Quality of water using multivariate analysis

Adriano Mendonça Souza[1], Roselaine Ruviaro Zanini[1],
Anaelena B. de Moraes[1], César O. Malavé[2]

[1]Department of Statistics
Federal University of Santa Maria - Santa Maria - RS - Brazil
amsouza@smail.ufsm.br; rrzanini@ccne.ufsm.br; anaelena@smail.ufsm.br
[2]Department of Industrial Engineering
Texas A & M University - College Station - TX - USA
malave@tamu.edu

Resumo

O objetivo deste artigo é analisar a qualidade da água distribuída para a população de Santa Maria - RS - Brasil, por meio de gráficos de controle multivariados e análise de componentes principais. Utilizando estas técnicas conjuntamente foi possível avaliar a estabilidade do processo e identificar as variáveis causadoras de instabilidade no processo.

Palavras chaves: Controle de qualidade, gráficos de controle multivariados, componentes principais

1. Introduction

The preoccupation with the quality is present in every aspect that involves goods and/or service productions. With the technological advances and with the increasing population rate must be pointed out the importance in evaluating also the potable water quality, considered for ingestion, assessing its physical, chemical and biological features, becoming the quality standard an important link in the study and control mechanism in public health.

In this way, it is searched to analyze, together, some variables involved in the water treatment process consumed in Santa Maria City - RS, with the objective to investigate the stability of this process and to verify if some features responsible for the potable water quality are under statistic control. The observed variables, each time, for thirty days, are: turbidity, color, pH, alkalinity, residual Clorox and Fluor.

The turbidity, expressed in units of turbidity (uT) is caused by the presence of insoluble particles of clay, thin sand, mineral matter, organic residues, plankton and other microscopic organism in suspension and that prevent the light to pass through the water. The color is measure in scale unit of Hazen (uH). The alkalinity measures the total quantity of alkaline substances present in the water, in parts by million of Calcium Carbon. The presence of residual Clorox measured in mg/l is the bacteriological quality guaranty on the water supply. The Fluorides measured in mg/l happened naturally in the supply systems, considering that they are important and essential components, helping in avoiding teeth decay.

Initially, the observations will be analyzed using Hotelling's T^2 chart, to evaluate the process stability. Since the multivariate chart has detected a point out of control, a diagnosis held through invariable charts, together with principal component analysis, must help to decide which component is out of control. Most of the diagnoses quoted in the literature suggest to use together invariable p-charts in the principal components derived from the original data (TRACY *et al*., *apud* GHOSH *et al*., 1996; LOWRY & MONTGOMERY, 1995; TSUI & HAYTER, 1994 e TIMM, 1996).

After it is realized a correlation study between the original variables and the components out of control, identifying the variable groups that will be possible responsible for the instability in the process and that must be monitored.

2. Control charts

The control charts are used, generally, to achieve a statistic control state and for monitoring them. They are useful to distinguish between common causes and special causes of variability. The first comes from the natural variability of the process, having a random behavior, indicating that it is under control. The second revels the special standard formation, accusing that the outside variables are influencing the process, and must be identified and removed for the whole production not to be affected.

2.1 Hotelling´s T^2 multivariate control chart

The statistic of Hotelling's T^2 performs an important role on multivariate control quality assessing the system stability, when there is *p-variables* that need to be analyzed, together. According to RYAN (1989) and TRACY *et al*., (1992), the multivariate charts are more sensible in detecting points out of control, when the variables have correlation.

If the population values are unknown, their values can be estimated. The vector μ will be estimated by $\bar{\bar{X}}$ e and the matrix Σ will be estimated by the S, hence:

$$T^2 = n(\bar{X} - \bar{\bar{X}})' S^{-1} (\bar{X} - \bar{\bar{X}}) \qquad 2.1$$

According to LOWRY & MONTGOMERY (1995) and TRACY et al. (1995), Hotelling's chart presents two distinct phases of process evaluation: In the first phase the control limits are used to test if the process was under control when the sample was withdrawn.

$$LSC = \frac{p(m-1)(n-1)}{mn - m - p + 1} F_{\alpha,p,mn-m-p+1} \qquad 2.2$$

$$LIC = 0$$

where: *p* is the number of analyzed variables; *m* is the number of observations in each variables; *n* is the total number of observations; *F* is the tabled statistic value; and α is the specific significance level.

Many times, it is necessary to establish the statistic control when the sample subgroup is of equal size to (n = 1). In this case, the statistic of T^2, in (2.3) becomes:

$$T^2 = (X - \overline{X})' S^{-1} (X - \overline{X}) \qquad 2.3$$

Then, it must be used the following control limits, for individual observations:

$$LSC = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p} \qquad 2.4$$

$$LIC = 0$$

When the signal is captured out of the control, using a multivariate chart, the characteristic or the characteristic groups that cause the signal can not be visually identified.

2.2 EWMA statistic applied to control chart

The Exponentially Weighted Moving Average (EWMA) chart has a mechanism that incorporates the information of every previous observation and also the present information, being updated recursively as showed (2.5).

$$Z_i = \lambda X_i + (1-\lambda) Z_{i-1} \qquad 2.5$$

The Z_i series is smoothed through a weighted constant λ that multiples the values of the X_i original series plus the pondering constant complement multiplied by the value Z_{i-1}. This initial value in general is represented by the process average. The control limits for the EWMA chart are:

$$LSC = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$$

$$LC = \mu_0 \qquad 2.6$$

$$LIC = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$$

where: L is the length of f the limits: λ is the weighted constant that must be in the interval of $0 \leq \lambda \leq 1$; μ_0 is the target value that is desired to reach and σ is the deviation standard of the Z_i´s . The process will be considered under control, if every point is in the established limits.

2.3. Principal component analysis

Nowadays, one of the main uses of Principal Component Analysis (PCA) happens when the variables are originated from processes in which several characteristics must be observed simultaneously. This technique is discussed by many authors like MORRISON (1976), REINSEL (1993), JACKSON (1956, 1981) e JOHNSON & WICHERN (1992). This technique makes possible the transformation of original variables in new variable groups that maintain the maximum, the set variability. The new variables are *independents and non-correlated*, denominated as Principal Components (PC) and are linear combinations of original variables.

Suppose that X is a vector of random *p*-variables and that the variance and correlation structure between the variables is of interest. In this study, it was used a sample set, with the matrix Σ that is estimated through the variance-covariance sample matrix S and the average vector $\bar{\bar{X}} = [\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p]$. From S matrix is possible to find the values $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq \ldots \geq \hat{\Lambda}_p \geq 0$, that are characteristic roots, all distinct and presented in decreasing order of values and, the total variance of the system is *S*.

According MORRISON (1976), the jth principal component extract from a sample of p-variables is a linear combination, such as $\hat{Y}_j = \hat{\ell}_{1j} X_1 + \cdots + \hat{\ell}_{pj} X_p$. The sample variance of jth component is $\hat{\Lambda}_j$ and the whole variance is $\hat{\Lambda}_1 + \cdots + \hat{\Lambda}_p = tr\, S$. The explanation degree supplied by jth component is supplied by $\dfrac{\hat{\Lambda}_j}{tr\, S}$.

The PCA is useful in identifying the variables out of control with the multivariate data, being more efficient than the T^2 chart to detect small changes of the target (WOODALL & NCUBE, 1985). The component number definition can be made by the chart method, (CATTEL (1966) *apud* PLA (1986) and by the method that consider the components which the eigenvalues are superior to 1 (KAISER (1960) *apud* MARDIA (1979)). In general, it was used the components that synthesize an accumulated variance around 70%.

3. Application

The quality control is defined and applied in several fields, as in industries as in render services. In this work, the use of control chart uni and multivariate was fundamental to find the possible variables out of control in the water treatment process consumed by the city population of Santa Maria - RS. The analyzed variables are: turbidity, color, pH, alkalinity, residual Clorox and Flour, compounded of water 673 observations. Initially, the last 200 variables involved in the water treatment process were analyzed through Hotelling's chart, according to what the Figure 1 shows.

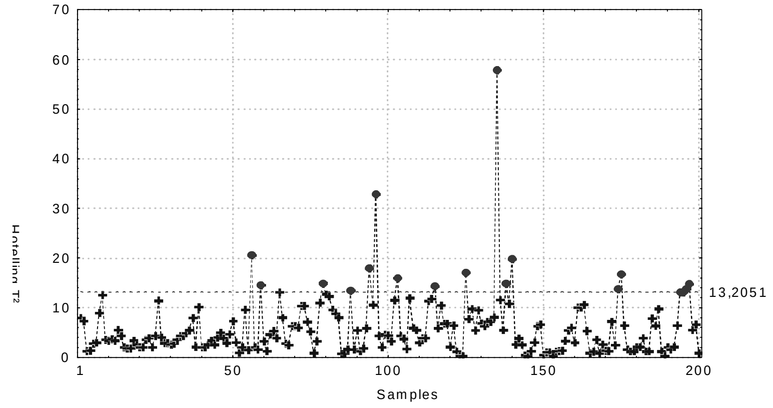

Figure 1. Hotelling's T^2 Chart to 5% for the six variables compounded of 200 observations involved in the water treatment process in the city of Santa Maria

It was observed that there are 18 points out of control, what revels an unstable situation. But, as previously discussed, the T^2 chart just signalizes this lack of control, without indicating the possible responsible variables. To investigate more in details it is needed to use the principal component analysis, writing the original variables in linear combination, which are studied by exponentially weighted moving average (EWMA).

After verifying that the process is unstable, it was determined the principal components by the R correlation matrix, standardizing in this way the measured units of the different variables, eliminating the magnitude influence of one above the other (JACKSON, 1981).

Afterwards, it was determined the component numbers which the selection followed CATTEL (1966) and KAISER method (1960). Thus, it was used the first three components that accomplish a total of 67,1431% of total explication. The principal components and other results can be visualized on the Table 1.

The first three CP selected were analyzed by EWMA chart and were identified out of control. In that case, it will be investigated which are the variables that are more representative in each component. Therefore, it will be used the correlation analysis between the principal component and the original variables, identifying that, on the first component, the color is the variable more representative, followed by turbidity; on the second component, the Clorox was the variable more representative, followed by alkalinity and on the third component the Flour showed itself more representative.

On the first component the color is monitored to be possible to obtain a quality potable water, the same must be done with the Clorox and Flour. This monitoring is made, building EWMA chart, applied to original variables, according to what is showed in Figure 2, 3 and 4.

Through the original variable charts analysis was verified that the process presents itself unstable, what recommends a careful investigation on the work routine in the water treatment station of Santa Maria - RS.

Table 1. Weighted values of each variables that will form the linear combinations, found by Varimax rotation, eigenvalues and explained variance percentile

Variables	***CP1***	***CP2***	***CP3***	***CP4***	***CP5***	***CP6***
Turbidity	0.276197	0.105838	-0.048839	0.051217	0.015776	0.952500
Color	0.959646	0.060520	0.07182	0.025106	0.045227	0.269610
PH	0.023391	-0.013597	-0.040535	0.997308	-0.031005	0.045180
Alkalinity	-0.042253	-0.164422	0.075040	0.032420	-0.981973	-0.015056
Clorox	0.058718	0.979333	0.003807	-0.01468	0.165900	0.098583
Flour	0.005437	0.003310	0.995640	-0.040755	-0.071948	-0.042686
Eigenvalues	1.739544	1.218386	1.070658	0.878854	0.649590	0.442968
Explained Variance Percentile	28.99240	20.30644	17.84429	14.64756	10.82651	7.38280

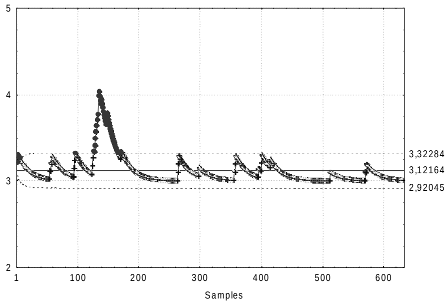

Figure 2. EWMA chart to color with L = 2,5 e λ = 0,05

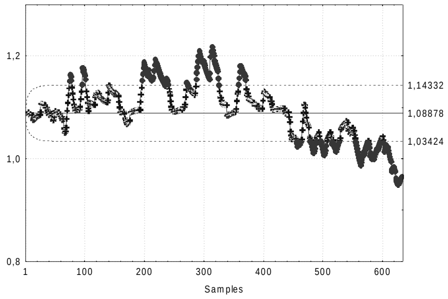

Figure 3. EWMA chart to Clorox with L = 2,5 e λ = 0,05

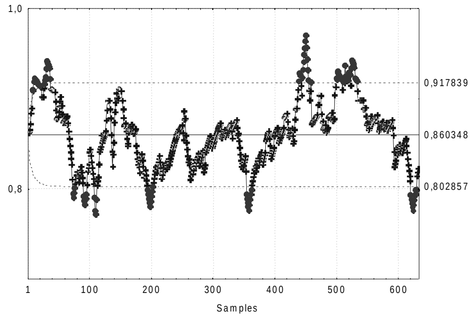

Figure 4. EWMA chart to Fluor with L = 2,5 e λ = 0,05

4. Conclusion

In this work was searched to analyze some variables involved in the water treatment process for human consume, in Santa Maria - RS. The variables analyzed were turbidity, color, pH, residual clorox, alkalinity and flour. Initially, it was constituted Hotelling's control chart, observing the variables together and verifying an instability in the process certified by the great number of points surpassing the control limit. Afterwards it was selected, among the variables, three principal components, which accumulated 67,1431% of the explication of total variance.

These components were analyzed, individually, through EWMA chart and it was proved that they were out of statistic control. Thus, it was sought to identify which ones were the representative variables in each component. It was observed that the color is more representative on the first component, on the second component it was identified the residual Clorox and on the third component it was identified the Flour component as the most representative.

It is highlighted, however that yet the charts have showed points out of the control statistic limits, this fact does not bring consequences to the consumer, because besides they have unilateral aspects of observation they maintain themselves in the pre-established limits of specification. The possible causes that could have led to a lack situation of control in this process will be in order of importance: the change of reservoir filters, washing filters, occasional pipe leaking that connects the water tanks and the low water pouring.

Acknolegements

We thanks to the "Companhia Riograndense de Saneamento de Santa Maria - RS" and to the Nuclei of Quality and Productivity of Department of Statistics - Federal University of Santa Maria and the referres.

References

GHOSH, S.; SMITH, W.; SCHUCANY, W. R. **Statistics of quality**. Statistics: textbooks and monograph series, vol. 153. Ed.: Gosh & Smith & Schucany, 1996.

JACKSON, J.E. (1956).**Quality control methods for two related variables**. IQC, January, pp. 4 - 8.

____ . (1981). **Principal components and factor analysis: Part I - principal components**. JQT, October, v.12, n.4, pp.201 - 213.

JOHNSON, R.A., WICHERN, D.W. Applied multivariate statistical analysis. 3 ed. Prentice-Hall. New Jersey, 1992.

____ . **Applied multivariate statistical analysis**. 4 ed. Prentice-Hall. New Jersey, 1998.

LOWRY, C.A. and MONTGOMERY, D.C. (1995). A review of multivariate control charts. IIE Transaction, v. 27, pp.800 - 810.

LOWRY, C.A; WOODWALL,W.H.; CHAMP, C.W.; RIGDON, S.E. (1992). **A multivariate exponentially weighted moving average control chart**. Technometrics, February, v.34, n.1, pp.46 - 53.

MARDIA, K.V.; KENT, J.T. and BIBBY, J.M. **Multivariate analysis**. Academic, London, 1979.

MORRISON, D.F. **Multivariate statistical methods**. 2. Ed., New York, NY. Mc Graw Hill. (1976).

PLA, L.E. **Analysis multivariado**: Metodo de componentes principales. Universidad Nacional Experimental Francisco de Miranda. Coro, Falcón, Venezuela, 1986.

REINSEL, G. C. **Elements of multivariate time series analysis**. Springer-Verlag. New York, 1993.

RYAN, T.P. **Statistical Methods for quality improvement**. John Wiley & Sons, Inc. New York, NY, 1989.

TIMM, N.H., (1996). **Multivariate quality control using finite intersection tests**. JQT, April, v. 28, n.2, pp. 233 - 243.

TRACY, N.D.; YOUNG, J.C.; MASON, R.L. (1992). **Multivariate control charts for individual observations**. JQT, April, v.24, n.2.

TRACY, N.D.; YOUNG, J.C.; MASON, R.L. (1995). **A bivariate control chart for paired measurements**. JQT, v.27, pp. 370 - 376.

TRACY, N.D.; YOUNG, J.C.; MASON, R.L. (1997). **A practical approach for interpreting multivariate T2 control chart signals**. JQT, October, v. 29, n.4, pp. 396 - 406.

TSUI, K., HAYTER, A. J., (1994). **Identification and quantification in multivariate quality control problems**. JQT, July, v. 26, n. 3, pp. 197 - 208.

WOODALL, W.H. e NCUBE, M. (1985). **Multivariate CUSUM quality control procedure**. Technometrics, August, v. 27, n. 3, pp. 285 - 292.