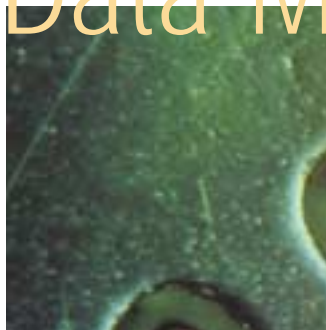




# Data Mining

Quality in Manufacturing Data



Best Practices  
Approach  
To The  
Manufacturing  
Industry

*Data Mining:*  
*Quality in Manufacturing Data*  
was written by

Ken Collier  
**KPMG Consulting**  
and  
Gerhard Held  
**SAS**

*Consultants*  
Curt Marjaniemi  
Don Sautter  
**KPMG Consulting**  
and  
Mohan Nambodiri  
**SAS**

*Technical Reviewers*  
Knowledge Management Solutions Group  
**KPMG Consulting**  
and  
Development Group  
Business Solutions Division, Knowledge  
**SAS**

---

## Table of Contents

List of Exhibits . . . . .	ii
1. Manufacturing in a Rapidly Changing Market . . . . .	1
2. The Role of Quality . . . . .	2
3. Enterprise Quality and Data Mining . . . . .	3
4. Case Study 1: Printing Process Out of Control . . . . .	6
5. Case Study 2: Failures in Hard Disk Drives . . . . .	11
6. Summary . . . . .	17
7. References . . . . .	17
8. Companion Document. . . . .	18
9. Recommended Reading . . . . .	18
Credits . . . . .	19

---

## List of Exhibits

Figure 1. Quality Data Warehouse . . . . .	3
Figure 2. Three-level Quality Strategy . . . . .	4
Figure 3. P-chart for the Proportion of Banding . . . . .	7
Figure 4. Pareto Diagram of Press Type . . . . .	7
Figure 5. Data Mining Flow for Band Data . . . . .	8
Figure 6. Data Replacement Node . . . . .	9
Figure 7. Variable Selection Node . . . . .	10
Figure 8. Lift Chart for Banding Models . . . . .	10
Figure 9. Tree for Banding Data . . . . .	11
Figure 10. Data Mining Flow for Hard Drive Failure Analysis . . . . .	15
Figure 11. Lift Curves for the Best Neural Net, Decision Tree, and Regression Model . . . . .	16
Table 1. Potential Causes for Bands in the Printing Process . . . . .	6

---

## 1. Manufacturing in a Rapidly Changing Market

The manufacturing industry has become more and more complex and has grown to include a variety of sub-sectors. Some of the sub-sectors in the manufacturing industry are the following:

- exploitation of mineral resources (coal, oil, gas)
- high technology equipment such as computers
- consumer-oriented mass production such as food processing
- highly specialized capital goods production such as turbines
- process-oriented production such as chemicals
- discrete consumables such as cars.

Although these sub-sectors are very diverse internally, they all have in common production, research, and development facilities, which often can be large. Most if not all of them also face very similar production issues such as new product development, quality management, capacity planning and tracking, preventive maintenance of production facilities, health, safety, and environmental protection, inventory optimization, and supply chain management.

The marketplace for manufacturing companies has changed drastically over the past 10 to 15 years. Production has become more complex. Most companies face increased competition both at home and abroad. Mergers and acquisitions create a series of opportunities and threats through economies of scale, integration, and globalization of operations. For example, in the engineering industry, the value of all cross-border business leaped from US\$43 billion in 1997 to \$78 billion in 1998, while many companies merged: Daimler/Chrysler, Ford/Volvo (both automotive), Honeywell/Allied Signal (automation systems/power generation), Veba/Viag, and Alstom/ABB (power generation) to name just a few.<sup>1</sup> In addition, manufacturing efficiency, quality control, and faster time-to-market influence competitive advantage.

---

<sup>1</sup>Marsh 1999

Manufacturing industries have reacted with increased investment in information technology to streamline production processes and to assemble data about their customers. These investments include spending on company staff as well as software. In some companies, such as ABB (mechanical and electronic engineering) and Cummins Engine (diesel engines), up to one-third of their development engineers are software engineers. Enterprise resource planning (ERP) systems from companies such as SAP and Baan have been installed as a standard to run the everyday business; however, such systems provide little help in adjusting to changes in customer demand.

Often going hand in hand with the drive for economies of scale is the move to restructure companies into smaller and more efficient sub-units with a large service component capable of reacting more quickly to ever-increasing customer expectations. Information delivery systems have been or are being introduced to track customer loyalty and market trends as early as possible.<sup>2</sup> Meanwhile the e-commerce revolution has already reached manufacturing. For example, volume car manufacturers such as General Motors and Ford are planning to coordinate their relationships with suppliers through online systems.<sup>3</sup>

---

<sup>2</sup>SAS, Inform 23, 1998

<sup>3</sup>Tait, Kehoe, and Burt 1999

On the production side, processes have been automated to the finest detail, and production is surveyed by measurement systems that collect a huge amount of data. Usually these data are only collected to signal if something has gone out of control so that operators are informed immediately to stop production and identify the potential root cause of the failure.

This best practices paper discusses quality-related aspects of the enterprise and explains some of the ways in which information technology can help solve quality problems in manufacturing data. These solutions are set in the context of developing quality efforts over time. The quality issue is discussed as one requiring management attention and an enterprise-wide solution approach. This paper focuses on the contribution that modern analytical techniques such as data mining can make to this approach and is substantiated with two case studies, one of a comparatively simple printing process and another from a more complex hard disk drive production process.

---

## 2. The Role of Quality

The interest in quality as a business topic was inspired by the success of Japanese production techniques in the 1960s and 1970s and in later years in other East Asian countries. Notable contributors such as W. Edwards Deming, Kaoru Ishikawa, and Joe Juran helped along the Quality Movement.<sup>4</sup> Inspection of crucial quality characteristics of manufactured goods became a widespread practice. Given mass production and the lack of automated measurement systems, inspections were initially done through *acceptance sampling*, that is, the inspection of a random sample from which conclusions were drawn about the underlying production lot or batch.

With the introduction of automatic measuring devices, this early phase of quality measurement led to continuous quality control online (during the production process). Widespread use of statistical process control (SPC) systems became standard, and control charts could be found everywhere on factory floors. When errors in production exceeded control limits, the root causes of failures were identified. The next step was to identify factors for problems in production through experiments designed offline. Often, production had to be stopped to run these experiments, so great care was taken to minimize the number of experimental settings (or runs) to restrict the cost of out-time. The heavy emphasis on statistical quality control is often referred to as the *first generation* quality initiative.

The first generation quality initiative is still in current practice but only as a baseline in the manufacturing industry. As Lori Silverman has noted, "The basic tools of quality are no longer sufficient to achieve the performance levels that today's organizations are seeking to maintain market leadership and competitive advantage."<sup>5</sup> Measuring quality continuously is a requirement, but the drawback is that SPC/experimental design only concentrates on individual processes. In modern production settings, manufacturing consists of numerous interrelated steps. For example, semiconductor manufacturing involves treating wafers of silicon in more than 100 steps. Typically, 100,000 of these are produced per day, which means a gigantic amount of process data is produced by manufacturing execution systems. Other sources of data cover other aspects of quality. SPC systems calculate quality-related *metadata*<sup>6</sup> such as statistics of subgroup samples or capability indices from process data. Laboratory information management systems (LIMS) include research and testing data, while ERP systems might add data about material resource planning or non-production-related data.

Modern quality implementation is therefore no longer restricted to controlling individual processes but has moved into a *second generation*, which considers the quality management of the whole enterprise. In this second generation, quality has become a top management issue. From an IT perspective, the quality initiative is now required to build up quality data warehouses,<sup>7</sup> which cover the whole production process including other quality-related data in an analysis-ready form. A quality data warehouse links supplier data, process data, data from other manufacturing plants, and human resource data to address questions such

---

<sup>4</sup>Deming 1986

---

<sup>5</sup>Silverman 1999

---

<sup>6</sup>*Metadata* are descriptive data or other information about data entities, such as field names and types, and are typically stored in data dictionaries and data warehouses.

---

<sup>7</sup>SAS, "The Quality Data Warehouse," 1999

as comparing quality across products, manufacturing lines, or plants, linking warranty problems to internal process data, or predicting product quality before the product reaches the customer (Figure 1).

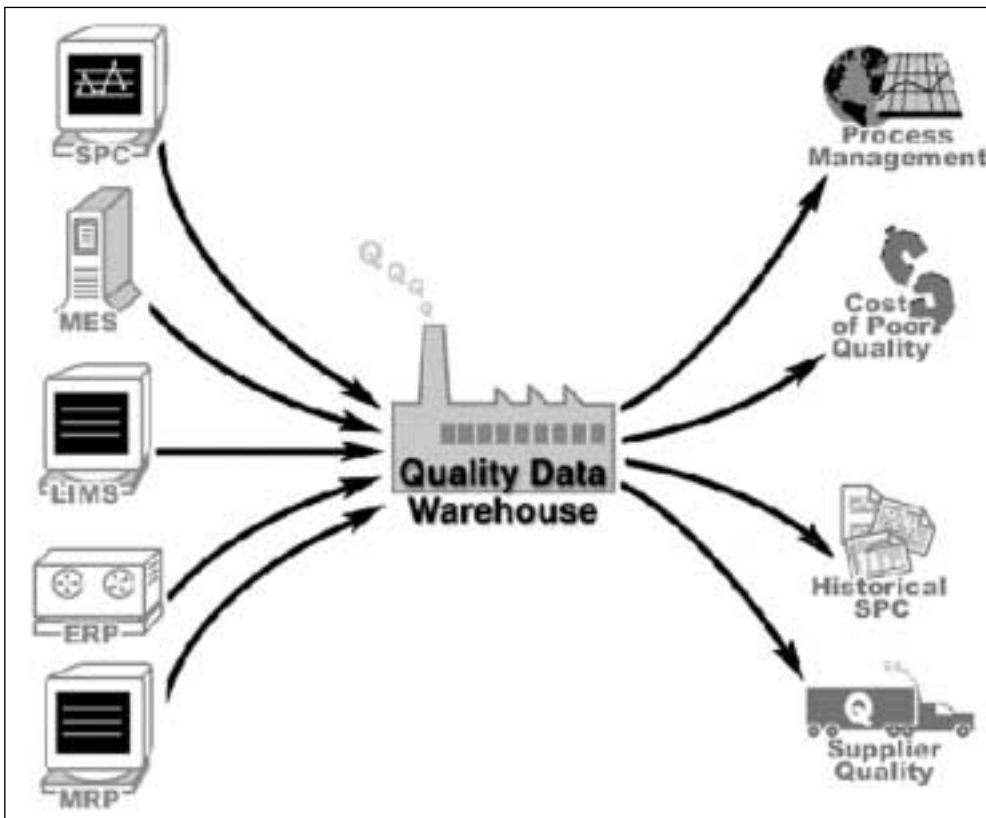


Figure 1. Quality Data Warehouse

Quality solutions that take into consideration the entire enterprise need to enable decision makers at various levels of the organization to make effective decisions that impact the quality of a process or product. For example, the introduction of an enterprise quality system at Gerber Products, the baby food company, has enabled floor operators to know exactly when to adjust the process and, just as importantly, when to leave the process alone.<sup>8</sup> At the same time, plant managers can track quality performance through a process flow reporting feature and identify immediately a root cause failure in the production upstream, while corporate management will receive standardized reports about key quality metrics on a regular basis.<sup>9</sup>

<sup>8</sup>SAS, SAS Communications, 1999

<sup>9</sup>SAS, SAS Communications, 1999

### 3. Enterprise Quality and Data Mining

Data warehouses populated with historical quality data serve to address questions of a more predictive nature, such as when a particular machine component is likely to break, and what combination of causes tend to lead to a malfunction in the production process. Questions of this nature require analytical modelling and/or data mining, which is a *third generation* of quality initiatives. *Data mining* is defined as the process of selecting, exploring, and modelling potentially large amounts of data to uncover previously unknown patterns for business advantage.<sup>10</sup> In contrast, more traditional decision support techniques like online analytical processing (OLAP) usually provide descriptive answers to complex queries and assume some explicit knowledge about the factors causing the quality problem.

<sup>10</sup>SAS, "From Data Mining to Business Advantage: Data Mining, The SEMMA Methodology and SAS Software," 1998

Analytical modelling can range from descriptive modelling using statistical analysis or OLAP to predictive modelling using advanced regression techniques and data mining methods. While data mining can generate high returns, it requires a substantial investment. Effective data mining requires well-defined objectives, high quality data in a form ready to be mined, and generally some amount of data pre-processing and manipulation. This technology is not a fully automated process. Data mining assumes a combination of knowledge about the business/production processes and the advanced analytical skills required to ask the right questions and interpret the validity of the answers. Typically data mining is done as a team effort to assemble the necessary skills. A feedback loop to deploy data mining results into the production system ensures that a return on investment can be realized together with some clues on how to repeat this exercise for the next problem to be addressed.

Thus, a three-level quality strategy can be employed in which each level serves as a precursor to the next, and each new level generates increased knowledge about the production process and additional return of investment (Figure 2).

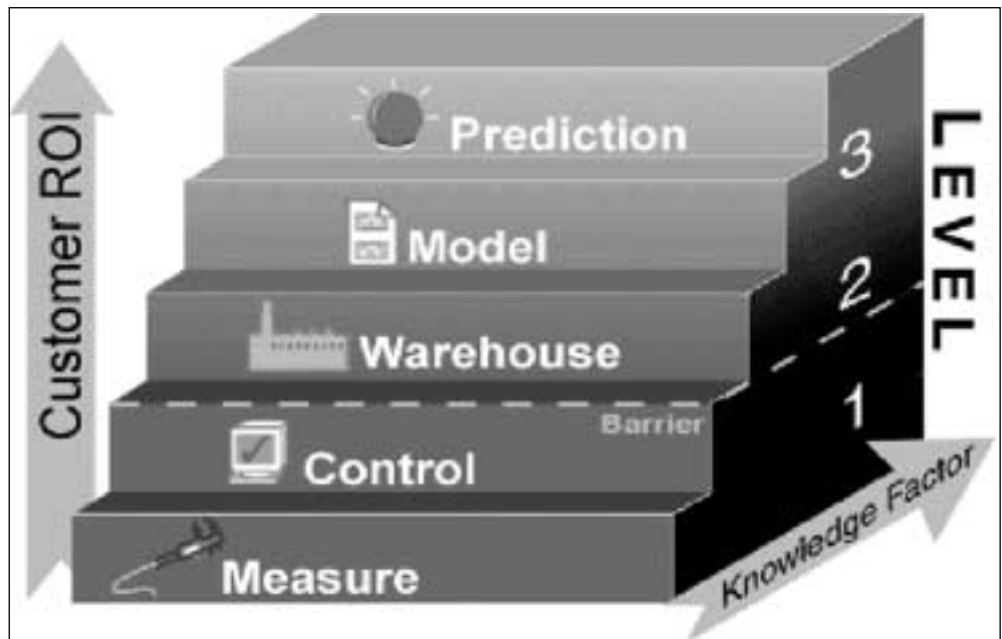


Figure 2. Three-Level Quality Strategy

Fortunately, manufacturing data lends itself well to advanced analytics and data mining. There is an abundance of data that are usually of high quality because their acquisition is automated. What is required is to establish a habit of storing historic data for mining analysis.

In the first generation of quality management, the quality control approach, data are typically only used for online SPC and then discarded or else archived but never analyzed.

In the second generation of quality management, the enterprise quality solution approach, data are also generated about research, suppliers, customers, and complaints. Such data are vital if the production data are to be enriched and exploited intelligently.

Decision support or data mining has been used successfully to streamline processes in manufacturing. The following are a few examples:

- Honda Motor Company in the United States is using Weibull analyses to predict at what age or mileage various components of cars are likely to fail. The resulting information allows engineers to plan maintenance schedules and design cars that will last longer. This careful analysis and the feedback of its findings into production have enabled Honda to achieve some of the highest resale values for cars in the United States.<sup>11</sup>
- A major South African power generating station experienced problems with tube failures in a re-heater. Tube failures are very costly; the material costs to replace the damaged tubes, the labor cost to perform the scope of work, the cost of lost production, and the external costs required to replace the lost production all add up. The company sought a method that would enable it to predict the potential tube failures to plan maintenance.

---

<sup>11</sup>SAS,  
SAS Communications, 1999

Data mining and multidimensional visualization techniques showed that the problem was due to a high local wear rate of a certain tube. Further investigation revealed that the inlet header disturbed the airflow, which caused the high local wear rate. A different setting of the inlet header reduced the wear significantly. Increasing the tube life by just one year delivers an estimated return on investment of 480 percent, an estimate that considers only the tube itself. Taking into account the damage and costs incurred for the re-heater and the wider effects for the entire plant, the return on investment is considerably greater.

- Data mining is used in semiconductor manufacturing to predict the likelihood that a micro-processor die will fail after packaging. It is often more cost-effective to discard defective die packages than to rework them. By pre-classifying each die with a probability of failure, the manufacturer can discard those with high probabilities very early in the assembly cycle. This analytics-based selection process eliminates unnecessary manufacturing costs and increases the percentage of good parts exiting the assembly/test process.
- Computer hard disks are produced at mass quantities (100,000 parts per day) with a current failure rate of 1 percent. With a cost of \$25 for each failure, even an improvement of 0.25 percent in the failure rate results in cost savings of \$2,281,250 per year.<sup>12</sup> Case Study 2 covers the details of this example.

---

<sup>12</sup>Collier 1999

There are many more examples where data mining has proven to be extremely useful for process control applications, maintenance interval prediction, and production and research process optimization. These examples include reducing inventory by 50 percent without any loss in service levels, optimizing filling operations in the food industry, optimizing yield in car engine testing, predicting peak loads in telecommunication networks, forecasting utility demand (water, gas, electricity), reducing energy consumption at power stations, and identifying fault patterns in gas drilling.

Data mining is also widely employed in sales and marketing operations, for example, to calculate the profitability of customers or to find out which customers are most likely to leave for the competition. Forrester Research reported in a recent study of Fortune 1000 companies comparing current (1999) and planned (2001) usage of data mining that while marketing, customer service, and sales will remain as the major business application areas for data mining, process improvement applications will experience the highest relative increase from 2 percent in 1999 to 22 percent of all data mining application areas in 2001 (multiple responses accepted).<sup>13</sup>

---

<sup>13</sup>Forrester Research Inc. 1999

## 4. Case Study 1: Printing Process Out of Control

<sup>14</sup>Evans and Fisher 1994

This case study was one of the earliest published examples on the use of data mining techniques to address a process-related problem. Bob Evans and Doug Fisher<sup>14</sup> discussed the problem of “banding” in rotogravure printing occurring at R.R. Donnelly, America’s largest printer of catalogues, retail brochures, consumer and trade magazines, directories, and books.

Rotogravure printing involves rotating a chrome-plated, engraved, copper cylinder in a bath of ink and pressing a continuous supply of paper against the inked image with a rubber roller. Sometimes a series of grooves – called a band – appears in the cylinder during printing and ruins the finished product. Once a band is discovered, the printing press needs to be shut down, and the band needs to be removed by polishing the copper cylinder and re-plating the chrome finish. This process causes considerable downtime, delaying time-critical printing processes, which wastes time, money, and resources. Banding became a considerable cost factor at R.R. Donnelly, and a task force was appointed to address the problem. In brainstorming sessions, the task force discussed a number of possible reasons for banding to avoid the problem in the first place, but the task force came up with a large list of factors, which could have potentially contributed to this problem. There were 37 factors being selected, some of which are listed in Table 1.

Name	Model Role	Measurement	Type	Format	Informal	Variable Label
CUSTOMER	id	nominal	char	#14.		CUSTOMER
JOBNUMB	id	interval	num	BEST9.		JOB NUMBER
ANODESPC	input	interval	num	BEST9.		ANODE SPACE RATIO
BLADEHFG	input	nominal	char	99.		BLADE HFG
BLADCPRS	input	interval	num	BEST9.		BLADE PRESSURE
CALIPER	input	interval	num	BEST9.		CALIPER
CHROMECC	input	ordinal	num	BEST9.		CHROME CONTENT
CURDENS	input	ordinal	num	BEST9.		CURRENT DENSITY
CYLINDIV	input	binary	char	99.		CYLINDER DIVISION
CYLSIZE	input	nominal	char	99.		CYLINDER SIZE
DIRECTSTA	input	nominal	char	99.		DIRECT STERN
ESAMP	input	ordinal	num	BEST9.		ESA AMPERAGE
ESAVOLT	input	interval	num	BEST9.		ESA VOLTAGE
GRAINSCR	input	nominal	char	99.		GRAIN SCREENED
HARDNER	input	interval	num	BEST9.		HARDNER
HUMIDITY	input	interval	num	BEST9.		HUMIDITY
INKCOLOR	input	nominal	char	99.		INK COLOR
INKPCT	input	interval	num	BEST9.		INK PCT
INKTEMP	input	interval	num	BEST9.		INK TEMPERATURE
INKTYPE	input	nominal	char	99.		INK TYPE
MILLLOC	input	nominal	char	912.		PAPER MILL LOCATION
PAPERTYP	input	nominal	char	99.		PAPER TYPE

Table 1. Potential Causes for Bands in the Printing Process

The task force studied conditions under which bands occurred and the settings of the potential causes (inputs) at that time. For control purposes, a number of settings were also recorded when the printing process was in control (no bands). For organizational reasons, the task force was not in a position to assemble a lot of data, and data were recorded when it was convenient, not in a controlled environment setting. In total, the data consisted of 541 records with 255 in 1990, 223 in 1991, 37 in 1992, and 16 in 1993. Bands occurred in 227 cases (about 42 percent). There were no bands in 312 cases (57.7 percent), and there were missing values for band in two cases.<sup>15</sup>

<sup>15</sup>These data are publicly available at <http://www.ics.uci.edu/pub/machine-learning-databases/>

The task force tried to analyze the data through a series of graphs but failed. In this re-analysis of the data, a P-chart (proportion of defectives) was applied to the target variable BAND/NO BAND (Figure 3). The chart was restricted to data points from April 1990 to November 1991, and data were summarized by month to have an acceptable number of points per month to calculate the proportion of defectives.

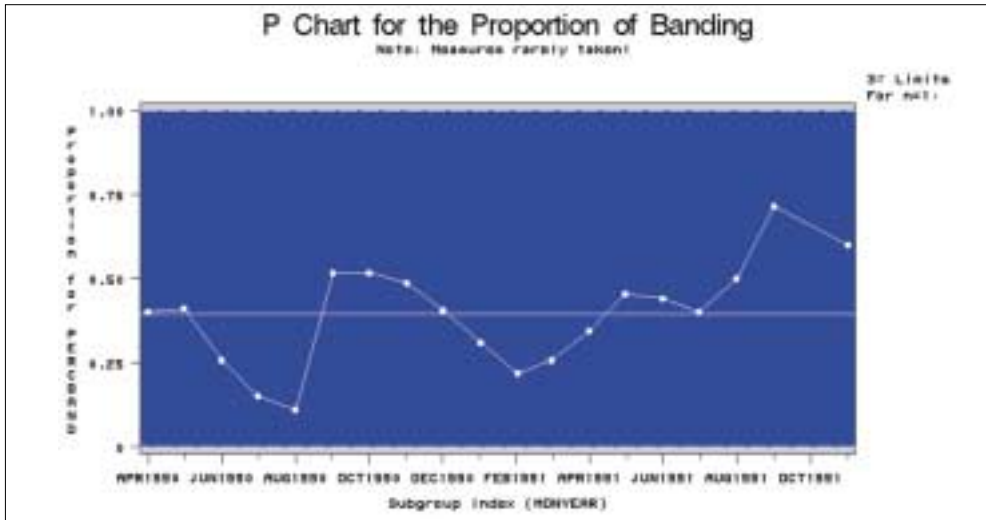


Figure 3. P-chart for the Proportion of Banding

At first sight, the printing process seems to be in control. The chart shows the proportion of defectives to be in the admissible range (Figure 3, shown in blue). However, because the data were not generated in a controlled experiment, the mean proportion of bands is artificially high (mean proportion for banding .395), and the admissible range covers the whole data range from 0 to 1 or no band to 100 percent banding. Figure 4 also shows a slightly upward trend of bands occurring, so the printing problem gets worse over the time frame considered.

One way to identify potential causes is with Pareto diagrams. Figure 4 shows a Pareto diagram of press types (four different types were used) against banding problems in percent. The first two printing machines with highest occurrence of banding already accounted for 80 percent of all banding problems, a clear indication that press type is an important influential factor for bands.

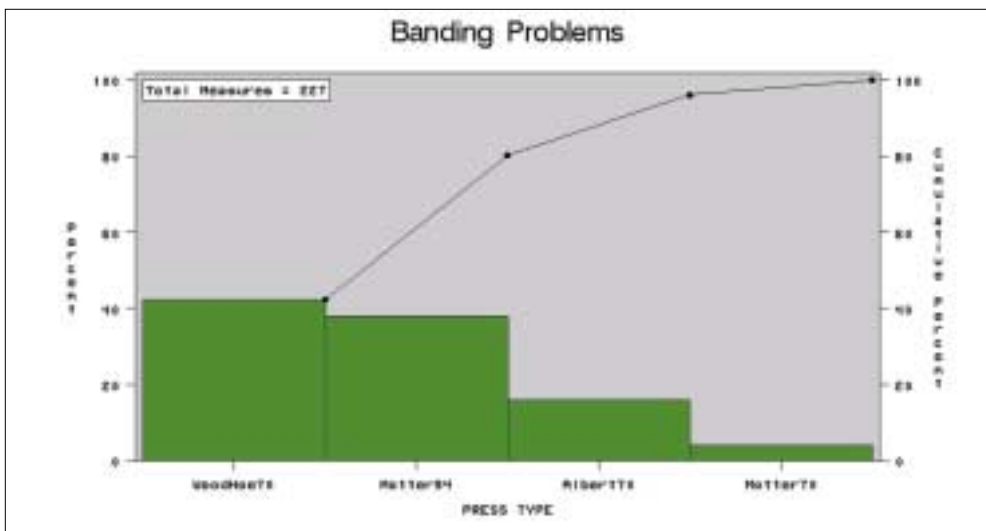


Figure 4. Pareto Diagram of Press Type

The value of Pareto charts is limited when numerous potential factors need to be considered. Moreover, Pareto charts are not suited to explore potential interactions between factors. In fact, the task force at R.R. Donnelly also did discover the impact of press type early on; however, although this factor was taken into account, the banding problem continued to exist at a lower rate. Clearly, control charts and Pareto diagrams (first generation of quality implementation) were not adequate to explain the banding problem fully.

Evans and Fisher therefore decided to use specific data mining methods (a decision tree algorithm) available at that time. These data were re-analyzed using SAS' data mining solution Enterprise Miner.<sup>16</sup> Enterprise Miner implements data mining analysis as a process. Data mining tasks are represented as icons, which can be dragged and dropped onto a workspace, arranged as nodes in sequence, and connected to form process flow diagrams. Data mining tasks are grouped according to an underlying data mining methodology called SEMMA, which stands for Sample, Explore, Modify, Model, and Assess.<sup>17</sup>

<sup>16</sup>For more details about Enterprise Miner, see the "Companion Document" section.

<sup>17</sup>SAS, "From Data Mining to Business Advantage: Data Mining, The SEMMA Methodology and SAS Software," 1998

Figure 5 shows a data mining flow using Enterprise Miner for the Donnelly banding data in the Diagram Workspace (the right side region of the graphical user interface). The Input node reads the data and automatically assembles a number of statistical metadata. These are descriptive statistics about each variable such as the role of the variable in the modelling process (input, target, identification, or a few other choices), and the variable's measurement level (nominal, binary, ordinal, or interval).

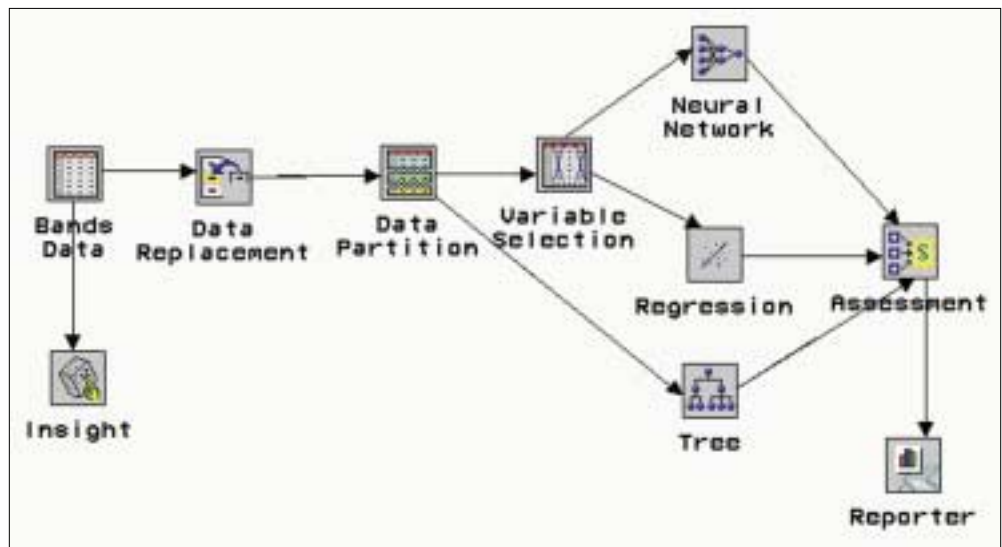


Figure 5. Data Mining Flow for Band Data

The Data Replacement node is one of the icons for data modification. As the name indicates, it allows replacing invalid data through user-selected values or the imputation of missing values using a wide range of imputation methods. Figure 6 displays an opened Data Replacement node with the Class Variables tab selected. As it appears, there were a number of inconsistencies (mainly between upper and lower case letters) in the original data, which blurred the analysis. A user-specified replacement value eliminates this problem with the data.

Also there were a number of missing values in the original data (up to 12 percent for a given variable). A typical strategy would be either to disregard cases with missing values for analysis or replace missing values with a representative value, such as the mean or the mode (the most frequent value). Instead, a strategy to impute missing values through *tree imputation* was chosen. Tree imputation uses all available information except the one from the imputed variable (all

information from user-selected variables) from this specific record as input to calculate the value of the imputed variable with a tree algorithm. This approach ensures that a maximum of information is used for imputation and the imputation itself is done in a very flexible way.

Name	Status	Imputation Method	Replace Value	R
CVLINDER	don't use	none		r
GRAINSCR	use	none		i
INKCOLOR	use	user specify -	KeV=KEV, koy=KEY	i
PROCFCTD	use	none		i
SLAGE#FG	use	none		i
CVLINDIV	use	user specify -	gallatin=GALLATIN	i
PAPERSTYP	use	user specify -	coated=COATED, uncoated=UNCOATED	i
INKTYPE	use	user specify -	coated=COATED, cover=COVER, uncoated=UNCOATED	i
DIRECTSTH	use	user specify -	no=NO	i
SOLUTYPE	use	none		i
TYPECYL	use	user specify -	no=NO, yes=YES	i
PRESSTYP	use	none		i
PRESS	use	none		i
UNITNUM	use	none		i
CVLSIZE	use	user specify -	catalog=CATALOG, spiegel=SPIEGEL, tableid=TABLEID	i
RILLLOC	use	user specify -	CANADIAN=CANADIAN	i
PLATINGT	use	none		i
ESAMP	use	none		i
CURDENS	use	none		i
CHROMECD	use	none		i
BANDTYPE	use	none		t

Figure 6. Data Replacement Node

Partitioning of the data is the next task in the data mining flow. Partitioning provides mutually exclusive data for *training*, that is, calculating the model for explanation of bands and subsequent assessment (comparison) of the models. As a result, assessment of the models is done on data independent of those used for model generation.

From data partitioning, the flow divides into a branch pointing to a Tree node and another branch that points into a Variable Selection node first and then into a Regression and a Neural Network model (see Figure 5). Variable selection is a very useful tool with a large number of model inputs. It assists analysts by dropping those variables that are unrelated to the target and retaining those that are useful for predicting the target (bands) based on a linear framework.

The remaining significant variables are then passed to one of the modelling nodes such as the Regression or Neural Network nodes for more detailed evaluation. Figure 7 shows results from the Variable Selection node. A number of variables are rejected because of their low relationship with the target. In this process, the Variable Selection node always tries to reduce the number of levels of each class variable to groups to test if that strengthens relationships with the target. If that is the case, then the original variable is rejected, and the grouped variable is selected instead.

Assuming for the moment that modelling has been completed, assessing all of the models would reveal which model explains banding most effectively. For example, double clicking on the Assessment node enables the analyst to select each of the three models and display a lift chart (Figure 8). For each model, records of the validation data are scored with the result (the formula) from the model, ordered from highest to lowest score, and then separated into deciles. In its first decile, the tree model classified 95.5 percent records correctly as bands. The baseline (Figure 8 in dark blue) shows the average of bands in the assessment data

(about 36 percent). The larger the distance between a model and the baseline, the better the model explains the data. As is apparent from the lift chart, the tree model outperforms the regression and neural network models.

Variables	Log	Output	Code	R-Squared	Effects	Notes
Name	Role	Rejection Reason	Dependencies	% Missing	# of Levels	
RMODESPC	rejected	Low RE w/ target		3%	68	
CURRDENS	rejected	Group variable Q_CURRDE preferred		2%	6	
HARDENER	rejected	Low RE w/ target		1%	21	
ESRAMP	rejected	Group variable Q_ESRAMP preferred		1%	3	
SDLVPCT	rejected	Low RE w/ target		1%	21	
INMPCT	rejected	Low RE w/ target		1%	76	
BLADEPRE	rejected	Low RE w/ target		1%	33	
ROUGHNES	rejected	Low RE w/ target		0%	17	
HURIDITY	rejected	Low RE w/ target		1%	35	
INMTEHP	rejected	Low RE w/ target		1%	53	
CN_LIFER	rejected	Low RE w/ target		0%	18	
PRODFCUT	rejected	Low RE w/ target		1%	23	
HILLLOC	rejected	Group variable Q_HILLD preferred		0%	6	
CVLSIZE	rejected	Group variable Q_CVLSIZ preferred		1%	3	
UNITNUM	rejected	Group variable Q_UNITNU preferred		0%	8	
PRESS	rejected	Group variable Q_PRESS preferred		0%	8	
SDLVTYPE	rejected	Low RE w/ target		1%	3	
CYLINDIV	rejected	Low RE w/ target		0%	1	
PRODFCTD	rejected	Low RE w/ target		1%	2	
INMCDOR	rejected	Low RE w/ target		0%	1	
Q_UNITNU	input			0%	7	
Q_PRESS	input			0%	6	
Q_HILLD	input			0%	3	
Q_ESRAMP	input			0%	3	
Q_CVLSIZ	input			0%	3	
Q_CURRDE	input			0%	6	
CHROMECCO	input			2%	3	
CH_LCDDI	input			1%	1	

Figure 7. Variable Selection Node

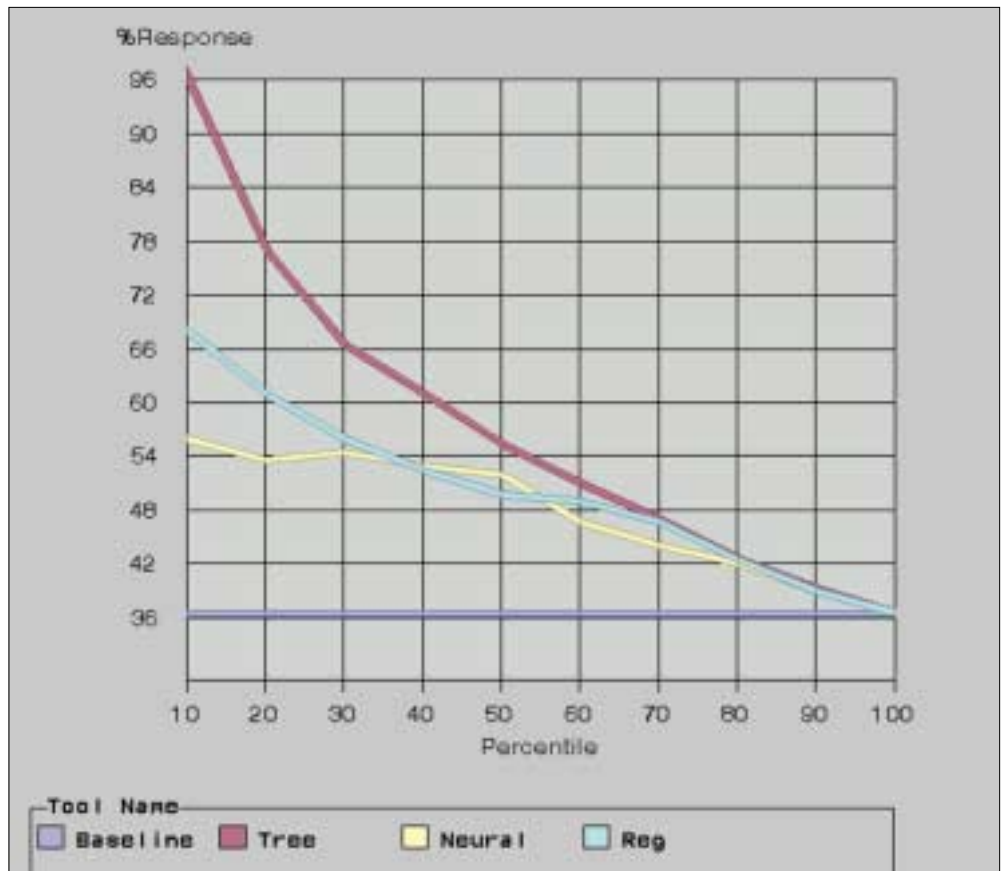


Figure 8. Lift Chart for Banding Models

Given its performance, it makes sense to take a closer look at the tree model results (Figure 9). A tree (also called a decision tree) is so called because the predictive model for banding can be represented in a tree-like structure. A decision tree is read from top down starting in the root node. Each internal node represents a split based on the values of one of the inputs with the goal of maximizing the relationship with the target. Consequently, nodes get purer (more or fewer bands depending on the split) the further down the tree. As is apparent, the percentage of solvent explains the most variation for bands. Whereas bands for the whole training data occur in 46.1 percent of cases, bands *always* result when the solvent percent is less than 31.35. Where the solvent percentage is equal to or greater than 31.35, press type is the next most important classifier. As noted earlier with the Pareto diagram (Figure 4), the press type Woodhoe 70 would generate the most bands, but the analysis has shown that this relationship would only be of interest for solvent values larger than 31.35 percent. The terminal levels of a tree are the “leaves.” For the branch Woodhoe 70, the humidity <78.5 indicates again a very high probability that bands will result. The sequence of steps in a tree generates rules by which each record can be classified. A sample rule would be: “If solvent percent >31.35 and press type is Woodhoe 70 and humidity <78.5 then the outcome will be band.”

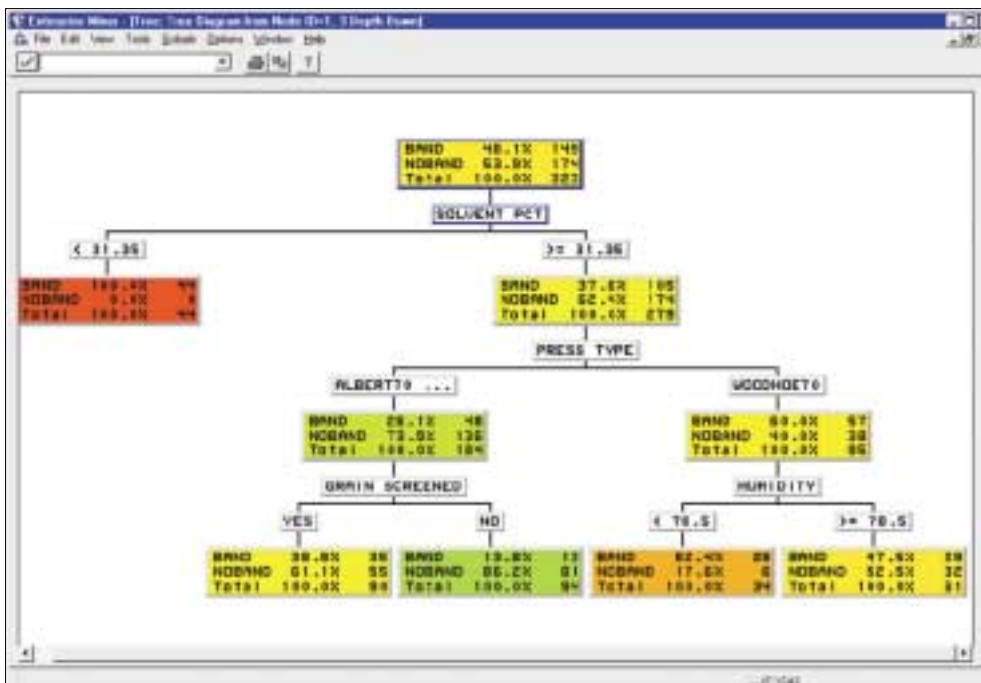


Figure 9. Tree for Banding Data

Obviously, obtaining these results would be difficult using trial and error or simple experimentation (first generation of quality implementation). The structured data-driven data mining process has simplified in a seamless fashion the process of finding root causes for bands, their relationships, and cut-off points. The next action then would be to implement findings according to the tree results, re-record bands and settings, and see if the process is now under control. If not, then recommence with the data mining process as described.

## 5. Case Study 2: Failures in Hard Disk Drives

In 1997, Western Digital, a leading hard disk drive manufacturer based in Irvine, California, engaged the consulting firm KPMG LLP to help design and implement a quality assurance data warehouse. The hard disk industry is highly competitive, product lifecycles are short, and profit margins are relatively small. Addressing the competitive challenge, Jerry Hill, Vice

---

<sup>18</sup>KPMG Consulting LLC 1999

President of Data Warehousing at Western Digital, commented that, "Increased competition and reduced product life cycles have eroded profit margins across the industry – to survive in this business, you've got to find new ways to reduce product development costs, improve quality, and deliver superior customer service."<sup>18</sup> Building a manufacturing quality data warehouse was the first step toward conducting extensive quantitative analyses and understanding better the factors affecting quality.

One of the issues for Western Digital was determining the root causes of field failures. Western Digital designs and manufactures hard drives for personal and enterprise-wide computing and markets them to systems manufacturers and resellers under the Western Digital brand name. Western Digital has long been recognized as a leader in hard drive quality assurance methods. A combination of traditional quality improvement methods and the new quality information system (QIS) data warehouse enabled Western Digital to reduce its field failure rates to approximately one percent, an enviable quality level across the manufacturing industry. However, to maintain the competitive advantage, Western Digital sought to reduce the rate of field failures even further by using data mining to identify root causes.

The company required a product-centric data storage system that would house information about the full life cycle of each of its hard drives – from manufacturing and testing to shipment and returns – individually traceable by serial number. The QIS data warehouse was the solution that provided these capabilities. (The QIS warehouse was the winner of the Data Warehouse Institute's coveted *Best Practices in Data Warehousing* award for 1999.) Once the warehouse was built and populated, Western Digital engaged KPMG's Center for Data Insight to demonstrate the value of the warehouse through the use of data mining for field failure root cause analysis.

---

<sup>19</sup>KPMG Consulting LLC 1999

"Western Digital's hard drives each comprise more than 200 components manufactured by outside suppliers, and the company manufactures well over 100,000 units per day. Finding the root of quality issues requires the ability to trace separate parts not only to their vendors, but also to their lot."<sup>19</sup> The further down the manufacturing stream that defects are detected, the greater the cost to the manufacturer in terms of dollars, as well as customer confidence. For example, suppose a field failure costs the company, on average, \$25 to resolve. At 100,000 units per day, a reduction in field failures of 0.25 percent translates into annual cost savings of \$2.3 million per year. This estimate does not even consider the value gains realized by increased customer satisfaction. Unfortunately, traditional data analyses such as the combination of online transaction processing (OLTP) and OLAP are insufficient for identifying root causes of field failure. More advanced analytics such as data mining are required for this objective.

## Data Mining Approach

In addition to the Center for Data Insight methodology, this case study used SAS' SEMMA methodology with Enterprise Miner™ to achieve meaningful, actionable results. Each phase is described below.

### Sample

Joining all relevant tables from the QIS warehouse resulted in a data set with approximately 700 variables. Most of these variables represented various quality parameters collected during the drive manufacturing cycle. To ensure meaningful data mining results, a single drive family was isolated to create a homogeneous data sample from which to conduct the analysis.

An equalized stratified sample of drive data was taken to balance the proportion of good drives to bad drives. This was done using the stratification option in the Sample node so that the sample contained 22.2 percent field failures and 77.8 percent good drives.

## Explore

The objective during the exploration phase of analysis is to develop a deeper understanding of the data and identify areas for further evaluation and analysis.

Using Enterprise Miner<sup>4</sup> and SGI Mineset,<sup>5</sup> the characteristics of the refined data set were explored. Enterprise Miner's Distribution Explorer node provided useful histograms and helped visualize statistical distributions. Mineset provided additional visualization capabilities that helped analysts focus on a few drive head suppliers who were associated with higher than normal field failures.

## Modify

Often the raw data in the database requires modification and refinement before the mining algorithms can make effective use of them. Such modifications include the derivation of new variables, the binning of numeric data into more manageable groups, and other data manipulation, which generally result in an enhanced data set.

This project required a variety of these data modifications. These data primarily consisted of quality assurance measurements taken during the manufacturing process. The measurements are generally very precise, real numbers within a given range. Many of these variables were binned into equidistant or equally balanced groups to enhance the analysis.

Once a valid, homogenous sample was extracted, variable reduction methods were used to discard all key attributes, variables known to be insignificant, variables that are artificially correlated, and independent variables that are derived from or are highly correlated with other independent variables, the result of which was to reduce redundancy.

Due to the size of the data and the number of variables with missing values, pre-processing was necessary before models could be built. The Enterprise Miner Variable Selection node was used for reducing the columns with a high percentage of null values. This step reduced the number of columns dramatically.

A Data Set Attributes node was used to remove insignificant or artificially correlated variables. Combined with the previous variable reduction, this reduction left 38 variables (from an original 700) for modelling activities.

Some of the remaining 38 variables contain null values that were actually valid values. While Enterprise Miner's decision tree algorithms can handle these missing values, the regression and neural network algorithms cannot. Therefore, all null values in this data set were replaced with the value "Missing" using the Data Replacement node.

## Model

For the modelling phase, the data were partitioned into a 40 percent training, 30 percent *validation*, and 30 percent testing data sets so that the 20.2/77.8 stratified population distribution was maintained in each partition.<sup>20</sup>

---

<sup>20</sup> *Validation data sets* are used to fine tune and/or select the best model. After the best model is selected and tested, it can be used to score the entire database (Ripley 1996, p. 354).

The partitioned data set was then sent in parallel into neural network, decision tree, and regression algorithms. Four different configurations of each algorithm were used. The settings for each algorithm are briefly described below:

- forward logistic regression with Newton-Raphson with line search optimization
- backward linear regression with conjugate gradient optimization
- stepwise logistic regression with quasi-Newton optimization
- stepwise logistic regression with conjugate gradient optimization
- decision tree with chi-square splitting criteria with a significance level of .200
- decision tree with chi-square splitting criteria with a significance level of .050
- decision tree with entropy reduction splitting criteria
- decision tree with Gini reduction splitting criteria
- neural network with three neurons in the hidden layer, hyperbolic tangent activation function
- neural network with ten neurons in the hidden layer, logistic activation function
- neural network with three neurons in the hidden layer, Arc Tangent activation function
- neural network with one forty neurons in the hidden layer, hyperbolic tangent activation function.

### Assess

All of these algorithms were connected into an Assessment node to evaluate the relative performance of each. Model performance and discoveries are described in the “Results” section of this case study. Overall validation and verification of models must include the following methods:

- Inspection - Do the results make sense to the business experts?
- Accuracy - Does the model accurately predict unseen data?
- Cross-validation - Do other algorithms and tools reach the same conclusions?
- Evaluation – Can the results be deployed in actionable ways to improve business practices?

Figure 10 contains the overall Enterprise Miner data flow that shows each of the steps used in this analysis.

Enterprise Miner™ was used in three different configurations listed below.

1. *Client/server with computations on the server.* This configuration involves data residing on a Sun Enterprise™ 3000 server and data mining computations performed on the server as well. A SAS view of the Oracle® field failure table allows Enterprise Miner to connect directly to the database. This configuration minimizes disk usage on both the server and client since a view of the data table is used rather than importing the data. This configuration was slow due to the server load combined with network congestion.
2. *Client/server with computations on the client.* This configuration involves data residing on a server and the computations performed on a client workstation. A SAS view of the target data table allows Enterprise Miner to connect directly to the database. This configuration minimizes disk space but takes advantage of the dedicated processor on the client workstation. Network communications remain a bottleneck in this configuration.
3. *Client only.* This configuration involves both data and computations on the client. The field failure data were converted into a SAS data set format and transferred to the client workstation's disk drive. Customized SAS programming is required to complete the data transfer process successfully. This configuration is best for improving runtime performance during modelling. However, this increase comes at the expense of client-side storage space.

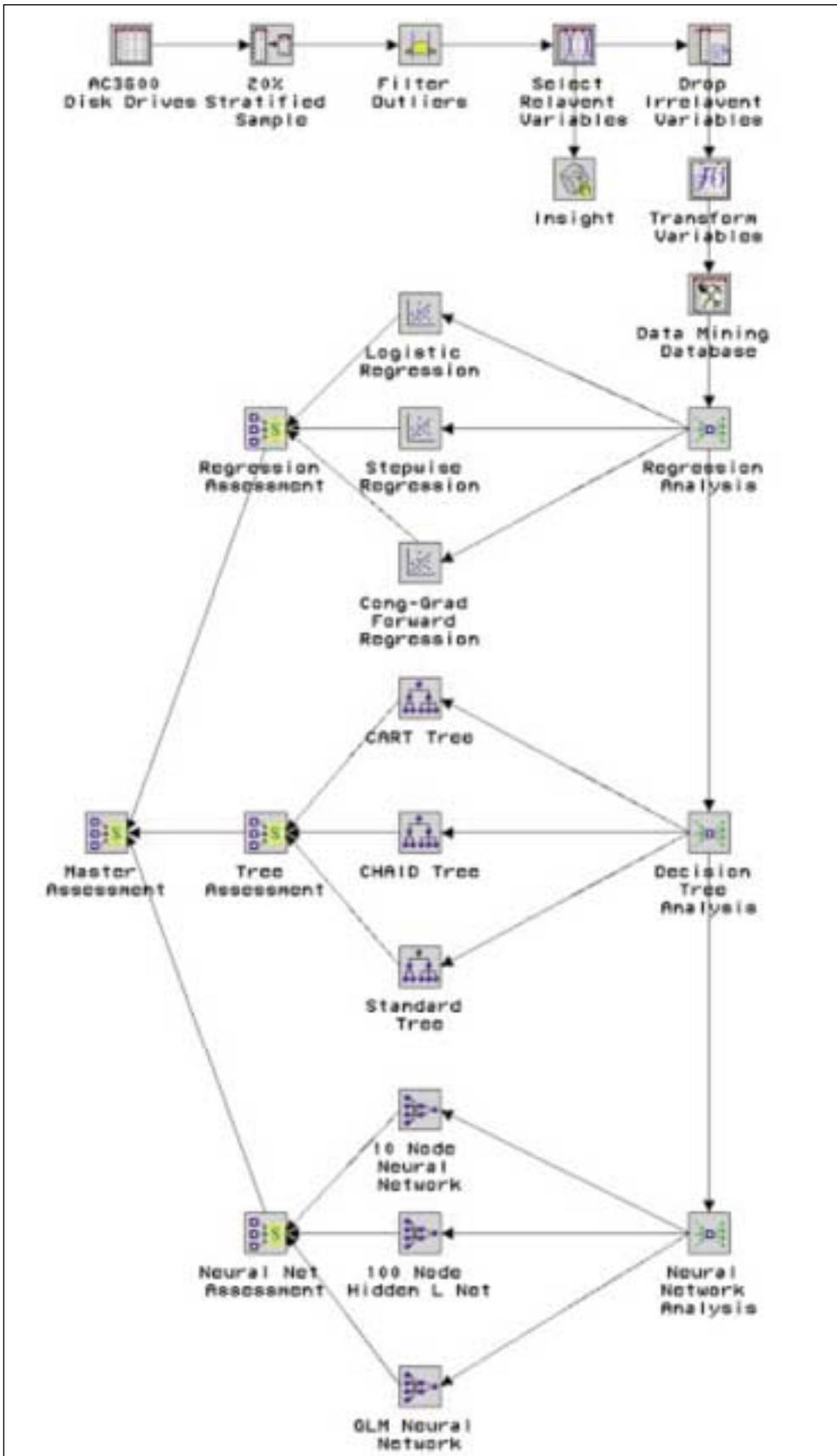


Figure 10. Data Mining Flow for Hard Drive Failure Analysis

## Results

The total run time for Enterprise Miner to manipulate the field failure data set, sample it, partition it into three data sets, and to run four regression models, four decision trees, and four neural networks was approximately five hours on a 450 MHz Pentium with 256 MB RAM.

The top decision tree, top regression analysis, and top neural network were connected into an Assessment node. The lift curve of this node is shown in Figure 11. The neural network provided the best predictive power. The configurations of the best models and the column importance of each model are discussed below for each algorithm.

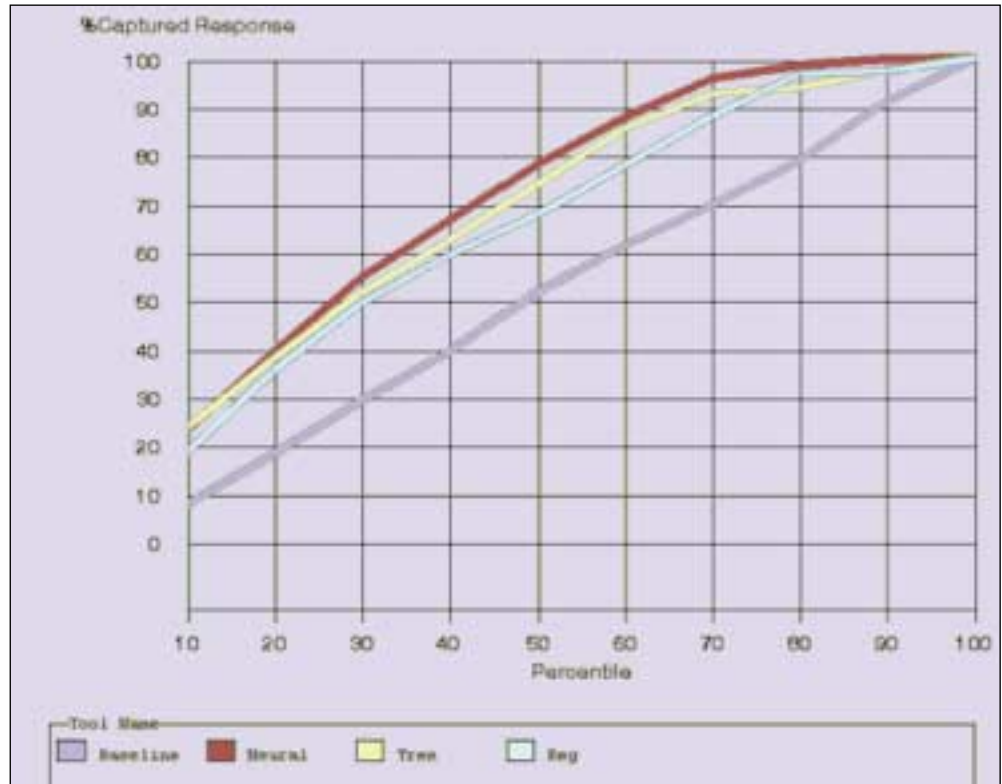


Figure 11. Lift Curves for the Best Neural Net, Decision Tree, and Regression Model

The stepwise logistic regression with the Quasi-Newton optimization performed better than all other regression methods. The decision tree with entropy reduction splitting criteria and a maximum of two branches per node outperformed other decision trees. In addition, the top performing neural network had three nodes in the hidden layer and used a deviance objective function. It is notable that all three of these models produced significant lift and that the most significant predictors found by the regression and decision tree algorithms were consistent with one another.

As a result of this analysis, Western Digital focused efforts on its component suppliers to improve component quality. Additionally Western Digital began analyzing its financial data to assess the cost and value of each of its supplier relationships. These measures are expected to result in a series of cost saving measures by focusing on high value suppliers.

## Recommendations

The following list shows some suggestions to get the best performance from Enterprise Miner:

1. *Data cleanup.* Any columns with a high percentage of null values must be removed. If a null value is an actual valid value, for example, null might equal false, then the null values must be changed to another value such as "Missing" or "False" so that Enterprise Miner does not ignore that record. Data cleanup is most important when modelling with the Neural Network or Regression nodes.
2. *Filter columns based on importance.* Use the Variable Selection node to prune the data set until only the columns with a certain degree of correlation to the target are used. Not only will this decrease the run times of the modelling nodes, it also will filter out a lot of noise that might impair the predictive performance of the modelling nodes.
3. *Weight the target.* With a training data set of 20.2 percent field failures and 79.8 percent good drives, simply predicting good drives all the time would provide an accuracy of 79.8 percent, which tells nothing about predicting field failures. Therefore, it is necessary to weight drives with field failures.

---

## 6. Summary

Effective quality solutions need to keep up with the complexity of the manufacturing processes. In this paper, a scalable three-level quality strategy was presented that has proven to successfully address the needs of the modern manufacturing enterprise. Data of individual processes should be combined to quality data warehouses to model the whole production process. Quality data warehouses should be designed to consider the needs of the final users (online supervisors, engineers, or even higher-level managers).

Increasingly, failures in manufacturing processes can no longer be attributed to a single root cause failure but are associated with a combination of causes somewhere downstream of the process that lead to a malfunction. Sophisticated reporting techniques like OLAP are needed to describe where the problem occurs, and with data mining, analysts can identify which combination of causes were responsible for the problem. The case studies, a printing process and an analysis of failures in hard disk drives, demonstrate the unique value of data mining solutions to manufacturing problems and the return of investment associated with implementing an enterprise-wide data mining solution.

---

## 7. References

- Collier, Ken (1999), "Is There Room Left to Change Your Processes? – Data Mining in Manufacturing," paper presented at the M'99 Data Mining Technology Conference, Cary, NC.
- Deming, W. Edwards (1986), *Out of the Crisis*, Mass: Massachusetts Institute of Technology, Center of Advanced Engineering Study, 1986.
- Evans, Bob and Fisher, Doug (1994), "Overcoming Process Delays with Decision Tree Induction," *IEEE Expert*, 9(1), 60-66.
- Forrester Research Inc. (1999), "Business-Centered Data Mining," February, 1-20.

- KPMG Consulting LLC (1999), KPMG Consulting, LLC Case Study, "Western Digital Rewrites Industry Standards for Quality with Global Data Warehouse," Information, Communications, and Entertainment, August 5.
- Marsh, Peter (1999), "Industry's Subtle Shift of Focus," *Financial Times Survey Engineering*, June 30.
- Ripley, B.D., (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.
- SAS Institute Inc. (1998), "Inform 23: A Natural Fit - SAS Solutions in Manufacturing," Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1998), SAS Institute White Paper, "From Data Mining to Business Advantage: Data Mining, The SEMMA Methodology and SAS Software," Cary, NC: SAS Institute Inc.
- SAS Institute (1999), "Gerber Products Company Selects SAS Institute's Enterprise-Wide Quality Solution," *SAS Communications*, 3Q, 19-21.
- SAS Institute Inc. (1999), SAS Institute White Paper, "Finding the Solution to Data Mining," Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1999), SAS Institute White Paper, "The Quality Data Warehouse," Cary, NC: SAS Institute Inc.
- Silverman, Lori (1999), "Quality Today: Recognizing the Critical Shift," *Quality Progress*, February, 53-60.
- Silverman, Lori (1999), "Recognizing the Critical Shift," *Quality Progress*, February, 53-60.
- Tait, Nikki, Kehoe, Luise and Burt, Tim (1999), "US Car Monoliths Muscle in on the Internet Revolution," *Financial Times*, November 8th.

---

## 8. Companion Document

- SAS Institute Inc., (1999), SAS Institute White Paper, "Finding the Solution to Data Mining: A Map of the Features and Components of SAS® Enterprise Miner™ Software," Cary, NC: SAS Institute Inc.

---

## 9. Recommended Reading

### Data Mining

- Adriaans, Pieter, and Zantinge, Dolf, (1996), *Data Mining*, Harlow, England: Addison Wesley.
- Berry, Michael J. A., and Linoff, Gordon, (1997), *Data Mining Techniques*, New York: John Wiley & Sons, Inc.

---

Sarle, W.S., ed. (1997), "Neural Network FAQ," periodic posting to the Usenet newsgroup comp.ai.neural-nets, <<ftp://ftp.sas.com/pub/neural/FAQ.html>> (accessed Dec 9, 1999).

SAS Institute Inc., (1997), SAS Institute White Paper, "Business Intelligence Systems and Data Mining," Cary, NC: SAS Institute Inc.

Weiss, Sholom M, and Indurkha, Nitin, (1998), *Predictive Data Mining: A Practical Guide*, San Francisco, California: Morgan Kaufmann Publishers, Inc.

## Data Warehousing

Berson, Alex, and Smith, Stephen J. (Contributor), (1997), *Data Warehousing, Data Mining and OLAP*, New York: McGraw Hill.

Inmon, W. H., (1993), *Building the Data Warehouse*, New York: John Wiley & Sons, Inc.

SAS Institute Inc., (1995), SAS Institute White Paper, "Building a SAS® Data Warehouse," Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1996), SAS Institute White Paper, "SAS Institute's Rapid Warehousing Methodology," Cary, NC: SAS Institute Inc.

Singh, Harry, (1998), *Data Warehousing Concepts, Technologies, Implementations, and Management*, Upper Saddle River, New Jersey: Prentice-Hall, Inc.

---

## Credits

*Mining for Quality in Manufacturing Data* was a collaborative work. Contributors to the development and production of this paper included the following:

### Authors

KPMG Consulting LLC  
Ken Collier

SAS  
Gerhard Held

### Consultants

KPMG Consulting LLC  
Curt Marjaniemi  
Don Sautter

SAS  
Mohan Namboodiri

### Technical Reviewers

KPMG Consulting LLC  
Knowledge Management Solutions Group

SAS  
Business Solutions Division, Knowledge  
Development Group



Tyson's Corner  
1676 International Dr.  
McLean, VA 22102  
Tel: (703) 747 3000  
Fax:(703) 747 8500

Silcon Valley  
500 East Middlefield  
Mountain View, CA 94043  
Tel: (650) 404 5000  
Fax: (650) 960 0566



SAS World Headquarters  
SAS Campus Drive  
Cary, NC 27513 USA  
Tel: (919) 677 8000  
Fax: (919) 677 4444  
U.S. & Canada sales:  
(800) 727 0025

SAS Europe, Middle East & Africa  
PO Box 10 53 40  
Neuenheimer Landstr. 28-30  
D-69043 Heidelberg, Germany  
Tel: (49) 6221 4160  
Fax: (49) 6221 474850

SAS Asia/Pacific & Latin America  
SAS Campus Drive  
Cary, NC 27513 USA  
Tel: (919) 677 8000  
Fax: (919) 677 8144

[www.sas.com](http://www.sas.com)