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

Rita M. B. Alves* and Claudio A. O. Nascimento

CESQ- Center of Engineering of Chemical Systems
Department of Chemical Engineering - 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

ABSTRACT

This article describes the analysis of industrial process data in order to detect outliers and systematic errors. Data reconciliation is an important step of the work in adjusting mathematical model from plant data since the quality of the data affects directly the quality of adjustment of the model for modeling, simulation and optimization purposes. To detect outliers on a multivariable system is not an easy task. 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 was used a neural network based approach to represent the process in order to make 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.

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.

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

* To whom all correspondence should be addressed

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.

In addition, 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.

For these reasons, 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. Then, these article demonstrate the aplicability of different techniques in data analyses, such as statistical models, first principal equations, neural network based approache and clusters analysis.

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, 2002). 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, 2002). 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 work, besides these techniques above, uses a neural network based approach which makes possible to classify similar inputs and outputs in order to identify clusters and then proceed with the elimination of the gross errors. This approach is a simple and easy way to detect outliers and requires much less knowledge of the underlying physical-chemical process.

For detection of systematic errors, first principles and stastical procedures were used and 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 methodology applied in this work follows the structure shown in Figure 1.

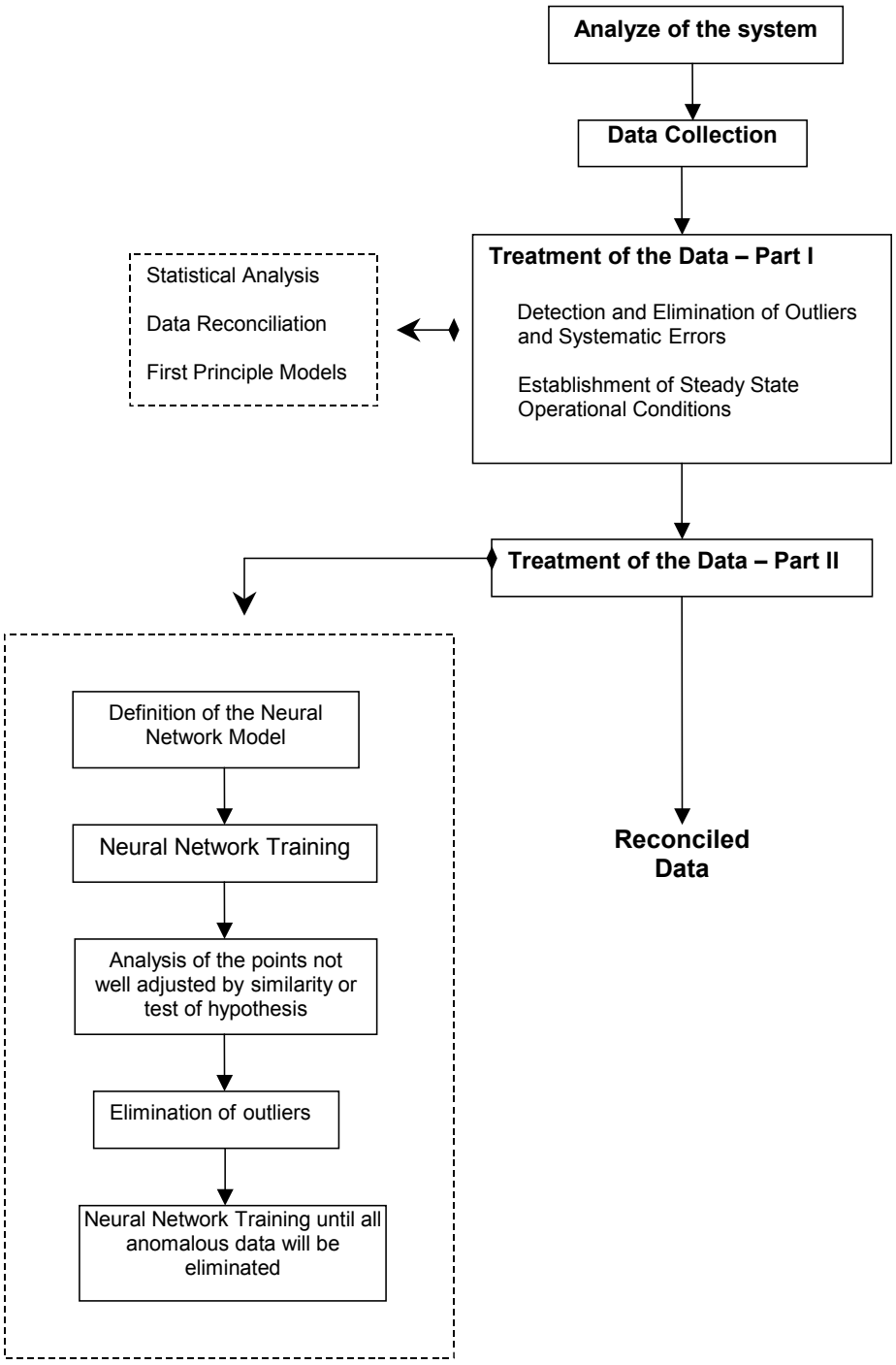


Fig.1. Data analysis methodology

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

- Selection of variables of interest

- Gross Errors detection
- Establishement of Steady State Operation
- 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, first, a preliminary analysis for abnormal values , which may subject to rejection, was carried out. Then, 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 graphical observations of the variables as a time function, experience whether statistical features as much az the process, material and energy balances. 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.

After this initial analysis, the methodology employed involves the construction of a reliable neural network model to represent the process and its training with few iterations. All the resulting data set from the previous step were included in the training data set. The construction and training of the NN, used in this work, were carried out using an in-house software developed by Nascimento, 1991.

A simple measure to assess the quality of fitness of the chosen neural network to the experimental data is usually a comparison of the calculated values by the neural network with the original experimental data. The scatter of data points around the ideal 45^{o} –line can be used to judge the fit of the neural network to the experimental data. The idea to use neural networks for the purpose of outlier detection is based on this kind of diagram (Büllau et al., 1999). Then the points in wich errors between the experimental and calculated data appear to be scattered far away from the majority of values are indication of consisntency problems or probably outliers not identified in the preliminary analysis. Hence, it only has to be shown that the probable outliers of the experimental data correspond to the outliers from this curve. Thus, the neural network was first of all trained for the entire data set and afterwards for the filtered data set. To decide if these points must or must not be eliminated were used some statistical analyses as clusters analysis and hypothesis tests for means. This procedure was repeated several times until the scattered data did not show abnormal points. Since the training of the network with the filtered data leads to different results from calculated data in comparison to the original data set, the input data base changes due to the filtration procedure. Figures 2 a-c show the results of this methodology for the first, second and final runs.

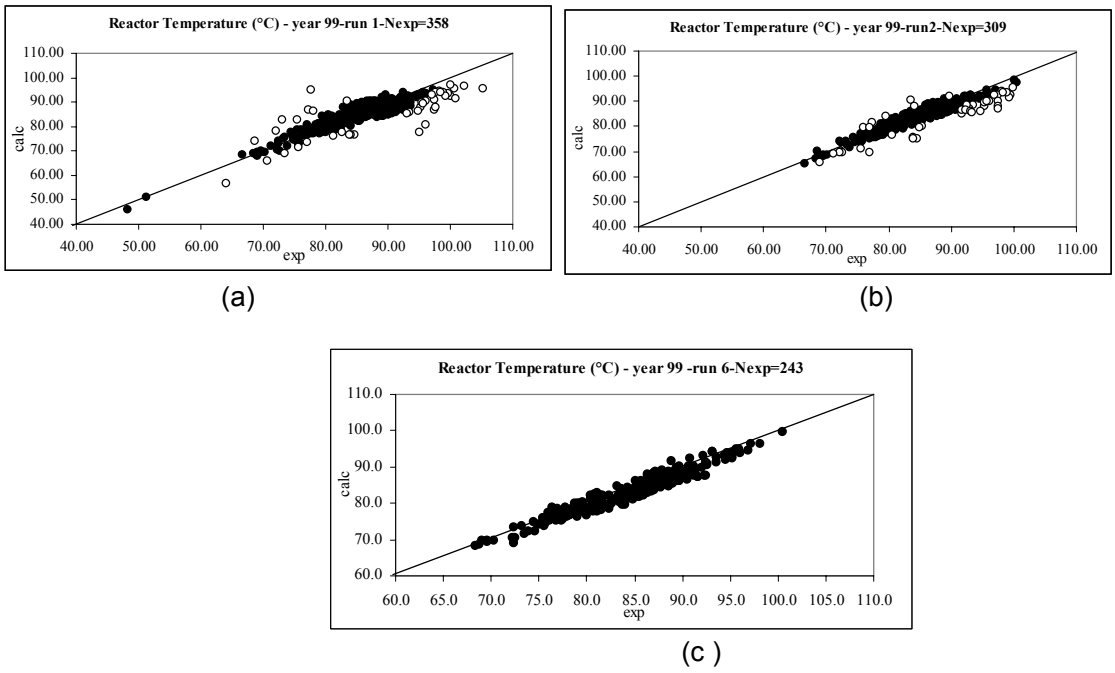


Fig. 2. Comparison of calculated and measured data: (a) before analysis; (b) intermediate results; (c) final result

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. Table 1 shows two exmples of cluster analysis: it can be observed, for the variable out2, that the values 25.24 and 20.85, in the first and second group of data respectively, must be rejected.

In some cases a simple and direct analysis is not possible, e.g. when a given data set is unique or when there are only two data sets for comparison with some distinct information, then it is not possible to determine which one is correct. In these cases, the domains of the variables are extended when compared with the previous group. Although these new groups are less accurate, usually it is possible to discriminate the abnormal point. For this step, the hypothesis test for means was employed, which involves a confidence interval estimate and a hypothesis test, with 95% as the confidence level (Himmelblau, 1970).

Table 2 shows the application of this methodology on the data plant analysis. The values in bold in groups 1 and 2 were not well adjusted during neural network training and it was not possible identify groups of similar data set for the cluster analysis, then, a hypothesis test for means analysis was performed. Its was observed that the value 25.81 is inside the interval of confidence and the null hypothesis is accepted and this data must not be eliminated. On the other hand, the value 1.62 is outside the confidence interval and the null hypothesis is rejected and this data set must be eliminated.

From this way, detection of outliers or gross errors was not difficul to achieve mainly because the data treated were collected every 15 minutes. Figures 3 e 4 show, respectively, data before and after elimination of outliers as described above.

Table 1. Cluster Analysis – Examples

Input Variables								Output Variables			
in1	**in2**	**in3**	**in4**	**in5**	**in6**	**in7**	**in8**	**out1**	**out2**	**out3**	**out4**
12.79	15.34	1.86	63.80	39.30	59.90	8.89	4.37	8.35	**25.24**	2.50	97.4
12.80	15.40	1.88	63.80	39.20	59.90	8.87	4.35	8.35	27.21	2.63	97.3
12.80	15.25	1.79	63.80	39.20	59.90	8.85	4.38	8.35	26.88	2.53	97.4
12.75	15.07	1.71	63.80	39.10	59.80	8.79	4.36	8.35	27.21	2.49	97.3
13.99	16.03	2.40	64.50	55.10	59.80	11.19	4.83	9.08	**20.85**	2.68	102.6
13.82	16.15	2.93	64.30	55.50	59.30	10.79	5.06	9.16	28.84	2.89	97.2
13.80	16.09	2.91	64.30	55.10	59.40	10.81	4.97	9.23	28.57	2.85	97.8
13.70	16.02	2.74	64.20	55.60	58.60	11.13	4.91	9.07	30.50	2.66	95.0
13.70	15.86	2.54	64.20	55.80	58.60	11.15	5.07	8.92	30.92	2.54	94.6

Table 2. Hypothesis Test for Means

Group	Input Variables								Output Variables			
	in1	**in2**	**in3**	**in4**	**in5**	**in6**	**in7**	**in8**	**out1**	**out2**	**out3**	**out4**
1	13.00	14.50	1.95	66.00	51.10	59.90	10.51	5.55	7.62	26.73	**1.62**	93.5
	13.00	14.57	2.77	65.90	50.60	60.00	10.25	5.37	7.72	26.15	2.68	91.6
2	13.20	16.20	2.43	64.30	50.40	59.70	9.94	4.41	8.68	**25.81**	3.17	100.1
Sample	12.00	13.58	2.38	66.00	50.20	61.20	10.40	3.77	8.78	19.91	2.77	97.2
	12.00	14.74	2.14	65.90	48.40	61.30	10.08	3.77	8.89	20.77	2.70	98.5
	12.09	14.13	1.75	66.50	48.60	61.00	10.23	4.04	8.58	23.51	2.07	97.5
	13.30	14.99	2.46	66.80	48.80	60.90	10.32	5.15	8.73	24.35	2.70	94.8
	13.48	15.74	2.37	66.80	49.60	60.70	10.47	5.35	8.76	24.60	2.47	93.8
	13.50	15.64	2.23	66.80	50.30	60.80	10.57	5.48	8.72	25.22	2.30	93.8
	13.51	15.02	2.01	66.70	50.50	60.80	10.56	5.51	8.76	25.23	2.26	94.0
	13.50	15.87	1.96	66.40	49.10	60.60	10.34	5.49	8.79	25.91	2.17	94.5
	13.00	14.50	1.95	66.00	51.10	59.90	10.51	5.55	7.62	26.73	**1.62**	93.5
	12.98	13.52	2.22	65.70	51.60	60.00	10.65	5.38	7.74	26.09	2.57	93.1
	13.00	14.57	2.77	65.90	50.60	60.00	10.25	5.37	7.72	26.15	2.68	91.6
	13.20	16.20	2.43	64.30	50.40	59.70	9.94	4.41	8.68	**25.81**	3.17	100.1
	13.00	14.06	2.05	64.30	51.20	59.70	10.27	4.28	8.65	20.93	2.66	100.5
	13.38	14.13	1.98	64.30	50.10	59.40	10.38	4.41	8.84	20.98	2.52	100.8
	14.01	13.99	2.60	64.50	52.30	59.60	10.97	4.40	9.45	17.98	2.94	100.9
	14.00	14.24	2.64	64.40	52.30	59.50	11.00	4.61	9.30	18.49	2.89	99.6
	14.00	14.11	2.57	64.40	52.30	59.50	11.00	4.62	9.30	19.09	2.89	99.9
	14.01	15.67	2.59	64.30	50.70	59.70	10.81	4.56	9.41	19.65	2.82	99.8
Minimum	12.00	13.52	1.75	64.30	48.40	59.40	9.94	3.77	7.62	17.98	1.62	91.60
Maximum	14.01	16.20	2.77	66.80	52.30	61.30	11.00	5.55	9.45	26.73	3.17	100.90
mean	13.22	14.71	2.28	65.56	50.45	60.24	10.49	4.79	8.71	22.86	2.57	96.88
Std dev	0.65	0.82	0.29	1.04	1.23	0.66	0.31	0.63	0.54	3.06	0.37	3.20
t/n-1=18 2.11												
Mean+std dev*t	14.60	16.44	2.90	67.75	53.05	61.63	11.13	6.10	9.84	29.31	3.35	103.63
Mean-std dev*t	11.84	12.97	1.67	63.37	47.85	58.85	9.84	3.47	7.57	16.41	1.78	90.13

Fig. 3. Data before elimination of outliers

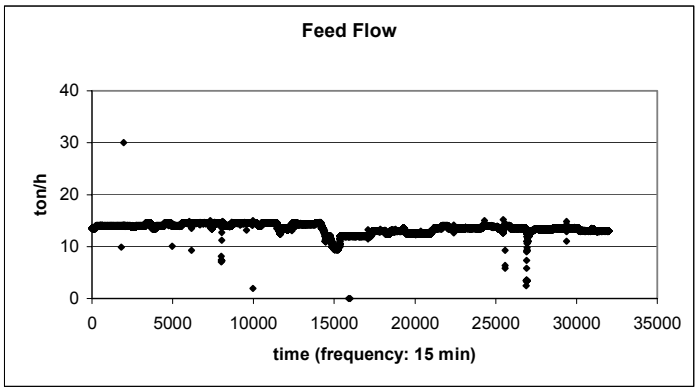
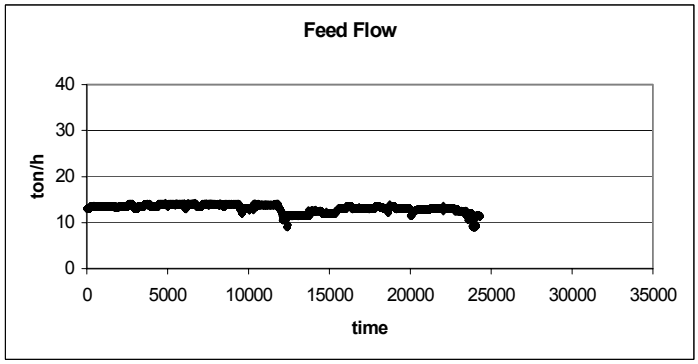


Fig. 4. Data after elimination of outliers



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 ton/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.

The task of detecting systematic errors is was more complicated and time-consuming than the detecting outliers, 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 5 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 5, the only difference is the mean equal to zero.

Fig. 5. Histogram- Data before adjusting

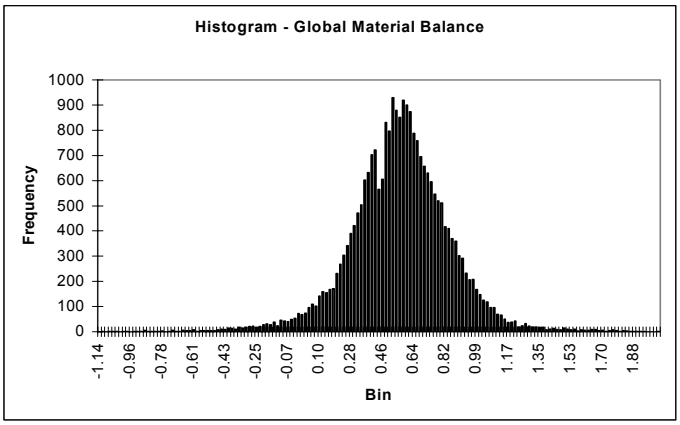
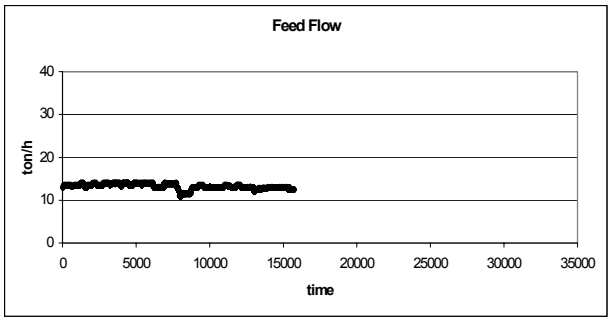


Figure 6 shows data after elimination of ouliers and systematic errors.

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.

Fig. 6. Feed flow after elimination of outliers and systematic errors



CONCLUSIONS

Analysis of data reconciliation is an important step of the work in adjusting mathematical model from plant data 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. Moreover, the application of a neural network approach is a very attractive tool to detect outliers.

Acknowledges

The authors wish to thank FAPESP for its financial support and BRASKEM for providing the industrial data used in this work.

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.

Alves, R.M.B. and Nascimento, C.A.O. (2002). Gross Errors Detection of Industrial Data by Neural Network and Cluster Techniques. Braz. J. Chem. Eng., vol.19, no.4, p.483-489.

Bülau, H. C., Ulrich, J., Guardani, R., Nascimento, C. A. O., "Application of Neural Networks to Data from a Melt Crystalization Process for the Detection of Outliers". Em Proceedings of AIDA, International Seminar on Advances in Data Analysis, 1999 – Washington, DC

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.

Himmelbalu, D. M. (1970). Process Analysis by Statistical Methods. John Wiley & Sons, Inc.

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

Nascimento, C.A.O. (1991). *Programa de Redes Neurais*, Neuro14, v.14, Departamento de Engenharia Química da Escola Politécnica da Universidade de São Paulo - Brazil.