Chapter 3

Quality Control for Multi-variable Problems

Eric Agyekum[2], Andreea Amariei[3], Brenda Caughlin[7], Jie Cheng[3], Ella Huszti[3], Veselin Jungic[4], Emmanuel Ngembo[5], Grace So[6], Paul Wiebe[1],

Report prepared by Luz Palacios[1]

Introduction, Discussion and Editing by Rita Aggarwala [1]

3.1 Introduction

Chemex Laboratories measures concentrations of various minerals in land samples provided by their clients. The problem presented by Chemex is that they would like to be able to detect when their measurement apparatus needs re-calibration due to a shift in process parameters, so as to minimize rework.

Currently, calibration of the measuring machine is monitored by measuring mineral concentrations in control samples about every 40 minutes or 20 samples. Chemex would ideally like to make a decision about the calibration at every control sample assessment, that is, in real time, with a low error rate. Therefore the tool implemented must be quite simple, powerful and readily interpretable. This is complicated by the fact that multiple dependent measurements are made on each control sample, and only one set measurements is made before the decision is made. At present, the re-calibration decision is made using traditional, univariate control charts for a few key mineral measurements.

Briefly, the measurement process can be described as follows: Some compound, whose composition is of interest is prepared and is passed to a machine, the ICP, which sprays it through a plasma flame exciting the atoms. Photons of characteristic wavelengths are emitted as the atoms return to their ground states. These are counted to determine the concentrations of the atoms present in the compound.

There are many sources of error in this process but, the three with the largest contributions are: the sample is susceptible to contamination at the preparation stage, the spray nozzles often get plugged and the plasma flame may not heat to precisely the correct temperature.

Each sample contains more than 20 elements whose atom concentrations are to be measured. In consultation with the industrial mentor, 16 of these elements were identified to be of interest for this problem: Ag, Al, As, Ca, Co, Cr, Cu, Fe, Mg, Mn, Ni, P, Pb, Sr, V and Zn.

[1]University of Calgary
[2]University of Victoria
[3]University of Alberta
[4]Simon Fraser University
[5]University of Montreal
[6]University of Toronto
[7]Chemex Laboratories

Chemex has considered univariate quality control analyses, that is, they have considered errors in measurement of each individual characteristic one at a time, but have not considered all of the characteristics together, thereby potentially losing information about dependencies between mineral measurements and inflating the chances of false alarms. In addition, the large number of measurements involved renders this a cumbersome procedure. The decision rule used by Chemex at present is that recalibration is performed if more than 3 of the 16 mineral concentrations for the control sample are outside the traditional univariate Shewhart "two-sigma control limits."

Different multivariate approaches are presented here for the 16 elements of interest, considering the characteristics together and taking advantage of computing capability available on even the most basic systems in order to arrive at effective decision rules which accurately and precisely flag truly out-of-control measurements on the process. Three very contrasting approaches to the problem were taken during the PIMS 1999 Industrial Problem Solving Workshop. They are: considering the process as a Markov Chain; use of Bayesian Belief Networks; and Multivariate control chart implementation. Preliminary data analysis for data provided by Chemex was also performed in order to initially assess some of the models considered.

Approaches

3.2 The Markov Chain Model

Paul Wiebe and Eric Agyekum

3.2.1 Introduction

The Chemex Labs multivariate quality control problem is to decide the 'best' rules for controlling their process of determining atomic concentrations. However before a quality tool can be implemented the mathematical foundations of the process of interest must be uncovered and compared with the mathematical assumptions associated with the tool. The suitability of a vector Markov Chain model for the Chemex measurement process is studied in this context. It is a good model for a simulator which is easy to implement and on which potential control rules can be tested without tying up the machine.

3.2.2 The Model

The physics of the interactions in the plasma flame are not well described and the other two sources of error (contamination at the preparation stage and spray nozzle plugging) seem to be purely stochastic in nature except that the plugging should effect a trend. This led us to a Markov chain description.

Let,

$$X : \Omega \times N \to S$$

be the discrete–time process

$$X := (X^{\mathtt{Pb}}, X^{\mathtt{Au}}, \cdots, X^{\mathtt{Zn}})$$

with state space

$$S := S^{\mathtt{Pb}} \times S^{\mathtt{Au}} \times \cdots \times S^{\mathtt{An}}.$$

The decision to 'exert control', or the **control rule**, on the ICP is modelled as a stopping time (with respect to the filtration generated by X),

$$T : \Omega \to N.$$

This is a function on the sample paths of the process X into discrete time: each sample path is associated to a time T, the time to 'exert control'. It may be argued that T is a stopping time and that X has the Markov property.

Let $\sigma(X(\cdot,1),\cdots,X(\cdot,n))$ be the element of the filtration generated by X up to time n.

Since it is impossible to exercise a control rule which depends on future states of the process (if it were then this problem would not exist), it must be the case that $1_{\{T\leq n\}} \notin \cup_{m>n}\sigma(X_1,\cdots,X_m)$. Hence a stopping time is a good model for the control rule.

The Markov property,

$$P(X(\cdot,n+1)\in B|\sigma(X(\cdot,1),\cdots,X(\cdot,n)) = P(X(\cdot,n+1)\in B|\sigma(X(\cdot,n))$$

where $B \in \sigma(S)$, is interpreted as follows: to determine the probable state of the process at time $n+1$ the information about the complete history of the process is no more (or less) useful than the information about its current state (at time n). So it is argued that the transition from one state to another (possibly the same state) in one time step does not depend on states other than the current state. It is conceivable that the transition does depend on previous states; for example if the nozzle is plugged there may be a time after which it becomes unplugged. If this time is short than we can ignore the plugging. In the event that this isn't the case, plugging will arrest the process. In any event the Markov property remains intact. Such arguments can be prolonged *ad infinitum* but the real reason for considering the Markov chain model is its well–developed theory, and the fact that Markov chain models are often useful in approximating physical processes.

Immediately, two aspects of Markov chain theory are useful: the transition probability matrix and the chain's equilibrium distribution. The transition probability matrix describes the dynamics of the process. For any given current state the matrix gives the conditional distribution of advancing to 'next' states. This is important because it is information on the future of the process. For example, if the conditional distribution is 'nearly degenerate' then it can be acted upon with confidence. The equilibrium distribution of the Markov chain is the long-term proportion of time each element is in each state. So, if an element has far exceeded its expected time in a certain state, the operator may suspect that a shift has occurred in the process parameters.

In the multivariate case, if a subset of elements are correlated, these should be grouped. This means, at worst, taking the Cartesian product of their respective state spaces; if the structure of the multi–correlation can be determined then a smaller state space is feasible. Overall we would like to have a partition of the full set of elements into statistically unrelated groups whose dynamics can be determined independently.

The Markov Chain model is also an excellent candidate for simulation. One only needs to determine the transition matrices of each of the groups of related elements. These can be estimated most simply (but there are probably more elegant and robust ways) by counting the state transitions. To simulate, begin the process in a known state (the expected state); the transition matrix gives the probability distribution of 'next' states given the current one, so draw from it. Now the process is in another known state and the transition matrix gives the (new) distribution of 'next' states, so draw from it, etc. Based on the history of the simulated process, control rules can be tested and the future states of the chain compared with other future states based on different control rules.

3.2.3 Postscript

A Martingale model was also considered but discarded since the Martingale property implies that the expectation of the process at any time is the same as it was in the beginning. In fact, this appears to be a part of the notion of a process being 'in control'. So it seems initially that a control rule monitoring the mean of a Martingale should never demand that control be exerted. A similar statement could be made about the (time) cross–sectional variance of the process: as soon as it ceases to be constant (probabilistically) but starts to expand (perhaps like a Brownian motion) current Shewhart rules object and demand that control be exerted. This idea brings to mind the notion of an **ergodic theorem**.

An ergodic theorem is a tool which tells the practitioner that statistical properties of a process, under some stationarity [8] condition, usually can be approximated by the statistics of a single sample path. It seems that control rules are dictated by ergodic theorems, or in particular by their contrapositives. That is, given an ergodic theorem on some statistic `S` of the process, as soon as it can be determined with some certainty that the estimate of `S` based on the sample path is deviating then it is time to exert control.

[8] A process Y is called **stationary** if its finite dimensional distributions are invariant under shifts of time ; it is called **weakly stationary** if its mean and autocorrelation are constant in time.

3.3 Using Bayesian Belief Networks for Quality Control

Jie Cheng

The Bayesian belief network (BN) is powerful tool for knowledge representation and reasoning under conditions of uncertainty. BNs have been used successfully in fault diagnosis, decision support and classification. Unlike rule-based knowledge base systems, the BN is based on the solid foundation of probability theory and graph theory - it can express uncertainties in a natural way. Unlike a neural network, which is often interpreted as a "black box", the nodes in a BN are domain variables and the links are causal connections among the variables. Therefore, human experts can create a BN by hand. In the case that the BN is developed from training data, human experts can easily understand it and make modifications. A BN usually requires less training than a neural network does. Evidence shows that BNs often outperform other tools like decision trees, neural networks and statistical programs in classification tasks. For each node in a BN, there is an associated probability table that specifies the probability distribution of the variable given its parents.

For the Chemex problem, we can create a group of BNs, one for each particular physical problem of the instrument. After getting the readings from the instrument as input, each BN will give the probability that the physical problem occurs. (This computation will take less than 0.01 second on an average PC.) Then we can set (and fine tune) a rule that will trigger the alarm when a probability appears to be too high.

There are three ways to get these BNs.

1. Learn both the structure and the parameters (probability tables) of the BNs from data.
2. Let human experts specify the structure of BNs and determine the parameters from data.
3. Let human experts specify both the structure and the parameters of the BNs.

Based on the information from the workshop, we can probably construct the BNs by hand, i.e., the diagnostic (problem) node is the parent of all elements. For example, in the power failure BN, the power supply node is the parent of all elements and the elements are independent given the status of the power supply. For the power supply node, the probability might be P(failure)=0.05, P(normal)=0.95. For an element node, the probability might be:

$$
\begin{aligned}
P(-s < Ag < s \mid failure) &= 0.2 \\
P(-2s < Ag < 2s \ and \ s > Ag > -s \mid failure) &= 0.3 \\
P(Ag < -3s \ or \ Ag > 3s \mid failure) &= 0.5
\end{aligned}
$$

$$
\begin{aligned}
P(-s < Ag < s \mid normal) &= 0.8 \\
P(-2s < Ag < 2s \ and \ s > Ag > -s \mid normal) &= 0.15 \\
P(Ag < -3s \ or \ Ag > 3s \mid normal) &= 0.05
\end{aligned}
$$

(Strictly speaking, this way of constructing BNs assumes that each problem occurs independently and the problems do not occur at the same time. If the assumption is unacceptable, we need a more complex BN structure instead of a group of simple BNs.)

If such probability distributions cannot be given, data collection and experiments will be necessary. The data can be collected through everyday operation by recording the readings of the normal condition and the readings when something is wrong. (The diagnostic results should also be recorded.) Alternatively, the data can be collected by doing experiments - create a physical problem intentionally and record the readings.

3.4 Multivariate Control Charts

Chemex provided data containing 109 'in-control' data points (that is, sample measurements which appeared to be within control limits as determined by Chemex's current rules), estimates of population mean concentrations of the various substances of interest when the process is operating within natural variation, and 348 'out-of-control' samples (that is, measurements which included both in-control and out-of-control points as per Chemex's current rules). It should be noted that good, valid data collection, although not the focus of this report, is another issue

which should be addressed before implementing any control tool. The data provided by Chemex were used in order to assess two multivariate algorithms, which are discussed in the following two subsections. The final subsection on multidimensional Shewhart control charts explores the mathematical generalization of univariate control rules to the multidimensional case. The use of two or more of the tools presented in this section can potentially provide a very strong, practical tool for both pinpointing when and assessing why the process of interest appears to be out of control.

3.4.1 Principal Component Analysis

Grace So

The key idea of principle components is dimension reduction. Our goal is to reduce the dimension of the problem to be smaller than 16, with enough explanation of variability.

The correlation matrix of the in-control data was used to carry out the principal components analysis.

The first 8 eigenvectors were chosen, which explain about 85% of the variation.

1. A matrix U consisting of the eigenvectors $u_1, u_2, \cdots, u_8$ was formed. Also, $U'SU = L$ was obtained, where L is an 8×8 diagonal matrix consisting of the eigenvalues $l_1, l_2, \cdots, l_8$ of S. The principle components of x can be formed by the following transformation:

$$z = U'(x - \bar{x}) \tag{3.1}$$

2. We use the Hotelling T^2 for the (8 dimensional) principal components to determine whether each observation is out of control. If a particular observation is out of control, we should be able to check which of the 8 Principal Components caused the problem in further analysis. We are now ready to calculate the Hotelling T^2. Before doing this, the principle components were required to be scaled. The formula used for scaling the principle components is

$$w_i = \frac{u_i}{\sqrt{(l_i)}} \tag{3.2}$$

such that a scaled matrix, W was formed and

$$W'W = L^{-1} \tag{3.3}$$

The Hotelling T^2 is defined as

$$T^2 = z'L^{-1}z \tag{3.4}$$

The 109 values of T^2 were obtained from the in-control data. The 95% quantile of the distribution of the T^2 statistic can be easily found, since[9]

$$T^2 \sim \frac{p(n-1)}{(n-p)} F_{p,n-p} \tag{3.5}$$

In this example, $p = 8,\ n = 109$, and $\alpha = .05$. $F_{8,109,0.05} = 2.0314$ so $\frac{8(109-1)}{(109-8)} F_{8,109,0.05} = 17.3774$. *All* of the 109 observed T^2 values were greater than this value, indicating a problem either in the data or in the application of the methodology. As the problem was further pondered, the difficulty of interpretation was discussed, since the individual entries of the main principal components were similar. This method was therefore discarded as the best approach this problem, and another approach was considered, as presented in the following subsection.

[9] Assumptions of multivariate normality and independence between samples must be checked and verified.

3.4.2 Confidence Region and Simultaneous Bounds

Luz Palacios and Emmanuel Ngembo

Since the dimension of the problem could not be reduced efficiently through principal component analysis, further analyses were done with the 16 characteristics, using Confidence Regions and Simultaneous Bounds. [10]

In order to compare the assessment of which points are classified as in-control by Chemex's current rules and the Hotelling multivariate approach considered here, Hotelling's T^2 for the original (16-dimensional) data was calculated for each of the 109 Chemex in-control samples, using the means and covariance matrix from this data.

$$T^2 = (\mathbf{X} - \mu)^T S^{-1} (\mathbf{X} - \mu)$$

The upper control limit is

$$\chi^2_{16} = 26.296.$$

The lower control limit is 0.

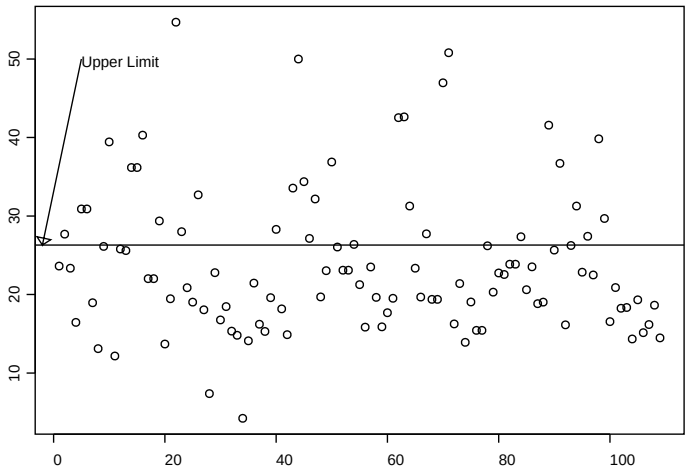

Figure 3.1: T^2 in-control data

As seen on Figure 3.1, many points fall outside of the limits. This is due to the fact that Chemex used a different method for determining in-control samples.

Hotelling's T^2 was also calculated for each of the out-of-control samples, using the mean and covariance matrix mentioned above from the in-control data. The upper and lower limits are the same as before.

As seen on Figure 3.2, many samples fall outside the limit.

At this point, we would like to be able to determine the cause (particular element measurements) contributing to the process falling outside control limits, if the process parameters have actually shifted. This would be done for each point falling outside the control limits.

In order to illustrate the rest of the analysis, sample 18 was chosen as an example. The T^2 value for this sample is approximately 80, which is greater than 26.296, which means that a problem might exist in this sample, so simultaneous bounds for each characteristic were calculated. These are based on Bonferroni Simultaneous Confidence Intervals, and have an overall confidence level of about 100(1-α)% provided associated assumptions are met. (Here, p is the dimension of the problem, in this case, 16.)

[10] Assumptions of multivariate normality and independence between samples must be checked and verified.

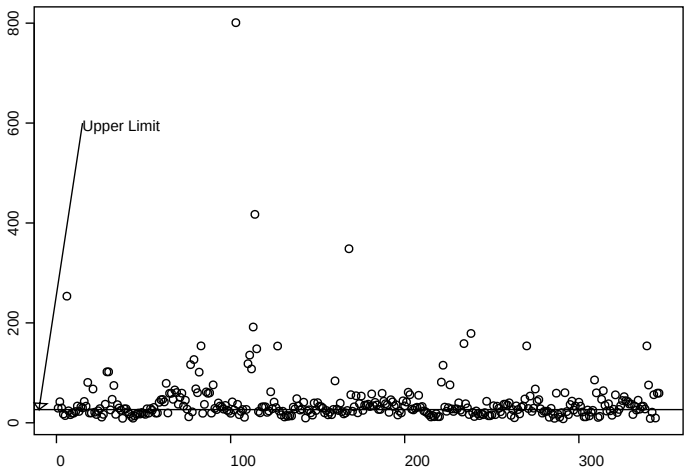

Figure 3.2: T^2 out-control data

$$\mu_i \pm Z_{\frac{\alpha}{2p}}\sigma_i \qquad i = 1, ..., 16$$

A table is presented showing the Silver (Ag) and Manganese (Mn) characteristics using $\alpha = 0.05$.

Characteristic	Upper limit	Lower limit	Observation
Ag	5.877	2.922	6.8
Mn	1,079	775	1,020

Figure 3.3 and Figure 3.4 are the graphs for Silver and Manganese.

The value for Silver falls outside of the limits, so we can say that a problem might exist in this characteristic. The value for Manganese does not fall outside of the limits, which means there is no signal for saying that a problem exists in this characteristic.

These analyses can be done with each sample associated with a point falling outside the control limits, and each characteristic, while maintaining the overall confidence level for each point.

3.4.3 Multidimensional Shewhart Control Charts

Veselin Jungic, Ella Huszti, and Andreea Amariei

This note is inspired by the fact that Shewhart control charts work well in the univariate (one characteristic) case. The effectiveness of Shewhart's idea in years of practice and its well defined and simple constraints give another reason for this exercise. Also, a motive for this note is that one can easily imagine a situation that if an instrument that measures n characteristics simultaneously fails, then it rarely does so for only one characteristic.

Over the years, the more rules for out-of-control detection have been implemented in analyzing control charts. For example, in addition to the decision rule based on a point falling outside the 3-sigma control limits, the following rules are used for univariate control charts in many industrial settings to alert the operator that the system appears to be operating in an out-of-control state:

7 successive points up or down

2 successive points outside the 2-sigma control limits

4 successive points outside the 1-sigma control limits

The mathematical formulation for extending these rules to the multivariate situation follows.

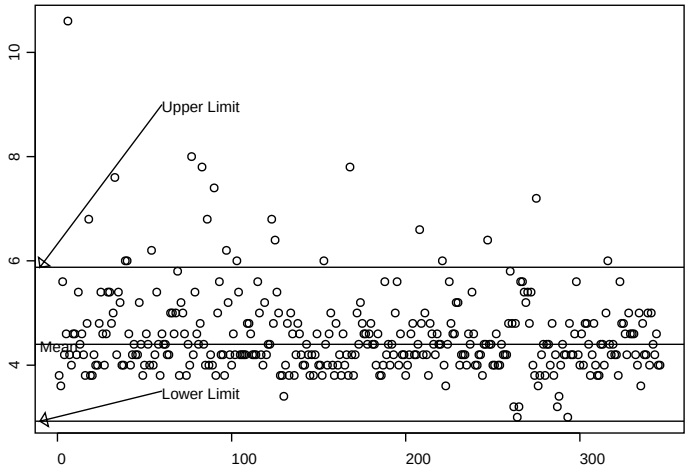

Figure 3.3: Silver

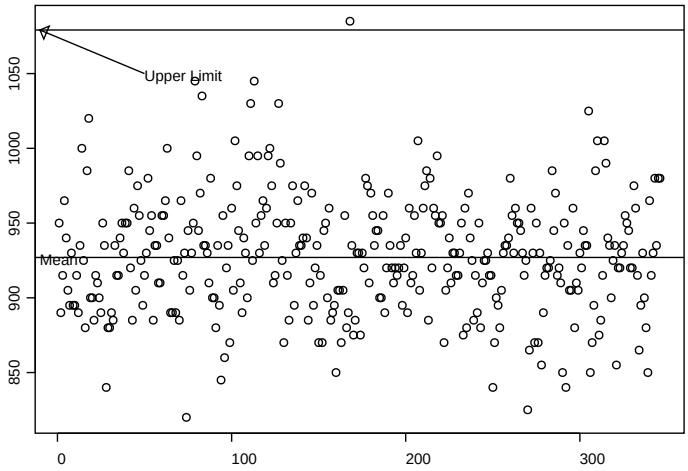

Figure 3.4: Manganese

Let $m, n \in \mathbf{N}$. Let, for $i \in [1,3]$, $(a_1^{(i)}, \ldots, a_n^{(i)}) \in \mathbf{R}^n$ be such that

$$(\forall j \in [1,n])\ 0 < a_j^{(1)} < a_j^{(2)} < a_j^{(3)} \ .$$

For the polyhedra defined by

$$P_i = \{(x_1, \ldots, x_n) : |x_j| \leq a_j^{(i)}\},\ i \in [1,3]$$

we have that

$$P_1 \subseteq P_2 \subseteq P_3 \ .$$

Suppose that a function

$$f : [1,m] \to \mathbf{R}^n$$

is given and let us write

$$f(l) = (x_1^{(l)}, \ldots, x_n^{(l)}) \ .$$

Also, for $l \in [1,m]$

$$S_l = \{j \in [1,n] : |x_j^{(l)}| > a_j^{(2)}\}$$

and

$$T_l = \{j \in [1,n] : |x_j^{(l)}| > a_j^{(1)}\} \ .$$

Note that

$$S_l \subseteq T_l \ .$$

Let k be the maximum of all $j \in [1,m]$ such that

$$(\forall i \in [1,j])\ f(i) \in P_3$$

(this is analogous to the univariate rule of 1 point outside the 3-sigma control limits) and with the following three properties.

1. For all $i \in [7,j]$,

$$(\exists l \in [i-6, i-1]) \min_{p \in [1,n]} (a_p^{(1)} - x_p^{(l)}) \leq \min_{p \in [1,n]} (a_p^{(1)} - x_p^{(l+1)}) \ .$$

(this is analogous to the univariate rule of 7 successive points up or down)

2. For all $i \in [2,j]$,

$$\{f(i), f(i-1)\} \subseteq P_3 \backslash P_2 \ \Rightarrow\ S_i \cap S_{i-1} = \phi \ .$$

(this is analogous to the univariate rule of 2 successive points outside the 2-sigma control limits)

3. For all $i \in [4,j]$,

$$\{f(i-p) : p \in [0,3]\} \subseteq P_3 \backslash P_1 \ \Rightarrow\ \cap_{l=0}^{3} T_{i-l} = \phi \ .$$

(this is analogous to the univariate rule of 4 successive points outside the 1-sigma control limits)

If $k = m$, we are done.

Otherwise, do the adjustment, i.e., form a new function, call it f again, so that, for $i \leq k$, $f(i)$ is the same as before and that $f(k+1) \in P_1$ with

$$\min_{p \in [1,n]} (a_p^{(1)} - x_p^{(k)}) \leq \min_{p \in [1,n]} (a_p^{(1)} - x_p^{(k+1)}) \ .$$

Continue till $k = m$.

Other univariate control rules may be similarly generalized.

3.5 Discussion

The problem posed by Chemex laboratories can be attacked in a variety of ways. In all the models considered, the common approach is to define a model, and discuss its suitability to the Chemex problem. The next step is then to determine a suitable measure on which to base a decision on whether or not to recalibrate. A measure which is often used in multiple-component systems and which has not been touched upon here is 'defects per unit'. This has the nice property of a very natural dimension reduction, but may not be a very powerful technique since it is an attributes (specifically, a count) measure. Once a measure has been chosen, real data may be used to determine whether or not the model actually performs satisfactorily. Model and distributional assumptions, data dependence and forms of data dependence are often factors which are overlooked and may render a tool useless or needing some adjustments. Natural data dependence such as running samples in batches, as is done by Chemex, can often be used to the advantage of the practitioner in specifying a more precise model. Actual data collection may differ from model requirement and therefore, if the study is a retrospective one, it is very important to know exactly how the data being analyzed have been collected. For example, in the Chemex problem, the 'in-control' data may or may not have been representative of the entire process, since operators and times were not known. The 'out-of-control' data may have been a series of points over time or a gathering of points from wherever they were available. These data may have been such that adjustments had been made by operators, if they were collected in real time. If the study requires data to be collected, they should be collected in a way such that a suitable model can be built or used. Since no mathematical model mirrors reality, simulation is a commonly used technique which allows experimenters to compare theoretical confidence levels with actual obtained confidence levels.

Bibliography

[1] Alt, F. B. (1985). Multivariate Quality Control, *Encyclopedia of Statistical Sciences*, **6** edited by N. L. Johnson & S. Kotz, John Wiley and Sons, New York.

[2] Caughlin, B. (Chemex Laboratories) - Dialogue and Correspondence during and following PIMS 1999 Industrial Problem Solving Workshop.

[3] Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*, 3rd ed. John Wiley and Sons, New York, NY.