ANALYSIS AND DETECTION OF OUTLIERS AND SYSTEMATIC ERRORS FROM AN INDUSTRIAL DATA PLANT

R. M. B. Alves *and C. A. O. Nascimento
LSCP- Laboratory of Process Simulation and Control www.lscp.pqi.ep.usp.br
Chemical Engineering Department, Polytechnic School, University of São Paulo
Av. Prof. Luciano Gualberto 380, tr3, cep 05508-900, São Paulo, SP, Brazil.
E-mail: rita@lscp.pqi.ep.usp.br; oller@usp.br

W.L.Carneiro
BRASKEM
Rua Eteno 1561, Polo Petroquímico de Camaçari, cep 42810-000
Camaçari, Bahia, Brazil. E-mail: williane@copene.com.br

Abstract

This article describes the analysis of industrial process data in order to detect outliers and systematic errors. Analysis of data reconciliation is an important step of our work since the data quality directly affects the quality of adjustment of the model for modeling, simulation and optimization of processes. For some cases, outlier points can be easily detected, but for others, it is not so obvious. If the origin of the abnormal values is known, these value are immediately discarded. On the other hand, if an error or an extreme observation is not surely justified, the judgment in discarding or not these values must be based on some kind of statistical analysis. In this work, besides the knowledge of the process, the employed methodology involves an approach based on either statistics or first principle equations or a composition of both. In addition, it is possible to classify similar inputs and outputs in order to identify clusters and then proceed with the elimination of the gross errors by the similarity principle or by hypothesis testing for means. The system studied is the Isoprene Production Unit from BRASKEM, the largest Brazilian petrochemical plant. The analysis of the process was undertaken by using a one-year database. The frequency of the data collection of the monitoring variables was 15 minutes.

Keywords

Gross error, Systematic error, Data analysis, Outliers, Cluster analysis.

Introduction

Multivariate data analysis is not easy to define. Broadly speaking, it refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. Any simultaneous analysis of more than two variables can be loosely considered multivariate analysis. One reason for the difficulty of defining multivariate analysis is that the term multivariate is not used consistently in the literature. (Hair et al., 1998). To be considered truly multivariate all the variables must be random and interrelated in such ways that their different effects can not meaningfully be interpreted separately.

* To whom all correspondence should be addressed

The use of multiple variables and the reliance on their combination in multivariate techniques also focuses attention on a complementary issue – measurement error. Measurement error is the degree to which the observed values are not representative of the "true" values. Measurement error has many sources, ranging from data entry errors to the imprecision of the measurement to the inability of respondents to accurately provide information. Thus all variables used in multivariate techniques must be assumed to have some degree of meaurement error. Statistical analysis provides the methods for stating the degree of precision of our measurements, when those measurements represent an estimate of the "true" but unknown value of a characteristic (Kachigan, 1991). The impact of measurement error is to add "noise" to the observed or measured variables. Thus, the observed value obtained represents both the "true" level and the "noise". When used to compute correlations or means, the "true" effect is partially masked by the measurement error, causing the correlations to weaken and the means to be less precise. The impact of measurement error and poor reliability can not be directly seen because they are embedded in the observed variables. The researcher must therefore always work to increase reliability and validity, which in turn will result in a "truer" portrayal of the variables of interest. Poor results are not always due to measurement error, but the presence of measurement error is guaranteed to distort the observed relationships and make multivariate techniques less powerful.

Multivariate data analyses require a rigorous examination of the data because the influence of outliers, violations or assumptions, and missing data can be compounded across several variables to have quite substantial effects.

Outliers and Systematic Errors

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations. Outliers can not be categorically characterized as either beneficial or problematic, but instead must be viewed within the context of the analysis and should be evaluated by the types of information they may provide. When benefical, outliers – although different from the majority of the sample – may be indicative of characteristics of the population that would not be discovered in the normal course of analysis. In contrast, problematic outliers, not representative of the population, are counter to the objectives of the analysis, and can seriously distort statistical tests (Hair et al., 1998).

Gross errors or anomalous measurements of the data set may arise due to changed conditions during plant operation, or due to errors with the operation of measurements and recording devices, or simply due to errors in the information register, which may contaminate the valid data. On the other hand, the outlier may be simply one of the extreme values in a probability distribution for a random variable, which occurs quite naturally but not frequently and should not be rejected (Alves and Nascimento, 2001). The researcher must decide whether the extraordinary event should be represented in the sample. If so, the outlier should be retained in the analysis; if not, it should be deleted. Another class of outlier contains observations that fall within the ordinary range of values on each of the variables but are unique in their combination of values across the variables. In these situations, the researcher should retain the observation unless specific evidence is available that discounts the outlier as a valid member of the population.

If the researcher knows the origin of the abnormal values, he does not hesitate to discard such an observation. On the other hand, when he is not sure about the error or he does not have enough practice to either accept or reject an extreme observation, he must base his judgment on some kind of statistical analysis. The question to be analyzed is how probable it is that the observed differences are due solely to random sampling errors in order to reject or not the information. This task becomes especially complicated for complex processes where not all of the influencing parameters are directly accessible or where large stochastic deviations of the process variables lead to a considerable scattering of the measured data (Alves and Nascimento, 2001). For this reason, a large variety of approaches were proposed in the past, which tackle this problem. These are commonly based on either statistics or first principle equations or a composition of both. Sometimes, this procedure may become extremely complicated both if the underlying physics and chemistry of the process are not very well understood and if the application of a sharp statistical criterion for the separation of the data into one set of valid and another of non-valid values is impossible. This article, besides these techniques above, classifies similar inputs and outputs in order to identify clusters and then proceed with the elimination of the gross errors. Moreover, based on a normal distribution of variables it was possible to correct some wrong values due to fail on measurement instruments by comparison with laboratory data analysis.

Methodology

The available monitoring variables from the industrial process studied (Isoprene Production Unit) were collected every 15 minutes. According to the average time considered for the data treatment, data fluctuation could be incorporated in the results. Many times, this could lead to unreliable information. In cases of errors with the measurement instruments over a long period of time, the average reflects this error. The higher frequency of data collected allowed to identify periods of steady state operation and possible errors of measurement instruments. The analysis of the process was undertaken by using a one-year database. The primary database consisting of about 34500 observations of 244 variables

The treatment of the data was performed at the folowing steps:

1. Selection of variables of interest
2. Gross Errors detection
3. Establishement of Steady State Operation
4. Systematic Erros detection

Selection of Variables of Interest

The variables of interest were defined by considering the available process data and their importance for the process and plant operation. Then, the minimum, maximum and mean values were identified as well the variance for each selected variable. The variables whose operational range were too close to the instrument's limits (e.g. wind-up measurements) were not included for analysis.

Gross Error Detection

At this step were evaluated and eliminated the following data: null and negative values, values with different magnitudes, possible flat lines as those at the instrument's limits (wind-up measurement) and abrupt changes of the variables along the time line. This analysis was carried out through observations of the variables as a time function, by verification of their minimum, maximum and mean values. The knowledge of the process also allowed the elimination of some points based on possible process values or acceptable operational range for the corresponding variable.

Another tool used to detect outliers was cluster analysis. Cluster analysis is an analytical technique for developing meaningful subgroups of individuals or objects. It is based on the similarity principle among several data sets. For this work, a data set was formed by the input and output variables chosen for each process unit, corresponding to information from one operation register. It is expected that for a series of similar input variables, the process must yield similar output variables (dependent variables). When a different input or output variable is observed among a series of similar data, the corresponding data set may be rejected..

Establishement of Steady State Operation

The higher frequency of data collected allowed to identify periods of steady state operation. The criterion adopted was a constant feed flow for a period of two or three days. A data flutuaction of 0.2-0.3 t/h was acceptable.

Systematic Errors Detection

At this step, first principles procedures were used in order to detect systematic errors. Knowledge of the process is also important at this step in order to evaluate these kind of error. Once carried out global and components material balances was possible to identify some distortion in the final results. At this point it was very informative to make a graphical representation of a frequency distribution.

Knowing that the distributions were normal in form, we could further interpret the values in terms of what percent of the total number of observations fall below or above the given value. Although real-life data distributions, due to their finite size, can never be perfectly normal in form, the approximation is often close enough to allow us to use the theoretical normal distribution as a model for interpreting empirical populations of data.

It is also well known that mathematic operations can help in adjusting data, i.e., the addition or subtraction of a constant value from a set of observation affects the mean but not the variation of the data; whereas the multiplication or division by a constant affects both the mean and variation of the original distribution (Kachigan, 1991). These fundamental relationships will be very useful in developing and understanding subsequent statistical concepts.

Thus, to identify the relative location of an observed value in a data distribution, besides the knowledge of the arithmetic average, i.e, the mean, it is necessary to know not only its deviation from the mean, but that deviation must be translated into standard deviations.

Based on these concepts above, we were able to correct systematic errors instead of deleting them by shifting the mean. Another criterion used at this step was the comparison between the plant data and the more reliable laboratory data analysis. This procedure allowed identification of possible errors in the measurement instruments at certain periods of plant operation and correction of the wrong values.

Results and Discussion

Detection of outliers or gross errors was not difficult to achieve mainly because the data treated were collected every 15 minutes as seen in Figures 1and 2. These figures show, respectively, data before and after elimination of ouliers as described above.

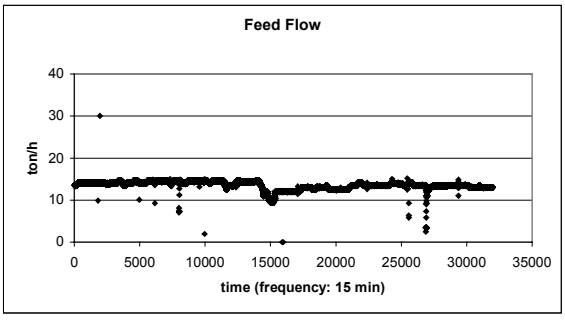

Figure 1. Data before elimination of outliers

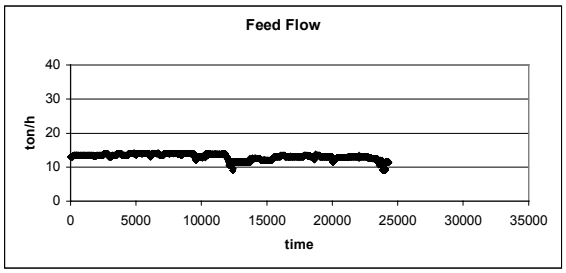

Figure 2. Data after elimination of outliers

In the case of systematic errors the task was more complicated and time-consuming, because first it was necessary to carry out global and components material balances for each process unit individually, then to establish periods of steady-state operation and after this to build histograms for each one of the balances and for each these periods, to calculate means and standard deviations. By analyzing these plots and results, we were able to detect systematic errors and delete or correct them. One way to correct them was adding or subtracting the variable by the mean value, which, as shown above, did not affect the shape of the curve nor the variation of the data. Figure 3 shows the histogram before analysis for the global material balance. In this figure we can see the mean equal to 0.55, which signifies that once the distribution is normal, the only problem was the shift of the mean and based on the fact that the global material balance must be equal to zero we were able to correct it. Then by analyzing the total and components balance we had two options: to decrease the feed flow rate or increase the output flow rates. Once the input flow rates show higher values, our decision was for the first option., i.e., decrease the input flow rate by the mean value. The final histogram has the same shape as the figure 3, the only difference is the mean equal to zero.

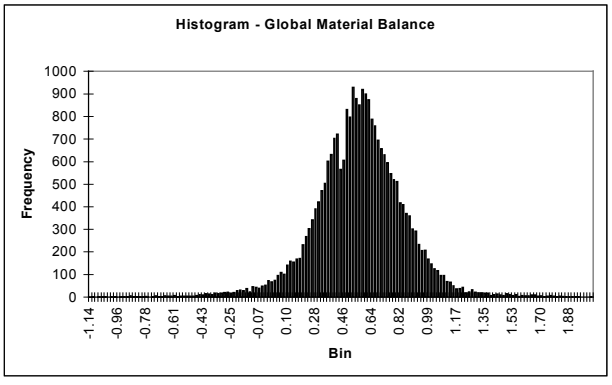

Figure 3. Histogram- Data before adjusting

Figure 4 shows data after elimination of ouliers and systematic errors.

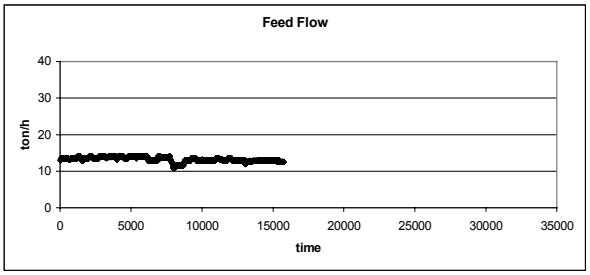

Figure 4 – Plot of feed flow after elimination of outliers and systematic errors

Another way to correct these errors was by observing the results of the histograms for the material balances for each component of interest and comparing the plant data analysis with the laboratory data analysis. For these we divided the data into range of steady state operation and then by verifying if the problem was in the input or output analysis, we tried to correct them by supposing that the flow rate was corrected. In some cases, when both are corrected, it was necessary to re-correct the flow rate again. In these case the correction was carried out by multiplying or dividing the data by a factor of correction.

Conclusions

Analysis of data reconciliation is an important step of our work since the quality of data affects directly the quality of adjust of data to modeling, simulation and optimization of processes, thus reducing measurement error, although it takes effort, time, and additional resources, may improve weak or marginal results and strengthen proven results as well.

Acknowledges

The authors wish to tanhk FAPESP for its financial support and BRASKEM to provide the industrial data used.

References

Alves, R.M.B. and Cláudio Nascimento (2001). Gross Errors Detection of Industrial Data by Neural Network and Cluster Techniques. Proceedings of the ENPROMER 2001, Santa Fé-Argentina.

Hair Jr., J.F., R.E. Anderson, R.L.Tatham and W.C. Black (1998). Multivariate Data Analysis, 5th ed. Prentice Hall,Inc., Upper Saddle River, New Jersey.

Kachigan, S.K. (1991). Multivariate Statistical Analysis-A Conceptual Introduction, 2nd ed., Radius Press, New York.