Specific features of potable water purification process

A Alekseeva\textsuperscript{1}, I Karpunina\textsuperscript{2} and V Klyachkin\textsuperscript{1}

\textsuperscript{1} Ulyanovsk State Technical University, Severny Venets street 32, 432027, Ulyanovsk, Russia
\textsuperscript{2} Ulyanovsk Civil Aviation Institute, Mozhaiskiy Street, 8/8, 432071, Ulyanovsk, Russia,

E-mail: v_kl@mail.ru

Abstract. Under this consideration there is a process of potable water purification, where the chemical and physical characteristics (water turbidity and color, aluminum and chlorine content, etc) are monitored. To work with correlated characteristics it is necessary to apply to multivariate methods. The peculiarities of Hotelling charts plotting for individual observations are well known, but generalized variance charts plotting causes some problems connected with the assessment of covariance matrix for individual observations. Solving this problem one question arises: how often is it necessary to update the limit lines on the charts? We suggest applying to the test methods of the hypothesis on the equality of the mean.

1. Specifying the task

Potable water purification process is under this consideration, the process where chemical and physical characteristics \( p \) (water turbidity and color, aluminum and chlorine content, etc) are monitored. It is assumed that a learning sample of these characteristics is selected, taken within a certain period of time (taking into account that the readings are taken once in 24 hours, this certain period of time may be 30 days, as example ). To detect the process anomaly on time, in order to prevent the quality indexes going beyond the limit, statistical monitoring is executed.

The correlation matrix is plotted on the basis of the received result, which assists to divide all the characteristics into two groups. If the correlation between two characteristics is very little (it’s value is assessed with Student criteria), than to monitor the procedure with these characteristics Shewhart control charts are applied for individual observations and moving range [1-2]. Before that it is necessary to test the normality of distribution. Taking into account the small size of the sample, Shapiro–Wilk’s test is applied here. If the normality is violated, the data are normalized with Johnson transformation (very often it is possible to get the correct result with the simplest kind of transformation: taking logarithm from the initial data. In this case the control charts are plotted on the transformed data [3-4].

The other group is with correlated characteristics, for which multivariate methods are mandatory: a Hotelling algorithm to control mean level of the procedure and generalized dispersion to control dispersion. The peculiarities of Hotelling control charts plotting for individual observations are well known [5-7], but generalized variance charts plotting causes some problems connected with the assessment of covariance matrix for individual observations. Solving this problem one question arises: how often is it necessary to update the limit lines on the charts? We suggest applying to the test methods of the hypothesis on the equality of the mean.
2. Shewhart control charts

Figure 1 represents Shewhart control charts for individual observations and moving scales as per “alkalinity”, for example, non-correlated with other characteristics, plotted in Statistica system [8-9]. In the event we are considering now the distribution normality hypothesis does not contradict to the experiment data, so the data transformation was not done.

On the chart of individual observations the average value is equal to 0.275, upper control limit is at 0.340, lower one is at 0.210. It is worth noting that the position of upper limit is much lower than the allowed alkalinity value (4.4 mmol/dm³). On the charts of moving scales the average value is equal to 0.024, upper limit is 0.080. It is well seen that there is no chart where the points go beyond the limit lines i.e the process is steady.

![Shewhart control charts](image.png)

**Figure 1.** Shewhart control charts as per “alkalinity “result.

3. Hotelling multivariate chart

To monitor the stability of the process as per the correlated characteristics multivariate control charts are used: to control the process mean level Hotelling chart is used, to control multivariate dispersion generalized dispersion chart ii used. Both Hotelling control chart for individual observations and Shewhart control chart could be plotted in Statistica system.

Figure 2 represents such a chart applied for the group of three characteristics: color, pH, oxidizability. Hotelling average value is equal to 3.845, upper limit is 17.476. No anomalies in the process: all the points are below the lower limit line.
Figure 2. Hotelling chart with three correlated characteristics: colour, pH, oxidizability.

4. **Multivariate generalized variance control chart plotting specific features.**

The approach application is suggested, similar to standard deviation assessment for moving scales in one-dimensional control. To assess the standard deviation in this case the following function is applied

$$\sigma = \frac{\bar{R}}{d},$$

where moving range mean value is

$$\bar{R} = \frac{1}{m-1} \sum_{t=1}^{m-1} MR_t,$$

and the moving range is calculated as per the following formula

$$MR_t = |x_{t+1} - x_t|.$$

The table coefficient $d$, whose value is based on the ranges distribution, is selected when the sub-group volume is equal to $n = 2$ (though the actual value is equal to 1), as for the moving range calculation two neighbouring observations are applied.

The peculiarity of generalized variance assessment is as follows.

To calculate the position of limit lines it is necessary to select the sample size equal to $n$, i.e. one unit bigger than the number of the monitored characteristics: $n \geq p + 1$ [10-11]. For independent indicator the moving range is assessed as per two neighbouring observations $x_i$ and $x_{i+1}$ ($n = 2$). While monitoring three correlated characteristics ($p = 3$) it is possible to average three moving ranges, i.e. apply to 4 observations ($n = 4$).

In this case we get:

$$MR_{jt} = |x_{jt} - x_{j,t-1}|.$$
\[ MR_{j,t-1} = |x_{j,t-1} - x_{j,t-2}|; \]
\[ MR_{j,t-2} = |x_{j,t-2} - x_{j,t-3}|; \]
\[ R_{jt} = \frac{MR_{jt} + MR_{j,t-1}M + R_{j,t-3}}{3}; j = 1,2,3; t = 4...m \]

(where \( m \) is the number of observations in the learning sample).

Then the assessment of the covariance matrix elements \( S_{ij} \)
\[ S_{jj} = \sigma_j^2, \]
\[ S_{jk} = \rho_{jk} \sigma_j \sigma_k, \]

where the standard deviation approximate value is
\[ \sigma_j = R_0 d; \]

and the correlation coefficient between the characteristics \( \rho_{jk} \) are assessed as per all observations \( m \).

Figure 3 depicts the control chart of generalized variance plotted as per the studied algorithm in Excel table. The status of central line is equal to 1.20 \( 10^{-3} \), the upper line is equal to 2.67 \( 10^{-3} \), the lower line is equal to zero, as the result of the calculations was negative. Figure 3 brings us to the conclusion that the process is steady as per multivariate dispersion criteria too.

**Figure 3.** The chart of generalized variance with three correlated characteristics: colour, pH, oxidizability.

5. **Control limit updating**

One more question arising during the problem solving is: “How often is it necessary to update the control limit on the control chart?” The January data learning sample is apparent to give false results in April. Here we may apply to the test methods of the hypothesis on the equality of the mean: is the difference between the data for the latest 10 days and the data of the preceding 10 days significant as an example?

Such an approach is possible. The data for the first 10 days of the following month are collected. The control charts for this period are plotted with the same limits as in the previous month. The test of the hypothesis on the equality of the mean is done for each characteristics monitored within the monitoring period and within the period of 10 days of the preceding month. If this hypothesis is accepted, no need to update the limits.
Firstly, the hypothesis on the equality of the variances by F-test is done. For that the following tools are applied: Excel (Data / Data analysis / Two-sample F-test for variance): 10 observations are compared each month.

According to the result of F-test for the hypothesis on the equality of the mean we apply to two-sample t-test with same or different variances. If the sample value of Student’s t-statistic happens to be bigger than the critical value, than the hypothesis on the equality of the mean is rejected, and it is necessary to update the limits.

The test for the hypothesis on the equality of the mean is required for each monitored characteristic, as the time of the limits updating on the different types of control charts may differ. For multivariate charts (Hotelling and generalized variance) the difference of at least in one correlated characteristics means the necessity to update the limits.

It is worth noting that the considered algorithm of the test of the hypothesis on the equality of the mean are accepted only for normally distributed data, so if necessary the transformation, rescaling the input results of observations is used.

6. Conclusion
During the potable water purification monitoring process with application of multivariate statistical methods for individual observations, there comes a problem to assess the covariance matrix in order to plot the control chart for generalized variance. There was offered an approach similar to the one applied for the monitoring process of independent characteristics, based on moving ranges.

Acknowledgment
The investigation is carried out supported by the joint research grant from Russian Foundation for Basic Research and Ulyanovsk region government, project 18-48-730001.

References
[1] Wheeler D and Chambers D 1992 Understanding Statistical Process Control (Tennessee: SPC Press) p 409
[2] Ryan T 2011 Statistical Methods for Quality Improvement (New York: John Wiley and Sons) p 687
[3] Montgomery D 2009 Introduction to Statistical Quality Control (New York: John Wiley and Sons) p 738
[4] Lowry C and Montgomery D 1995 A review of multivariate control charts Technometrics 27 800-10
[5] Klyachkin V, Zhukov D and Karpunina I 2020 Hydraulic Unit Vibration Monitoring Using Machine Learning International Multi-Conference on Industrial Engineering and Modern Technologies 9271439
[6] Alekseeva A, Karpunina I and Klyachkin V 2020 Analysis of hydraulic unit operation stability according to its vibration monitoring results CEUR Workshop Proceedings 2667 46-9
[7] Klyachkin V, Zhukov D and Zentsova E 2019 Analysis of stable functioning of objects using machine learning CEUR Workshop Proceedings 2416 19-25
[8] Borovikov V 2001 STATISTICA: Art of Data Analysis Advanced (Petersburg: SPb) p 656
[9] Franceschin F, Galetto M and Genta G 2015 Multivariate control charts for monitoring internal camera parameters in digital photogrammetry for LSDM (Large-Scale Dimensional Metrology) applications Precision Engineering 42 133-42
[10] Klyachkin V, Shirkunova K and Bart A 2019 Analysis of the stability of the chemical composition of wastewater in the production of printed circuit boards Ecology and Industry of Russia 23(5) 47-51
[11] Carlos García-Diaz J 2007 The ‘effective variance’ control chart for monitoring the dispersion process with missing data Industrial Engineering 1(1) 40-5