HUNGARIAN JOURNAL
OF INDUSTRIAL CHEMISTRY
VESZPRÉM
Vol. 35. pp. 85-93 (2007)

APPLICATION OF EXPLORATORY DATA ANALYSIS TO HISTORICAL PROCESS DATA OF POLYETHYLENE PRODUCTION

J. ABONYI

University of Pannonia Dept. of Process Engineering, H-8201, Veszprém, P.O.Box 158, HUNGARY

In modern chemical process systems huge amount of data are recorded. These data definitely have the potential to provide information for product and process design, monitoring and control. This paper presents a brief survey of simple Exploratory Data Analysis procedures have been found to be useful in the qualitative analysis of historical process data. The presented Box plots and quantile-quantile plots are applied to an industrial polyethylene plant to analyse different productions of a given product and to explore the relationships between different operating and product quality variables.

Keywords: Exploratory data analysis, Box plot, Quintile-quintile plot

Introduction

The major aims of monitoring plant performance are the reduction of off-specification production, the identification of important process disturbances and the early warning of process malfunctions or plant faults [1]. In modern production systems huge amount of process operational data are recorded with distributed control systems (DCS). These data definitely have the potential to provide information for product and process design, monitoring and control [2]. Process monitoring based on multivariate statistical analysis of process data has recently been investigated by a number of researchers [3]. The aim of these approaches is to reduce the dimensionality of the correlated process data by projecting them down onto a lower dimensional latent variable space where the operation can be easily visualized. These approaches use the techniques of principal component analysis (PCA) or projection to latent structure (PLS). Beside process performance monitoring, these tools can be used for system identification [3], ensuring consistent production [4] and product design [1]. For these classical data analysis approaches, the collection of the data is followed by the imposition of a model and the analysis, estimation, and testing that follows are focused on the parameters of that model.

Most of operational process data may be characterised as historical in the sense that is was not collected on the basis of experiments designed to test specific statistical hypothesis. Consequently, the resulting databases are likely to contain unexpected features (e.g. outliers from various sources, unexpected correlation between variables, etc.) This observation is important for two reasons: first, these data anomalies can completely negate the results obtained by standard analysis procedures, particularly those based on squared error criteria (a large class includes many SPC and chemometrics techniques, like PCA). Secondly and sometimes more importantly, an understanding of these data anomalies may lead to extremely valuable insights [5].

Pearson suggested using Exploratory Data Analysis (EDA) tools for both of these reasons [5]. For Exploratory Data Analysis (EDA), the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate. EDA is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models, and determine optimal factor settings. The seminal work in EDA is written by Tukey, [6]. Over the years it has benefited from other noteworthy publications such as Data Analysis and Regression by Mosteller and Tukey [7], and the book of Velleman and Hoaglin [8].

Most EDA techniques are graphical in nature with a few quantitative techniques [9]. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as data traces and histograms).
2. Plotting simple statistics such as mean plots, standard deviation plots and box plots.

3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

The aim of this paper is to present an application relevant survey of some of Exploratory Data Analysis procedures that have been found to be particularly useful in the qualitative analysis of historical databases of production systems. The rest of this paper is organised as follows. The next section deals with the description of the problem used through the paper to illustrate the presented exploratory data analysis approach. Section 3. deals with the description of the *Box plot* and shows its application in the comparison of different production of a given product. In the third section the application of Quantile-quantile plot is proposed for the exploration of the differences of the different productions and for the analysis the relationship among different operating variables. The examples illustrate that the proposed EDA based tools are useful to identify the similar behaviour of operating and model quality variables.

Problem Description

Formulated products (plastics, polymer composites) are generally produced from many ingredients, and large number of the interactions between the components and the processing conditions all have the effect on the final product quality. If these effects are detected, significant economic benefits can be realized. This consideration lead the "Optimization of Operating Processes" project of the VIKKK Research Center at the University of Veszprém supported by the largest Hungarian polymer production company (TVK Ltd., www.tvk.hu). The aim of the project is to work out a methodology for the data-driven improvement of process. Hence, in this paper the monitoring of a medium and high-density polyethylene (MDPE, HDPE) plant of the TVK Ltd. in Hungary is considered. HDPE is versatile plastic used for household goods, packaging, car parts and pipe. A brief explanation of the Phillips license based low-pressure catalytic process is provided in the following.

Fig. 1 represents the Phillips Petroleum Co. suspension ethylene polymerization process. The polymer particles are suspended in an inert hydrocarbon. The melting point of high-density polyethylene is approximately 135 °C. Therefore, slurry polymerization takes place at a temperature below 135 °C; the polymer formed is in the solid state. The Phillips process takes place at a temperature between 85-110 °C. The catalyst and the inert solvent are introduced into the loop reactor where ethylene and an α-olefin (1-hexene) are circulating. The inert solvent (isobuthane) is used to dissipate heat as the reaction is highly exothermic. A cooling jacket is also used to dissipate heat. The reactor consists of a folded loop containing four long runs of pipe 1 m in diameter, connected by short horizontal lengths of 5 m. The slurry of HDPE and catalyst particles circulate through the loop at a velocity between 5-12 m/s. The reason for the high velocity is because at lower velocities the slurry will deposit on the walls of the reactor causing fouling. The concentration of polymer products in the slurry is 25-40% by weight. Ethylene, α-olefin comonomer (if used), an inert solvent, and catalyst components are continuously charged into the reactor at a total pressure of 450 psig. The polymer is concentrated in settling legs to about 60-70% by weight slurry and continuously removed. The solvent is recovered by hot flashing and distillation. The polymer is dried and pelletized. The conversion of ethylene to polyethylene is very high (95%-98%), eliminating ethylene recovery. The molecular weight of high-density polyethylene is mainly determined by the type of the catalyst and the temperature of the catalyst activation [10]. The main properties of polymer products (e.g. Melt Index (MI) and density) are controlled by the reactor temperature, monomer, comonomer and chain-transfer agent concentration.

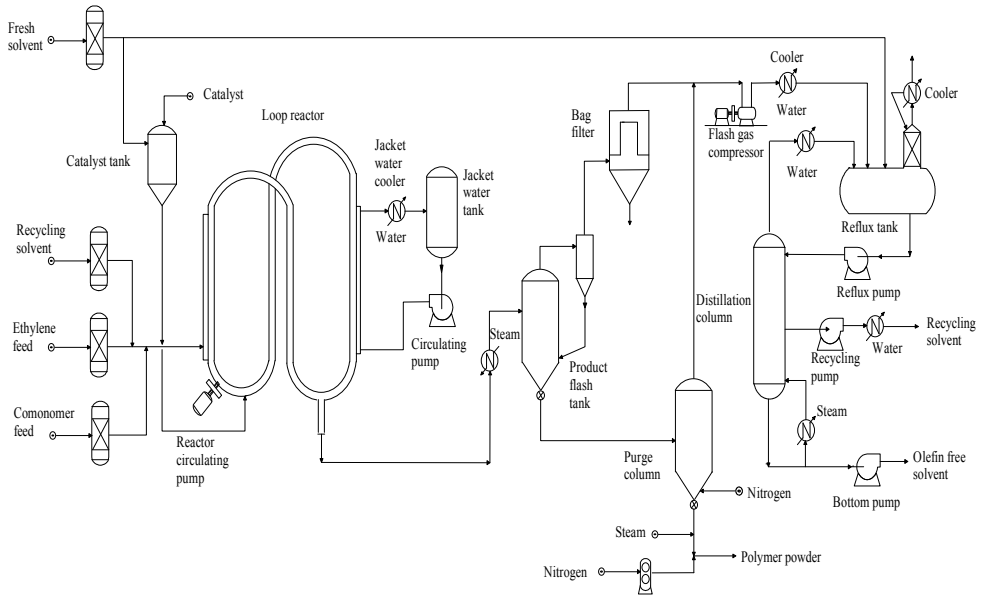

Figure 1: Scheme of the Phillips loop reactor process [10]

An interesting problem with the process is that it is required to produce about ten product grades according to market demand. Hence, there is a clear need to minimize the time of changeover because off-specification product may be produced during transition. The difficulty of the problem comes from the fact that there are more than ten process variables to consider. Measurements are available in every 15 seconds on process variables $\mathbf{z}_k$, which are the $z_{k,1}$ reactor temperature (T), $z_{k,2}$ ethylene concentration (C2) and $z_{k,3}$ hexene concentration (C6) in the loop reactor, $z_{k,4}$ the ratio of the hexene and ethylene inlet flowrate (C6/C2in), $z_{k,5}$ the flowrate of the isobutane solvent (C4), $z_{k,6}$ the hydrogen concentration (H2), $z_{k,7}$ the density of the slurry in the reactor (roz), $z_{k,8}$ polymer production intensity (PE), and $z_{k,9}$ the inlet flowrate of the catalyzator (KAT). The product quality $\mathbf{y}_k$ is only determined later, in another process. The interval between the product samples is between half and five hours. The $y_{k,1}$ melt index (MI) and the $y_{k,2}$ density of the polymer power (ro) are monitored by off-line laboratory analysis after drying of the polymer that causes one hour time-delay.

Since, it would be useful to know if the product is good before testing it, the monitoring of the process would help in the early detection of poor-quality product. There are other reasons why monitoring the process is advantageous. Only a few properties of the product are measured and sometimes these are not sufficient to define entirely the product quality. For example, if only rheological properties of polymer are measured (melt index), any variation in end-use application that arise due to variation of chemical structure (branching, composition, etc.) will not be captured by following only these product properties. In these cases the process data may contain more information about events with special causes that may effect the product quality [11].

The modelling and monitoring of processes from data involve solving the problem of data gathering, pre-processing, model architecture selection, identification or adaptation and model validation. The process data analyzed in this paper have been collected over three months of operation. The data have been extracted from the distributed control system (DCS) of the process. An SQL server has been installed to store and merge this data with the product quality database. According to the data warehousing methodology, the application relevant data have been extracted from this SQL database. As one of the objectives is to infer the values of product quality from process data obtained at different operating regions, a set of transition-free data is used that covers the whole range of specifications of the quality properties and the process variables over all the possible operating regions. The data were pre-processed by normalization performed on single variables.

The aim of the following sections is to present exploratory data analysis tools that can be applied for the previously presented problem.

Box plot of operating and quality variables

Suppose that X is a real-valued random variable for the experiment. In our research work, the analysis of process and product quality variables is considered. Hence, the variables are $X \in \{\mathbf{z}_k, \mathbf{y}_k\}$.

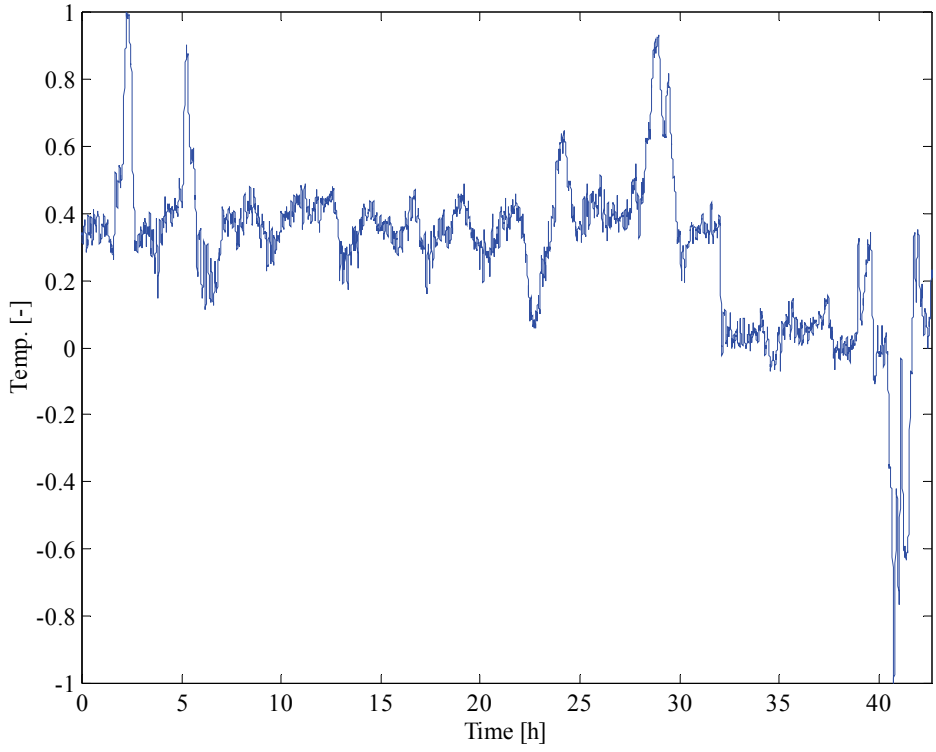

Figure 2: Example of the change of a process variable (T)

An example of the behaviour of a process variable is given in *Fig. 2*, where the change of the dimensionless reactor temperature is shown. At our industrial partner's request most of the data shown in this paper are normalized.

The (cumulative) *distribution function* of X is the function F given by $F(x) = P(X \leq x)$ for x, which is a function giving the probability that the random variable X is less than or equal to x, for every value x. For a discrete random variable, the cumulative distribution function is found by summing up the probabilities. For a continuous random variable, the cumulative distribution function is the integral of its probability density function. Suppose that $p \in [0, 1]$. A value of x such that $F(x^-) = P(X < x) \leq p$ and $F(x) = P(X \leq x) \geq p$ is called a *quantile* of order p for the distribution. Roughly speaking, a quantile of order p is a value where the cumulative distribution crosses p. Hence, by a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.25 (or 25%) quantile is the point at which 25% percent of the data fall below and 75% fall above that value. Note that there is an inverse relation of sorts between the quantiles and the cumulative distribution values. A quantile of order 1/2 is called a *median* of the distribution. When there is only one median, it is frequently used as a measure of the *center* of the distribution. A quantile of order 1/4 is called a *first quartile* and the quantile of order 3/4 is called a *third quartile*. A median is a second quartile. Assuming uniqueness, let $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ denote the first (lower), second, and third (upper) quartiles of X. Note that the interval from $q_{0.25}$ to $q_{0.75}$ gives the middle half of the distribution, and thus the *interquartile range* is defined to be IQR = $q_{0.75} - q_{0.25}$, and is sometimes used as a measure of the *spread* of the distribution with respect to the median. Let q_0 and q_1 denote the minimum and maximum values of X, respectively (assuming that these are finite). The five parameters q_0, $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, q_1 are often referred to as the *five−number summary*. Together, these parameters give a great deal of information about the distribution in terms of the center, spread, and skewness. An example of a cumulative distribution function and quantile is given in *Fig. 3*, where the distribution of the reactor temperature shown in Fig. 2 is depicted.

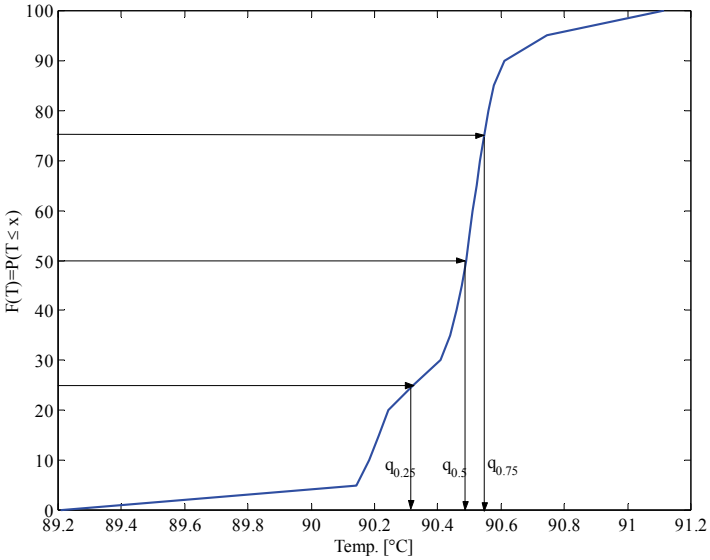

Figure 3: Example of a cumulated distribution function of a process variable (T), the $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ quintiles are also depicted

Tukey's five number summary is often displayed as a *boxplot.* Box plots are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data [9,12].

A box plot consists of a line extending from the minimum value q_0 to the maximum value q_1, with a rectangular box from $q_{0.25}$ to $q_{0.75}$, and tick marks at the q_0, the median $q_{0.5}$, and q_1 . Hence, the lower and upper lines of the "box" are the 25th and 75th percentiles of the sample. The distance between the top and bottom of the box is the interquartile range. The line in the middle of the box is the sample median. If the median is not centered in the box that is an indication of skewness. Thus the box represents the body (middle 50%) of the data. There is a useful variation of the box plot that is more specifically identifies outliers. To create this variation:

1. Calculate the median and the lower and upper quartiles.
2. Plot a symbol at the median and draw a box between the lower and upper quartiles.
3. Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.
4. Calculate the following points:
   L1 = q_0 – 1.5 IQ
   L2 = q_0 – 3 IQ
   U1 = q_1 + 1.5 IQ
   U2 = q_1 + 3 IQ
5. The line from the lower quartile to the minimum is now drawn from the lower quartile to the

smallest point that is greater than L1. Likewise, the line from the upper quartile to the maximum is now drawn to the largest point smaller than U1.

6. Points between L1 and L2 or between U1 and U2 are drawn.

The "whiskers" are lines extending above and below the box. They show the extent of the rest of the sample (unless there are outliers). Assuming no outliers, the maximum of the sample is the top of the upper whisker. The minimum of the sample is the bottom of the lower whisker. By default, an outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box. The plotted outlier points may be the result of a data entry error, a poor measurement or a change in the system that generated the data.

An example of a box plot is given in *Fig. 4*, where the distribution of the reactor temperature given in Fig. 1 is shown.

A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. Such a comparison is given in *Fig. 5*, where the melt index (MI) distribution of five different production of the same product is shown. Hence, box plot has a significant effect on the response with respect to either location or variation and the box plot is also an effective tool for summarizing large quantities of information.

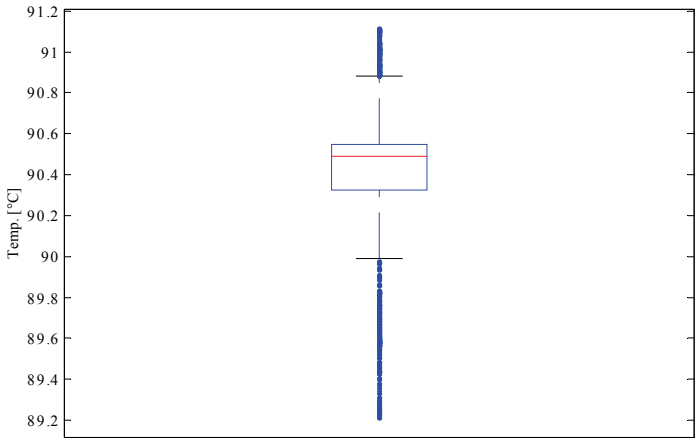

Figure 4: Example of a cumulated distribution function of a process variable (T), the $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ quintiles are also depicted

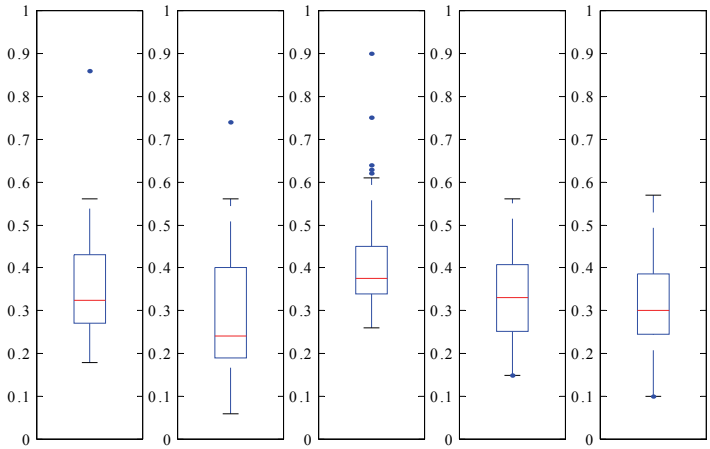

Figure 5: Melt Index (MI) of five different production of the same product

Quantile-quantile plot of process and product quality variables

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The Quantile-quantile (q-q) plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests [5,9].

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is. If the data sets have the same size, the q-q plot is essentially a plot of sorted data set *A* against sorted data

set *B*. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

A diagonal reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. If the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the diagonal reference line. The q-q plot is similar to a probability plot, where the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

The q-q plot can be used to answer the following questions: Do two data sets come from populations with a common distribution? Do two data sets have common location and scale? Do two data sets have similar distributional shapes? Do two data sets have similar tail behaviour?

These questions arise at the qualitative analysis of historical databases of production systems. Firstly, an example for comparison of two production of two different productions of the same product is given in the first column of *Fig. 6*.

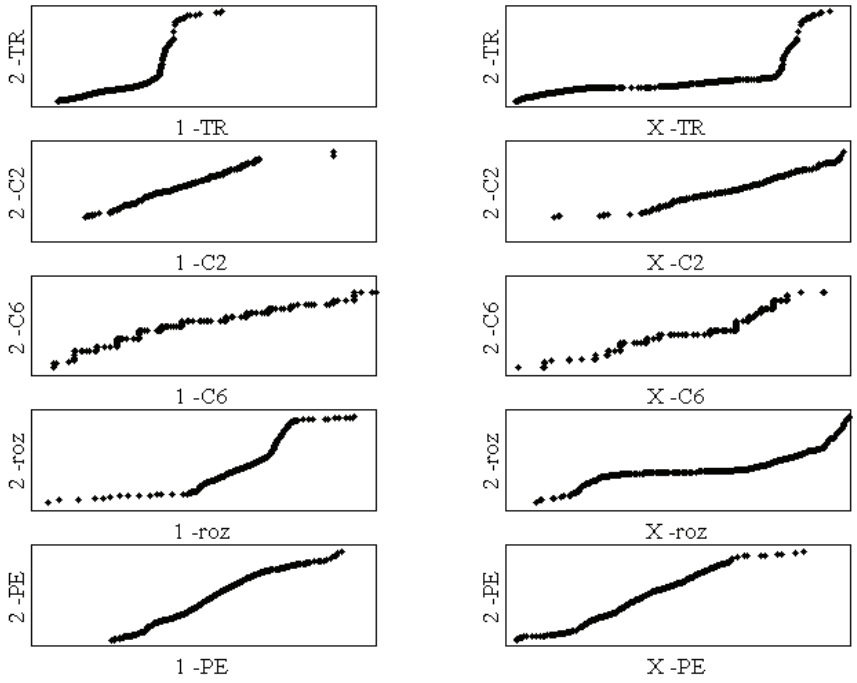

Figure 6: Example of a quantile-quantile plots of process variable distributions related to two different productions (first column: same products, second column: different products).

This plot shows that the distributions of the temperature are different, while the distributions of the concentrations are much more similar to each other. This difference is much bigger if the temperature related to the production of two different products are compared (see the second column of Fig. 6). The difference between the production of the same and different products is much more characteristic if we compare the distributions of the quality properties (see *Figs 7* and *8*).

This small application example suggests that quantile-quantile plots can be effectively used to compare different productions.

Another type of application is given in *Figs 9* and *10*, where the similarities between the distributions of different process and quality variables are analysed. This analysis could be extremely useful to detect dependencies between the operating parameters of the process.

Based on the application of the proposed tools and the analysis of the presented figures several rules have been extracted. Most of these rules found to be useful for our industrial partner, since the extracted knowledge and the resulted plots can be effectively used to summarise trends of the process variables and estimate the quality of the products.

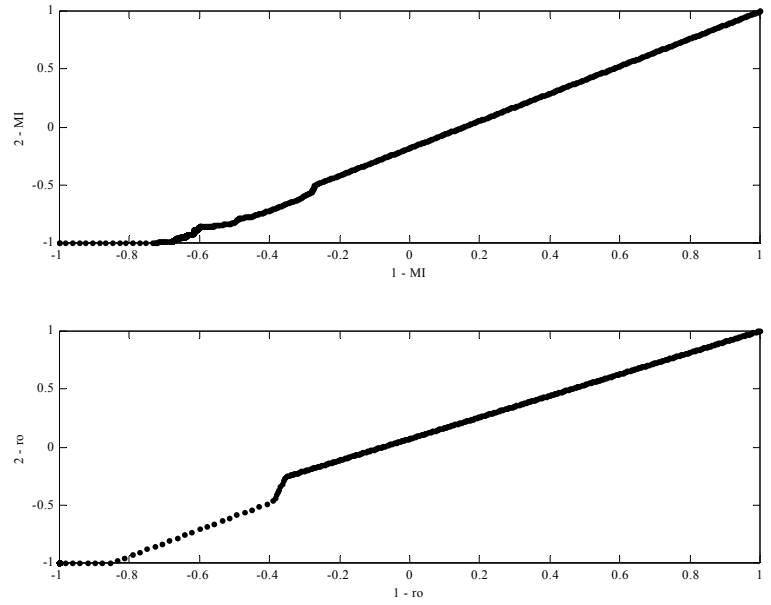

Figure 7: Example of a quantile-quantile plots of quality (Melt index - MI polymer density – ro) distributions related to two different productions of the same product

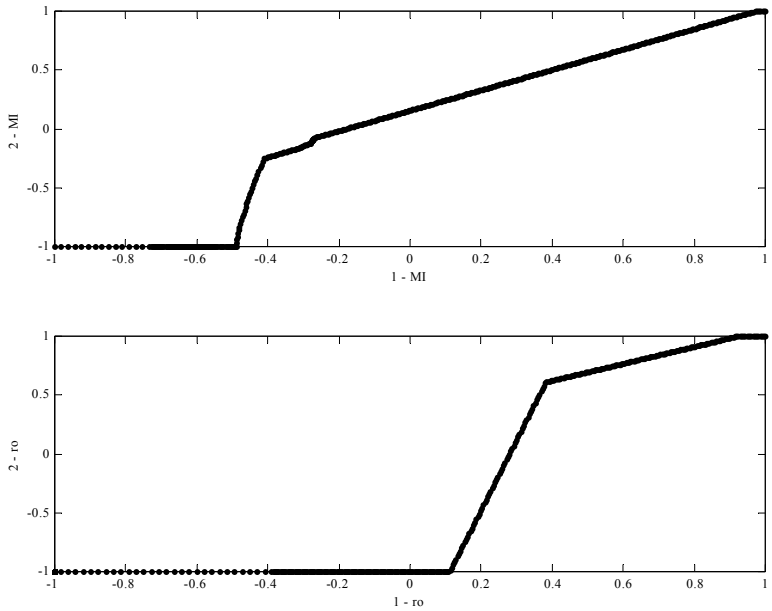

Figure 8: Example of a quantile-quantile plots of quality (Melt index - MI polymer density – ro) distributions related to two productions of different products

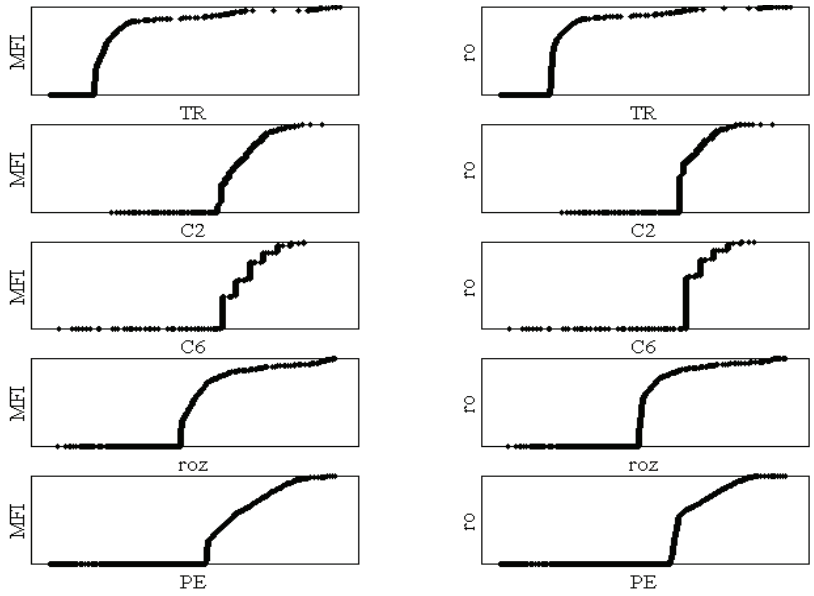

Figure 9: Example of a qunatile-qunatile plot of a process and quality variable distributions related to the same production of a product.

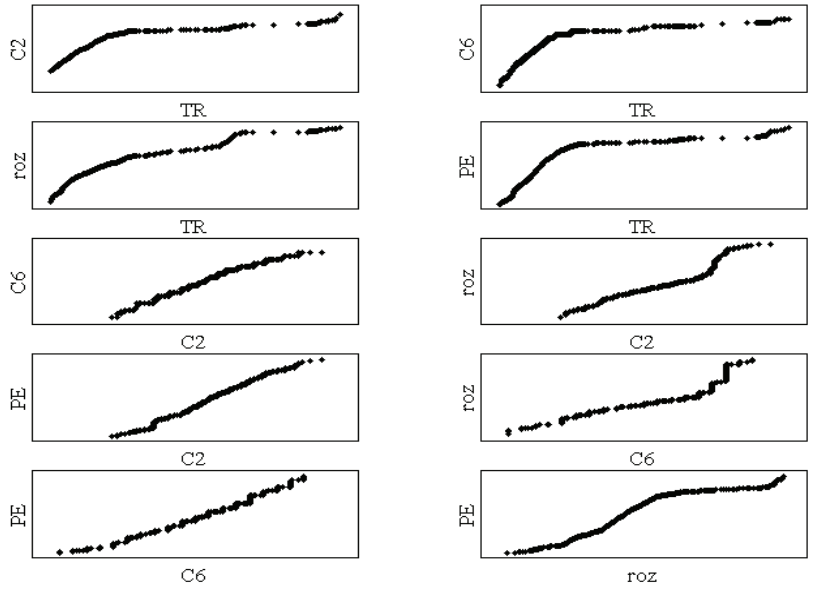

Figure 10: Qunatile-qunatile plot of a process variable distributions related to the same production of a product.

The behaviour of the control algorithm of the advanced model-based control system can be also identified from the analysis of these plots. E.g. since the density of the slurry in the reactor (roz) is controlled by the hexene concentration (C6), these two variables have similar distributions as it is shown in Fig. 10. Furthermore, the ratio of C2-C6 is also controlled, which makes the behaviour of these two variables also similar. Because of these relations, the C2-roz distributions become also similar.

The quantile-quantile plot of the production rate (PE) and ethylene concentration (C2) is also close to a straight line. This is because the highest ethylene concentration results in highest reaction speed. The previously presented rules are only illustrative, but they show that the proposed tools can be effectively used to detect relationships between process and product quality variables and compare different productions

Conclusions

The paper presented a brief survey of simple Exploratory Data Analysis procedures that have been found to be particularly useful in the qualitative analysis of historical databases of production systems. It has been showed that box plot is an important EDA tool for determining if a factor has a significant effect on the response with respect to the quality of a given productions. To analyse the relationships between different production, different products, and different operating variables quantile-quantile plots are proposed.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the Cooperative Research Center (VIKKK) (project KKK-II-1A), and founding from the Hungarian Ministry of Education (FKFP-0073/2001). János Abonyi is grateful for the financial support of the János Bolyai Research Fellowship of the Hungarian Academy of Science and OTKA (Hungarian National Research Foundation), No. T037600. The support of our industrial partners at TVK Ltd., especially Miklós Németh, Lóránt Bálint and dr. Gábor Nagy is gratefully acknowledged.

REFERENCES

1. LAKSHMINARAYANAN S., FUJII H., GROSMAN B., DASSAU E., LEWIN D. R.: New product design via analysis of historical databases, *Computers and Chemical Engineering* 24 (2000) 671-676
2. YAMASHITA Y.: Supervised learning for the analysis of the process operational data, *Computers and Chemical Engineering* 24 (2000) 471-474
3. MACGREGOR J. F., KOURTI T: Statistical process control of multivariate processes, *Control Eng. Practice*, Vol.3, No. 3, (1995) 403-414
4. MARTIN E. B., MORRIS A. J., PAPAZOGLOU M. C., KIPARISSIDES C.: Batch process monitoring for consistent production, *Computers Chem. Eng.* Vol. 20. (1996), pp. S599-S605
5. PEARSON R .K.: Exploring Process Data, Journal of Process Control, 11, (2001), 179-194
6. TUKEY J.: Exploratory Data Analysis, Addison-Wesley, (1977)
7. MOSTELLER F., TUKEY J.: *Data Analysis and Regression*, Addison-Wesley, (1977)
8. VELLEMAN P., HOAGLIN D.: The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis, Duxbury, (1981)
9. MILITKÝ J., MELOUN M.: Some graphical aids for univariate exploratory data analysis, *Analytica Chimica Acta,* 277(2), (1993), 215-221
10. NAGY G.: The polyethylene, *Magyar Kémikusok Lapja (MKL)*, 52(5), (1997) 233-242, In Hungarian
11. JEACKLE C. M., MACGREGOR J. F.: Product design through multivariate statistical analysis of process data, *American Institute of Chem. Eng. J.*, 44(5) (1998) 1105-1118
12. CHAMBERS J., CLEVELAND W., KLEINER B., TUKEY P.: Graphical Methods for Data Analysis, Wadsworth, (1983)