

# DATA ANALYSIS TECHNIQUES IN SUGAR FACTORY SURVEYS

<sup>1</sup>K JHUNDOO AND <sup>2</sup>M GOBLET

<sup>1</sup>Mauritius Sugar Industry Research Institute, Réduit, Mauritius

<sup>2</sup>Constance & La Gaieté Sugar Factory, Mauritius

## Abstract

A method of analysis for the determination of the characteristics of a system from real time data is described. Examples based on surveys carried out in the Mauritian sugar factories over the last three years are given to illustrate the techniques involved. An application is described where this method has been used to measure the effectiveness of the modifications carried out over two crop seasons at Constance & La Gaieté sugar factory in order to improve the steam output of its two boilers.

Keywords: Data, analysis, logging, on-line, survey

## Introduction

A survey is conducted when the technical condition of a system is required. This happens whenever engineering problems are anticipated, prior to changes in the manufacturing process, or modifications of existing equipment. The characteristics of the existing system are first determined before setting the specifications of the proposed system. Many parameters are usually monitored on-line for a reasonably long period to ensure that sufficient data are obtained at different operating conditions. Once the modifications are done, the investigation is carried out again to verify whether the target specifications have been achieved.

In the past, it was very common to look at the average values. However, this approach is very limited when dealing with a system of numerous parameters, as it does not give a true picture of the steady state condition of the system. Factory stoppages or frequent abnormal conditions affect the results if not taken into account. In addition, the operating range or ranges of the system, and the interaction of the different parameters are difficult to obtain. By converting time plots of the different parameters to a data frequency distribution plot, a global view of the system's behaviour is obtained.

## Data analysis techniques

### Data collection

The output signals from sensors installed in the factory are fed to the different channels of a data logging system. A data logging unit, having 24 channels with a scan rate of 5 channels every second, has been used in conjunction with a custom made programme written in 'C'. Some of the monitored variables are obtained directly, while others have to be calculated from the channel values. For example, pressure is readily obtained from the output signal of a transmitter, whereas, if an orifice plate is used to monitor the flow of steam, its pressure and temperature, and the differential pressure across the plate would be required to calculate the flowrate.

As soon as all the selected channels are scanned, the data values, obtained either directly or indirectly from the channel values, are stored in a buffer. After a user-selectable time interval, consisting of at least one scan period, averages of the

data values are transferred from the buffer to a file together with the time and the date. Each line in the data file represents a set of all the average data values at a particular time, and each column represents a set of average values of a particular variable at all successive time intervals. A typical example is shown in Table 1.

Table 1  
Typical on-line data collected in a survey

Date	Time	Boiler steam output (t/h)	Rotational speed of ID fan (rpm)	Suction pressure of ID fan (Pa)	ID fan shaft power (kW)
18/9/95	15:32:05	70,3	648	2650	227,6
18/9/95	15:33:05	68,0	651	2667	232,8
18/9/95	15:34:05	67,9	611	2310	208,5
18/9/95	15:35:05	66,8	594	2180	197,8

### Relationship between variables in the time domain

From the collected data, time graphs can be plotted as shown in Figure 1, corresponding to the data given in Table 1.

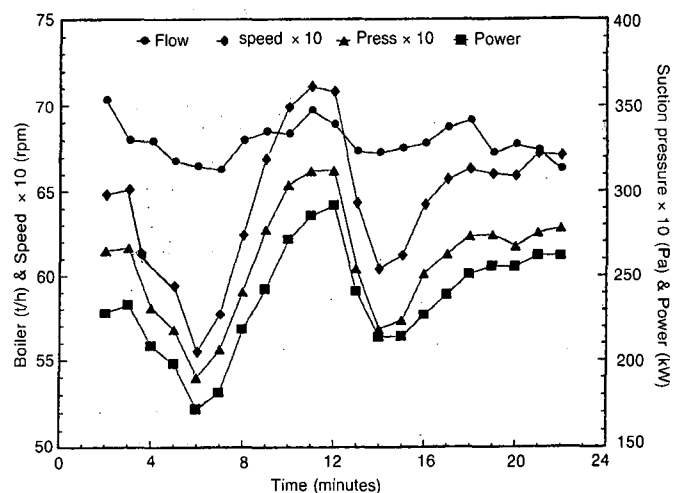


FIGURE 1: Time graphs of raw data

It is very difficult to extract relevant information from time graphs. Their clarity rapidly deteriorates with the amount of data plotted, and time lag within the system, added to the poor response time of the sensors, make it very difficult to establish relationships between different variables.

Given that variable  $y$  is related to variable  $x$  by a function  $f$ , then

$$y = f(x)$$

or,

$$y(t) = f(x(t)) \quad (1)$$

However, due to time lag, a change in the variable  $x$ , at a time  $t$ , would cause a corresponding change in the variable  $y$  at a time  $t+\tau$  where  $\tau$  represents the total lag time which is relatively small. Therefore

$$y(t+\tau) = f(x(t)) \tag{2}$$

Assuming  $y(t)$  is an estimate for  $y(t+\tau)$ , then

$$y(t) = y(t+\tau) + e(t) \tag{3}$$

where  $e(t)$  is the error in the estimate.

Hence, from equations 2 and 3,

$$y(t) = f(x(t)) + e(t) \tag{4}$$

From Equation 4, it is seen that corresponding values  $(x(t), y(t))$  cannot be easily related to each other because of the unknown error,  $e(t)$ .

*Establishing a time independent relationship between variables*

For a large amount of data, the expectation of  $e(t)$  can be reasonably assumed to be zero for all times. Then, the expectation of (4) gives,

$$E[y(t)] = E[f(x(t))] \tag{5}$$

that is,

$$\bar{y} = f(\bar{x}) \tag{6}$$

The time lag problem has been eliminated in the following manner. A large set of random values  $\{x_1, x_2, \dots, x_n\}$  of the variable  $X$ , observed to be bounded within a small selected interval  $(x_a, x_b)$ , at random times  $t_1, t_2, \dots, t_n$  will correspond to a set of random values  $\{y_1, y_2, \dots, y_n\}$  of a variable  $Y$  whose mean value from Equation (6) is given by,

$$\bar{Y} = [f(x_1) + f(x_2) + \dots + f(x_n)] / n \tag{7}$$

For  $i = \{1, 2, \dots, n\}$ ,  $x_i$  can be expressed as,

$$x_i = \bar{X} + \sigma_i \tag{8}$$

where  $\bar{X}$  is the mean of  $x$  over the small interval  $(x_a, x_b)$  and  $\sigma_i$  is a small deviation of  $x_i$  from the mean. Equation (7) can then be expressed as

$$\bar{Y} = [f(\bar{X} + \sigma_1) + f(\bar{X} + \sigma_2) + \dots + f(\bar{X} + \sigma_n)] / n$$

which can be approximated as

$$\bar{Y} \approx [f(\bar{X}) + \sigma_1 f'(\bar{X}) + \sigma_2 f'(\bar{X}) + \dots + \sigma_n f'(\bar{X})] / n \tag{9}$$

Simplifying,

$$\bar{Y} \approx f(\bar{X}) + f'(\bar{X})(\sigma_1 + \sigma_2 + \dots + \sigma_n) / n$$

As  $n$  gets large, the term  $(\sigma_1 + \sigma_2 + \dots + \sigma_n)/n$  becomes negligible, and

$$\bar{Y} \approx f(\bar{X}) \tag{10}$$

where,

$$\bar{X} = (x_1 + x_2 + \dots + x_n)/n \tag{11}$$

and,

$$\bar{Y} = (y_1 + y_2 + \dots + y_n)/n \tag{12}$$

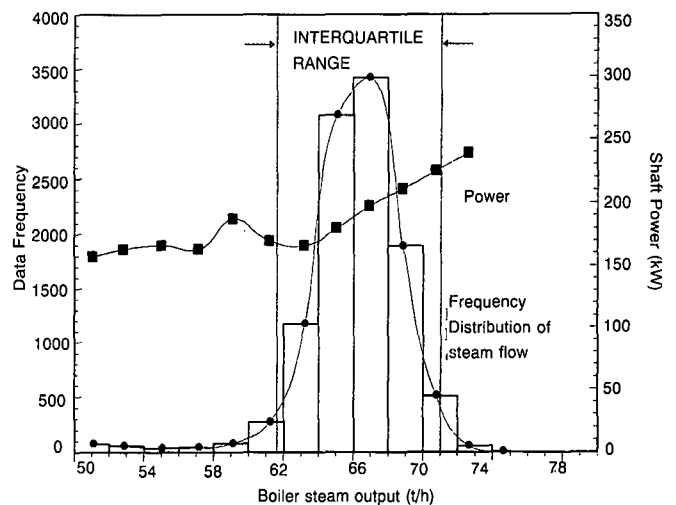
A computer program has been written to calculate the interval  $(x_a, x_b)$ , which takes the general form  $(i.s, i.s+bw)$  for  $i = \{0, 1, 2, \dots, N\}$  where  $s \leq bw$  and  $N, s$ , and  $bw$  are user selectable values. The program then scans through the data file in steps of  $s$ , and for each interval  $(x_a, x_b)$ , the data frequency  $n$  is recorded, and the values  $\bar{X}$  and  $\bar{Y}$  are calculated from Equations (11) and (12) respectively. For example, raw

data, such as those found in Table 1, are processed and then stored in an ASCII file format as described in Table 2.

**Table 2**  
Processed data

$(x_a, x_b)$ Boiler steam flow range (t/h)	$\bar{X}$ Mean flow (t/h)	$\bar{Y}_1$ ID fan speed (rpm)	$\bar{Y}_2$ ID fan suction pressure (Pa)	$\bar{Y}_3$ ID fan shaft Power (kW)	$n$ Data frequency
(50, 52)	51,1	568	1828	157,1	76
(52, 54)	52,8	566	1905	163,0	56
(54, 56)	55,0	576	1975	165,8	34
(56, 58)	57,1	564	1860	162,9	43
(58, 60)	59,1	587	2103	187,0	80
(60, 62)	61,2	563	1903	169,7	270
(62, 64)	63,2	558	1870	166,0	1174
(64, 66)	65,1	575	2015	179,7	3076
(66, 68)	67,0	600	2203	197,3	3420
(68, 70)	68,9	619	2350	210,5	1892
(70, 72)	70,8	640	2508	225,1	509
(72, 74)	72,6	659	2655	239,1	58
(74, 76)	74,6	676	2790	253,2	4

In this example,  $x$  represents the boiler steam output, and for  $s=bw=2$ , the range  $(x_a, x_b)$  has been calculated for  $i = \{25, 26, \dots, 37\}$ , and for each range the computer program has calculated the values  $X$  and  $Y$ , where the latter can, for example, represent the expected value of the ID fan shaft power. The coordinates  $(\bar{X}, \bar{Y})$  are then plotted together with the frequency distribution of  $x$  as shown in Figure 2.



**FIGURE 2:** Distribution of  $\bar{X}$ , superimposed on graph relating the shaft power of the ID fan to the boiler steam output.

When it is required to establish a relationship between  $x$  and  $y$ , the distribution curve enables the user to select visually those points that lie within the interquartile range where the amount of data is relatively high. In this example, a power regression has given  $\bar{Y} = 3\bar{X}^{2.64}$  with an  $R^2$  value of 0,99 where  $\bar{Y}$  is the shaft power of the fan in watts and  $\bar{X}$  is the boiler steam output in t/h. From equations (1) and (10), it can be inferred that  $y = 3x^{2.64}$  is a valid relationship that can be used to predict the value of the ID fan shaft power when the boiler steam output is known.

Several other variables such as the speed of the ID fan and its suction pressure can also be plotted on the same graph, as shown in Figure 3.

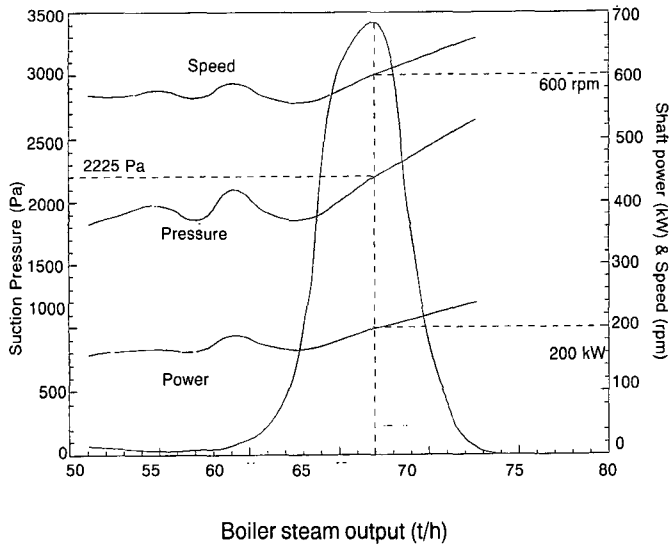


FIGURE 3: Distribution graph with several variables

The operating parameters of the system are thus obtained, and in this case, the boiler operates in the range 58 – 74 t/h, with a most probable steam output of 67 t/h corresponding to an ID fan speed of 600 rpm, a suction pressure 2225 Pa, and a shaft power of 200 kW. The operating ranges of the speed, the suction pressure, and the power are 555 – 650 rpm, 1850 – 2650 Pa, and 160 – 240 kW respectively.

Technical information on the relationships between variables and the operating characteristics of a system can be used by the engineer either to improve the existing system or to design a new one.

Multimodal systems

Some systems having many operating ranges are very difficult to analyse from time graphs. For example, Figure 4 shows the interaction between the juice flowrate and the torque produced by the juice pump when an operator manually controls the flowrate by setting a butterfly valve at three discrete openings.

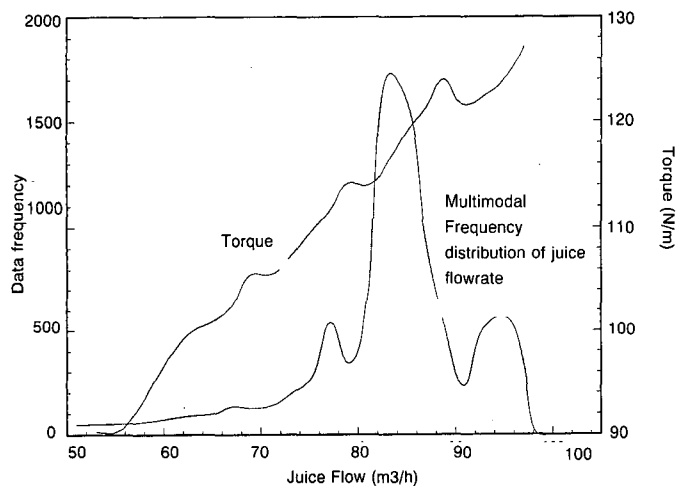


FIGURE 4: Manual juice flow control system

Information, such as the flowrate at the different valve settings, the variation of the torque with respect to the flow, the

extent of variations in the flow at the different valve settings, and the amount of time the valve is set at a particular opening, are easily obtained.

Reference variable

A system can be analysed from different perspectives, i.e. X can represent any variable. For instance, in the above example, X can represent the pump’s motor power as shown in Figure 5.

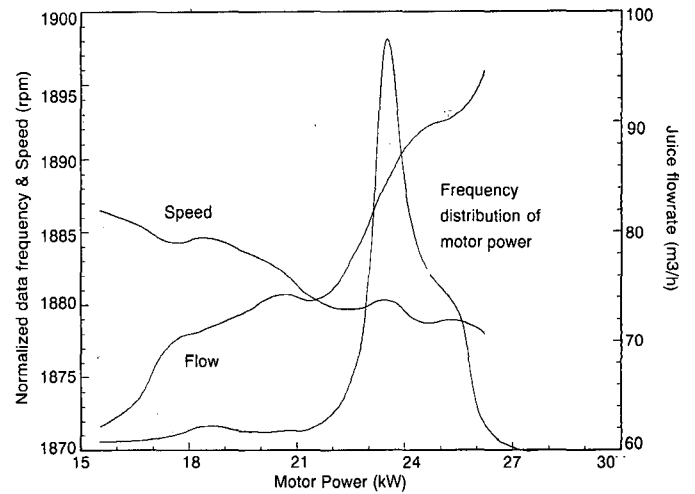


FIGURE 5: Juice flow control system from the pump’s motor power perspective

Matching the y axis scale of the distribution with that of a variable

Sometimes it is not possible to use the same scale to represent the frequency of the distribution as well as the y value of a variable. Since only the shape of the distribution is of interest, the different scales can be matched as follows.

Given that n is the frequency of the distribution curve and n<sub>max</sub> is its maximum value, Y<sub>max</sub> and Y<sub>min</sub> are the maximum and the minimum y axis scale values of a variable y, and  $\bar{n}$  is the matched or normalised value, then

$$\bar{n} = n ( k Y_{max} - Y_{min} ) / n_{max} + Y_{min} \tag{13}$$

where

$$( Y_{max} / Y_{min} ) < k \leq 1$$

Equation (13) has been used to match the frequency scale of the motor power distribution with that of the motor speed as shown in Figure 5.

Application

Constance sugar factory has two boilers, namely boiler 1 and boiler 2, rated at 17 and 45 t/h respectively. Together, they can potentially produce 62 tons of steam per hour. However, because of the interaction between the two boilers, which are connected at one end of the steam circuit main line, a total of only 55 t/h was produced in 1993. This amount was just sufficient for running the factory. A survey was carried out to evaluate the existing system. It was found that the boilers were operating at 19 and 36 t/h respectively as shown by the distributions (shown dotted) in Figure 6.

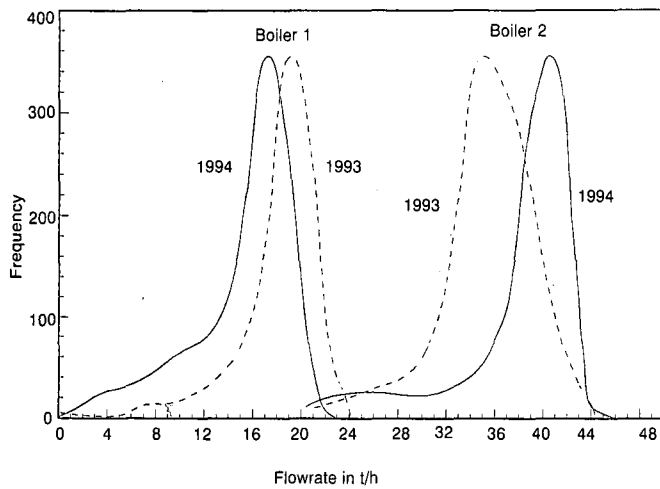


FIGURE 6: Steam flowrates in 1993 and 1994 at Constance sugar factory

There was a need to reduce the steam output of boiler 1 and increase that of boiler 2. Analysis of the existing system was done from the perspective of boiler 1 steam flowrate as shown in Figure 7.

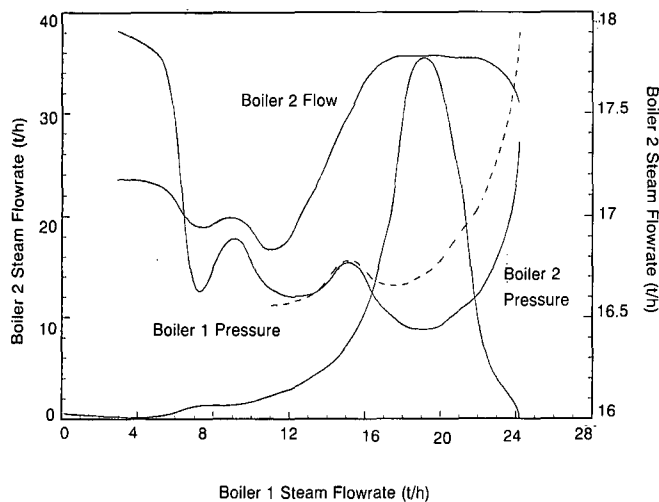


FIGURE 7: The operating conditions of the boilers in 1993

It was noted that the pressure of boiler 1 was always higher than that of boiler 2. Consequently boiler 1 was dominating boiler 2. A decision was taken to increase the induced draught of the latter in order to improve its combustion efficiency. The system was analysed again in 1994, after the modifications were completed and the results have been summarised in Figure 8.

The pressure of boiler 2 has been found to be slightly higher than that of boiler 1, and this has caused the distributions to shift accordingly, as depicted in Figure 6. The boilers now operate at 17 and 41 t/h respectively, and furthermore, it was noted that there were less fluctuations in the steam output of boiler 2 in 1994 as seen from the spread of its distribution curve. Thus, the steam output of boiler 1 has been reduced to its rated value, that of boiler 2 has been increased considerably, and their total output has increased from 55 to 58 t/h.

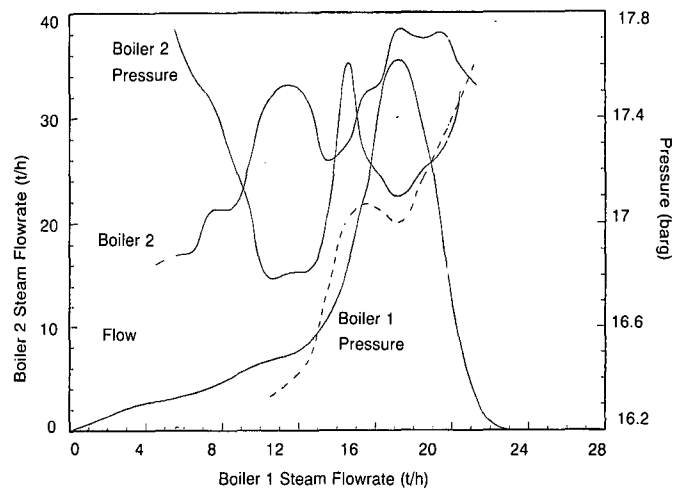


FIGURE 8: The operating conditions of the boilers in 1994

**Discussion**

The data analysis technique has the following advantages:

- Massive data can be displayed without affecting the clarity.
- Time lag does not affect the correspondence between variables.
- The operating ranges and the most probable values of the variables are easily obtained. The most probable value is in fact more relevant than the average value, especially when the distribution is skewed.
- Factory stoppages and temporary abnormal operating conditions do not affect the results.
- The standard deviation of the distribution can be used to measure the extent of variations of a particular parameter. In some cases, it can be used to measure the effectiveness of a control system.
- Multimodal systems can be easily analysed.

**Conclusions**

This method of analysing massive data proves to be a powerful tool that has been used with great success in Mauritius during the past few years. It has been of great help in the determination of the operating parameters of much factory equipment, and in the evaluation of the effectiveness of modifications carried out in the sugar factories.

**Acknowledgements**

The authors wish to thank Dr R Julien, the Director of the MSIRI for the opportunity of presenting this paper, Mrs L Wong Sak Hoi, the Administrator of the Sugar Engineering Division for her advice and constant support, Mr J Lim, Head of the Biometry Division for his valuable suggestions during the writing of this paper, and the staff of the Sugar Engineering Division for their assistance during the factory surveys.

**Bibliography**

Mack, C (1969). *Essentials of statistics for scientists and technologists*. Heinemann Educational books Ltd. London: pp 30-37.