina, Ulyanovsk Civil Aviation Institute , Ulyanovsk, RUSSIA

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

with n denoting a sample size , $\bar{\mathbf{x}}_t$ is a vector of means in instantaneous samples , $\bar{\mathbf{x}}_t = (\bar{x}_{t1}...\bar{x}_{tp})^T$, $\bar{x}_{tj}$ is an average value in t-instantaneous sample in terms of j-data (j = 1, …, p); $\boldsymbol{\mu}_0$ is a vector of objective means , $\boldsymbol{\mu}_0 = (\mu_1...\mu_p)^T$,

$$\mu_j = \frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} x_{ijt}$$

(with x_{ijt} denoting the result of i- observation in terms of j-data in sample t).

S components assessments of p by p covariance matrix Σ are determined according to the following formula:

$$s_{jk} = \frac{1}{m(n-1)}\sum_{t=1}^{m}\sum_{i=1}^{n}(x_{ijt}-\mu_j)(x_{ikt}-\mu_k), \quad j, k = 1, \dots, p. \quad (2)$$

The process is considered to be steady if $T_t^2 < T_{\kappa p}^2$, where $T_{\kappa p}^2$ - critical domain limit.

If the covariant matrix $\boldsymbol{\Sigma}$ is known, then Hotelling statistics has χ^2 – distribution. In this case the critical value (Hotelling chart Upper Control Limit) with set significance level equal to α can be found with the use of quantile table:

$$T_{\kappa p}^2 = \chi_{1-\alpha}^2(p). \quad (3)$$

Assessing the covariance matrix with instantaneous samples as per formula (2), the size of which is more than one, the position of control limit may be calculated as per the following formula:

$$T_{\kappa p}^2 = \frac{p(m-1)(n-1)}{mn-m-p+1} F_{1-\alpha}(p, mn-m-p+1) \quad (4)$$

(with $F_{1-\alpha}(k_1,k_2)-$ denoting Fisher inverted distribution with k_1, k_2 degrees of freedom. The given formulas are used for multivariate statistical control of process mid - range with the instantaneous samples. If the monitoring is done with individual observations, it is necessary to perform the corrections.

3.2 Specific features of individual observations monitoring

Computation as per formulas (1) - (4) is correct if we assume that there is data normalcy of distribution. Monitoring of unsteadiness by individual observation can lead to errors

In case of stability violation the application of normalizing transformation is good, the monitoring is to be done with the transformed data. There are many ways of data normalization. The simplest one is taking logarithms (the transformed value of x' = ln x), that often leads to the required result (primary distribution is logarithmically normal distribution, the transformed data are normal). More common is the transformation of Jonson [2].

Let's admit that the monitoring is done with m individual observations for p data , with **S** denoting the assessment of covariant matrix , and $\bar{\mathbf{x}}$ is a vector of mean values of the controlled parameters. Then Hotelling statistics for observation t (t = 1… m) is determined as per the following formula:

$$T_t^2 = (\mathbf{x}_t - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}_t - \bar{\mathbf{x}}), \quad (5)$$

Carrying out the individual observations brings the problem of covariant matrix assessment. Sullivan and Woodall [4] offered 2 variants. Assessing by all the sample:

$$\mathbf{S} = \frac{1}{m-1}\sum_{t=1}^{m}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T. \quad (6)$$

Tracy, Young, and Mason showed, that at this, the critical value T2 $_{cr}$ is based on beta distribution [3]:

$$T_{kp}^2 = \frac{p(m+1)(m-1)}{m^2-mp}\beta_{1-\alpha,p/2,(m-p-1)/2}, \quad (7)$$

with $\beta_{1-\alpha,p/2,(m-p-1)/2}$ denoting the inverted beta-distribution of order (1- α) with p/2 degrees of freedom and (m – p – 1)/2.

The other variant is covariant matrix assessment by the range: the rolling range is determined as $v_t = x_{t+1} - x_t$, t = 1,…, m – 1; then vector range is $\mathbf{V}^T = (v_1\ v_2\ \dots\ v_{t-1})$, and the assessment of covariant matrix is as follows:

$$\mathbf{S} = \frac{1}{2}\frac{\mathbf{V}^T\mathbf{V}}{(m-1)}. \quad (8)$$

Williams with his co - authors [5] showed that in this case Hotelling statistics will have distribution $\chi^2(p)$. To estimate the position of the limits, chi- square inverted distribution is used. (3).

To answer the question which approach (6) or (8) will have a priority from the point of view of the efficiency in the detection of violations in a certain process is possible only judging by the results of the statistic tests.

3.2 Deliverables interpretation

Hotelling chart does not show directly, which data (or data combination) are the reason for the process violation. The chart interpretation task is to reveal these data, causing the process violation. Let's assume that at a certain *t equal to t_0* the chart indicates the violation in the process. To check the hypothesis that the violation was caused by j-data (or some data combination), we may apply to partial H-test:

$$T_j^2 = \left[\mathbf{c}_j^T(\mathbf{x}_{t_0} - \bar{\mathbf{x}})\right]^2 / \left[\mathbf{c}_j^T \mathbf{S}\mathbf{c}_j\right] > T_{kp}^2$$

where $\mathbf{c}_j$ is a column vector , consisting of zeros in all the lines , except j, and one in j line, $\mathbf{x}_{t0}$ is an ordinate vector of the monitored data in observation t_0.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Another approach is the chart plotting with element -by-element data removal. For example in case of violation detection with the help of Hotelling chart , monitoring 3 data , one can make 3 Hotelling charts using 2 data , removing one of them in sequence.

3.4 Multivariate dispersion control

Besides the mid-range level monitoring it is reasonable to check the stability of its dispersion. For independent data the rolling range charts are used, for multivariate control we use the generalized dispersion chart. The generalized variance chart is the analog of the rolling range chart in a multivariate case. The generalized variance stands for covariant matrix determinant [1, 7].

Let's monitor p correlated data: m times the figures are taken for reference for p data in the instantaneous samples of size n. We plot general dispersion figures $|\mathbf{S}_t|$ taken at random for each t-sample. The elements of covariant matrix $\mathbf{S}_t$ are found as per the formula:

$$s_{jkt} = \frac{1}{n-1}\sum(x_{ijt} - \overline{x_j})(x_{ikt} - \overline{x_k}), \quad (9)$$

with $\overline{x_j}$ and $\overline{x_k}$ denoting average meanings of j- (k-) data correspondingly .

The same way is used to estimate covariant mean for all the samples, which plot covariant matrix $\mathbf{S}$, the determinant of which is used as the assessment of objective generalized variance $|\mathbf{\Sigma}_0|$:

$$\overline{s_{jk}} = \frac{1}{m}\sum_{t=1}^{m} s_{jkt}.$$

The mean line of the generalized dispersion is estimated by formula:

$$m_{|\mathbf{s}|} = b_1 |\mathbf{\Sigma}_0|.$$

The upper and lower limits are found on the normal distribution [1]:

$$\left.\begin{matrix} UCL \\ LCL \end{matrix}\right\} = |\mathbf{\Sigma}_0|(b_1 \pm 3\sqrt{b_2}) \quad (10)$$

If the value for the lower meaning is negative, it is taken as zero value.

The coefficients b_1 and b_2 are found as per formulas:

$$b_1 = \frac{1}{(n-1)^p}\prod_{j=1}^{p}(n-j), \quad (11)$$

$$b_2 = \frac{1}{(n-1)^p}\prod_{j=1}^{p}(n-j)[\prod_{k=1}^{p}(n-k+2) - \prod_{k=1}^{p}(n-k)]. \quad (12)$$

From the dependences (9) - (12) the conclusion follows , that it is impossible to plot the generalized variance chart based on the individual observations. It is necessary to have instantaneous samples, the size of which exceeds the number of the monitored parameters at least by one. For example, monitoring two correlated data, it is necessary to connect three observations as minimum in one sample. But such an approach significantly reduces the monitoring efficiency. As an alternative one may use the control method with the use of rolling mid-range: in formula (9) instead of x_{ijt} ,the smoothed values are used , e.g. three dots: $x'_{ijt} = (x_{ij,t-1} + x_{ijt} + x_{ij,t+1})/3$.

IV. THE EXPERIMENTS' RESULTS

The experiments were performed on the examples of potable water quality and heat sink cone processing parameters test

The quality of the potable water is assessed according to its physical and chemical data : turbidity , colority , aluminium and chloride contents , pH, residual chlorine , oxidability and alkalinity [8]. The observations are made once a day (once per 24 hours), so the charts of individual values were used. The correlation in two groups turned out to be significant : between aluminium content and oxidability, and also between pH, residual chlorine alkalinity. To monitor these data multivariate methods were used: Hotelling charts and generalized variance in the mentioned above variants (for the rest of data the standard charts of Stewhart were applied) .

The similar problem was solved with the geometric parameters of heat sink cone , used for reducing heat in IT-hardware of special functions: the correlated values happened to be the length of the cone and its angle of skew.

To plot Hotelling chart, both covariant matrix calculation variants with the help of formulas (6) and (8) were used. The best (i.e. the fastest one in the simulated violation detection) was the variant at which the calculation at the stage of the process analysis is based on formulas (5) - (7), and during monitoring dependence (9) is in use. During the process dispersion control the use of generalized variance chart with the smoothed mean values indicated the reduction of the average length of the set by 3 to 5 times in comparison with chart of instantaneous. Both Hotelling chart and generalized variance chart had the process stability violation data in the terms of the monitored statistics getting beyond a limit as well as the availability of non-random structures on the charts .

V. THE DISCUSSION RESULTS

On the one hand, the received results turned out to be similar to each other in two different examples (that gives us the right to believe that they are universal), on the other hand, to prove this conclusion it is necessary to perform statistics tests, simulating different possible types of process violations and set average length assessment.

VI. SUMMARY

To perform the statistics control of multivariate process by individual observations the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

following algorithm can be offered:

1. Taking the results of the preceding analysis into account , the normalcy of the data distribution is checked, if necessary the data normalization is done. To assess the data association, covariant matrix is used with significant correlation revealed.
2. To assess the process stability using the independent data, Stewhart charts for individual observations and rolling ranges are used.
3. At the stage of correlated data analysis, one uses Hotelling chart on beta –distribution and generalized variance chart with smoothed data.
   Monitoring the process mid – range level, Hotelling chart with the assessment of rolling range covariant matrix is applied , to increase its efficiency warning limits and non random structures availability analysis are used..
4. In case of any violations on the chart, for its interpretation partial H -test or element -by-element eliminating chart plotting are used.
5. The similar approach is good for variance monitoring
   It is worth mentioning that for more objective conclusion for stage 3 it is necessary to perform statistics experiments.

VII. ACKNOWLEDGEMENT

The research is done supported financially by Russian Foundation of Fundamental Research, grant № 18-48-730001.

REFERENCES

1. Montgomery D C, 2009. Introduction to Statistical Quality Control. New York: John Wiley and Sons, 754 p.
2. Klyachkin V N, 2011. Models and Methods of Multivariate Statistical Control for Engineering Process. M.:FIZMATLIT (ФИЗМАТЛИТ) p 196.
3. Tracy N D, 1992. Multivariate Control Charts for Individual Observations / N. D. Tracy, J. C. Young, and R. L. Mason // *Journal of Quality Technology*, **24(2)**. pp. 88–95.
4. Sullivan J H, 1996. A Comparison of Multivariate Quality Control Charts for Individual Observations / J. H. Sullivan, and W. H. Woodall // *Journal of Quality Technology*, **28(4).** pp. 398-408.
5. Williams, J D., W H Woodall, J B Birch, and J H Sullivan., 2006, Distribution of Hotelling's T2 Statistic Based on the Successive Difference Estimator // *Journal of Quality Technology*,. **38**, pp. 217–229.
6. [6] Klyachkin V N. and Kravtsov Yu A, 2016. The Detection f Violations during Multivariate Statistical Control of Engineering Process *// Software products and systems*, № 3. pp. 192-197.
7. Klyachkin V N and Karpunina I N, 2017.The Analysis of Technical Object Functioning Stability as per the Data of Monitored Parameters Multivariate Dispersion // CEUR Workshop Proceedings,**1903**. *Data Science. Information Technology and Nanotechnology* / DS-ITNT, pp.28-31.
8. Klyachkin V N. and Bart A.D., 2017.The Assessment of Potable Water Purification Process Quality Data Stability // *Ecological systems and green instruments*, №8. pp. 14-20.