DATA MINING IN MANUFACTURING QUALITY CONTROL:

A CASE STUDY OF A COLOUR TV PANEL PRODUCTION LINE

by

Feng Fang

B. Engineering, Beijing University of Aeronautics and Astronautics, 1989

PROJECT SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMNTS FOR THE DEGREE OF
MASTER OF BUSINESS ADMINISTRATION

in the Faculty
of
Business Administration

SIMON FRASER UNIVERSITY
April 2003

APPROVAL

NAME: Feng Fang

DEGREE: Master of Business Administration

TITLE OF PROJECT: Data Mining In Manufacturing Quality Control: A Case Study of a Colour TV Panel Production Line

SUPERVISORY COMMITTEE:

Dr. Andrew Gemino
Senior Supervisor
Assistant Professor
Faculty of Business Administration
Simon Fraser University

Dr. Drew Parker
Associate Professor
Faculty of Business Administration
Simon Fraser University

Date Approved: Apr. 28, 2003

ABSTRACT

Product quality is critical for today's manufacturers. Data mining is one component of Knowledge Discovery in Databases that has established broad application in industry. The goal of this paper is to blend two areas of research and explore the business value data mining can provide to process quality control.

Several causes and remedies for product defeat exist. Practitioners apply the metrology method, software and analysis method and quality management method to improve product quality. Statistical process control is one of the common techniques in traditional quality control. Data mining is statistical in nature, but differs from traditional statistical techniques by addressing some of the disadvantages inherent in statistical techniques. For example, the exploding size of quality data from a variety of sources make data mining close to a necessity for many firms. High power data mining tools can also handle complex and non-linear relationships and provide better decision support for management.

This paper presents a case study of a part of a panel production line to elaborate how data mining can be used to solve quality control problems. The objectives of the case study were to verify existing knowledge and to discover unknown relationships or patterns. The data mining software used was SAS Enterprise Miner 3.0. The results show that identified significant variables were consistent with existing knowledge. In addition the findings of a time series effect and the discovery of business rules are useful in estimating defeat based on previous defeat and simple combinations of values of predictive variables. Moreover, the data mining approach can be magnified to quality control of the entire production line because all production conditions can be incorporated in tree nodes. The manufacturer can then identify significant variables

along the whole line and even the external variables, e.g. the weather temperature and humidity. The scientific and systematic model approach also facilitates quality knowledge management in organizations.

ACKNOWLEDGEMENTS

I would like to acknowledge the contribution made to this paper by Dr. Andrew Gemino, my senior supervisor, who not only gave me guidance and advice, but also contributed significantly to the organization and interpretation of data analysis. His enlightenment deepened my understanding of technology and his encouragement made the work a pleasant one. His patience and carefulness were most appreciated.

I would also like to thank Dr. Drew Parker, my project supervisor, for sharing with me his insights on the project. My appreciation is also extended to Dr. Albert Wong, Professor of Decision Support System, and Dr. Robert Krider, Professor of Marketing Model. Their enthusiastic and dedicated teaching aroused my interests in data mining technology and brought the production of the project.

A special thanks to Mr. Wei Dang and Mr. Shengzhi Zhao who are my previous colleagues and provided me with the data needed for the project. Without them, I would not have had the chance to experience the power of information and expand my view of the technology.

I would like to thank my MBA colleagues. They made my learning experience enjoyable and unforgettable.

Finally, I would like to thank my family. They supported me to pursue a MBA degree outside my home country and gave me the strength and endurance to face challenges.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Nothing is more important than quality for manufacturing companies in the 21^{st}. century. Quality determines a company's profitability and productivity, the competitiveness of its products, customer satisfaction and often the fate of the company (Bhote, 1991; Deming, 1975; Escalante, 1999). Globalization brings more opportunities for outsourcing, but raises the concern of standardized quality (Adams, 2001). Quality control in manufacturing covers several disciplines including industrial engineering, statistics, control theory and computer science. For a manufacturer, every step in producing and delivering products affects quality: from the preparation of raw materials, the development of product designs, the maintenance of equipment and tools status, to production process control, inspection, employee training etc (Bhote, 1991; Deming, 1975; Hinckley & Barkan, 1995; Kane, 1989; Shewhart, 1989). This paper focuses on the process control aspect of quality.

Technological changes have brought new dimensions to quality improvement. Data mining (DM) is one component of Knowledge Discovery that has established broad applications in industry (Adams, 2002; Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Marakas, 2003). Pattern finding and prediction have enabled manufacturers to detect and predict defects and improve quality (Adams, 2001). Recent research has emphasized the technical comparison of DM technologies, e.g. the variety of algorithms, and finding new application areas for these algorithms (Apte, Liu, Pednault & Smyth, 2002; Becerra-Fernandez, Zanakis & Walczak, 2002; Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996; Bradley, Gehrke, Ramakrishnan & Srikant, 2002; Han, Altman, Kamrani, Rong & Conzalez, 2001; Hirji, 2001; Kantardzic, Djulbegovic & Hamdan, 2002; Kumar, Mannila & Pregebon, 2002; Port, 2001; Smyth, Pregibon &

Faloutsos, 2002; Trafalis, Richman, White & Santosa, 2002). The practical use of DM has drawn little attention, however, from academics (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996; Hirji, 2001) probably because the knowledge discovery process is performed mainly by researchers whose profession is in statistics and data analysis (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). A fundamental question is what business value data mining technology can bring.

Among the limited practical papers of data mining technology, even fewer touch on the application of DM in manufacturing processes. This paper addresses this issue by integrating two areas of research, process quality control and data mining, in a manufacturing context. The research question is whether data mining techniques can be used to potentially improve process quality control. The research method is case study. A case will be presented where DM technology was applied to find defeat causes in a color-TV panel production line. Production control data as well as product quality data collected within a period of more than one month were analyzed using an industry leading data-mining tool, SAS Enterprise Miner (SAS, 2003a). This case study provides a first hand look at the application of data mining to process control of quality.

The paper is structured in six chapters. The second chapter reviews literature on quality control. Definitions of quality, causes and remedies of quality variation and defeat, quality control approaches and methodologies, and process control in manufacturing are discussed. The third chapter provides definitions of data mining, justifies why data mining is needed in practice, compares data mining technology with traditional statistical process control and discusses popular DM techniques, applications, and benefits. The fourth chapter of the study is devoted to the case study. This section includes a brief description of the production line. A discussion of data mining methodologies leads to a process of problem definition, data selection, data pre-processing, data analysis and interpretation of analysis results. Comments and evaluation of the results from the production line's managers and technicians were recorded. Chapter 5 discusses implications of the analysis results and looks at the strength and weaknesses of data mining tools from the process of mining. The paper concludes in chapter 6 with a discussion of limitations of the paper and suggestions for future research in the cross-disciplinary field.

CHAPTER 2. QUALITY

2.1 Definitions of Quality

Businesses and industries have different categories of quality (Weiss, 2002). The American Society of Quality (ASQ) defines quality as the "*increase of customer satisfaction, the reduction of cycle time and cost and the elimination of error and rework*" (ASQ, 2003b). According to ASQ, the performance and financial result of a company is the natural consequences of quality management. From a manufacturing perspective, quality is related to nonconformity and variation from standards (Escalante, 1999). Shewhart (1989) defined quality as conformance to specified standards; Deming (1975) and Box (1988) characterized quality as being associated with minimizing variation. Quality can be measured as product performance, options and features, maintainability, customer satisfaction, reliability, durability and reputation of a product and services (Tellier, 1978; Weiss, 2002). Six Sigma, a Motorola-invented quality control methodology, defines quality as the value realized for both customers and producers in every aspect of business, which reflects the needs of utility of the two parties (www.asq.com). Quality refers to both product quality and the quality of processes through which products are made and delivered (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996; Tellier, 1978). Quality improvement is one important goal for all product and service providers. To achieve the goal, a basic understanding of potential causes and remedies of poor quality from both theoretical and practical perspectives is required.

2.2 Causes and Remedies of Variation and Defeat

In his article *Quality and Productivity Improvement,* Escalante (1999) studied the philosophy of the most influential experts in quality control and provided a good summary of causes and remedies of variation and defeat that deteriorate product quality. He grouped the ideas from Deming (1975), Shewhart (1989), Bhote (1991), Kane (1989), and Hinckley and Barkan (1995) into seven variation and defeat causes: measurement, methods and standards, non-robust design, workforce, machines, quality policies and uncomfortable working conditions.

Deming (1975) argues that variation of product quality is caused by the faults of the system and managerial responsibility. He emphasizes the importance of understanding variation and identifying whether it is produced by special causes or common causes. Processes with common variations are considered stable and predictable. Quality improvement of these processes must come from the system. Processes with special variations are perceived to be out of control and have to be adjusted to improve the output quality. Deming suggests statistical methods should be applied in a process measurement to find the causes behind mistakes and to reduce the chances of misrepresentations of real causes. This statistical method is commonly referred to Statistical Process Control (SPC).

Box (1988) tends to find the sources of variation and defeat and build sound quality from preliminary stages of production, e.g. the design of products and processes. His idea is to use experiments that bring the mean to design products and processes, which is called Design of Experiments (DOE). Schmidt and Launsby (1994) also hold a traditional method of DOE and propose that it is important to understand relationships between output variables and input variables.

Taguchi (1986) classifies two types of quality control: off-line and on-line quality control. Design and production engineering departments can use off-line methods that are systems, parameters and tolerance designs to improve quality whereas production departments can use on-line methods that include process control, inspection and customer relations to improve quality. Taguchi's tools are orthogonal arrays, linear graphs and the quality loss function.

2.3 Approaches to Quality Control

From a practical point of view, there are three approaches (www. Qualitymag.com) to improve quality, all of which build on each other.

2.3.1 Metrology methods

Metrology methods seek to improve quality by using precise measurement or inspection devices and techniques to assure production accuracy and compliance with standards. Examples of the approaches are gage calibration, optical inspection, material testing, leaking testing and machine tool alignment and maintenance (www.qualitymag.com). Production processes that involve a variety of instruments need calibration to make sure that these instruments do not weary after daily use. Calibration procedure consists of determination of the accuracy of instruments, setting up calibration standard, keeping records of maintenance and calibration. Depending on the complexity of problems, calibration can simply mean calibrating instruments or deploying a calibration program (Bremmer, 1991).

2.3.2 Software and analytical tools

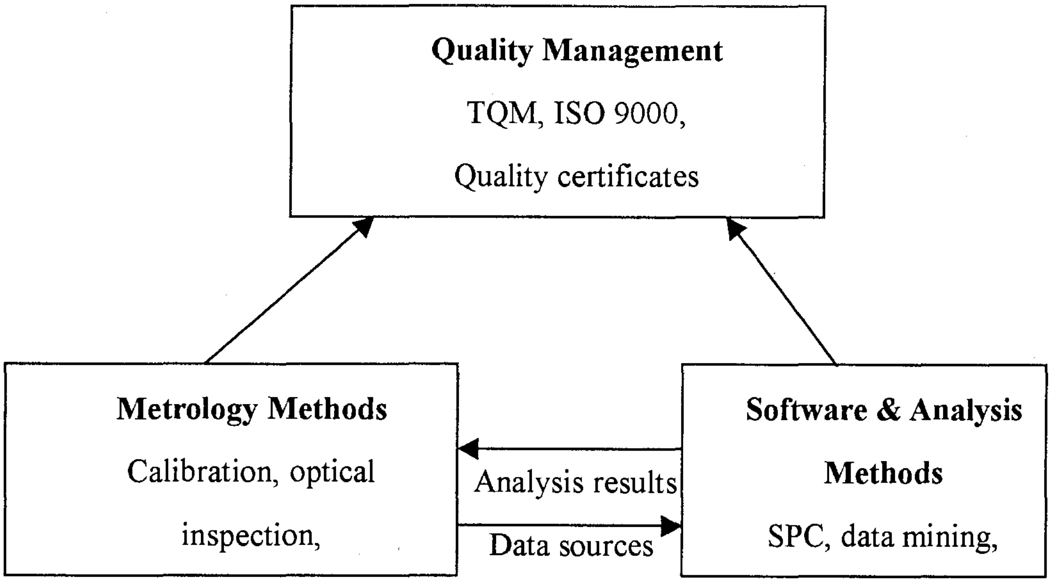
This approach intends to analyze defeat causes, find patterns and correct errors by using statistical methods and knowledge discovery techniques. Examples are statistical process control (SPC), data mining, process simulation, document control, and gage management. SPC and data mining have the same roots in statistics (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Hirji, 2001,). More comparison of the two techniques will be given in the later discussion. On one hand, analysis will not conclude plausible suggestions without proper data input that comes from metrological instruments and methods. On the other hand, the effectiveness of analytical results can only be realized through the use of process and inspection metrological instruments.

2.3.3 Quality management

This approach ties quality enhancement with employment of quality programs. An example of enterprise-wide quality programs is total quality management (TQM). Though there are different definitions of TQM, TQM generally means that good product quality relates to all organizational aspects (Packard, 1995). To build quality products or service, organizations need to improve leadership, human resource management, production and delivery processes, supplier relationship, and information management. Organizations should also be customer-focused and measure business results carefully (Sun, 2001). TQM is widely used in all industries as well as research and educational institutions and has achieved tremendous success (Galapon, E. A. & Norton, J. S. 2001; Imbellone & Lopez, 1997; Lowery, C. M., Beadles, N. A. II & Carpenter, J. B., 2000).

Well-recognized quality management certificates and standards are used as benchmarks of quality management programs. The most popular standard is the International Organization for Standardization ISO/QS 9000 (Sun, 2001). Other certificates include Quality Engineer/Technician/Audit/Manager Certificate, Six Sigma Certificate, and Calibration Technician Certificate etc (www.qualitymag.com). Consistent with Deming's argument (1975) that quality relies on the responsibility of management, quality management focuses more on the "soft" managerial side rather than the "hard" instruments and data analysis. Ideally quality management should be a high-level quality improvement approach incorporating metrological methods and software analysis and tools. Relations of the three practical approaches can be illustrated in Figure 1.

Figure 1. Relations of three practical quality control approaches



2.4 Process Control in Manufacturing

Process control has become a significant approach of quality improvement in manufacturing for a number of reasons. First, a process-centered view of products and service is essential to product quality (SAS, 2003c). Though manufacturers can utilize final inspections to prevent defeat products from being deliveried to customers, inspections can be costly and inefficient (Weiss, 2002). It has been found that the cost of remedying a problematic finished product is much higher than that of producing a quality product. Sometimes there is no way to correct mistakes but waste time and resources on faulty products. The best way to reduce a product's cost is to make it right the first time (Weiss, 2002). Instead of blocking the defect at the final inspection, process control at the factory floor level is less costly and more effective in that it reduces the causes of the defect from upstream. Second, manufacturing such as chemical, pharmaceutical, petroleum, and food industries usually involves large equipment and sequential processes. These processes may belong to different subsystems, e.g. water-supply subsystems, wind-supply subsystems, power-supply subsystems etc.

Effective process control can yield lower energy consumption in addition to high quality-stability (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996).

The number of variables is enormous within manufacturing processes. Raw materials quality, equipment and tools maintenance status, operators' skills and experience all play a role in forming product quality (Bhote, 1991; Deming, 1975; Hinckley & Barkan, 1995; Kane, 1989; Shewhart, 1989). Moreover the output of a process may become the input of the next process, which creates non-linear correlations between processes (Kang, Choe & Park, 1999). The complex nature of manufacturing processes demands the application of data mining techniques to analyze causes of variation and defeat and predict production patterns. The case study of this paper shows how data mining can be used to identify important process parameters in a complex production situation. These process parameters were found vital to product quality and should be monitored closely.

CHAPTER 3. DATA MINING

3.1 Defining Data Mining

Data mining is a multidisciplinary field related to areas such as artificial intelligence, database theory, data visualization, mathematics, operation research, pattern recognition and statistics (Berry & Linoff, 1997; Hirji, 2001; Smyth, Pregibon & Faloutsos, 2002). George Marakas (2003) and Larry Adams (2002) defines data mining as a component of Knowledge Discovery in Databases (KDD) that uses statistical analysis and modeling techniques to uncover patterns and relationships hidden in large databases and beyond the analytical scope of human beings. Fayyad, Madigan, Piatetsky-Shapiro and Smyth (1996) define data mining as a step in the KDD process which consists of applying computational techniques that produce a particular enumeration of patterns or models over the data. Adriaans and Zantige (1996) generalized different definitions into a basic tune that data mining is a process of searching through details of data for unknown patterns or trends. Furthermore, DM employs querying tools to enable users to send IF-THEN questions, and visualization tools that make analytical results easier for business users to understand (Kang, Choe & Park, 1999; Marakas, 2003).

3.2 Why Bother With Data Mining?

The huge size of datasets generated by manufacturers' quality control efforts makes it frustrating for people to capture useful information underlying the data (Adams, 2001). Data volumes grow exponentially with the number of records and fields in many domains. For example, databases in astronomical science may contain 10^9 objects. Data manipulation has gone beyond human capability (Fayyad, Piatetsky-Shapiro & Smyth, 1996). The second factor that makes data mining a necessity is that it is difficult necessity is that it is difficult to use traditional statistical software to analyze trends and patterns when data comes from different sources and is stored in different places (Adams, 2001). As in this case study, quality data and process data are gathered and stored by different operational databases. A separate section below will discuss the similarities and differences between traditional SPC and data mining and explore situations demanding data mining techniques. Because data mining is objective and fast compared to human analysis and is scalable and robust compared to SPC (Apte, Liu, Pednault & Smyth, 2002; Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996; Fayyad, Piatetsky-Shapiro & Smyth, 1996), it is not difficult to understand that this technology is appealing to more and more practitioners.

3.3 Data Mining and Statistical Process Control

Problems to be solved through data mining are inferences of patterns and models from data. Thus data mining is statistic in nature (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Glymour, Madigan, Pregibon & Smyth, 1996; Hirji, 2001). Statistics offers a great deal to evaluate the assumptions/hypothesis of data mining, to assess estimation results and to help in understanding the uses of the results (Glymour, Madigan, Pregibon & Smyth, 1996). An important feature of data mining results is the uncertainty associated with the finite sample size (Glymour, Madigan, Pregibon & Smyth, 1996). In other words, one cannot be sure of estimating accuracy from a model not extracted from the whole population sample frame. Statistical theory, however, *"provides measurements of uncertainty (e.g. standard errors) and methods of calculating them for various families of estimators"* (Glymour, Madigan, Pregibon & Smyth, 1996), which enables users to apply the results with certain degrees of confidence.

Nevertheless most SPC techniques comprise only a small number of *variables*, usually the quality variables in a manufacturing environment (Hirji, 2001; Kang, Choe &

Park, 1999). The factors that influence quality variables, e.g. parameters of machine and tools, are not clear, which means SPC cannot provide useful information to operators (Kang, Choe & Park, 1999). Thus statistical results alone are insufficient in modern manufacturing process control (Adams, 2001; Hinckley & Barkan, 1995; Kang, Choe & Park, 1999). In order to monitor and diagnose complex operational correlations, both process and product data should be extracted. Data mining techniques uses more parameters in its model and can provide better predictive results than SPC (Adams, 2001; Apte, Liu, Pednault & Smyth, 2002). An inductive algorithm of data mining, for example, can induce rules from processing data to help users understand the relations that SPC cannot disclose and to provide early warning of exceptional processes (Kang, Choe & Park, 1999).

The limited numbers of *datasets* that can be managed by statistical techniques also become a bottleneck when analysis involves large amount of data (Hirji, 2001). Data mining can handle large and multidimensional (multiple variables) data automatically (Fayyad & Uthurusamy, 2002). Moreover data mining is a step beyond SPC in that it not only looks for what happened in the past, but also tries to predict the future thought a process called *predictive modeling* (Adams, 2001). Traditional statistical methods are good at forecasting linear relations; data mining technology has higher predictability because it handles both linear and non-linear relations.

Traditional statistical data analysis relies on humans to formulate models, which requires prior knowledge of variable relations (Fayyad & Uthurusamy, 2002). However, data mining identifies interesting trends and relations about which people do not have precedent experience and knowledge through crunching data. Since sometimes *we do not know what we do not know* (Krider, 2003), the unknown area is where data mining works. In this sense, data mining techniques are "*dirtier*" than SPC (Andrew, 2002).

3.4 Data Mining Technologies

The underlying principle of DM technology lies three steps: (1) to use a training sample to develop models/decision rules; (2) to use a validation sample to confirm the fitness of the models; and (3) to use a testing sample to verify the models' predicting ability. Several data mining technologies prevailing in the market are discussed below.

3.4.1 Statistical analysis

Statistical analysis is the most mature DM technology. Regression is a commonly used statistical approach to search for linear relationships between variables and to predict the future (Marakas, 2003). A regression model can be represented as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + e$$

Where X is independent variables (predictors) and Y is an independent variable.

Compared to other DM technologies, statistical analysis is relatively easy to understand and grasp. However, the disadvantage is that traditional statistical methods can be problematic when the data is not described in a linear model. Solving for non-linearity requires advanced statistical methods that are fairly obscure for business users.

Another significant problem associated with statistical analysis is that the models are calibrated from historic data that may be invalid for dynamic business situations. To overcome this problem, neural network modeling with its non-parametric nature and its ability for continuous learning as new data is presented, serves as an evaluation mechanism for business decision (Walczak, 2001).

3.4.2 Artificial neural networks

Artificial neural networks (ANN) are the technology that simulates the human brain's cognitive learning process by trial and error (Marakas, 2003; www.sas.com). Trafalis et al. (2002) refer to ANN models as intellectual algorithms for tasks such as learning, classification, recognition, estimation and optimization that are based on the concept of how the human brain works. ANN is especially useful in prediction and commonly applied in credit risk assessment, direct marketing, and sales prediction (SAS, 2003b).

ANN consists of three components: input units, hidden units and output units. Input units obtain the values of input variables; hidden units perform internal computations and provide the nonliterary that makes neural network powerful; and output units compute predicted values and compare those values with the values of target variables (SAS, 2003b).

ANN reduces the need for input from domain experts (Marakas, 2003). A trained ANN model can be treated as an expert capable to answer WHAT-IF questions in new daily situations (Trafalis, Richman, White & Santosa, 2002). This is valuable in discovering new knowledge or relationships ahead of the capabilities of current technologies. ANN is capable of dealing with complexity and with non-linear relationships (Trafalis, Richman, White & Santosa, 2002; www.sas.com). It can also process noisy and incomplete datasets (Trafalis, Richman, White & Santosa, 2002). From a user's perspective, the avoidance of explicit programming and of detailed IF-THEN rule base in ANN facilities broad applications of the technology. On the other hand, ANN cannot be operated with explicit inferences processes (Marakas, 2003). It is hard to determine from its outputs which input variables are found to be significant. The fact that one cannot explain the results with confidence makes communication problematic (Becerra-Fernandez, Zanakis & Walczak, 2002). The development of ANN also depends on the advancement of existing hardware (Marakas, 2003).

3.4.3 Decision trees

Decision trees work by continuously breaking datasets into separate smaller groups according to predefined rules. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node.

The original group which contains the entire dataset is called the root mode of the tree whereas the final nodes are called leaves. These rules can be built on statistical significance or practices (Wong, 2002). In predictive modeling, a rule is simply a predictive value (SAS, 2003b). Decision trees can be useful in prediction and classification with mixed continuous and categorical data (Sarkis & Reimann, 1996; www.sas.com).

Bradley et al. (2002) believe the decision tree technique is attractive in a data mining environment. One reason is that human analysts can readily comprehend the resulting models. The other reason is that the construction of decision trees does not require input parameters. In contrast with ANN, rules derived from decision trees are written in English and can be used to communicate predictive models (Weiss, 2002; www.sas.com). By employing rule induction algorithms to build decision trees, one can easily inspect how variables can be effectively used to classify datasets (Weiss, 2002). Decision trees algorithms, for instance C5.0, accept missing values and do not require long training times for coming up with estimates. This results in an increase of the predictive ability of this data mining technique (Becerra-Fernandez, Zanakis & Walczak, 2002; SAS, 2003b). However, one limitation of the tree analysis is that it can become more complex when dealing with an increasing number of variables (Marakas, 2003).

3.4.4 Genetic algorithms

Genetic algorithms (GA) are à set of techniques based on the ideas of genetics and Darwin's evolutionary theory of natural selection (Marakas, 2003). GA searches and finds the optimal sets of parameters that can describe a predictive model. Similar to statistics, GA is a directed or supervised technique in that models need to be known before the search. GA is an optimization approach that can be applied to enhance neural network performance (Berry & Linoff, 1997).

3.5 Data Mining Applications

Data mining techniques have been developed in a variety of domains including marketing, finance, insurance, banking, customer services, manufacturing and telecommunications. Most of the applications use predictive models although a few of these applications use other methods (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996).

Marketing has been a long time user of knowledge discovery technology, and most marketing applications fall into a broad area called database marketing (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). This method employs customer databases, using interactive querying, segmentation and predictive models, to select target customers, design new products and predict market responses of marketing programs. Market basket analysis, a popular approach using pattern-finding techniques, studies retail consumer purchases based on point-of-sale information and directs targeted communication plans to improve the efficiency and effectiveness of marketing actions (Marakas, 2003). Statistical regression is widely used to optimize net profit and reallocate resources on sales force and advertisement (Lilien & Rangaswamy, 2003).

Many financial institutions employ predictive modeling techniques to optimize portfolio and estimate trading models (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). In this application, the predictive accuracy is vital; the need to explain a recommended portfolio is less important (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). Thus analysts use neural network to select investment instruments and to make business decisions blending domain experience with judgment.

Fraud detection systems are of special interest to banking and insurance. Neural networks help banks in detecting suspicious credit card transactions and assist insurance companies in identifying cheating on loss claims. Government sectors also use this technology to sense money-laundering activities. Fraud detection systems provide useful information by highlighting unusual patterns. Neural networks enable the systems to learn and combine detectors for optimal performance (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996).

In manufacturing, data mining techniques such as fuzzy logic, neural networks and decision rules are being used to improve the control of complex nonlinear dynamic processes (Fouhy, 1998). Processes that display nonlinear dynamics are difficult to model with rigorous approaches. Neural networks can build up a model upon existing data; fuzzy logic can learn from the experience of operations and technicians; and decision rules can find hidden patterns in data that are otherwise impossible to expose using human brain activities (Fouhy, 1998). Data mining can also help in the design of experiments (DOE). With the right factors found in historical data through neural net software, manufacturers of new product designs can get rid of testing data piece by piece in a laboratory (Adams, 2001). Rule-induction techniques can determine the relationships between control variables (e.g. quantities of raw material, heating temperatures etc), and quality variables (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). ' The value of this technology is elaborated in this case study where quality-indicators along a production line were recognized through data mining.

3.6 Benefits of Data Mining in Manufacturing

The benefits that data mining technology can bring to manufacturing are illustrated in Table 1. From a business perspective, data mining can easily handle large-sized datasets stored in the factories' operational databases or data warehouses. Aided by advanced data mining software, business users can benefit from the flexibility of digging and visualizing interesting information (Marakas, 2003). Commercial data mining software may also ease the workload of personnel managers in factories since the packages embed domain knowledge and may consequently require less experienced workers. The use of standardized platforms (e.g. Windows) ensures data mining software is potentially compatible with other organizational, operational or decision support systems. This has paved the way for more promising applications of the technology in manufacturing.

Table 1. Benefits of Data Mining in Manufacturing

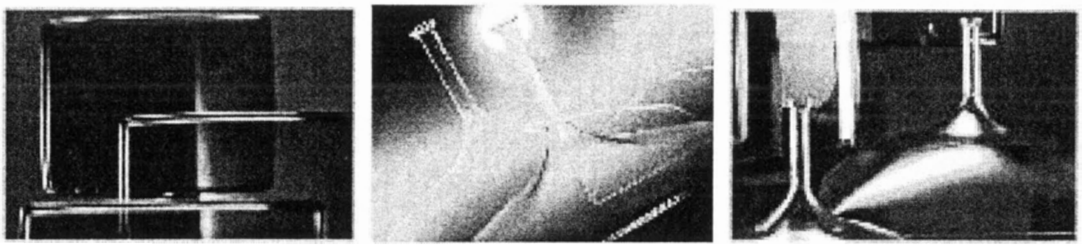
Availability of data	As more and more production, test and inspection equipment become more popular and tied to computer programs and enterprise-wide management software (e.g. ERP and MRP), there may be an increase in the amount of data that is ready to be extracted and analyzed in order to improve production performance.
Multi-sourced data	Data mining helps analyze the mountain of information stored in a variety of sources in a manufacturer's production line that may otherwise be difficult to grasp by human beings.
Advances of DM software	Data mining software allows users to drill down into data analysis for usable information and to generate reports hourly or daily rather than monthly (Adams, 2001). Data mining software often integrates domain knowledge and experience from many manufacturers, thus greatly reducing the time spent on learning curves and training costs.
Management	It is easier to manage machines or software than to retain people. Many manufacturers count on experienced technicians and workers to supervise production and inspect products. Their experience may be discontinued or lost due to lack of documentation and the resignation of employees. It is also true that tacit knowledge such as judgments and intuition may be hard to record and to store.
Integration	The emergence of Windows NT as the new operation platform allows tight integration of data mining techniques and expert systems with all parts of the process control environment (Fouhy, 1998).

CHAPTER 4. CASE STUDY

4.1 The Production Line

The production line belongs to a glass plant of a leading manufacturer of colour television tubes in Asia. The manufacturer mainly produces cathode ray tubes (CRT), whose sizes range from 37cm full square, 54cm full square, 64cm full square, 64cm super planar to 74cm super planar. The manufacturer also makes 40cm colour displays and 16cm projection displays. Ten million products are sold annually in both domestic and international markets. A CRT is the most important component of a colour television set and results in garnering about one third of total production costs (Jiang, 1999). CRT is a vacuum glass bulb equipped with electronics and consists of a panel (the front portion), a funnel (the back portion), an electronic gun, a neck linking the electronic gun and the funnel, and a shadow mask which is a metal screen that focuses the electrons on the back of the panel (Corning Asashi Video Inc., 2003). Figure 2 shows what panels, funnels and necks look like.

Figure 2. TV panels, funnels and necks

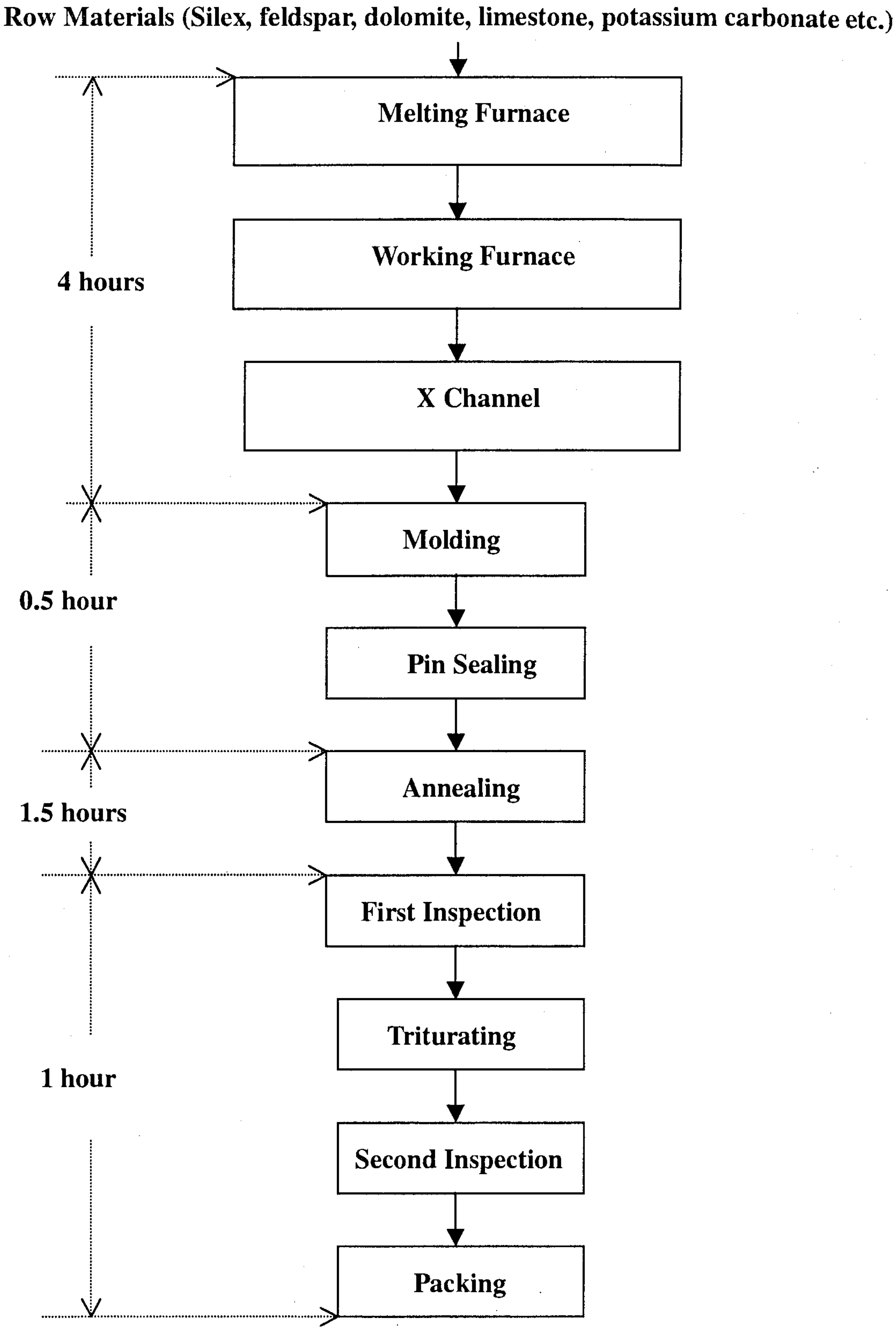


(Source: http://www.schott.com/display/english/products/crt/
courtesy of SCHOTT GLAS)

The glass plant, from which a 54cm panel line was chosen as the research object, is a key department of the manufacturer. The panel line's production process as well as the time slots desired for each process is showed in Figure 3.

The production process is divided into two big sections: the hot section which consists of a melting furnace, a working furnace, an X channel (for 54cm panel); and the cold section which includes a compressor (molding), a pin-sealing machine, annealing equipment and two inspections. The two sections are so called because the working temperature is over 1000 centigrade at the hot section and is lowered to the normal room temperature at the cold section. Line staffs are assigned into two groups accordingly. The production is non-stopped from the time the first batch of raw materials was poured into the melting furnace. The line staff works 24 hours per day divided into 8 hours per shift. The whole production process takes about 7 hours to produce ready-made panels with 4 hours in the hot section, 2 hours in the cold section and 1 hour for the two inspections and triturating. The following sections explain the different functions of individual process.

Figure 3. Panel production process



4.1.1 The hot section

As indicated by the name, the melting furnace is where pre-blended raw materials including smashed glass, silex, feldspar, dolomite, limestone, potassium carbonate etc. are melted. The melting furnace is divided into six burning zones, each of which has a burner firing mixed gas and air. The fire usually raises the inside temperature to a high of 1600 centigrade, which is the glass liquidizing temperature. The furnace's wall is made of silica bricks that resist heat, acid and alkalescency and can last for a minimum of ten years. The melting furnace creates two types of quality variation. *Stone* is the crystal mineral of residuals from raw materials and corroded bricks while *bubbles* are created from the air inside the glass when it is heated to high temperature. Stone and bubbles are the main causes of defeat produced from the hot section. They mostly come from the melting furnace even though the working furnace and the X channel also increase the chance of defeat if these are not well operated.

After the glass is turned into liquid, it flows into the working furnace whose sole purpose is to reduce the amount of bubbles. Theoretically, the longer the liquid stays in the working furnace, the less bubbles remain, which signifies fewer defeat. After several hours of de-bubbling in the working furnace, the glass liquid flows to the X channel.

The main function of the channel is to cool the glass flow down from 1600 centigrade to around 1000 centigrade, a suitable temperature for molding, which is done in the next process. The channel is divided into six zones, C, B, M, ST, R1, R2 and R3 zones, for easy administration and control. The space, bottom and side parts of each zone are supervised through temperature monitors, flow gauges and pressure meters embedded in the channel. There are a total of 54 spots within the channel where temperature, flow speed or pressure is collected on an hourly basis. Cooling air is blown into the channel to help lower the temperature. Two poles located in the C and B zones blend the glass flow continuously to make the temperature even. Though the channel does not help to reduce the amount of stones and bubbles, it may produce more defeat if the temperature is not held constant. This case study investigates how the channel parameters affect the defeat, stone and bubble rate from the 54-spot input and the first inspection output.

4.1.2 The cold section

The cold section starts with a compressor that presses the melted glass to panels. The compressor has an opening from which huge glass liquid drops one by one into six precise metal molds moving in a circle. The speed of dropping is constantly held under control. The compressor works under normal room temperature; thus, hot liquid drops are turning colder and solidifying in the molds. The solid panels are delivered to the pin sealing stage where four pins are inserted into four corners of a panel before it turns completely hard. Molding and pin-sealing can cause surface defeats on panels such as hollows, nicks and dot marks.

Annealing comes next in the sealing stage, cooling the panels entirely to normal temperature and providing them with designed physical and chemical characteristics. Annealing takes 1.5 hours and is the longest process in the cold section. Panels are almost completed after annealing and then sent to the first inspection. Before the panels are triturated, their appearances are not clear and some inside defeats are invisible in the first inspection. This is why the second inspection comes subsequent to triturating. Following the second inspection, good products are packed and delivered. Quality variables, which include numbers of total products, defeat panels and defeat products, are caused by different reasons (bubbles, stone, hollows etc.), and are recorded respectively at the two inspections.

4.1.3 Quality control practices

The panel line is automatically operated by industrial robots and controlled by man-machine interactions through several subsystems. There are the X channel control system, the mixing pole supervision system and the wind motor supervision system in the channel. An operational database, Distribution Control System (DCS), gathers data from the hot section and transfers it to a historical Oracle database. The Oracle database collects data from both the cold and the hot sections. Measurements in the whole process include temperatures, flows, pressures, levels (solid and liquid), gas, speed, power, viscosity and weight. Line technicians watch operational parameters through instruments like sensors, thermometers, gages and displays to ensure that each part of the line works normally. The hot section-caused defeat is the main source of defeat and therefore the DCS group receives timely reports from the two inspections and adjusts operational parameters through control devices.

Once the line starts to work, the glass flow is non-stop unless part of the line is broken down or big maintenance is needed to replace bricks in the melting furnace, which usually happens ten years after its start date. Once either inspection shows abnormal outputs, the DCS group has to look back at the line parameters 3 to 7 hours prior to the problematic output and make some adjustments. The outcome of these adjustments may not appear until after another three to seven hours. This time lag makes production control difficult. The daily supervision of the line status and occasional adjustments are based on the line staff's experience and historical data. For example, charts drawn from historical data may be used to show trends of variable variations. A normal approach employed by the quality control department of the glass factory in order to manage quality, is to summarize defeat causes and to set hard defeat limits from experience means for each process. Process technicians are responsible to keep their defeat under the limits.

4.2 Data Analysis

Researchers suggested assorted process methodologies for data mining that have no fundamental difference (Adams, 2002; Cabena et al., 1998; Champman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2000; Conzalez, 2001; Edelstein, 1999; Fayyad, Piatetsky-Shapiro & Smyth, 1996, Kamrani, Rong & Berry & Linoff, 1997). Kamrani et al.'s data mining methodology (2001), which includes problem definition, acquisition of background knowledge, selection of data, pre-processing of data, analysis and interpretation and reporting and use, was regarded as appropriate for this paper and was chosen to lead the case study.

4.2.1 Problem definition

Even though data mining techniques are the core of knowledge discovery, most researchers agree that this core only takes 20% of the effort of the whole process while the other 80% effort is endeavored into data preparation or problem definition (Becerra-Fernandez, Zanakis & Walczak, 2002; Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Hirji, 2001). Problem definition determines the goals of data mining, techniques and algorithms to achieve these goals (Fayyad, Piatetsky-Shapiro & Smyth, 1996), and the type of raw data that should be gathered and analyzed (Kamrani, Rong & Conzalez, 2001). The high-level goals of data mining fall into two categories: verification of users' hypothesis and discovery of new patterns (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). Furthermore, discovery incorporates description and prediction (Berry & Linoff, 1997; Fayyad, Piatetsky-Shapiro & Smyth, 1996). The descriptive goal focuses on presenting realities to users in an understandable way. The predictive goal does not care what happens in the past but provides users with future patterns.

In this case, 54 variables collected from different spots of the channel may affect final product quality. These variables are mainly temperature and flow pressure (glass, cooling air and combustion-supporting air). The high temperature working condition makes it impossible to directly observe changes happening inside the channel. From years of experience, line technicians know that the temperature variables around the mixing poles in Zone C and B are critical. These variables are FX9BSL, FX9BSR, FXBBNL, FXBBNR and FXAFC. Moreover, air pressures of the cooling air and the combustion-supporting air may influence these temperatures. FXGF, the total airflow of the natural gas in the channel is important as well. However, the mechanisms through which these parameters work jointly and whether other variables may affect the defeat rate are unknown. The goal of data mining is to (1) verify the existing knowledge, and (2) disclose other strong relationships or control rules between operational variables and quality variables. The discovery of new patterns focuses more on describing relationships than on predicting.

4.2.2 Acquisition of domain knowledge

This step includes learning prior knowledge and understanding data. Learning prior knowledge is necessary in any information technology project as it prevents data mining from relearning certain patterns or relations that already exist (Kamrani, Rong & Conzalez, 2001). Knowledge of the production line and the business problems were gained through the author's working experience with the glass factory, field trips, and direct and indirect interviews with the production line engineers and managers. During the preparation of the project, international calls were made every week between the author and her previous colleagues. The main topics surrounded the production process, how operation staff keeps quality under control, what data mining may contribute, and what types of data are needed in order to do so. In the event the author's colleagues had been unable to answer a critical question, a call was made to line technicians or managers for clarification.

4.2.3 Selection of data

Subramanian et al. (1997) and Marakas (2003) proposed to use data warehouse as an ideal aid for selecting potential data in data mining. The production line's quality and operational data, as described, are stored in an Oracle database, which eased the task of data collection. Kamrani et al. (2001) suggested that selection of data should be open-minded because *the purpose of data mining is not for human to solve the problem, but rather to let the data speak for itself.* They further stated that selection should be narrowed down in order to approach the range where the potential conclusions go. Obviously, a trade-off should be made in this step to center on the target problems without losing useful information. The quantity of data should also be sufficient to generate meaningful knowledge. For example, a small sample size in categorization neural network modeling can reduce the quality of models (Becerra-Fernandez, Zanakis & Walczak, 2002). This occurs because the data has to be divided into a training sample and a validation sample; a low training/validation ratio normally decreases the performance of the models.

The quality data from the first inspection and the channel data were extracted for 44 days from the Oracle database into MS Excel. Since the goal is to find critical factors, all of the 54 channel variables were included in the data analysis.

4.2.4 Data pre-processing

The first step is to clean noise and to convert data, where necessary, into an alternative form for better analytical results (Fayyad, Piatetsky-Shapiro & Smyth, 1996). New attributes may be created in the dataset. Kamrani et al. (2001) held that data mining algorithms call for pre-processing data for a couple of reasons. Certain relations between data (the one-to-many relations in databases for example) may be difficult for data mining algorithms to assert. Algorithms may take longer time to verify certain assertions. Therefore if the knowledge is readily obtainable, it is efficient for analysts to add the relations as an attribute into the dataset.

Raw data

The raw data was included in two files. An INPUT worksheet consisted of 54 variables taken from the channel hourly. The OUTPUT worksheet had all kinds of quality-related variables including the number of defeat. Although there are many ways to measure quality in manufacturing, the defeat rate is the most popular. In this case, the OUTPUT data did not carry any defeat ratio. Thus, a new variable called DEFEAT was created by taking the proportion of hourly defeat products to the hourly total products. As the main sources of defeat from the hot section where the X channel positions are stones and bubbles, STONE and BUBBLE were also computed from the OUTPUT data to indicate the stone-defeat rate and the bubble-defeat rate. Thus, DEFEAT, BUBBLE and STONE are the three target variables in the analysis.

Data preparation

The raw data is relatively clean and has only 2% missing values since it was extracted from a database. However, the panel production nature indicates that data pre-processing is inevitable. There exists at least a two-hour lag between the channel, where the input variables were gathered, and the first inspection, where the target quality data was used in the analysis. The input variables and the three targets were collapsed into one worksheet by moving the targets two hours earlier than they were actually taken, e.g. 4:00pm targets correspond to 2:00pm channel variables. Based on the collapsed data, two new datasets were generated to facilitate the analysis.

BINARY (see Attachment 1) has all the raw data and three new binary targets BIDEFEAT, BIBUBBLE and BISTONE. For the clarification purpose, Attachment 1 only shows an incomplete part of BINARY with only two input variables (there are actually 54) and 23 records (there are actually 1063 records). The binary targets were produced to make decision trees easy 'to work because decision trees are good for classification problems. Moreover since regression model with an interval target cannot have a lift chart in SAS, a logistic regression with a binary target can be constructed to compare lifts with decision tree and neural network models. The three binary targets were made by initially calculating the mean of the defeat/bubble/stone rate and then setting the rates higher than the mean "BAD" and those lower "GOOD". As 75% of the defeat was caused by the presence of bubbles, BIDEFEAT and BIBUBBLE are mostly identical. This data conversion can be accomplished in both MS Excel and SAS Transformation node. In order to run preliminary analysis in Excel, a **CLEAN** dataset was produced by deleting the 2% missing value from the collapsed dataset.

4.2.5 Analysis and interpretation

The analysis starts with a preliminary analysis using simple statistical methods. A correlation matrix and autocorrelation were run in MS Excel. Then SAS Enterprise Miner was utilized to run regression, auto-regression, decision trees and neural network. Every data mining method was operated at least twice with different objectives, mostly for comparison purposes. The following discussion elaborates each analysis and findings.

Correlations in Excel

The correlation matrix run on the CLEAN dataset found the following variables had higher than 0.2 correlations with the targets: FX9BSL, FXAFR11 and FXAFR12 with DEFEAT; FX9BSL, FXAFR11, FXAFR12 and FXGF with BUBBLE; and FX9BSL and FX9BSR with STONE (see Appendix 2. Correlation results). The result may indicate these variables are important predictors for the defeat rate. In fact, FX9BSL and FX9BSR, the temperature in the C zone, and FXGF, the natural gas pressure, are perceived important by line technicians.

Auto regression in Excel

Auto regression is a method to test time effects by studying correlations among observations of a single variable. An autoregressive model is simply a linear regression of the current value of a series against one or more prior values of the series. Auto regression can be analyzed with various models including standard linear least square techniques (NIST, 2003a).

The production is a continuous, non-stop process, which suggests that there may be time effects among the defeat rate. The defeat rate at a time t may be influenced by that at time t-1, t-2, ... A time series model of defeat rate might be:

$$DEFEAT_t = \alpha + \beta_1 DEFEAT_{t-1} + \beta_2 DEFEAT_{t-2} + \beta_3 DEFEAT_{t-3} + \ldots + \beta_n DEFEAT_{t-n}$$

Where β is the autocorrelation coefficient estimate,

n is the time lag,

$DEFEAT_t$ is the defeat rate at the time t,

$DEFEAT_{t-1}$, $DEFEAT_{t-2}$, $DEFEAT_{t-3}$ and $DEFEAT_{t-n}$ are the defeat rates one hour, two hours, three hours and n hours earlier than t.

Autocorrelation is related to a time lag that indicates how often the time pattern repeats itself. For the above mathematic model, the defeat rate repeats the time pattern every (n+1) hour. If β_n is statistically significant, $DEFEAT_{t-n}$ influences $DEFEAT_t$. To verify the hypotheses that time effect exists, an autocorrelation coefficient was gained by running regression in Excel on a new dataset. This dataset called **BIAUTO** was created by rearranging the BINARY dataset. Please see Appendix 3. BIAUTO dataset. Similar to the BINARY dataset, this appendix only shows part of the actual dataset.

First, it was decided that 6-hour lags would be used to run the regression. The choice is not rigorously grounded. As the production line shifts three times a day and every shift is 8 hours long, choosing a lag shorter than 8 hours may help line managers look at possible time trends more easily from a single shift report. Meanwhile, since the analytical goal is to show the existence of the time effect and not to build a model, 6-hour time lags can make the analysis simple enough without losing useful information by setting a limit to the most adjacent time lags, e.g. t1 and t2.

Second, a new dependent variable called AUTODEFEAT was created by copying the DEFEAT and deleting the first five defeat rates. Then, several new variables called dt1, dt2, dt3, dt4 were created by removing the first four, the first three, the first two and the first defeat rates from the DEFEAT column. The new variable dt5 is identical to the DEFEAT. These new variables represent the defeats rate at t-1, t-2 ... and t-5. AUTOBUBBLE, AUTOSTONE and their related time variables are completed in the same manner. Lastly, the first five observations of input variables are removed in order to make everything start at the sixth time slot. Only AUTODEFEAT and dt1 to dt5 were used to run the autocorrelation regression in Excel. The result is showed in Table 2.

In using the 90% confidence level, the P-values show that almost all time-related variables are significant. In using 95% confidence level, t1 and t5 are significant. The results confirmed the existence of time effect in the defeat rate. The presence of time effect indicates that the dataset is not a simple random walk. Most statistical analysis is based on an assumption that sample data is collected randomly. The validity of the estimation result is subject to the validity of the randomness assumption. If randomness is violated, the estimation of coefficient is invalid (NIST, 2003b). Therefore, time series model, for example $DEFEAT = f(t, x)$ where x is input variables, should be more appropriate if a prediction model is to be built. Since the goal of this data mining is to discover relationships and not to predict, in-depth discussions about time series model are disregarded in this paper.

Table 2. Autocorrelation Regression Result

Regression Statistics	
Multiple R	0.357815
R Square	0.128032
Adjusted R Square	0.123887
Standard Error	0.03219
Observations	1058

ANOVA

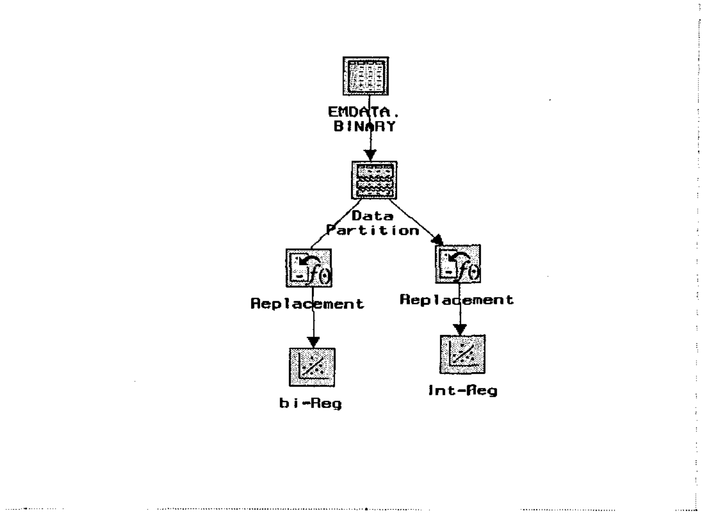
	df	SS	MS	F
Regression	5	0.160053188	0.032010638	30.89315802
Residual	1052	1.090053365	0.001036172	
Total	1057	1.250106554		

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.021292	0.00223858	9.511453665	**1.23367E-20**
dt1	0.284724	0.030759728	9.256392008	**1.15083E-19**
dt2	0.056591	0.031951569	1.771156818	**0.076824102**
dt3	0.052046	0.031944777	1.629236614	**0.103562319**
dt4	0.049894	0.031939186	1.562161646	**0.118550667**
dt5	0.063022	0.030756445	2.049069045	**0.040703003**

Regression in SAS

Two regressions from SAS Enterprise Miner were run on the same dataset **BINARY** (Figure 4). One regression, bi-reg, sets the binary BIDEFEAT as the target in order to generate a lift chart whereas the other, int-reg, chooses the interval DEFEAT as the target to compare with auto-regression, which will also use interval targets but run on the BIAUTO dataset.

Figure 4. Regression diagram

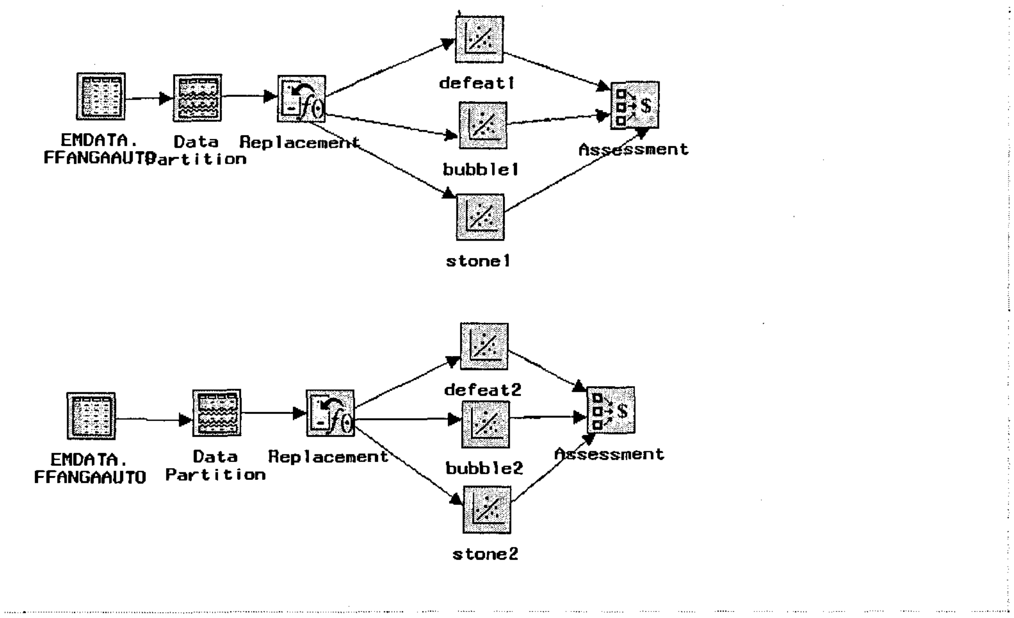


The comparison between the bi-Regression, trees and neural network will be discussed in the following section. The int-Regression results a very low adjusted R square 6.8%, which suggests that a linear regression model is not reliable to explain the defeat rate, probably because the time factor plays a role in causing defeat (see Appendix 4. Regression results).

Auto-regression in SAS

To further confirm the time effect on the defeat rate, the bubble rate and the stone rate individually, two sets of regression were run by the SAS Regression nodes on the **BIAUTO** dataset (see Figure 5). The time-related variables, dt1 to dt5 for the overall defeat rate, bt1 to bt5 for the bubble rate and st1 to st5 for the stone rate are incorporated in auto-regressions.

Figure 5. Auto-regression diagram



These auto-regressions and the linear regression mentioned above used the Replacement node to replace missing values in the target variables (there is no missing value in the input variables). These missing values resulted when the defeat rate was unusually high and people were not sure what was going on. In order to find possible causes and not waste raw materials and labour, all the line facilities were kept running without depositing raw materials. So the input data was available but the hourly production was zero, which makes the defeat rate function result to null values. Since the highest defeat rate in the dataset is 25.3%, an arbitrary number 30% was assigned to substitute the missing AUTODEFEAT. The missing AUTOBUBBLE and AUTOSTONE were replaced in the same manner to 20% and 10%. Although the replacement values are subjective, the assignment should not have great impact on results as only 2% missing value emerges in the dataset. Since regression tends to eliminate missing values, replacement is performed to keep as many observations as possible in the dataset. As the auto-regressions uses interval defeat/bubble/stone rate, no lift chart is available for the final assessment. The Assessment node is placed merely to make the three models run one by one automatically rather than compare their lifts.

The first set of auto-regression includes all input variables and the output of AUTODEFEAT or AUTOSTONE or AUTOBUBBLE respectively. These models can be expressed as below:

$$AUTODEFEAT = \alpha + \beta_1 dt1 + \beta_2 dt2 + \beta_3 dt3 + \beta_4 dt4 + \beta_5 dt5 + \delta_1 \chi_1 + \delta_2 \chi_2 + \delta_3 \chi_3 + ... + \delta_{54} \chi_{54}$$

$$AUTOBUBBLE = \alpha + \beta_1 bt1 + \beta_2 bt2 + \beta_3 bt3 + \beta_4 bt4 + \beta_5 bt5 + \delta_1 \chi_1 + \delta_2 \chi_2 + \delta_3 \chi_3 + ... + \delta_{54} \chi_{54}$$

$$AUTOSTONE = \alpha + \beta_1 st1 + \beta_2 st2 + \beta_3 st3 + \beta_4 st4 + \beta_5 st5 + \delta_1 \chi_1 + \delta_2 \chi_2 + \delta_3 \chi_3 + ... + \delta_{54} \chi_{54}$$

Where δ_m is the coefficient estimate of the input variable χ_m and m is from 1 to 54 (54 input variables).

The results of three models are shown in Appendix 4 Auto-regression result (1). The findings indicate that auto-regression has a much higher adjusted R square than the linear regression based on the raw data. This suggests that the auto regressive model more accurately predicts the defeat rate. The adjusted R squares are 74.6% for AUTODEFEAT, 71.3% for AUTOBUBBLE and 62.2% for AUTOSTONE. The time-related variables, e.g. t1 and t4, are statistically significant across the defeat/stone/bubble model. In particular, t1 affects the bubble rate more and t4 affects the stone rate more. Meanwhile, auto-regression indicated another significant variable across three models, which is FXAFC, the temperature of the combustion-supporting air in the C zone. These models also include the effect of FXGF, FXAFM and FXAFR11, but their coefficients are not significant.

To test the effects of the significant variables found from Correlation matrix, the second set of auto-regression includes only these variables with FXAFC as the input. These models can be depicted as:

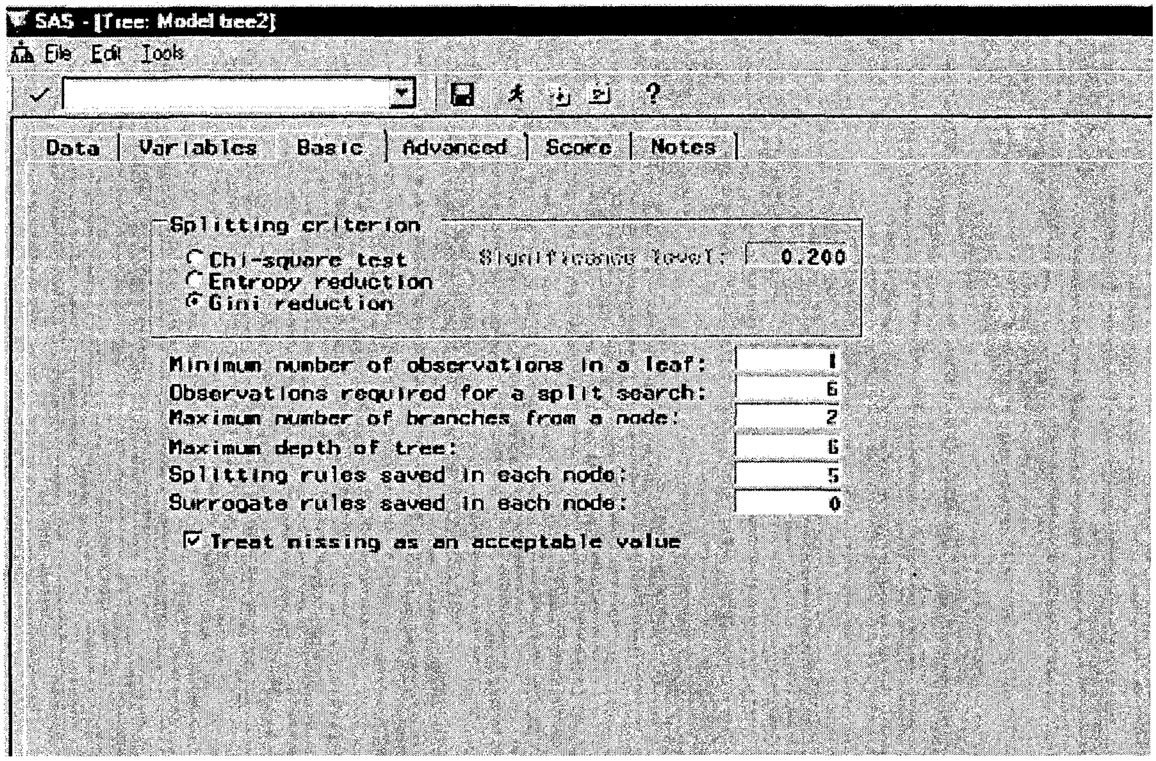
$$AUTODEFEAT = \alpha + \beta_1 dt1 + \beta_2 dt2 + \beta_3 dt3 + \beta_4 dt4 + \beta_5 dt5 + \delta_1 F3019BSL + \delta_2 F301AFC$$
$$+\delta_3 F301AFR11 + \delta_4 F301AFR12$$
$$AUTOBUBBLE = \alpha + \beta_1 bt1 + \beta_2 bt2 + \beta_3 bt3 + \beta_4 bt4 + \beta_5 bt5 + \delta_1 F3019BSL + \delta_2 F301AFR11$$
$$+\delta_3 F301AFR12 + \delta_4 F301GF + \delta_5 F301B_1 + \delta_6 F301B_2 + \delta_7 F301AFC$$
$$AUTOSTONE = \alpha + \beta_1 st1 + \beta_2 st2 + \beta_3 st3 + \beta_4 st4 + \beta_5 st5 + \delta_1 F3019BSL + \delta_2 F3019BSR + \delta_3 F301AFC$$

The three models show once again high adjusted- R squares 75.3%, 74.4% and 62.8% (see Appendix 5. Auto-regression results (2)). The significant variables are t1, t3, t4, FXAFC, FXAFR11, FXAFR12, FXGF and FXB2. The appearance of more significant variables in this auto-regression may be due to the possibility that the interactions among variables may be so strong in the first set of auto-regression that the strongest single variable is FXAFC. The finding is also consistent with that of the first set of auto-regression in that time plays a role in determining the quality.

Decision trees in SAS

A decision tree model can apply different splitting criteria and options that clearly distinguish results (see Figure 6). Several combinations of the splitting criteria and options were attempted. In an effort to find an optimal tree, the Transformation node was utilized to transform selective variables to make their distributions normal. The assessment confirmed the non-transformed tree with 60/20/20 partition for training/validation/testing sample, Gini reduction splitting rule, maximal 2 branches at each split, and maximal 6 layers produced the highest lift and thus had the best combination.

Figure 6. Decision tree splitting criteria and options



Splitting criterion

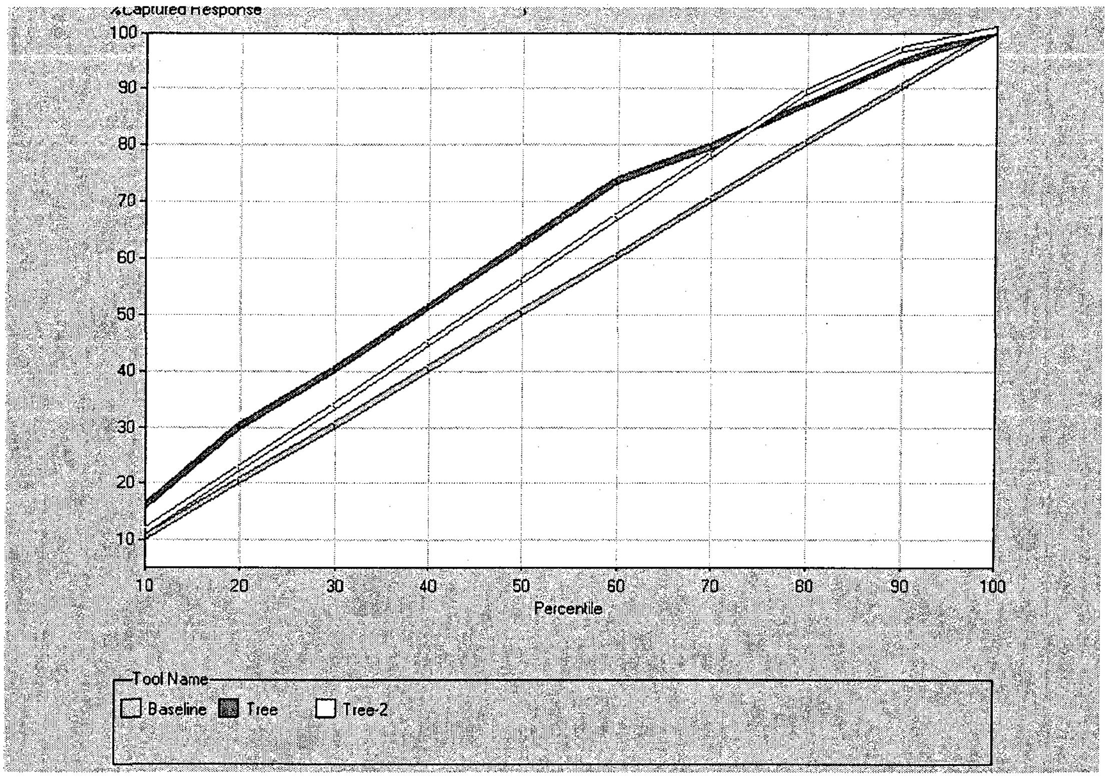
○ Chi-square test   Significance level: 0.200
○ Entropy reduction
● Gini reduction

Minimum number of observations in a leaf: 1
Observations required for a split search: 6
Maximum number of branches from a node: 2
Maximum depth of tree: 6
Splitting rules saved in each node: 5
Surrogate rules saved in each node: 0

☑ Treat missing as an acceptable value

Both BINARY and BIAUTO datasets are used to run decision trees, one with and the other without time adjustment. Their basic settings are the same as the best combination. The target variables are BIDEFEAT that is either "GOOD" or "BAD" depending on whether they are lower or higher than the mean of defeat rate. It is hypothesized that the BIAUTO tree should perform better than the BINARY tree because auto-regressions strongly support the time effect in the sample. The comparison of lifts of the two trees does not prove the hypotheses: the BINARY tree outperforms the BIAUTO tree overall with an exception at high percentile where BIAUTO behaves better than BINARY (see Figure 7, Tree 1 is the binary one.).

Figure 7. Lift chart for the BINARY and BIAUTO trees

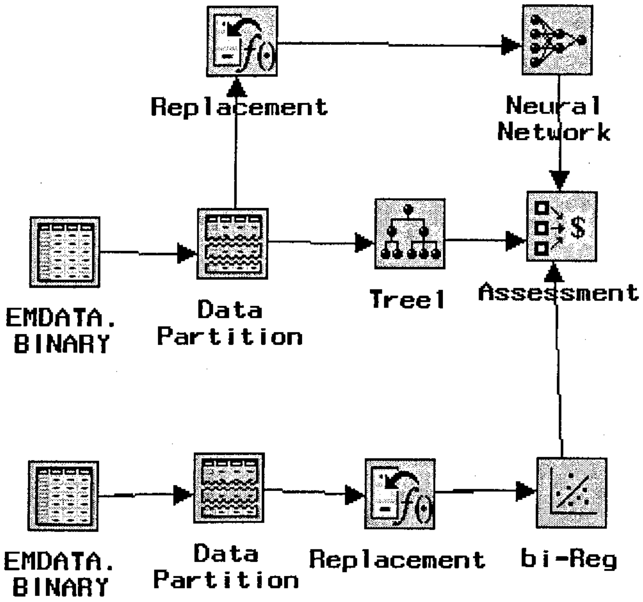


Neural network in SAS

The same method in which the best tree option is chosen is applied to decide the best neural network options. It is concluded that the default neural network with 3 hidden layers is the best. Replacement node is used to substitute the missing values, as neural network cannot handle them either. The BINARY dataset has a better lift in decision trees and is used to run the neural network.

To decide the best model across models, bi-regression, the BINARY tree and the neural network are assessed (Figure 8).

Figure 8. Diagram of decision tree, regression and neural network



Lift charts are generated as shown in Figure 9 and Figure 10. Obviously, the BINARY tree presents a higher lift and therefore is used as the base to find possible quality predictors and interpret business rules for quality control.

Figure 9. Capture rates of decision tree, regression and neural network

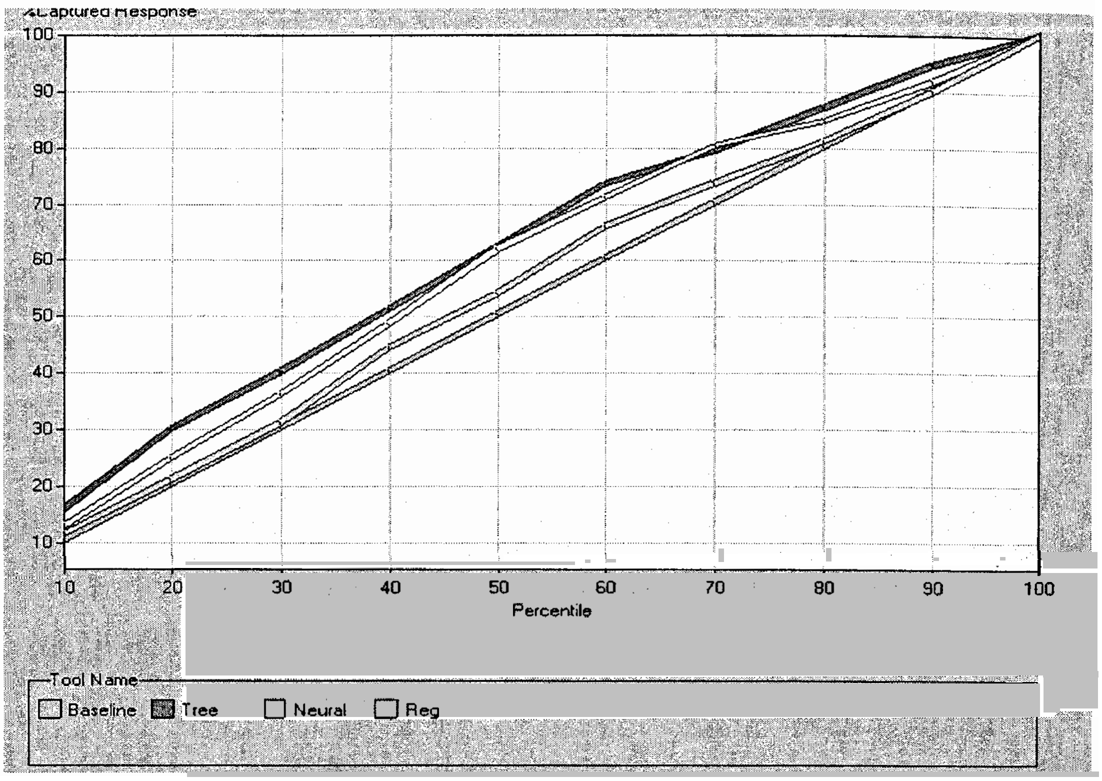
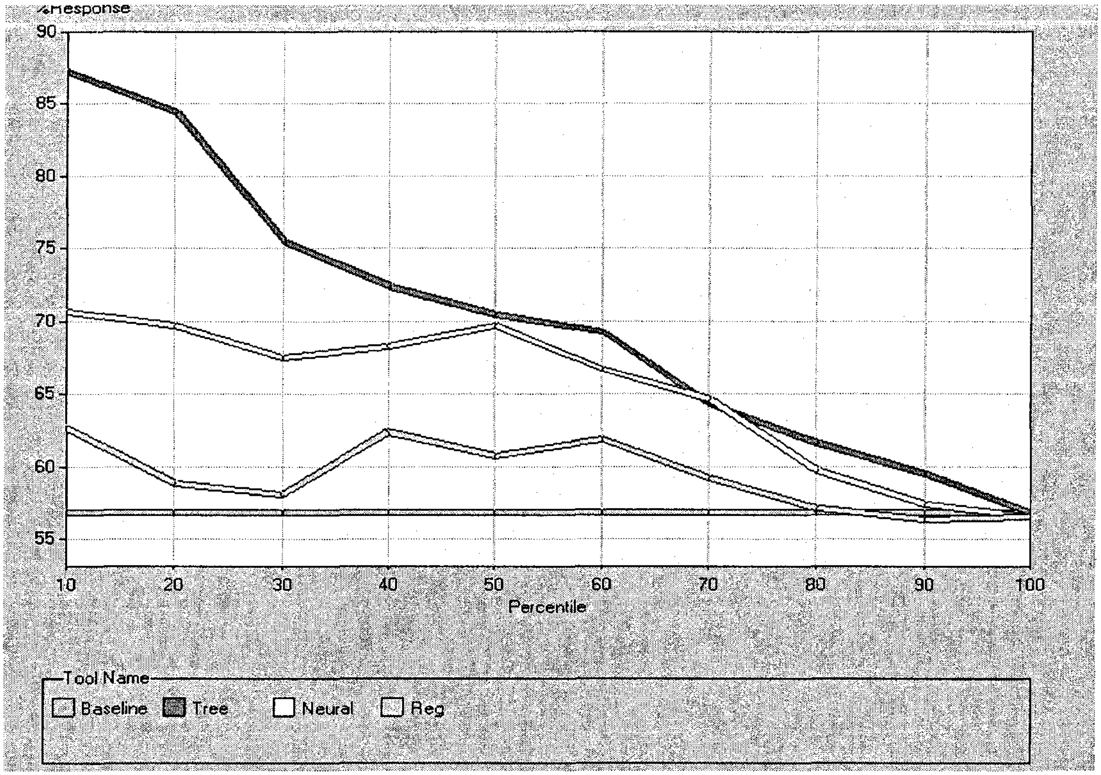


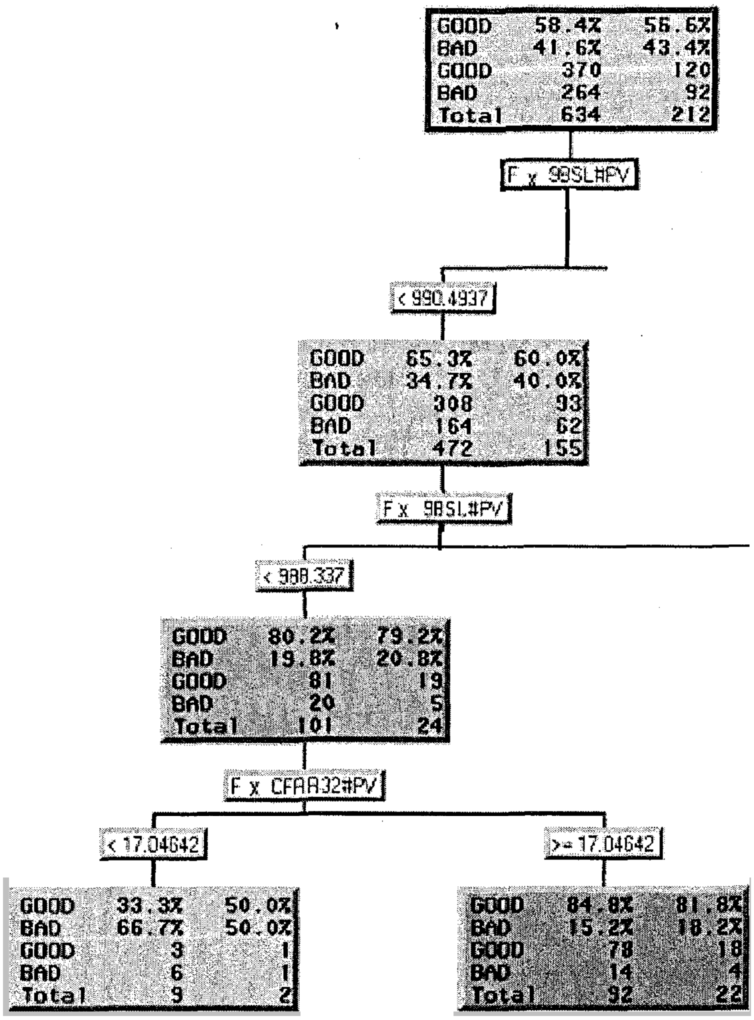
Figure 10. Response rates of decision tree, regression and neural network



Tree interpretation

The tree diagram (Figure 11) shows splitting variables and their levels to split. Each node contains two sets of sample, the training sample (the middle column) and the validation sample (the right column), and indicates the percentages of correct classification through splitting variables.

Figure 11. Tree diagram



The top node is called a root node which contains 634 observations in the training sample and 212 in the validation sample. The cutoff of BAD and GOOD in the root node is 54.4% and 41.6% because the mean of the target is used to separate the dataset. The first splitting variable is FX9BSL, the most significant variable from the correlation matrix. The level of the split is 990.4973. The final nodes are called leaves.

The optimal tree selected by SAS under predefined criteria has 21 leaves. The whole optimal tree was divided into four branches (Figure 12-1, 12-2, 12-3, 12-4.

The optimal tree). Please refer to the CD attached on the back cover of the project for a clearer picture. The splitting variables are FX9BSL, FX8BS, FXCFRR32, FX1A, FX6A, FXCFST, FXCFRST, FX3BS, FXCFM, FXAFM and FX2BS, among which FX9BSL, FX8BS and FXCFM are used twice at different layers.

Tree node statistics include splitting variables and IF-THEN rules that may be interpreted into business rules to help line technicians understand the relationships between them and to improve quality. An example of node statistics is shown below:

```
IF FX3BS#PV < 1011.242
AND FX6A#PV < 1063.976
AND 959.9789 <= FX8BS#PV
AND 990.4937 <= FX9BSL#PV
THEN
    NODE   :    22
    N      :    14
    GOOD   :   7.1%
    BAD    :  92.9%
```

Usually the redder the leaf's colour, the more predictive validity the leaf has. But if a leaf has a small number of observations in the training/validation sample, it may be less represented and ignored. It was decided that the number of observations of either training samples or validation samples should be more than 20 in order to keep a leaf for interpretation. Based on this criterion, four leaves are chosen to extract the business rules.

Figure 12-1. The first branch of the optimal tree

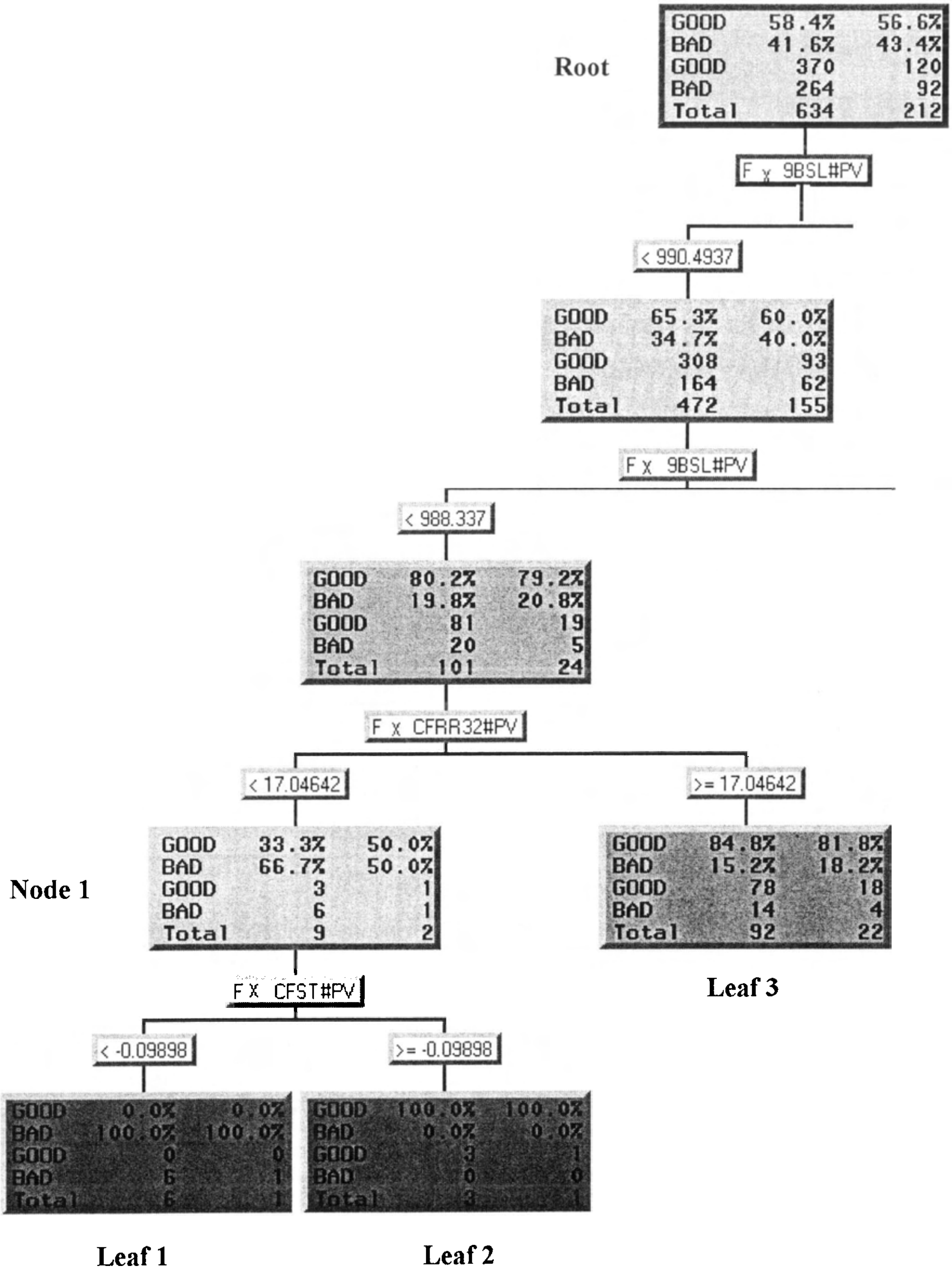


Figure 12-2. The second branch of the optimal tree

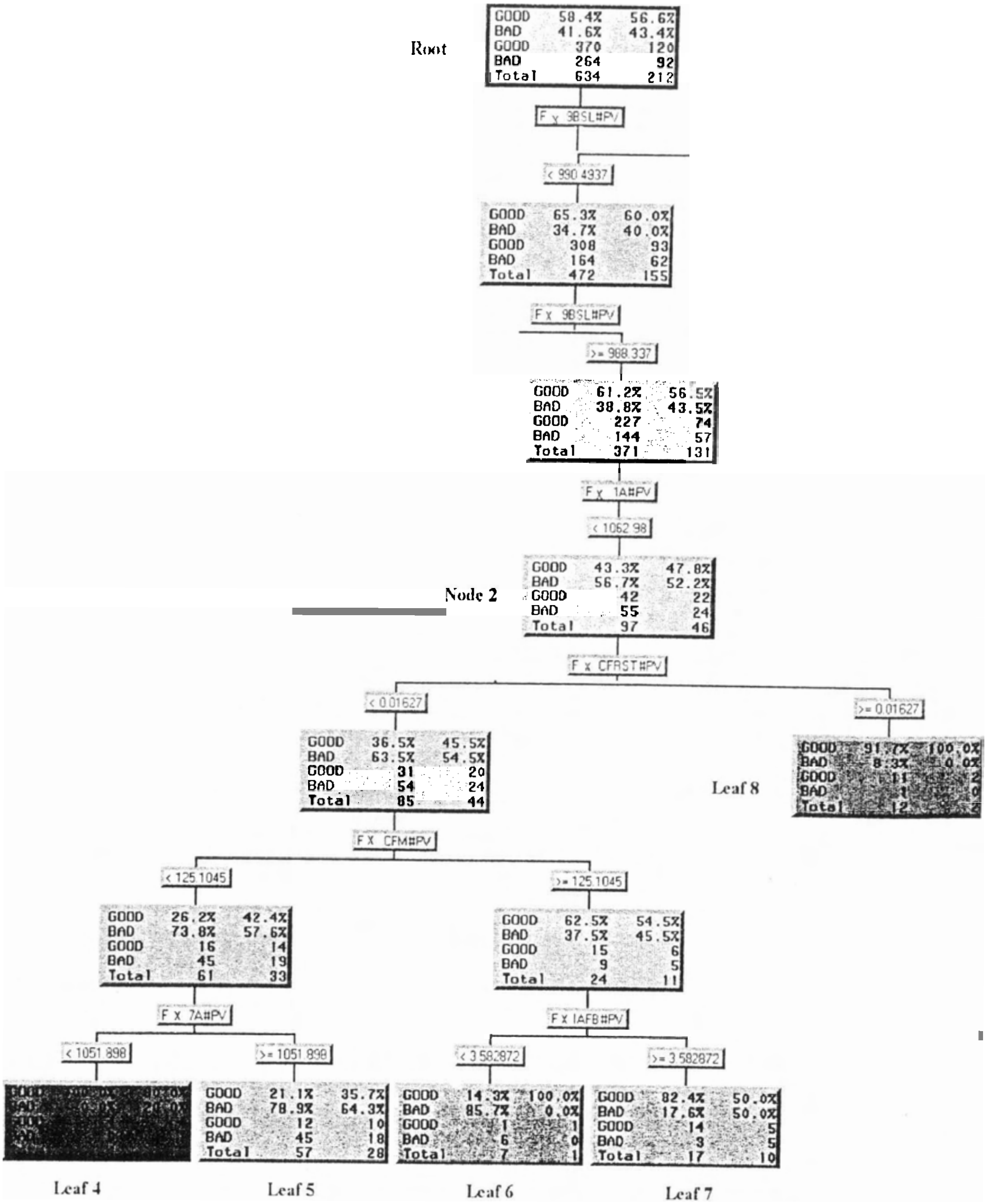


Figure 12-3. The third branch of the optimal tree

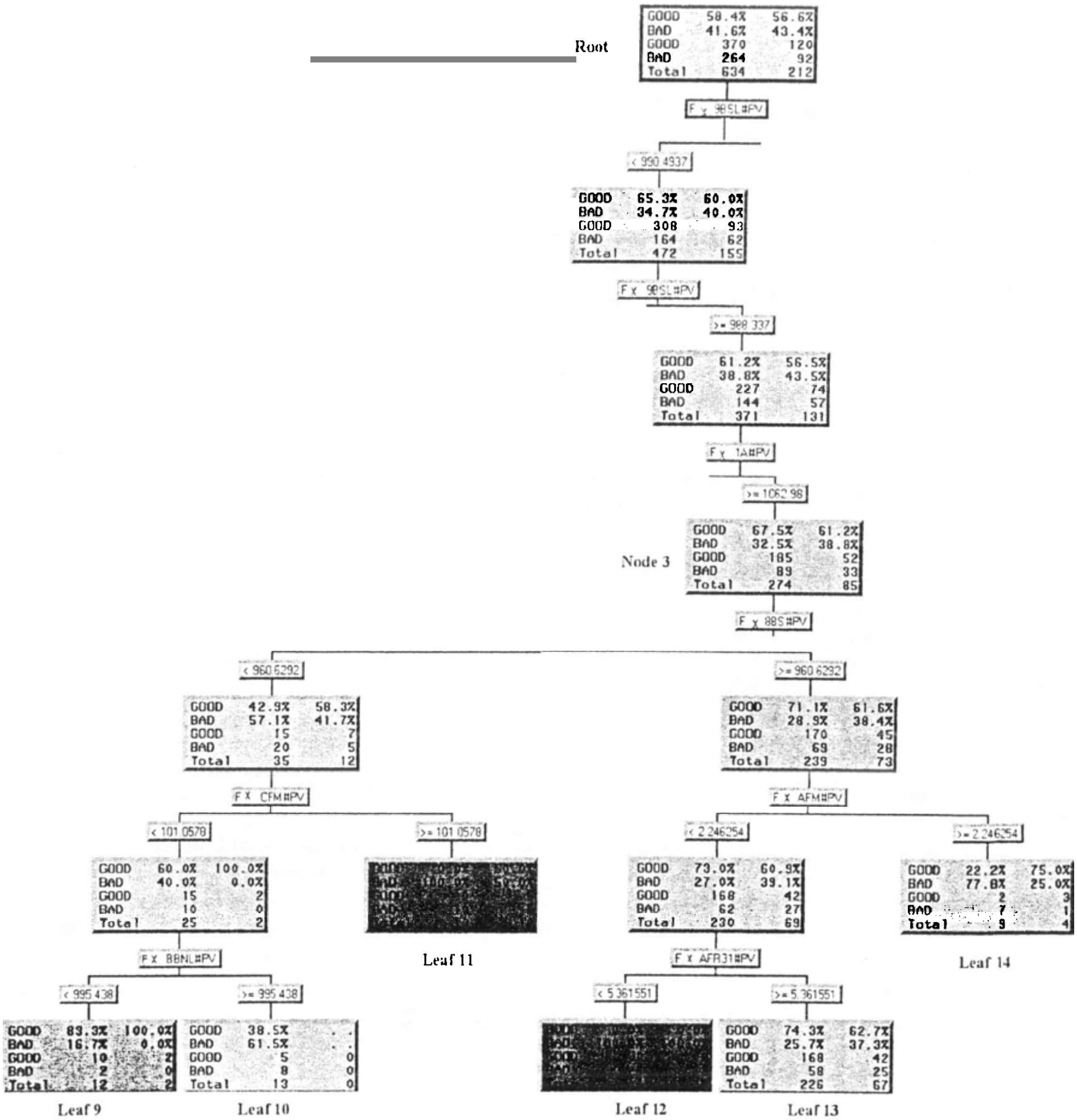
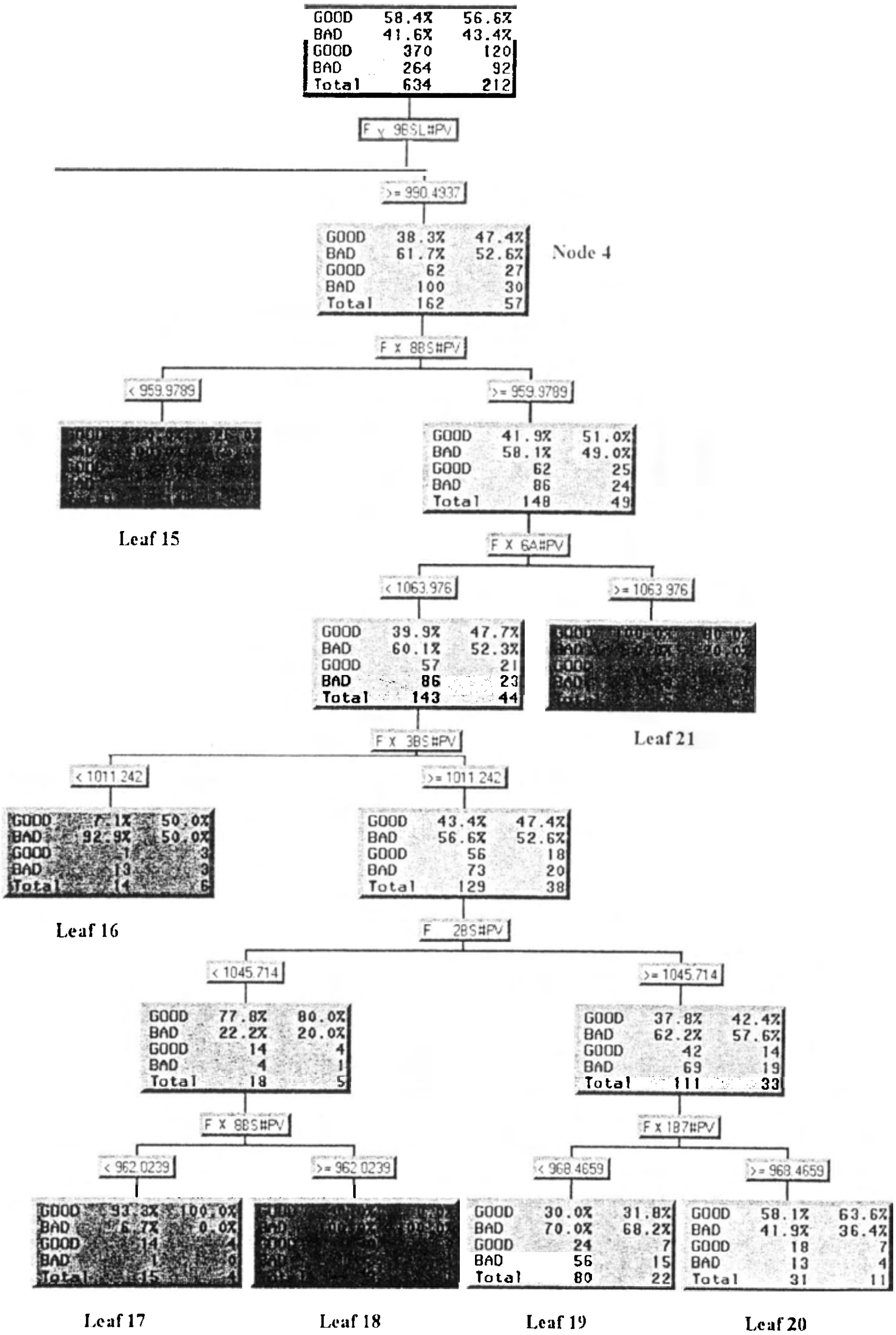


Figure 12-4. The fourth branch of the optimal tree



The first rule is in Leaf No. 3: "if FX9BSL < 988.337 and FXCFRR32 >= 177.04642, the probability of GOOD is above 80% (84.8% for the training sample and 81.8% for the validation sample)". Please refer to Figure 12-1. The first branch of the optimal tree. We can almost say "**if FX9BSL < 988.337, the GOOD probability is high**" because even if the BAD probability is a little higher when FXCFRR32 < 17.04642 in Node 1, the number of observations is small.

The second rule is in Leaf No. 5 and 13. These two leaves resulted when 990.4937< FX9BSL<= 988.337. Please refer to Figure 12-2, 12-3. The second and the third branches of the optimal tree. This is a very delicate range for a temperature in Zone C and the defeat rate is sensitive to the value of FX1A: if FX1A<1062.98, most observations are BAD while if FX11A >=1062.98, the GOOD rate is higher than the BAD rate. Although, the GOOD can be improved from 67.5% (Node 3) to 74.3% (Leaf 13) by following the rule of "IF FX8BS >= 960.6292, AND FXAFM<2.2462, AND FXAFR31>=5.3615", a judgment should be made if the incremental change of defeat rates justifies the sacrifice of parsimony. In this case, since two branches beneath FX1A have only two reliable leaves (Leaf 5 and 13) separately, the second rule can be simplified as in Node 2 and 3: **"when 990.4937 < FX9BSL <= 988.337, the results are mixed. If FX1A >= 1062.98, the result is GOOD; if FX1A < 1062.98, the result is BAD"**.

The third rule is in Leaf No. 19 where the BAD rate is 70%. Please refer to Figure 12-4. The fourth branch of the optimal tree. The above-mentioned simplification rule may apply in this branch such that a simplified business rule falls in Node 4 where FX9BSL >= 990.4937, and the BAD rate is 61.7%. Thus the third rule may be "**when FX9BSL >= 990.4937, the BAD likelihood is high**".

Analysis summary

The analysis through traditional statistical and data mining tools have three consistent findings. *Finding 1, time dynamics plays a role in determining defeat rates, especially t1, t3 and t4. Finding 2, the significant variables are FX9BSL, FXAFR11, FXAFR12, FXAFC and FXGF (see Attachment 7. Significant variables).* FX9BSL, FXAFC and FXGF are perceived important by line technicians. The finding suggests that more attention may be paid to FXAFR11 and FXAFR12 that are the CP pressure in Zone R1. Other perceived important variables, FXBBNL and FXBBNR, are not found significant in the analysis. *Finding 3, three control rules are proposed from decision tree analysis. They are "If FX9BSL < 988.337, the GOOD probability is high"; "when 990.4937 < FX9BSL <= 988.337, the results are mixed. If FX1A >= 1062.98, the result is GOOD; if FX1A < 1062.98, the result is BAD";* and *"when FX9BSL >= 990.4937, the BAD likelihood is high".* The finding stresses the importance of FX9BSL, the temperature of the left-side wall in Zone C. Line technicians can see the variable of FX9BSL as a control flag which points to different directions based on its values. 988.337 and 990. 493 are the two dividing temperatures that produce high likelihoods of GOOD, BAD or mixed products.

4.2.6 Evaluation

Berry and Linoff (1997) suggested that each data mining technique comes with its own set of evaluation criteria. Classification models, for example, are assessed using a ratio called a lift. However, in order to answer questions such as "is the model worth the time, effort, and money spent on building it?", data mining models should be evaluated tied in conjunction with business models.

The lift charts in the analysis part shows that the decision tree model provides a potential for increasing GOOD products by using the IF-THEN rules defined by the most predictive paths. The highest lift is at 60 percentile where the base capture rate is 60% and the decision tree capture rate is 75%. As the panel production is a continuous process; the yield for each line is about 400 pieces panels per hour; the annual production of the whole factory is more than 10 million pieces. A 15% increase of quality products translates to improved profitability.

The line technical managers were surprised that important variables were found so quickly. Their experience is that a new technician needs at least two years to understand the production techniques and the relationships among variables. However, they were suspect of the business rules extracted because they did not see the tree diagram but were just told the rules. The optimal tree diagram (Figure 12-1, -2, -3 and -4) with six layers could not be saved as a separate file but only put together by 10 pieces of letter-sized papers, which is hard to attach to an email or fax to the glass factory. They doubted whether a single value of FX9BSL dominates the determination of defeat. Since the production line has many processes and the X channel is merely one step of the hot section, a tree with all operational variables may give line managers a whole picture of the production process and increase confidence to use tree results. Another problem may be the glass factory has ever used any analytical software to help quality control. This unfamiliarity can prevent people from implementing the technology and realizing its potential benefits. It is believed that tree methods are easy to comprehend. Once the optimal tree diagram is presented to line technicians and managers and the basic tree mechanisms explained, they should be able to find the rules helpful.

One quality initiative that line managers are generating now is to use inspection devices to analyze the chemical and physical components of defeat products, to fine-tune the causes, and then to identify the problem-causing process or variation of parameters that brings the defeat. As discussed before, final inspection is costly. It is recommended both metrology method and software analysis method be deployed to supplement each other. A cost-and-benefit assessment can be done to decide on a proper data-mining tool. The project will be provided to the managers to enable them to test the rules, and more importantly to display the power of data mining. More effort is required to explain the data analysis method and communicate solutions based on the results, if action is going to take place in the real world.

CHAPTER 5. DISCUSSION

5. 1 Major Findings and its Implications to Quality Management

The data analysis, especially the data mining approach, identified several potentially useful insights. The regression identified the most significant predictive variables, and a time series effect was also found. In addition, three candidate business rules for operation control in the panel production line were also identified. These findings can potentially have an enormous impact on quality management practices in the glass factory.

5.1.1 Implications of time series prediction model

The presence of time effect suggests that two types of quality checks should be considered in the first and/or second inspection: static checks, single-time checks, and dynamic checks. Dynamic checks would take advantage of other information prior to the current check. Line managers can predict defeat rates by looking back and forward. A defeat rate can be estimated by looking back at defeat one hour, three hours or even four hours earlier if production parameters are unchanged. On the opposite side, a high or low defeat may indicate a similar pattern in the next first, third or fourth hour defeat rate.

If the two-hour time lag from the X channel to the first inspection is taken into consideration, this time pattern means a change of a temperature in the X channel may cause a change of defeat rate after two hours, and then probably three hours, five hours and six hours later. The presence of time effect increases the complexity of quality control in the panel production. Care should be paid to assess changes of defeat rate resulted from alteration in the X channel. Every adjustment of control parameters should be handled circumspectly, and control results should be monitored across different time frames.

5.1.2 Implications of decision tree modeling

The decision tree approach is a valuable method to process quality control. Decision trees can be used to identify predictors. The splitting variables are important predictors of the target variable because the combination of these predictors produces different defeat. Instead of being overwhelmed by hourly 54 values in the X channel, line technicians can merely monitor a few predictors, which greatly increase the efficiency and effectiveness of quality control.

The decision tree approach provides line managers a road map that includes all the possible production states. Because a tree has different splitting variables, technicians can decide which node the current production status belongs to by looking at combinations of values from a root node to leaves. For example, if a tree stresses the importance of FX9BSL, FX1A and FXCFRR by positioning them at the top layers of a tree in this case, a combination of FX9BSL<988.337 and FXCFRR32>17.04642 indicates that the production status falls in Leaf 3 (see Figure 12-1. The first branch of the optimal tree). Furthermore, Leaf 3 specifies a rate of 80% GOOD products. If the actual rate is greatly below the predicted value, there might be something wrong with the other part of the line, not the X channel. If the actual production falls into the adjacent Node 1 instead of Leaf 3, line technicians can identify the fastest way to turn the situation to a favorable one: to change the value of FXCFRR32 to lower than 17.0462.

The principle can be expanded to a tree including the whole line parameters. Should a tree be built based on the data from the whole production line (the cold and hot sections), at anytime of the day, technicians can map the current production condition on one of the nodes of the tree that includes all the instances of the operational variables.

Operators can then "angle" the production towards nodes that provide lower levels of defeat. In addition, a lower than predictive GOOD rate at any node might point to abnormal changes in factors outside the production parameters. These factors might include the outdoor temperature and humidity, the raw material property and even the quality inspection process. Line technicians at the glass plant have sensed the seasonal pattern of the production line, but are not clear about the impact of the outdoor temperature, humidity and pressure. By building a tree of all operational variables, line technicians should be able to narrow their findings to outsider variables and build a more flexible model for prediction and control. Finally, line technicians should be able to determine the most efficient and effective way to switch from a BAD node to a GOOD node by checking the adjacent nodes and the value of splitting variable.

Decision trees can be used to extract quality control rules by recognizing the most predictive paths. Simply looking at individual variables cannot assure that the quality is under control. Business rules from decision trees take into consideration the interactions between variables, and can be documented into quality regulations. These regulations are easy to follow and explicitly indicate predicted defeats. The ease of tree operation and the explicit IF-THEN tree rules also benefit the communication and application of data mining techniques and facilitate quality-related training in manufacturing. Running decision trees does not need a predefined model. Even people without statistical knowledge can easily grasp the meaning of decision trees. These features may indicate a promising future for the technology.

5.1.3 Implications of data mining approach

The data mining approach can contribute greatly to strategic quality management and quality-related knowledge management.

Currently, majority of line technicians rely on their years of first-hand working experience and endless trial and error to gain the knowledge of the relationships between the input and output variables. Even so, such knowledge is hard to record and difficult to share for two reasons. First, it is not developed through scientific methods and technicians may not be confident enough to communicate their knowledge. Second, some technicians are not willing to disclose their insights since they may want to take advantage of the unique and irreplaceable benefits of possessing this knowledge. The line quality-control managers interviewed admitted they do not have a complete handbook of quality management, which makes training difficult. Technicians gain knowledge by learning from skilled co-workers and from their own experience. When they leave or resign, their knowledge is gone. The data mining approach can discover knowledge in a short time and in a systematic way, which greatly shortens the learning curve. Business rules derived from decision trees are easy to comprehend and to communicate. This should enhance the knowledge sharing in organizations.

5.2 Data Mining vs. Traditional Prediction Models

Data mining can handle large datasets with missing values that are complex for normal statistical tools. For example, regression embedded in MS Excel can only run less than 16 independent variables while SAS has no limit. In this case, 54 operational variables call for a robust data-mining tool. In Excel, missing values (null or text) have to be substituted with some numbers before regression can be successfully run. In SAS, users can choose to ignore missing values or replace them with set values.

Most predictive models, for example regression and neural network, are sensitive to noisy data and easily produce overfitting models (Lilien, G. L. & Rangaswamy, A., 2003). The data mining techniques in SAS Enterprise Miner use the "Partition" node to separate a dataset into training, validation and testing samples. The training sample is employed to build initial models; the validation sample is to solve overfitting problems; and the testing sample is to verify the model on a new dataset. Yet data partition in traditional statistics can be such a time-consuming work that it is often disregarded. Thus data mining can build more reliable models with less effort.

Data mining has three very powerful advantages over traditional model building approaches. First, traditional approaches do not handle non-linearity easily. If the final model is not linear, data transformation has been done in order to use regression to get significant and useful coefficients. Decision trees and neural network can handle non-linear relationships very easily (Krider, 2003). Second, regression needs predefined models before coefficients can be calculated. This demands prior knowledge about relationships. Data mining, such as the decision trees in the case study, can get a better model without knowing the relationships between variables. Third, data mining makes it possible to compare the fit among different models. If regression models are to be compared, adjusted R square is the benchmark. However, decision trees do not have R square. SAS can provide a lift chart for comparing models. Moreover, the pictorial graphics and drawings are easier for business users to understand than statistical numbers.

However, these benefits do not guarantee the broad application of data mining techniques in the real world. Except for the justification of the cost associated with professional data mining tools, the software itself carries some shortcomings as well. For a business user, it may not be very user-friendly especially when something goes wrong. Error messages are written in computer or statistical languages and do not really help users to solve problems. When certain rules are violated, the software may crash without warning. Freezing happens so often that a "Find" window has to be kept open and "lock files" have to be deleted frequently to rerun the software. The biggest obstacle in being able to take full advantage of the technology might be the fact that a basic understanding of statistics and the training on mining software needs to be mandatory. Knowledge discovery process in industry is often performed by statistical and data analysis professionals (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). People with the blend of business, statistical and software knowledge are relatively scarce and represent a real resource constraint to the growth of this practice.

CHAPTER 6. CONCLUSIONS

6.1 Limitations

It should be noted that the interpretations of the results from decision trees were made under various assumptions. These assumptions include the accuracy of data from the X channel and a fixed level of variables from other production processes. The interpretations are therefore limited by these assumptions and the findings should be recognized as occurring only at the specific production circumstances.

Furthermore, the 44-day data sample may not be representative of the normal production because of its limited size. Successful data mining requires a large repository of accurate, subject-related and independent data (Fouhy, 1998). The size of raw data in the case may constrain the predictability of proposed variables and rules. According to Deming's theory (1975), defeat may be produced by common causes and special causes. The common defeats, which come from the system, are replicated and predictable. The special defeats are caused by the variations from system parameters, and are thus random and unpredictable. A large size of data repository can find predictable patterns and eliminate noisy random patterns by adjusting models through testing samples. If the sample size is small, "noise" may affect the validity of the models. An inherent problem associated with complex models built on small samples is instability. Trees using different datasets would grow from different samples (Statsoft Inc., 2003). Besides, model interpretations may be problematic. Take the case study as an example. The limited observations in some leaf nodes made the tree interpretation difficult even though these leaves may have practical implications.

6.2 Areas for Future Research

Using data mining to discover knowledge is an interactive and iterative process (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996; Escalante, 1999; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Hirji, 2001; Kamrani, Rong & Conzalez, 2001; www.sas.com). Consistent patterns will not be found without trial and error. Therefore, data from a longer period of time may be obtained and the analysis refined in order to differentiate common and special causes. True business rules should be acquired from repeatable patterns. Nevertheless, large data sizes are expensive to obtain and inefficient to operate, and having more data does not guarantee better business knowledge. Future research may seek the balance between the economic and technical efficiency of sample sizes and the representativeness of results from data mining.

An unsolved problem in the analysis is that even though auto-regression shows high reliability in running the time-adjusted data, decision trees use original data and produces a better lift. It might be that decision trees take care of time effects automatically. Furthermore, the predictive variables chosen from decision trees are not completely identical with those from auto-regression. A few variables perceived to be predictors by line technicians showed up at the bottom of decision trees and did not have sufficient sample size to support their significance. Thus, the interaction of time effect and decision tree modeling may be an area for further research.

In the data analysis process, little effort was put to integrate and clean the data because the warehoused data was relatively clean. Although databases or data warehouse provide a good basis for mining data, not every manufacturer is afford the expensive storages. An obvious advantage of data mining techniques is to deal with data from different sources and handle missing values. Future research may be directed to find whether it is easy and what additional knowledge and technology are requisite to integrate multiple sourced data through data mining tools. Findings from the research would aid potential data mining users to evaluate the technical infrastructures they must have before they proceed to purchase data mining tools.

The case study chose SAS Enterprise Miner to analyze data. There might be other data mining tools more appropriate to analyze the quality data than SAS Enterprise Miner. Moreover, there are many areas that can be done to improve the user-friendliness of data mining tools. Future research can study the suitability of tools and business problems, which might shorten the time that users would spend on comparing various results generated by different tools and reduce fatigue and other hinders to apply data mining technology.

6.3 Generalizability of Results

Manufacturing production is a process of planning, control and implementation to transform input into products and services (Teller, 1978). Efficient and effective management of production is crucial to the success of a firm. Quality management relates to the effectiveness of production management (McNamara, 2003). Continuous production plants have complex relationship between input variables and output variables. The quality of final products is determined by that of unfinished products from every step. Traditional statistical process control is widely used to produce stable quality by building linear models between input and output.

Nevertheless, quality control of continuous production should also take into consideration of time effects. The flow of operation produces circular defeat. Quality improvement is a constant endeavor to eliminate mistakes in continuous production. Every adjustment of operative variables should be monitored across time frames.

Current defeat can be anticipated by prior defeats. The case study utilized more than 50 variables and a continuous 44-day sample to illustrate the interactions among variables and the effect time played in product defeat. Time oriented models are suggested to other similar continuous processes in industries.

The value of the data mining technology is displayed not only by` the findings of predictors, time series and the control rules but also by the usefulness of the technology as a decision making tool. Fed with sufficient and accurate historical data, well-trained decision trees can be used as an aid of decision-making process for common business users who do not presses statistical knowledge. The display of nodes and the probabilities of GOOD and BAD in each node are intuitive and visualized. The IF-THEN rules inform users what results they can expect under certain production states. The splitting variables help users to choose an efficient way to change production states to desired ones. Although the use of data mining techniques is not necessarily intended to replace users' roles of decision-making, the resulting rules could serve as a basis for expanding insights and acting as a cognitive prosthesis.

This paper is an exploration of data mining techniques in a glass plant. The case study showed the potential benefits by confirming major predictors, considering time effects and recognizing business rules. To generalize the study, more research should be devoted to find out applications of data mining techniques in other continuous industrial processes. Chemical, petrochemical, pharmaceutical and food industries have enormous data and complex production processes. It is hoped that data mining can be tested in these fields to contribute to process quality control. The possible differences of the results can be compared to enhance the understanding of the technology. The potential benefits can speed the spread of the technology in manufacturing quality control. The research shall provide material for data mining theories from both academic and practical directions.

6.4 Conclusions

This paper aims to draw research attention to the cross-sectional study of quality control and data mining technology. It explores the value of this KDD method in process quality control. Quality control literature suggests that defeat comes from different variations of the production process and might be classified as common or special defeats. Common defeats are generated from the system and can be reduced by adjusting the production system parameters. Special defeats are generated from variation of parameters and can be reduced by choosing better equipment and instruments. Statistical process control should be used to identify common and special defeats. Quality control practitioners apply metrology method, software and analysis method and quality management method to improve product quality. These three methods supplement each other. Manufacturing calls for a robust quality control mechanism to manage complex processes. Data mining as a software method brings a new aspect to quality control. Compared to traditional statistical process control, data mining can handle large data, discover non-linearity, have robust functionality, and provide objective results. A variety of data mining tools have their own strengths and weaknesses. Their features were demonstrated by SAS Enterprise Miner, the analysis software devoted to a case study of complicated process control of a TV panel glass plant. The objective of the analysis was to use these data mining tools to confirm existing knowledge and to discover unknown patterns.

Data mining is a process that includes problem definition, acquisition of background knowledge, selection and pre-processing of data, analysis, interpretation, report generation and practical use. Following this methodology, regression, decision trees and neural network were utilized to analyze the data. The findings from the analysis verified existing knowledge in the panel production line, offered business rules for process control, and discovered time pattern in the defeat rate. Most importantly, the case showcased how data mining technology could help manufacturers to develop a systematic quality control practice and to enhance organizational quality knowledge management. Meanwhile, the knowledge discovered from data mining will enable industry to economically produce materials with improved quality, consistent properties, and enhanced functionality.

DM requires a mixture of statistical and computer knowledge. For example, database knowledge is needed to precede OLAP for better usage of data. Yet the trend of knowledge discovery technology is to turn the technology into the hands of users (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro & Simoudis, 1996). In the meantime, practitioners also foster the improvement of data mining technology. Although the visualization tools in SAS make data mining easy to learn and comprehend, enhancements can still be made to increase the stability and understandability of certain functions. Data mining helps explain the cause of known manufacturing problems discloses unknown patterns and supports management to make better business decision. More research should focus on the applied aspects of data mining in various industries in order to help practitioners understand these techniques to and generate more quality initiatives.

LIST OF REFERENCES

Adams, L. (2002). Mining factory data. *Quality*, May 2002, 28-31.

Adams, L. (2001). Mining the world of quality data. *Quality*, August 2001, 36-40.

Apte, C., Liu, B., Pednault, E. P. D. & Smyth, P. (2002). Business applications of data mining. *Communications of the ACM,* 45(8), 49-53.

Adriaans, P. & Zantinge, D. (1996). *Data mining,* Harlow: Addison-Wesley.

American Society for Quality (2003a). ASQ's Six Sigma Portfolio. *Six Sigma*, http://www.asq.org/sixsigmaportfolio/index.html (14 Feb. 2003).

American Society for Quality (2003b). Quality Glossary. *Information*, http://www.asq.org/info/ (14 Feb. 2003).

Becerra-Fernandez, I., Zanakis, S. H. & Walczak, S. (2002). Knowledge discovery techniques for predicting country investment risk. *Computer & Industrial Engineering*, 43(4), 787-800.

Bhote, K. (1991). *World class quality (*pp 4,5, 30, 32-35, 41, 55, 169, and 206). New York, MA: American Management Association.

Box, G. E. P. (1988). The R. A. Fisher memorial lecture. *Phil. Trans. Roy. Soc. London Series A.*

Brachman, R. J, Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. & Simoudis, E. (1996). Mining business databases. *Communications of the ACM* 3, 9(11), 42-48.

Bradley, P., Gehrke, J., Ramakrishnan, R. & Srikant, R. (2002). Scaling mining algorithms to large databases. *Communications of the ACM,* 45(8), 38-43.

Bremmer, B. (1991). Verify accuracy through calibration. *Quality Progress*, 24(3), 108-111.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovery data mining: from concept to implementation.* Eaglewood Cliffs, NJ: Prentice Hall.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Technical report. SPSS.

Cohen, M. D., Kelly, C. B. & Medaglia, A. L. (2001). Decision support with web-enabled software. Interfaces, 31(2), 109-129.

Corning Asashi Video Inc. (2003). What is a CRT? *Products & Service,* http://www.corningasahi.com/products__services/ (15 Mar. 2003).

Deming, W. E. (1975). On some statistical aids toward economic production. *Interfaces,* 5(4), 1-5.

Edelstein, H. (1999). Introduction to data mining and knowledge discovery (3rd. ed.) (p. Maryland) Potomac, Maryland: Two Crows Corporation.

Escalante, E. J. (1999). Quality and productivity improvement: a study of variation and defects in manufacturing. *Quality Engineering*, 11(3), 427-442.

Fayyad, U., Madigam, D., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17, 37-54.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* 39(11), 27-34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.), (1996). The process of knowledge discovery in databases: a human-centered approach. *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: AAAI Press,, Menlo Park, CA/MIT Press

Fayyad, U. & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11), 24-26.

Fayyad, U. & Uthurusamy, R. (2002). Evolving data mining into solutions for insights. Communications of the ACM, 45(8), 28-31.

Fouhy, K. (1998). Discover process plan gold. *Chemical Engineering*, 105(4), 151-155.

Galapon, E. A. & Norton, J. S. (2001). TQM Works for Lucent "QITS" in Saudi Arabia of TEP6. : *Annual Quality Congress Proceedings,* 55, 660-673.

Gemino, A. (2002). Lecture Notes: Data Mining. *Bus876 Decision Support Systems.*

Glymour, C., Madigan, D., Pregibon, D. & Smyth, P. (1996). Statistical inference and data mining. *Communications of the ACM,* 39(11), 35-41.

Grossman, R. L., Hornick, M. F. & Meyer, G. (2002). Data mining standards initiatives. *Communications of the ACM,* 45(8), 59-61.

Han, J., Altman, R. B., Kumar, V., Mannila, H. & Pregibon, D. (2002). Emerging scientific applications in data mining. *Communications of the ACM,* 45(8), 54-58.

Hinckley, C. M. & Barkan, P. (1995). The role of variation, mistakes, and complexity in producing nonconformities. *Journal of Quality Technology*, 27(3): 242-249.

Hirji, K. K. (2001). Exploring data mining implementation. *Communications of the ACM,* 44(7), 87-93.

Hui, S. C. & Jha, G. (2000). Data Mining for Customer Service Support. *Information & Management*, 38, 1-13.

Imbellone, O. & Lopez, A. M. (1997). A Successful Application of TQM in CPC Argentina. *Annual Quality Congress Proceedings,* 51, 703-712

Jiang, Z. L. (editor) (1999). *Colour-tube manufacturing technologies.* Beijing, China: Electronic Industry Press Company.

Kamrani, A. Rong, W. & Conzalez, R. (2001). A genetic algorithm methodology for data mining and intelligent knowledge acquisition. *Computers & Industrial Engineering,* 40(3), 361-377.

Kane, V. (1989). *Defect prevention.* New York: Marcel Dekker, Inc.

Kang, B. S., Choe, D. H. & Park, S. C. (1999). Intelligent process control in manufacturing industry with sequential processes. *International Journal of Production Economics*, 60-61, 583-590.

Kantardzic, K., Djulbegovic, B. & Hamdan, H. (2002). A data-mining approach to improving polycythemia vera diagnosis. *Computers & Industrial Engineering*, 43(4), 765-773.

Kim, S. H. & Lee, C. M. (1997). Nonlinear prediction of manufacturing systems through explicit and implicit data mining. *Computer & Industrial Engineering*, 33(3/4), 461-464.

Kohavi, R., Rothleder, N. J. & Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM*, 45(8), 45-48.

Krider, Robert (2003). Lecture Notes: Neural Network. *Bus846 Marketing Model.*

Lilien, G. L. & Rangaswamy, A. (2003). Marketing engineering: computer-assisted marketing analysis and planning (2nd. Edition). Upper Saddle River, New Jersey: Prentice Hall.

Lowery, C. M.; Beadles, N. A. II & Carpenter, J. B. (2000). TQM's Human Resource Component. *Quality Progress*, 33(2), 55-59.

McNamara, C. (2003). Operations management. *The Management Assistance Program for Nonprofits,* http://www.mapnp.org/library/ops_mgnt/ops_mgnt.htm (29 Apr. 2003).

Marakas, G. M. (Ed.), (2003). *Modern data warehousing, mining, and visualization.* Upper Saddle River, New Jersey: Prentice Hall.

Michael, J. A. Berry & Gordon Linoff (1997). *Data mining techniques fro marketing, sales, and customer support.* New York: Wiley Computer Publishing.

National Institue of Standards and Technology (2003a). Autocorrelations. *Engineering Statistics Handbook,* http://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm (12 Apr. 2003).

National Institue of Standards and Technology (2003b). Common Approaches to Univariate Time Series. *Engineering Statistics Handbook,* http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc444.htm#AR (12 Apr. 2003).

Packard, T. (1995). TQM And Organizational Change And Development. *Total Quality Management in the Social Services: Theory and Practice.* Burton Gummer and Philip McCallion, Eds., Albany, NY: Rockefeller College Press.

Port, O. (2001). Virtual prospecting. *Business Week*, 50, 185-188.

Sarkis, J. & Reimann, M. (1996). Quality information systems in advanced manufacturing environments. *Quality Engineering*, 8(3), 419-431.

SAS Institute Inc. (2003a). Enterprise miner. *Products and Solutions*, http://www.sas.com/technologies/analytics/datamining/miner/index.html (14 Feb. 2003).

SAS Institute Inc. (2003b). On-line help. *Enterprise Miner 3.0,* http://www.sas.com/technologies/analytics/datamining/miner/ (14 Feb. 2003)

SAS Institute Inc. (2003c). Solution lines. *Products and Solutions,* http://www.sas.com/solutions/pri (14 Feb. 2003).

Schmidit, S. & Launsby, R. (1994). *Understanding Industrial Designed Experiments, 4th ed.* Colorado Springs, CO: Air Academy Press.

Shewhart, W. (1931). *Control of quality of manufactured products.* New York: D. Van Nostrand; reprinted in 1989 by the American Society of Quality Control, Milwaukee, WI.

Statsoft Inc. (2003). Data mining techniques. *Textbook,* http://www.statsoftinc.com/textbook/stdatmin.html (28 Apr. 2003)

Subramanian, A., Smith, L. D., Nelson, A. C., Campbell, J. F. & Bird, D. A. (1997). Strategic planning for data warehousing. *Information and Management,* 33, 99-113.

Smyth, P., Pregibon, D. & Faloutsos, C. (2002). Data-driven evolution of data mining algorithms. *Communications of the ACM,* 45(8), 33-36.

Sun, H. (2001). Comparing quality management practices in the manufacturing and service industries: learning opportunities. *Quality Management Journal,* 8(2), 53-71.

Taguchi, G. (1986). *Introduction to Quality Engineering. Designing Quality into Products and Processes.* Asian Productivity Organization, Tokyo/Unipub Kraus International Publication, White Plains, NY/The American Supplier Institute, Dearborn, MI.

Tellier, R. D. (1978). Operations management: fundamental concepts and methods. New York: Harper & Row, Publishers, Inc., 10 East 53rd Street, New York, NY.

Trafalis, T. B., Richman, M. B., White, A. & Santosa, B. (2002). Data mining techniques for improved wsr-88d rainfall estimation. *Computers & Industrial Engineering,* 43(4), 775-786.

Walczak, S. (2001). Neural networks as a tool for developing and validating business heuristics. *Expert Systems with Applications*, 21(1), 31-36.

Weiss, W. H. (2002). Issues related to quality control and its cost. *SuperVision*, 63(12), 3-7.

Wong, Albert (2002). Lecture Notes: Data Mining. *Bus 876.Decision Support Systems.*

APPENDIX 1. BINARY DATASET

TIME	FX1A.PV	FXGF.PV	defeat rate	bubble rate	stone rate	bidefeat
2001-8-8 20:00	1054.535	86.91342	0.036723	0.028249	0.008475	GOOD
2001-8-8 21:00	1054.444	86.9476	0.048048	0.03003	0.018018	BAD
2001-8-8 22:00	1054.429	87.08658	0.072072	0.036036	0.036036	BAD
2001-8-8 23:00	1054.38	87.10988	0.032641	0.023739	0.008902	GOOD
2001-8-9 0:00	1054.181	87.12437	0.011173	0.011173	0	GOOD
2001-8-9 1:00	1054.245	87.24268	0.023121	0.017341	0.00578	GOOD
2001-8-9 2:00	1054.244	87.33449	0.046961	0.027624	0.019337	BAD
2001-8-9 3:00	1054.227	87.31119	0.046196	0.029891	0.013587	BAD
2001-8-9 4:00	1054.13	87.30248	0.023324	0.011662	0.008746	GOOD
2001-8-9 5:00	1054.089	87.18609	0.039394	0.021212	0.018182	GOOD
2001-8-9 6:00	1054.066	87.10875	0.042105	0.028947	0.013158	BAD
2001-8-9 7:00	1054.157	87.33581	0.055215	0.052147	0.003067	BAD
2001-8-9 8:00	1054.143	87.41645	0.065753	0.054795	0.010959	BAD
2001-8-9 9:00	1054.146	87.4296	0.020896	0.020896	0	GOOD
2001-8-9 10:00	1053.955	87.45571	0.042493	0.036827	0.005666	BAD
2001-8-9 11:00	1054.013	87.28493	0.043988	0.041056	0.002933	BAD
2001-8-9 12:00	1054.016	87.19602	0.033426	0.022284	0.011142	GOOD
2001-8-9 13:00	1054.046	87.16055	0.021407	0.018349	0.003058	GOOD
2001-8-9 14:00	1053.922	87.14319	0.014837	0.014837	0	GOOD
2001-8-9 15:00	1053.96	87.06384	0.067797	0.059322	0.00565	BAD
2001-8-9 16:00	1054.149	86.95982	0.025862	0.020115	0	GOOD
2001-8-9 17:00	1054.061	87.04886	0.049689	0.043478	0.006211	BAD
2001-8-9 18:00	1054	87.1576	0.066852	0.058496	0.008357	BAD

APPENDIX 2. CORRELATION RESULTS

Significant correlations between input variables and the defeat rate

FX9BSL.PV	0.260250202
FXAFR12.PV	0.179012341
FXAFR11.PV	0.166443692

Significant correlations between input variables and the bubble rate

FX9BSL.PV	0.227918596
FXAFR11.PV	0.220349194
FXAFR12.PV	0.208842414
FXGF.PV	0.201981777
FXB1.PV	-0.189043652
FXB2.PV	-0.178735425
FX3A.PV	0.175751313
FXBC.PV	-0.171132422
FX4BS.PV	0.161598104

Significant correlations between input variables and the stone rate

FX9BSL.PV	0.204016728
FX9BSR.PV	0.182876763

APPENDIX 3. BIAUTO DATASET

TIME	FX1A.PV	FXGF.PV	autodefeat	biautodefeat	dt1	dt2	dt3	dt4	dt5
2001-8-9 1:00	1054.245	87.24268	**0.023121**	GOOD	0.011173	0.032641	0.072072	0.048048	0.036723
2001-8-9 2:00	1054.244	87.33449	0.046961	BAD	**0.02312**	0.011173	0.032641	0.072072	0.048048
2001-8-9 3:00	1054.227	87.31119	0.046196	BAD	0.046961	**0.02312**	0.011173	0.032641	0.072072
2001-8-9 4:00	1054.13	87.30248	0.023324	GOOD	0.046196	0.046961	**0.02312**	0.011173	0.032641
2001-8-9 5:00	1054.089	87.18609	0.039394	GOOD	0.023324	0.046196	0.046961	0.023121	0.011173
2001-8-9 6:00	1054.066	87.10875	0.042105	BAD	0.039394	0.023324	0.046196	0.046961	0.023121
2001-8-9 7:00	1054.157	87.33581	0.055215	BAD	0.042105	0.039394	0.023324	0.046196	0.046961
2001-8-9 8:00	1054.143	87.41645	0.065753	BAD	0.055215	0.042105	0.039394	0.023324	0.046196
2001-8-9 9:00	1054.146	87.4296	0.020896	GOOD	0.065753	0.055215	0.042105	0.039394	0.023324
2001-8-9 10:00	1053.955	87.45571	0.042493	BAD	0.020896	0.065753	0.055215	0.042105	0.039394
2001-8-9 11:00	1054.013	87.28493	0.043988	BAD	0.042493	0.020896	0.065753	0.055215	0.042105
2001-8-9 12:00	1054.016	87.19602	0.033426	GOOD	0.043988	0.042493	0.020896	0.065753	0.055215
2001-8-9 13:00	1054.046	87.16055	0.021407	GOOD	0.033426	0.043988	0.042493	0.020896	0.065753
2001-8-9 14:00	1053.922	87.14319	0.014837	GOOD	0.021407	0.033426	0.043988	0.042493	0.020896
2001-8-9 15:00	1053.96	87.06384	0.067797	BAD	0.014837	0.021407	0.033426	0.043988	0.042493

APPENDIX 4. REGRESSION RESULTS

Summary of Stepwise Procedure

	Effect		Number		
Step	Entered	DF	In	F	Prob>F
1	FX5A_PV	1	1	11.6568	0.0007
2	FXGF_PV	1	2	20.7380	<.0001

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 2.

It consists of the following effects: Intercept FX5A_PV FXGF_PV

Analysis of Variance

		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr>F
Model	2	0.019147	0.009573	16.47	<.0001
Error	422	0.245303	0.000581	.	.
Corrected Total	424	0.264450	.	.	.

Model Fitting Information

R-square	**0.0724**	**Adj R-sq**	**0.0680**
AIC	-3163.3736	BIC	-3162.2946
SBC	-3151.2173	C(p)	78.8715

```
The SAS System          14:54 Thursday, April 10, 2003    91

                     The DMREG Procedure

                Analysis of Parameter Estimates

                                               Standard

Parameter          DF     Estimate        Error       t Value      Pr>|t|

 Intercept          1       0.6295        0.1102          5.71      <.0001

FX5A_PV             1     -0.00059      0.000107        -5.45      <.0001

FXGF_PV             1     0.000440      0.000097         4.55      <.0001
```

APPENDIX 5. AUTO-REGRESSION RESULTS (1)

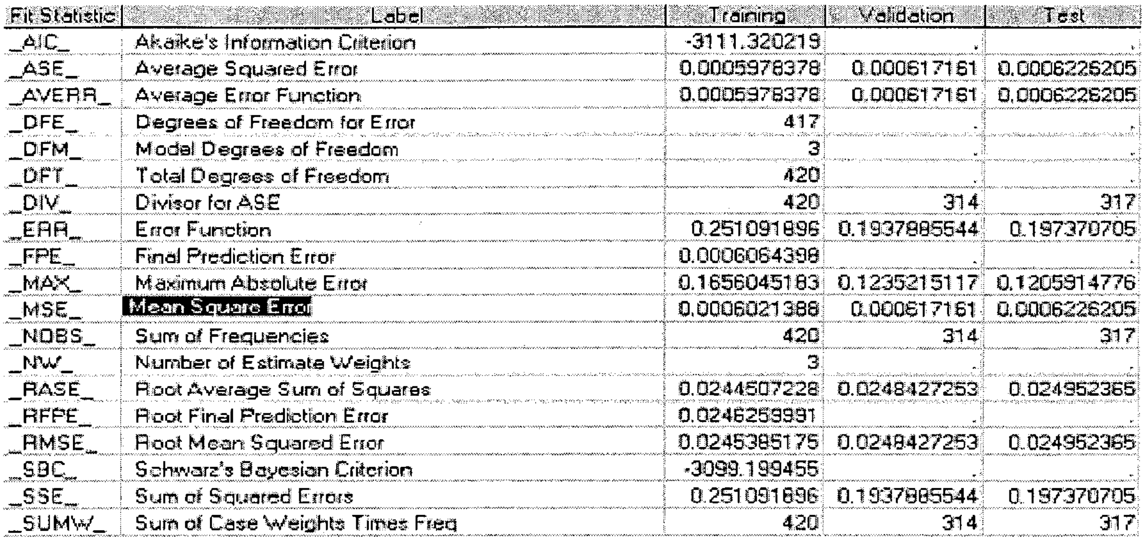
REGRESSION 1-AUTODEFEAT

Forward (if backward is chosen, the best model is the one at step 0 including every input variable)

Target: autodefeat

Input: all input variables + dt1/2/3/4/5

The chosen model only includes three variables but the adjusted R square is high.



Fit Statistic	Label	Training	Validation	Test
AIC	Akaike's Information Criterion	-3111.320219	.	.
ASE	Average Squared Error	0.0005978378	0.000617161	0.0006226205
AVERR	Average Error Function	0.0005978378	0.000617161	0.0006226205
DFE	Degrees of Freedom for Error	417	.	.
DFM	Model Degrees of Freedom	3	.	.
DFT	Total Degrees of Freedom	420	.	.
DIV	Divisor for ASE	420	314	317
ERR	Error Function	0.251091896	0.1937885544	0.197370705
FPE	Final Prediction Error	0.0006064398	.	.
MAX	Maximum Absolute Error	0.1656045183	0.1235215117	0.1205914776
MSE	Mean Square Error	0.0006021388	0.000617161	0.0006226205
NOBS	Sum of Frequencies	420	314	317
_NW_	Number of Estimate Weights	3	.	.
RASE	Root Average Sum of Squares	0.0244507228	0.0248427253	0.024952365
RFPE	Root Final Prediction Error	0.0246259991	.	.
RMSE	Root Mean Squared Error	0.0245385175	0.0248427253	0.024952365
SBC	Schwarz's Bayesian Criterion	-3099.199455	.	.
SSE	Sum of Squared Errors	0.251091896	0.1937885544	0.197370705
SUMW	Sum of Case Weights Times Freq	420	314	317

```
                 Summary of Forward Selection Procedure

               Effect              Number
       Step    Entered        DF      In          F     Prob>F


          1    FXAFC_PV        1       1      936.3     <.0001
          2    dt1             1       2    61.4497     <.0001
          3    dt4             1       3    29.1767     <.0001
          4    dt3             1       4    11.4347     0.0008
          5    FXAFM_PV        1       5     4.3877     0.0368
          6    FXGF_PV         1       6     4.8823     0.0277
```

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 3. It consists of the following effects: dt1 dt4 FXAFC_PV

```
                         Analysis of Variance

                                  Sum of          Mean
Source                   DF      Squares        Square    F Value     Pr>F

Model                     3     0.746720      0.248907     412.77    <.0001
Error                   417     0.251456      0.000603          .         .
Uncorrected Total       420     0.998176             .          .         .


                       Model Fitting Information

          R-square          0.7481      Adj R-sq        0.7463
          AIC           -3110.7115      BIC         -3109.0623
          SBC           -3098.5907      C(p)           31.5637

          The SAS System       13:58 Saturday, April 12, 2003   31

                        The DMREG Procedure

                  Analysis of Parameter Estimates

                                           Standard
     Parameter          DF    Estimate        Error    t Value    Pr>|t|

     dt1                 1      0.3252       0.0493       6.60    <.0001
     dt4                 1      0.2605       0.0482       5.40    <.0001
     FXAFC_PV            1     0.00332     0.000544       6.10    <.0001
```

REGRESSION 2-AUTOBUBBLE
Forward
Target: autobubble
Input: all input variables + bt1/2/3/4/5

```
                    Summary of Forward Selection Procedure

                Effect                Number
        Step    Entered        DF       In          F      Prob>F

           1    FXAFC_PV        1        1      793.5      <.0001
           2    bt1             1        2    87.9261      <.0001
           3    bt3             1        3    29.3553      <.0001
           4    bt4             1        4    13.1604      0.0003
           5    bt5             1        5     4.9441      0.0267
```

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 2. It consists of the following effects: bt1 FXAFC_PV

```
                           Analysis of Variance

                                     Sum of          Mean
    Source                  DF      Squares        Square   F Value     Pr>F
    Model                    2     0.340819      0.170410    523.00   <.0001
    Error                  418     0.136198      0.000326         .        .
    Uncorrected Total      420     0.477017             .         .        .


                          Model Fitting Information

        R-square        0.7145        Adj R-sq        0.7131
        AIC         -3370.2388        BIC         -3368.8012
        SBC         -3362.1583        C(p)           67.8114

                       Analysis of Parameter Estimates

                                            Standard
    Parameter        DF     Estimate          Error     t Value     Pr>|t|
    bt1               1       0.4401         0.0469        9.38     <.0001
    FXAFC_PV          1      0.00299       0.000314        9.51     <.0001
```

REGRESSION 3-AUTOSTONE

Stepward

Target: autostone

Input: all input + st1/2/3/4/5

```
                     Summary of Stepwise Procedure

              Effect            Number
      Step    Entered       DF    In          F       Prob>F

        1     FXAFC_PV      1     1        592.7      <.0001
        2     st4           1     2       25.8962     <.0001
        3     st1           1     3       16.5755     <.0001
        4     st3           1     4        7.4560     0.0066
        5     FXAFR11_PV    1     5        5.3334     0.0214
```

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 3. It consists of the following effects: FXAFC_PV st1 st4

```
                         Analysis of Variance

                             Sum of         Mean
   Source               DF   Squares        Square        F Value    Pr>F

   Model                  3  0.050987       0.016996      231.57     <.0001
   Error                417  0.030605       0.000073393     .          .
   Uncorrected Total    420  0.081592           .           .          .


                       Model Fitting Information

        R-square     0.6249        Adj R-sq      0.6222
        AIC         -3995.2754     BIC          -3993.6176
        SBC         -3983.1546     C(p)            30.9099


                   Analysis of Parameter Estimates

                                           Standard
     Parameter        DF    Estimate       Error       t Value    Pr>|t|

     FXAFC_PV          1    0.00121       0.000162       7.42     <.0001
```

st1	1	0.1974	0.0485	4.07	<.0001
st4	1	0.2532	0.0530	4.78	<.0001

APPENDIX 6. AUTO-REGRESSION RESULTS (2)

REGRESSION 1-AUTODEFEAT
Target: autodefeat
Input: time-related variables and those from CORRELATION AND AUTOCORRELATION 1. they are FX9BSL, FXAFR11/12, XAFC
Model : backward

```
                    Summary of Backward Elimination Procedure

              Effect              Number
    Step      Removed        DF        In          F     Prob>F

       1      FX9BSL_PV       1         8     1.2950     0.2558
       2      FXAFR12_PV      1         7     0.4463     0.5045
       3      dt2             1         6     1.8946     0.1694
       4      dt5             1         5     2.1400     0.1443
       5      FXAFR11_PV      1         4     2.8010     0.0950

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 5. It
consists of the following effects:
  dt1  dt3  dt4  FXAFC_PV

          The SAS System        13:58 Saturday, April 12, 2003   67

                         The DMREG Procedure

                         Analysis of Variance

                                    Sum of        Mean
Source                   DF        Squares       Square    F Value     Pr>F

Model                     4       0.753447     0.188362     320.18   <.0001
Error                   416       0.244729     0.000588          .        .
Uncorrected Total       420       0.998176            .          .        .
```

Model Fitting Information

R-square	**0.7548**	**Adj R-sq**	**0.7525**
AIC	-3120.1003	BIC	-3118.0925
SBC	-3103.9393	C(p)	7.5914

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr>|t|
dt1	1	0.2873	0.0500	5.75	<.0001
dt3	1	0.1606	0.0475	3.38	0.0008
dt4	1	0.2169	0.0493	4.40	<.0001
FXAFC_PV	1	0.00265	0.000572	4.64	<.0001

REGRESSION 2-AUTOBUBBLE
backward
Target: autobubble
Input: time-related variables and variables from both COREELATION and AUTOCORRELATION 1. They are FX9BSL, FXAFR11/12, XGF, XB1/B2 and XAFC.

Summary of Backward Elimination Procedure

Step	Effect Removed	DF	Number In	F	Prob>F
1	FX9BSL_PV	1	11	0.0366	0.8483
2	FXB1_PV	1	10	0.3748	0.5407
3	FXAFR12_PV	1	9	1.8912	0.1698
4	FXAFR11_PV	1	8	1.1684	0.2804
5	FXGF_PV	1	7	1.1271	0.2890
6	FXB2_PV	1	6	1.6305	0.2023
7	bt2	1	5	3.2214	0.0734

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 2. It consists of the following effects:

bt1 bt2 bt3 bt4 bt5 FXAFR11_PV FXAFR12_PV FXGF_PV FXB2_PV

```
FXAFC_PV

                          Analysis of Variance

                                  Sum of          Mean
     Source              DF      Squares         Square   F Value     Pr>F

     Model               10     0.357766       0.035777    123.01   <.0001
     Error              410     0.119251       0.000291       .        .
     Uncorrected Total  420     0.477017           . .        .        .


                        Model Fitting Information

          R-square         0.7500        Adj R-sq        0.7439
          AIC          -3410.0491        BIC         -3407.4832
          SBC          -3369.6466        C(p)            8.4106


                     Analysis of Parameter Estimates

                                          Standard
       Parameter      DF    Estimate        Error     t Value     Pr>|t|

       bt1             1      0.2739       0.0503        5.44     <.0001
       bt2             1      0.0894       0.0502        1.78     0.0755
       bt3             1      0.1561       0.0478        3.27     0.0012
       bt4             1      0.1284       0.0471        2.72     0.0067
       bt5             1      0.0909       0.0479        1.90     0.0581
       FXAFR11_PV      1    -0.00905      0.00541       -1.67     0.0949

          The SAS System        13:58 Saturday, April 12, 2003   75

                         The DMREG Procedure

                     Analysis of Parameter Estimates

                                          Standard
       Parameter      DF    Estimate        Error     t Value     Pr>|t|

       FXAFR12_PV      1     0.00615      0.00447        1.38     0.1698
```

```
FXGF_PV         1    0.000145    0.000093     1.55    0.1209
FXB2_PV         1    -0.00003    0.000016    -2.13    0.0341
FXAFC_PV        1     0.00449     0.00222     2.02    0.0442
```

REGRESSION 3-AUTOSTONE
backward
Input: time-related variables and significant variables from both CORRELATION and AUTOCORELATION 1. they are X9BSL/R, XAFC.

```
                 Summary of Backward Elimination Procedure

                  Effect              Number
          Step    Removed       DF      In          F      Prob>F

            1     FX9BSL_PV      1       7     0.1253     0.7235
            2     st5            1       6     1.0395     0.3085
            3     st2            1       5     1.6799     0.1957
            4     FX9BSR_PV      1       4     1.6099     0.2052
```

The selected model, based on the CHOOSE=VERROR criterion, is the model trained in Step 4. It consists of the following effects:

```
  FXAFC_PV   st1   st3   st4
```

```
           The SAS System        13:58 Saturday, April 12, 2003   61

                            The DMREG Procedure

                           Analysis of Variance

                                   Sum of          Mean
    Source               DF        Squares         Square    F Value     Pr>F

    Model                 4       0.051526       0.012881     178.23    <.0001
    Error               416       0.030066    0.000072274        .         .
    Uncorrected Total   420       0.081592              .        .         .
```

Model Fitting Information

R-square	**0.6315**	**Adj R-sq**	**0.6280**
AIC	-4000.7364	BIC	-3998.6683
SBC	-3984.5754	C(p)	4.4483

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr>|t|
FXAFC_PV	1	0.000989	0.000180	5.51	<.0001
st1	1	0.1881	0.0482	3.90	0.0001
st3	1	0.1307	0.0479	2.73	0.0066
st4	1	0.2279	0.0534	4.27	<.0001

APPENDIX 7. SIGNIFICANT VARIABLES FROM DIFFERENT MODELS

	Correlation	Regression	Auto-reg (1)	Auto-reg (2)	Decision Tree	Description of variables
FX9BSL	X				X	the temperature of the left-side wall in Zone C
FXBSR	X					the temperature of the right side wall in Zone C
FXAFR11	X			X		the CP pressure in Zone R11
FXAFR12	X			X		the CP pressure in Zone R12
FXAFC		X	X	X		the temperature of the combustion-supporting air in Zone C
FXGF	X	X		X		the total airflow of the natural gas
FX5A		X				the space temperature in Zone R22
FXB2				X		the bottom temperature in Zone R11
FX1A					X	the space temperature in Zone ST
FXCFRR32					X	the control feedback of the cooling air flow in Zone R31