STATISTICAL ANALYSIS OF AN INDUSTRIAL DISTILLATION PROCESS*

Patrizio Frederic

Dipartimento di Scienze Statistiche, Università di Bologna
frederic@stat.unibo.it

Mauro Catellani

Ciba S.C. S.p.A., Pontecchio Marconi (BO)

Abstract

This paper deals with the statistical analysis of a distillation process. The aim is to find statistical relations between the quality of final output and the variables measured during the process. Such relations cannot be analyzed with standard deterministic models because the presence of refluxes and recycled products tends to increase the complexity of the system. Since many variables are involved in the study, we implemented algorithms that can search for such functional relations automatically.

1 INTRODUCTION

For a few years, it has been conducting electronic monitoring of a distillation process. The aim of this paper is to study the process from a statistical point of view. The need to combine a statistical study with chemical-engineering considerations derives from the intrinsically dynamic nature of the process. The features of recycling and refluxing in the process increase the complexity of the system and necessitate an investigation of the process to search for relations of dependence among the variables that cannot be directly evaluated theoretically. The aim is to formulate a functional relation between the final product measurement and the variables that affect the process (temperatures, pressures, etc.) according to

Il presente lavoro è stato finanziato dal Progetto COFIN99 "Metodi di Inferenza Statistica per Problemi Complessi", responsabile prof. Fortunato Pesarin.

the following scheme:

$$y_t = f(X_{(1,t)}, X_{(1,t-1)}, \ldots, X_{(2,t)}, X_{(2,t-1)}, \ldots, X_{(i,t-j)}, \ldots) + \varepsilon_t \qquad (1)$$

where y_t is the value of the final analysis at time t (usually the titre percentage an indicator of the purity of the finished product) and $X_{(i,t)}$ is the system variable i evaluated at time t. To construct f, we will require three preliminary projects.

In section 2 we will summarize the process of distillation and reconstruct some missing information on the measurement of the final analysis. Although the variables X_i are recorded every 10 minutes, the final output analysis are carried out at irregular time intervals. The available information will be interpolated by the *smoothing spline* technique. In section 3 we will evaluate the waiting times elapsed before the system variables X_{it} affect the final output. We will devise an algorithm that non-parametrically assesses the probability distribution of the waiting times based on flow data. In section 4 we will select the set $\tilde{X}$ of variables $X_{(i,t)}$ highly correlated with the final output. A non parametric correlation index will be proposed. It will be shown that $\tilde{X}$ is composed of variables relating to a single distillation column. Finally in section 5 we will construct f using the appropriate multidimensional method *Projection Pursuit* regression.

2 THE DATA SET

Distillation is a process of separation of chemical mixtures. A detailed introduction can be found in Kirk (1998). The initial mixture is subjected to changes of pressure and temperature along a path consisting of three distillation columns. As a consequence of these changes, the heavier constituents tend to move toward the bottom of each column and the lighter ones toward the top. The excessively heavy (or light) constituents are eliminated as waste. Part of the remaining mixture becomes the input of the subsequent column, and the rest is recycled in the current column. The data used in this study are divided into two groups: process variables such as flows, temperatures, pressures and densities if mixture components in the column, and final analysis witch are measurement of the final product output.

The process variables were automatically recorded every 10 minutes from 07:00:00 hours on 12/03/2000 until 01:30:00 hours on 17/03/2000, yielding a total of 688 observations. The final analysis were recorded less

regularly (on average every 122.31 minutes) for a total of 54 observations. The discrepancy between the recording intervals was the first problem to tackle. We hypothesized that the phenomenon did not undergo abrupt changes between one recording and the next. Rather it passed smoothly from the recording at time t to the one at time $t+1$ in a *gradual* manner. For this reason, we chose the method of spline smoothing interpolation i.e. an interpolating function without any but smooth assumption. Figure 1 shows the result.

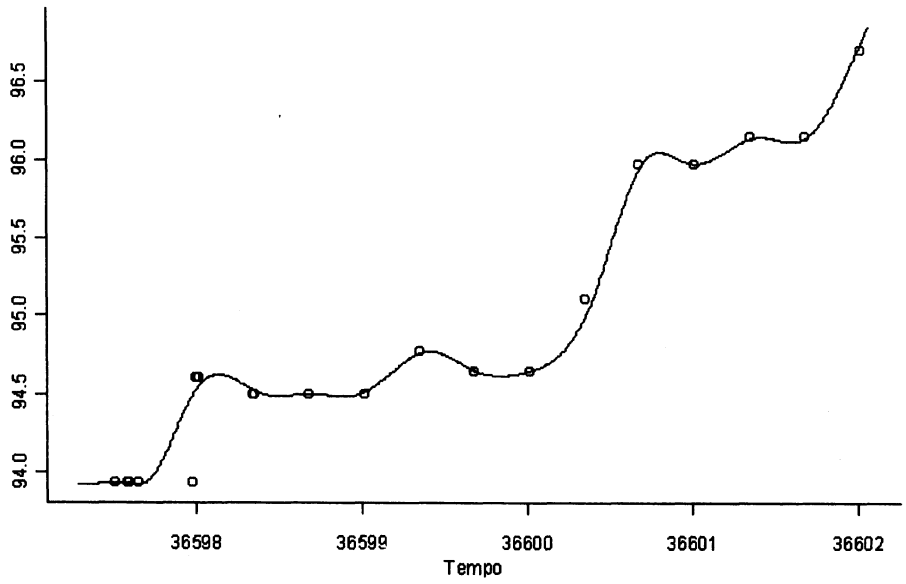

Fig. 1: Values of y interpolated with the spline smoothing method.

3 A NON-PARAMETRIC MODEL FOR WAITING TIME

The basic idea consists in assigning a probability distribution to all possible histories that a small quantity of mixture might have through the process. Clearly, the waiting times that the mixture in a given position reach to the end of the process depend on the state transition probabilities. These probabilities are not constant but change with time, which is

why we will use a non-parametric approach. Now let $i = 1, 2, 3$ be the i-th stripping column and $j = 0, 1, 2$ be the j-th state of the system ($j = 0$ the mixture enters column i, $j = 1$ the mixing is in the recycle state, $j = 2$ the mixture is a waste and is rejected). Let $p_{i',j';i,j}(t)$ be the probability that the small quantity of mixture in column i and state j at time t will be in column i' and state j' at time $t + 1$. Since the flow values (for each time t) and the capacity of the columns are known, the transition probabilities are obtained by dividing flow by capacity. A numerical simulation is used to compute the probability distribution of waiting times, according to the following algorithm:

Program to compute waiting times

set up number of simulations n
    for k in 1 to n {
        Initialization:
        $History(k) = [1, 0]$;
        Set up starting time t_0;
        $i = 1$;
        $j = 0$;
    Step t { $p_{i,j;i,j'}(t) = \frac{\text{Flow state } j' \text{ Column } i}{\text{Column capacity } i}$;
        r = random uniform $[0, 1]$;
        if $r < p_{i,j;i,1}(t)$ then {
                $j = 1$ (mixture is in state 1 and column i);
                $History(k)$ =link$(History(k), [i, j])$};
        else if $r < p_{i,j;i,1}(t) + p_{i,j;i,2}(t)$ then {
                $j = 2$ (mixture is in state 2 and column i);
                $History(k)$ =link$(History(k), [i, j])$;}
        else if $r < p_{i,j;i,1}(t) + p_{i,j;i,2}(t) + p_{i,j;i,3}(t)$ then
                STOP (mixture has been rejected);
        else
                $j = 0$;
                $i = i + 1$;
        if $i = 4$ then
                STOP (distillation process is finished)
        $t = t +$ 10minutes;
    }
}
Probability of history k^* = $\frac{\text{number of histories equal to history } (k^*)}{n}$

Simple modification of this algorithm allows one to compute $p_{c,t,s}$, the probability a variable measured in column c at time t affects the final product at time $t + s$.

4 IDENTIFICATION OF A SINGLE COLUMN

We shall now identify the set $\widetilde{X}$ of variables X that are highly related with the measure of the final output. Let y_t be the value of pureness of the product measured at time t. To take account of different time lags, we will use the mean effect that variable $X_{it}^{(c)}$ ($X_{it}^{(c)}$ denotes the generic i-th variable of column $c, c = 1, 2, 3$, and $x_{it}^{(c)}$ the observed value) has on y:

$$\bar{y}_{ct} = \sum_{k=1}^{\infty} y_{t+k} p_{c,t,k}. \tag{2}$$

Equation (2) has been computed for every c, i. It is now possible to search evidence of relations between variables simply looking at the clouds of data $(x_{it}^{(c)}, \bar{y}_{ct})_{t=1,\ldots,T}$. Since system variables are more than fifty this is very hard to do. For this reason univariate non-parametric regressions are performed to explore the clouds of data $(x_{it}^{(c)}, \bar{y}_{ct})_{t=1,\ldots,T}$. Many standard techniques are available such as Kernel smoothing, Locfit, B-spline regression etc. Here a roughness penalty approach is used (see Green and Silverman (1994)). Such approach is formulated by:

$$g_i^{(c)} = \underset{g \in \mathcal{S}_2}{\operatorname{arginf}} \left\{ \sum_{t=1}^{T} \left(\bar{y}_{ct} - g(x_{it}^{(c)}) \right)^2 + \lambda \int \left(g''(t) \right)^2 dt \right\} \tag{3}$$

where $\mathcal{S}_2$ is the *smooth* functions space (continuous functions with absolutely continuous first derivative) and λ represents the smoothing parameter. It is well known (Eubank (1982)) that (3) allows a unique solution and that this solution is a *natural cubic spline* (De Boor (1978)). The smoothing parameter λ was chosen by the *cross-validation* method (Green and Silverman (1994)). Furthermore, a goodness of fit index $R_i^{(c)}, \forall i, \forall c$, was created by dividing the explained variance (deduced from model (3)) by the total variance. It was observed that eight of the ten variables that had index $R_i^{(c)}$ greater than 0.6 belonged to column III. Therefore, the following analysis will deal only with column III.

5 PROJECTION PURSUIT REGRESSION

In the previous section, eight variables highly correlated to y were selected. All eight variables were measured in column III. In this section, the

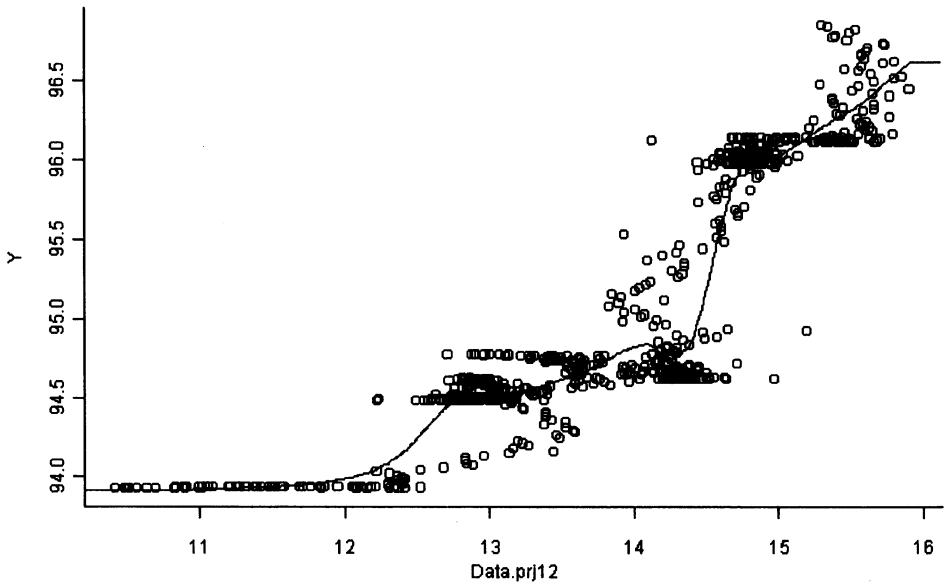

Fig. 2: Model 1: data projected over the real plane denoted by the direction $\mathbf{a}_1$

variables of column III will be used jointly to estimate the follow relation:

$$\overline{y}_{3,t} = f(x_{1t}^{(3)}, \ldots, x_{kt}^{(3)}) + \varepsilon_t \tag{4}$$

where f is a non-parametric regression surface. Since in the present situation k is a large value, certain problems associated with the *"curse of dimensionality"* arise, see Huber (1985). For this reason, standard non-parametric multidimensional techniques (such as multivariate kernel, thin-plate spline, ACE models, etc.) are not able to pick up small features of the data. To overcome this problem, a *Projection Pursuit Regression* (**PPR**) will be used (Friedman and Stuetzle (1981), Huber (1985) and Hall (1989)). Roughly speaking, **PPR** provides a linear decomposition of function f in terms of the sum of univariate non-parametric regression functions

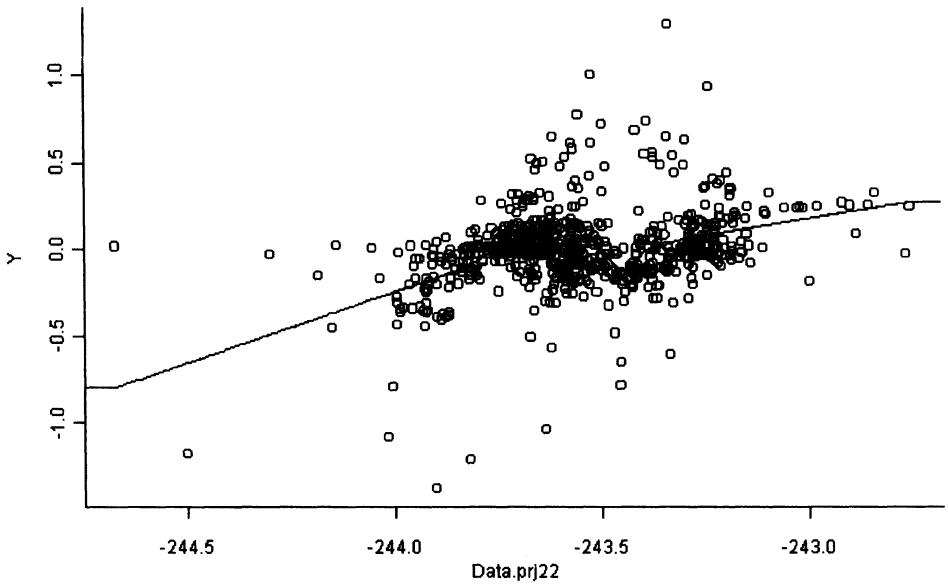

Fig. 3: Model 1: residuals of step 1 projected over the plane denoted by the direction $\mathbf{a}_2$

evaluated in linear combinations of regressors (projections). Formally:

$$f(x_{1t}^{(3)}, \ldots, x_{kt}^{(3)}) = \sum_{j=1}^{M} \phi_j \left(\sum_{i=1}^{k} a_{ij} x_{it}^{(3)} \right) \tag{5}$$

where M represents the number of terms of the model, k the dimension of the variables, and ϕ_j denote non-parametric univariate regression functions (such as smoothing spline or Nadarya-Watson, in the present work, super-smoother regression of Friedman and Stuetzle (1981) has been used). The projection coefficients a_{ij} are real values constrained so that: $\sum_{i=1}^{k} a_{ij}^2 = 1$ and $\sum_{i=1}^{k} a_{il} a_{ih} = 0$ if $l \neq h$. Huber (1985) discusses the class of regression functions that can be approximated by (5); actually this is a quite wide class of functions for the present problem. In other words, **PPR** consists in projecting the cloud of regressors along a real straight line orthogonal to y axes, computing the residuals, and using these residuals as new dependent variables. Clearly **PPR** is an iterative procedure. Let $\mathbf{a}_j = (a_{j1}, \ldots, a_{jk})$,

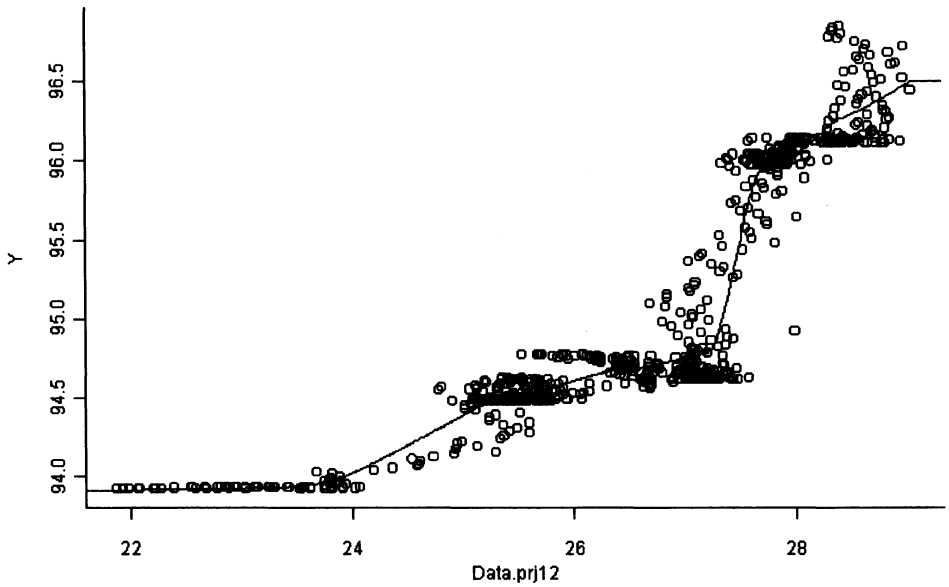

Fig. 4: Model 1: data projected over the real plane denoted by the direction $\mathbf{a}_1$

$\mathbf{x}_t = (x_{1t}^{(3)}, \ldots, x_{kt}^{(3)})$, $\mathbf{y} = (\overline{y}_{3,1}, \ldots, \overline{y}_{3,T})$, and $\bar{\mathbf{y}} = (T^{-1}\sum_t^T \overline{y}_{3,t})\mathbf{i}_T'$, where $\mathbf{i}_T$ is the unit vector of dimension T and $'$ is the transposition operator. The **PPR** algorithm can be written as follows:

Program PPR

Step 1 — setup the starting model

$\overline{y}_{3,t} = \phi_1(\mathbf{a}_1{}'\mathbf{x}_t^{(3)}) + e_{1t}$

choose vector $\mathbf{a}_1$:

$\mathbf{a}_1 = \underset{\mathbf{a}\in\mathbf{R}^k}{\text{arginf}}\{\sum_{t=1}^T(\overline{y}_{3,t} - \phi_1(\mathbf{a}'\mathbf{x}_t))^2\}$

Under constraint:

$\|\mathbf{a}_1\| = 1$

compute residuals:

$e_{t1} = \overline{y}_{3,t} - \phi_1(\mathbf{a}_1'\mathbf{x}_t)$

compute the division of the Residual Sum Square by Total Sum Square of y:

$E_1 = \mathbf{e}_1{}'\mathbf{e}_1/(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})$

Step $j+1$ — setup the model

$e_{jt} = \phi_{j+1}(\mathbf{a}_{j+1}{}'\mathbf{x}_t) + e_{j+1,t}$
choose the vector $\mathbf{a}_{j+1}$:
$\mathbf{a}_{j+1} = \underset{\mathbf{a}\in\mathbf{R}^k}{\text{arginf}}\{\sum_{t=1}^{T}(\overline{y}_{3,t} - \phi_{j+1}(\mathbf{a}'\mathbf{x}_t))^2\}$
Under constraint:
$\|\mathbf{a}_{j+1}\| = 1$
$\mathbf{a}_{j+1}{}'\mathbf{a}_l = 0, l = 1, \ldots, j$
compute residuals:
$e_{j+1,t} = \overline{y}_{3,t} - \phi_{j+1}(\mathbf{a}'_{j+1}\mathbf{x}_t)$
compute the division of the Residual Sum Square by Total Sum Square of y:
$E_{j+1} = \mathbf{e}_{j+1}{}'\mathbf{e}_{j+1}/(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})$
Stop Criterion **if** $\|E_{j+1} - E_j\| < \delta$ **then STOP** (δ is a given threshold)
the number of terms is $M = j$

Two **PPR** models have been implemented: in the first model, the independent variables are the eight variables of column III with index R_i greater than 0.6 (see section 4); while in the second model the independent variables are all fourteen variables of column III. We found that both models give a good explanation of the variability of y, and in both algorithms the stop criterion is satisfied for $M = 1$. Then the models have the simple form

$$y_t = \phi\left(\mathbf{a}'\mathbf{x}_t\right) + \varepsilon_t$$

Figure 2 shows the scatter-plot of the cloud after the first projection according to Model 1. The percentage of explained variability is around 91.8%. The improvement of performance at the second step was only 1.4%. Since this value is not large enough to include the second step in the final relation, $M = 1$ was chosen. Figure 3 shows the residual cloud evaluated after step 1, plotted along the projection of $\mathbf{x}$ data specified by the best $\mathbf{a}_2$ vector. A look at the square of coefficients a_j allowed us to identify the most important variables explaining the variability of y. Only two of the eight variables have a_j^2 greater than 0.3. These variables are a temperature variable with squared coefficient equal to 0.624 and a pressure variable with squared coefficient equal to 0.349. All the other variables have coefficient less than 0.02.

The second model explains 94.1% of the variability of y data at the first step, which is 2.3% better than Model 1. However, a look at the coefficients a_j^2 reveals that again only two of the 14 variables (exactly the same as above) have coefficients greater than 0.3. Figure 4 shows the results of the first projection in Model 2. It is easy to see that the result is not far different than the simpler Model 1

6 CONCLUSIONS

The methods used, in particular those described in Sections 3 and 4, seem to be very useful in exploring complex dynamical systems. Simply using the algorithm described in Section 3 and the regression methods described in Section 4, it is possible to calculate the desired index $R_i^{(c)}$. This index summarizes the goodness of fit of functional relation between any system variable and the final output. In other words, this index allows an automatic exploration of quantitative measurements in a large data set. In the particular example of the distillation process, the list of values of index $R_i^{(c)}$ gives evidence of important connections between some specific variables measured in column III of the production process and the final output, while relegating others as uninformative.

7 ACKNOWLEDGMENT

Authors thanks CIBA S.C. for technical consulting and partial financial support in fulfilling this paper. Such a paper deals with partial results of research involving the cooperation the Dipartimento di Scienze Statistiche dell'universita' di Bologna and CIBA S.C. SPA.

REFERENCES

DE BOOR C. (1978), *A practical guide to splines.* Springer, Berlin.

EUBANK R.L. (1982), *Spline smoothing and non parametric regression.* Marcel Dekker, New York

FRIEDMAN J., STUETZLE W. (1981), On Projection Pursuit Regression. *JASA*, 76, 817-824

GREEN B.J., SILVERMAN B.W. (1994), *Non parametric regression and generalized linear models. A roughness penalty approach.* Chapman and Hall, London.

HALL P. (1989), On Projection Pursuit Regression. *Ann. Statist.*, 13, 573-588

HUBER P. (1985), Projection Pursuit. *Ann. Statist.*, 13, 435-475

KIRK R. (1998), *The Encyclopedia of Chemical Technology.* 4th ed. Wiley, New York

ANALISI STATISTICA DI UN PROCESSO DI DISTILLAZIONE INDUSTRIALE

Riassunto

In questo lavoro si analizza un processo di distillazione industriale da un punto di vista statistico. L'utilizzo di modelli nello studio e nel controllo di sistemi dinamici complessi, come quello di un processo di distillazione, si rileva spesso inadeguato. Da qui l'esigenza di un'analisi esplorativa dei dati. Si è dunque cercato di trovare un insieme di variabili di processo (quali temperature, pressioni, ecc.) che fossero legate statisticamente alla qualità del prodotto finito. Poichè le variabili in gioco sono più di 50, sono stati sviluppati algoritmi che, utilizzando tecniche non parametriche di regressione, cercano in automatico le variabili che in media meglio spiegano la varibilità della qualità del prodotto finale.