4-2005

A Method for Classification Analysis on Simultaneous Product Defects

Jason S. Trahan

A METHOD FOR CLASSIFICATION ANALYSIS ON SIMULTANEOUS PRODUCT DEFECTS

by

Jason S. Trahan

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Master of Science in Engineering (Industrial)
Department of Industrial and Manufacturing Engineering

Western Michigan University
Kalamazoo, Michigan
April 2005

ACKNOWLEDGMENTS

I acknowledge my committee members, Dr. Butt, Dr. Engelmann, and Dr. Mihalko. Their flexibility, rapid responses, expertise and understanding have been invaluable. I especially thank Dr. Engelmann for his devotion to my well being. I will treasure his guidance.

Of course, this would not have been possible without the values instilled in me by my parents, Peggy and Roy. I will cherish any endeavor on their behalf.

I am forever grateful for the support from my wife, Julie. Her sacrifices have kept our rollercoaster on track more than I probably know. I look forward to our continuous ride upward.

I dedicate this manuscript to the memories of three family members who will always remain very dear to me. They are my great-grandmother, Agnes Davis, my grandpa, Don Trahan, and my aunt, Yvonne Misner.

Jason S. Trahan

A METHOD FOR CLASSIFICATION ANALYSIS ON SIMULTANEOUS PRODUCT DEFECTS

Jason S. Trahan, M.S.E.

Western Michigan University, 2005

There have been many advancements that share similar tools and techniques that help reduce the manufacture of nonconformities. These include computer-aided analysis, design reviews, total quality management, multivariate analysis, process monitoring and control, and root cause analysis to mention a few.

This work details the methodology developed for manufacturing companies to predict attribute defects. Injection molding was used to demonstrate the proposed methodology. Data were collected on a variety of tool design and construction attributes thought to affect the performance of a tool. The independent variables consisted of categorical and numerical data types. The dependent variable was a nominal four-tuple describing the types of defects that can coexist on one part.

A series of steps taken to prepare the data set for classification tree analysis can be categorized by the following: 1) variable screening and selection due to missing data and high dimensionality and 2) causal analysis and similarity computations for combining defects, thus reducing the number of classes in the four-tuple. A method was designed for classification tree analysis. The models provided a way for designers and engineers to assess the potential for success prior to production.

TABLE OF CONTENTS

Table of Contents—continued

Table of Contents—continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Product Development

Manufacturing is an enormous industry comprised of a variety of sectors such as appliance, electronic, automotive and furniture. These sectors are also served by an array of processes such as welding, fabrication, casting and molding; not to mention an assortment of materials such as wood, ceramics, ferrous and non-ferrous metals, natural rubbers and plastics. A common thread to all manufacturing sectors is the process of introducing new products to market, known as product development. Although the focus of this work is on plastic injection molding, it is intended for most manufacturers handling product development.

The business of injection molding is complicated at best. It requires the skill and tools within many departments. The transfer of knowledge between those departments is vital for successful operations. Most plastic products are developed through three fundamental departments: part design, mold design, and molding. It is not necessary that one company house all these departments. In fact, it is very common that the skills and tools of one or more of these departments are outsourced.

Part design is responsible for transforming a concept into a product that is capable of being injection molded. Whenever possible, it is important that designers adhere to sound guidelines set forth by the industry and/or company. Most of these guidelines

represent a compilation of years of industrial experience and best practice (Gyrna, 1988c). Part design is most often the first stage in product development. It sets the stage for downstream decisions. It is also considered to have the most influence on product and mold performance.

The next step, mold design, creates a tool that conforms to the geometry of the part(s). During production, the mold is used to solidify molten plastic into a given shape. It seeks to incorporate the injection molding machine's specifications while achieving the stages of the molding cycle. The cycle includes mold filling, pack/hold, cooling and ejection. The impact of mold design may be best summarized by some experts claiming "even the best product design can be spoiled by a poor mold design, but a poor part design cannot be compensated by even the soundest mold design" (Anonymous, n.d.). Very frequently the construction of the mold begins before mold design is completed.

The third and final step is molding. Molding is the process of solidifying molten plastic into a desired shape. The molding machine is used to plasticate the material, inject the material into the tool under pressure, pressurize the tool cavity until solidification occurs, and actuate the tool to eject the part(s). The times for a typical molding cycle are shown as percentages in Figure 1.

The goal for processors is to attain the shortest cycle possible while maintaining the quality standards set forth for the product (Rosato & Rosato, 1995). It is here that many of the deficiencies in the product and tool design are realized. Wherever it is cost effective to do so, these are corrected. An initial period of trial and error is so common that an intermediate step between tool design and molding, called mold try-out or mold

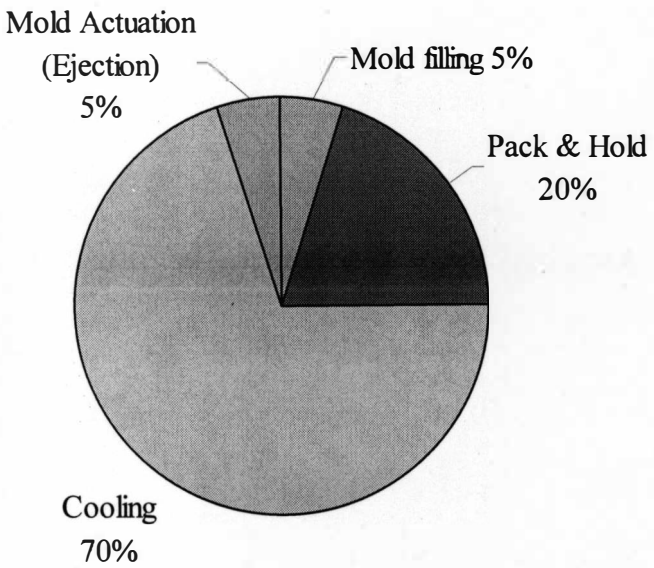


Figure 1. Steps of the Molding Cycle as a Percentage of the Overall Cycle Time

sampling is often scheduled beforehand. Figure 2 illustrates the sequence and overlap of the three fundamental steps in product development.

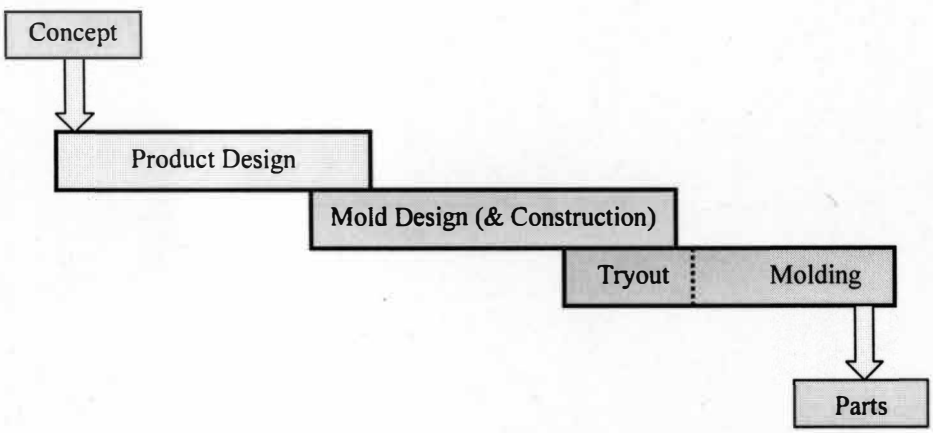


Figure 2. Fundamental Steps in Product Development

Problems in Injection Molding

The competitive nature within the industry and the onset of globalization has forced many companies to accept sub-par work. In other words, the given product design has violated guidelines and/or the tool has been designed and constructed poorly due to time and budget constraints. The end result is a narrow processing window for the molder, which often leads to narrow profit margins. A processing window is the range of molding parameters in which the product can be manufactured free of defects at an efficient cycle time (Moldflow, 2000). For example, a small processing window would be realized by a slight fluctuation in the tool temperature causing a short shot, glossy surface, or warping. A narrow processing window may also be realized when problems occur from the tool being run in different molding machines or the cycle time needs to be extended in one machine and not the other. In Figure 3, the flower represents a product that is having less room to grow as product development moves forward. Eventually the product may move outside the process window and be produced unsuccessfully.

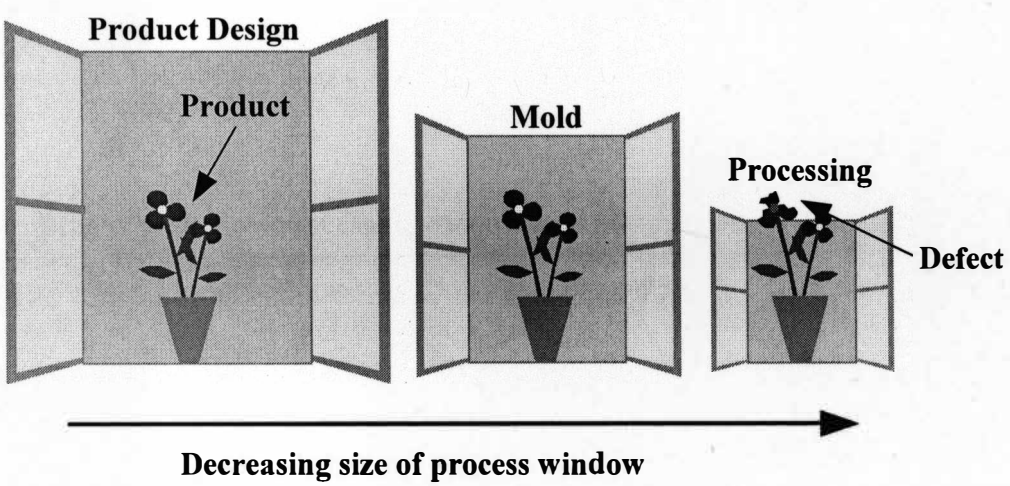


Figure 3. A Diminishing Process Window Results in a Poor Product (Flower)

Problems in injection molding can usually be traced to a flaw in at least one of the three fundamental phases in product development. For example, short shots, a lack of material in the product, can be a result of a long flow length constrained by the product design, insufficient venting designed in the tool, and/or a decrease in material temperature while processing. It should be noted that the properties of plastic material could be a source of many problems as well; however, this will be assumed to be a component of the part design process.

Tooling problems arise in all sorts of ways. The most common type of problem and the focus of this research is the attribute defect, meaning the presence of some undesirable (or absence of some desirable) characteristic in the product (American Society for Quality [ASQ], 2004). It is not uncommon for individual products to have several different defects.

Other non-attribute problems that occur in injection molding include tool damage, excessively lengthy startups, mechanical field failures, grease issues, hot runner failures, and blocked vents. These are often considered sporadic rather than chronic (Gyrna, 1988b, Latino & Latino, 1999). In other words, these types of problems are less frequent than those of attribute defects, but are generally more severe in terms of costs.

CHAPTER II

REVIEW OF RELATED LITERATURE

Defect Prevention

Designers are constantly introducing complicated parts that stretch the bounds of capability for the injection molding process. The industry continuously faces higher demands in terms of product quality, time to market and cost reductions. Furthermore, the competitive nature within the industry and the onset of globalization has made being a successful molder all the more challenging.

One of the ways companies are achieving success is by applying more resources up front in product development to prevent problems from occurring. Gyrna (1988a) illustrates this point in a cost of quality model (see Figure 4). He suggests that companies investing in prevention and appraisal have the opportunity to reap lower total costs while achieving lower levels of defects.

Gyrna (1988a) identifies four types of costs: internal failure, external failure, appraisal and prevention. Internal failure costs include scrap, rework, downtime and failure analysis. External failure costs consist of warranty charges, recalls and loss of sales. Appraisal costs are incoming inspection, in-process inspection, final inspection and maintaining accuracy of test equipment. Prevention costs include product reviews, process planning, process control, supplier evaluation and training. Although the nature of this work is in the form of an internal failure cost, its mission is to prevent defects,

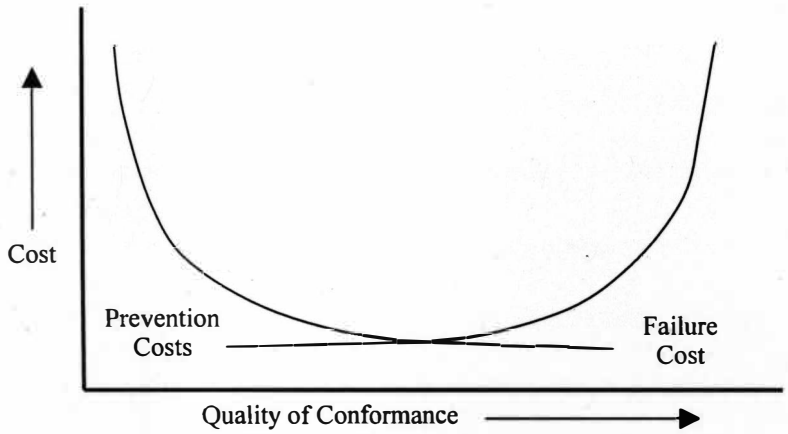


Figure 4. The Total Cost Optimum Compromises Prevention and Failure Costs (Gyrna 1988a)

which translates to fewer internal and external failure costs. It should also be noted that costs could be in the form of money, time, people, tools and other resources.

The following is an overview of some of the most popular trends in defect prevention found in literature and adopted in practice. Its purpose is to provide a framework for which this project exists. Many of the techniques listed are evolving daily and overlap each other. Some are very costly, some more time consuming and others are relatively new.

Design

Design Review- a systematic technique for evaluating a proposed design to assure that the product design quality reliably reflects and meets customer requirements within cost and time constraints (Gyrna, 1988c; Ichida & Voight 1996). Where appropriate,

formal and informal design reviews should be conducted throughout product development.

Design for Manufacture and Assembly (DFMA)- the process of applying three steps: 1) design for assembly, 2) selection of materials and processes and 3) design for individual part manufacture (Dewhurst, 2001). It serves as a basis for concurrent engineering studies so as to provide guidance to the design team in simplifying the product, reduce costs, and quantify the improvements. It removes the traditional "over the wall approach", where designers are one side of the wall throwing designs to manufacturing engineers, who have to handle the problems because of their lack of input (see Figure 5) (Boothroyd, Dewhurst & Knight, 2002).

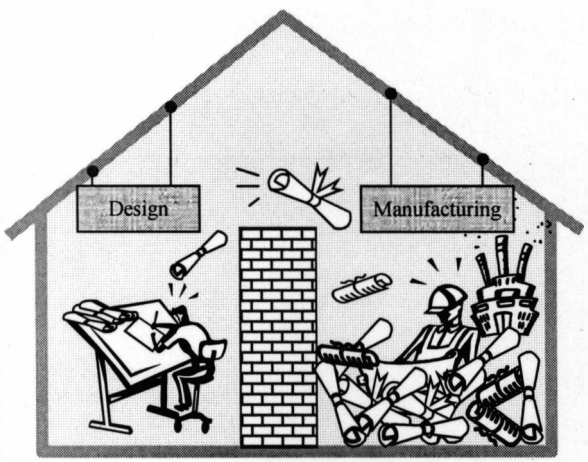


Figure 5. "We Design It, You Build It" Attitude Adapted from Boothroyd et al., 2002

Failure Mode, Effect, and Criticality Analysis (FMECA)- a methodical way to examine a proposed design for possible ways in which a failure can occur. Each failure is

reviewed by its impact on the entire system, frequency of occurrence, severity, likelihood of detection and action to minimize the failure (Gyrna, 1988c; Mobley, 1999).

Fault (Failure) Tree Analysis- similar to FMECA, but differs in three ways. It considers only serious failures, isolates failures by removing dependencies of other events and explicitly shows relationships between events. It is essentially the reverse of FMECA, which starts with the origins and causes and looks for any resulting negative effects (Gyrna, 1988c).

Prototypes and Validation- a prototype or trial version of a product provides engineers with a visual example and some insight into design and manufacturing. Some popular methods include stereolithograpy and laminated object manufacturing. In some cases, a sample tool with minimal features not only provides a prototype, but also serves to develop tool design data for building the production tool (Berins, 1991). Coupled with validation, a simulation of product use, a good indication of product performance can be realized. Other design methods for preventing defects worth mentioning are worst-case analysis, 8-D, lessons learned and deductive reasoning.

Computer-Aided Engineering

One of the recent technologies to affect injection molding has been computer-aided engineering (CAE). CAE includes computer-aided design (CAD) and computer-aided manufacturing (CAM). The rapid growth of CAE has transcended the "art" and "rules of thumb" governing plastic designs to a sophisticated science (Rosato & Rosato, 1995). The three most popular techniques in CAE with respect to defect prevention are flow analysis, cooling analysis and stress-strain analysis. Each method uses a finite

element analysis (FEA) or finite element difference (FED), depending on the application, to determine approximate solutions to physical problems described by differential equations. These analyses break up a structure into small elements, which are connected at points called nodes. When variables of interest (e.g., stress, temperature) are applied to the part, a series of equations can be solved to describe the distribution of values for the variable throughout the part (Rosato & Rosato, 1995; Groover, 2001).

Stress-strain Analysis- a way to simulate the potential for failure during use for products intended for mechanical applications. It enables designers to evaluate the effects and interactions of material properties, wall thicknesses and various types of forces. Although not as useful in predicting attribute defects, it is a very important tool in failure prevention.

Flow Analysis- simulates the flow of plastic throughout the tool cavity. The ability to calculate complex algorithms (flow and heat transfer equations) has provided rational solutions to many of the hard-to-understand problems (Rosato & Rosato, 1995). Flow analysis ties all three groups of product development together by addressing issues such as: material selection, processing parameters, gate location, runner sizes, wall thickness and fill time. It gives insight into potential problems such as warpage, residual stress, gas traps, weld lines, flash and short shots.

Cooling Analysis- simulates the heat exchanged between the plastic and mold/coolant (Rosato & Rosato, 1995). It considers the thermodynamic properties of the plastic, tool and cooling medium, the size and placement of cooling channels, the tool geometry, and processing conditions. It aids the designer in generating a uniformly

cooled part, thus preventing differential shrinkage, internal stresses, mechanical failures, and mold release problems. It also improves process economics by more accurately predicting cooling times. Moldflow Corporation has been an industry leader in both flow and cooling analysis while continuously adding features to their software such as the Moldflow Community Center, automated product usage feedback, gate optimization analysis, modeling tools and results visualization enhancements and enhanced interfaces (Anna-Reddy, 2003).

Statistical Methods

Statistical studies can typically be categorized in four ways: 1) controlled experiments, 2) controlled experiments with supplemental variables, 3) confirmatory observational studies, and 4) exploratory observational studies (Neter, Kutner, Nachtsheim & Wasserman, 1996). Controlled experimental research controls the variables for a predetermined analysis (StatSoft Inc., 2004). It focuses on the quantitative inputs and outputs of a model (National Institute of Standards and Technology [NIST], n.d.). Assumptions are usually made about the data to fit a model (NIST, n.d.). Controlled experiments with supplemental variables are used when variables are unable to be collected as part of the study. Instead, they are incorporated into the model to reduce error. Confirmatory observational studies monitor variables thought to influence a dependent variable. These studies use observational data to prove or disprove hypotheses. Exploratory observational studies cannot leverage previous studies or experiments. Often referred to as correlation (or classical) research, it utilizes intuition and the collection of many variables to draw a relationship to the response. There is no

intention of influencing variables or drawing early conclusions (StatSoft, Inc., 2004). The techniques used are often graphical and subjective, allowing the data to suggest a suitable model (NIST, n.d.). The independent variables are not controlled as in experimental research. For example, failures of a product may be caused by uncontrollable independent variables such as climate, service life and use. In general, experimental research yields statistically stronger conclusions. For example, an experimental study shows that higher light intensities produce taller plants, which implies a causal relation exists. Conversely, observational studies cannot irrefutably boast such causal relationships (StatSoft Inc., 2004).

Another and frequently used categorization of statistical practice is parametric and nonparametric. Nonparametric methods are used in cases when the researcher knows nothing about the parameters of the variable of interest in the population. Nonparametric methods are most appropriate when sample sizes are relatively small (e.g., $n < 100$). As samples become large, the sample means will follow the normal distribution. Intuitively, parametric methods, which are usually much more sensitive, have more statistical power than their counterparts (StatSoft Inc., 2004).

Traditional Methods

Design of Experiments (DOE)- once an opinion has been formed about which factors are most likely causing the problem(s), the next step is to test or verify the hypothesis. DOE is the branch of applied statistics employed to define and organize an experiment and analyze the results, so that the effect of each causative factor can be evaluated efficiently (Tsuyuki, 2001).

Multivariate Analysis- in the broadest sense refers to all statistical methods, which simultaneously analyze more than one variable. Coupled with advancements in computers, it is clear that these methods are required to efficiently and adequately study multiple relationships and obtain a complete realistic understanding for decision making (Hair, Anderson, Tatham & Grablowsky, 1979). Since most studies in defect prevention are not immune to this, univariate methods will not be discussed. "One researcher states: 'For the purpose of ...any...applied field, most of our tools are, or should be, multivariate. One is pushed to a conclusion that unless a ...problem is treated as a multivariate problem, it is treated superficially (Gatty, 1966).'"(Hair, et al., 1979, p. 4). Moreover, many multivariate techniques are an extension of univariate analyses (i.e. multiple vs. simple regression).

Mulitvariate analyses include regression, analysis of variance and covariance, discriminant analysis, principle components analysis and common factor analysis, canonical correlation analysis, cluster analysis, mulit-dimensional scaling and conjoint analysis. Each method has served a purpose in preventing defects and can be summarized in Figure 6. The fundamental splits for these methods depend on the nature of the variables and data types. The first division separates "dependence" from "interdependence". A dependence technique predicts or explains one or more dependent variables by other independent variables. An interdependence technique does not define variables as dependent or independent. Instead, all the variables are analyzed simultaneously in an effort to explain the entire set of variables.

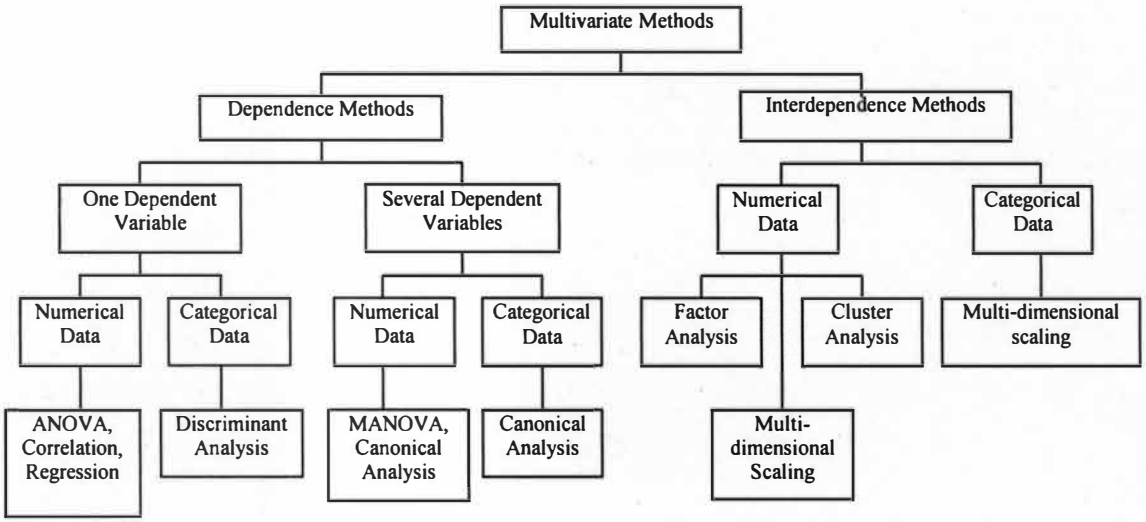


Figure 6. Categorization of Multivariate Techniques Adapted from Hair, et al. (1979)

Multivariate techniques can also be categorized in terms of their purpose or application. Functional multivariate methods are used for building predictive models or explaining relationships. Considerable understanding is necessary to conceptualize a realistic model. Structural multivariate methods, on the other hand, try to simplify complex relationships in a manner, which provides insight into an underlying and nonintuitive structure of relationships. They are more descriptive and less predictive in nature (Hair, et al., 1979).

Reliability Engineering- is the scientific process of developing a product that satisfactorily performs its prescribed function for a specified period of time under specified conditions. Measures of reliability include mean-time-between failures, failure rate, warranty claims, maintainability and availability. Although the terms reliability and quality are often used interchangeably, quality is ultimately defined by the customer and

usually includes reliability in its definition. In most cases, attribute defects are characterized as a quality defect, which can be located by conventional inspection techniques as opposed to a reliability defect, which requires some stress applied to create a detectable defect (Lamberson, 2000).

Non-Traditional Methods

Expert system- also known as knowledge-base system, it is an advanced computer program (a set of facts and heuristics) that can, at an acceptable level of competence, solve difficult problems requiring the use of expertise and experience. It deals with the processing of knowledge as opposed to data (Badiru, 1992). It is also the most applicable branch of artificial intelligence, machines mimicking human thinking (Ichida & Voight, 1996).

Neural networks- in contrast to expert systems, which implement knowledge from experts, neural networks learn by example (Burke, 2001). Just as humans learn associations between inputs and outputs via numerous examples, neural networks use large amounts of data to converge on a statistically accurate representation of relationships inconceivable to human experts. The network is then tested on new data to make sure it has not simply memorized the training set of data (Routh, 2001). This method of empirical modeling is an excellent complement to expert systems (Burke, 2001).

Management Approaches

From less tactical perspectives to more strategic ones, approaches by management can be very effective in reducing nonconforming product. These generally require change

throughout an entire organization. The most popular and recent philosophies are total quality management, ISO and QS standards, concurrent engineering, six sigma, and Kaizen. A brief description of each follows:

Total quality management (TQM)- is an approach to long-term success through customer satisfaction. TQM is based on the participation of all members of an organization in improving processes, products, services and the culture in which they work (Sorensson, 2001; ASQ, 2004).

ISO and QS standards- are a set of standards on quality management and quality assurance, developed to help companies effectively document the quality system elements to be implemented to maintain an efficient quality system. ISO standards published in 1987 by the International Organization for Standardization (ISO) are not specific to any particular industry, product or service. The standards known as QS (currently mostly replaced by ISO Technical Specification 16949) were developed by the Big Three Automakers for the automotive sector (ASQ, 2004).

Concurrent engineering (CE)- is a team approach to reduce costs, improve quality and shrink development time by simplifying a product's system of life cycle tasks during the early concept stages (Hartley, 1992; ASQ, 2004). By focusing efforts in the design stage, a successful product with few defects and engineering changes can be realized with relatively little cost (see Figure 7). CE requires four main elements to be successful: 1) customer's voice, 2) cross-functional teams, 3) automated tools and 4) process management (Walker, 2001).

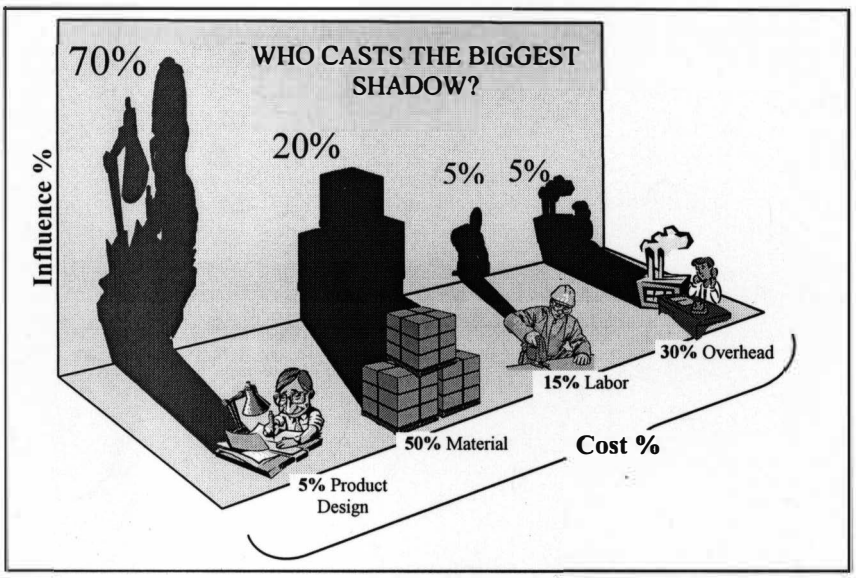


Figure 7. Influence of Product Development Stages on Cost Adapted from Boothroyd et al., 2002

Six Sigma- is a methodology that provides businesses with the tools to improve the capability of their business processes. This increase in performance and decrease in process variation lead to defect reduction and improvement in profits, employee morale and quality of product (ASQ, 2004).

Kaizen- is a Japanese term coined by Masaaki Imai that means gradual unending improvement by doing little things better and setting and achieving increasingly higher standards (ASQ, 2004).

Process Monitoring and Control

Although process monitoring and control is able to prevent defects from occurring, it is machine oriented and post-design. It has the ability to constantly fine tune the machine, maintain preset parameters and provide consistency and repeatability in the

operation (Berins, 1991). Devices can range from basic to extremely sophisticated. In some cases, an alarm in the form of a page or e-mail can be sent signaling a shift outside the acceptable molding window.

Summary

The aforementioned techniques are arguably entities of one another or simply repackaged concepts of early, well-known quality experts, such as Deming, Crosby, Juran, and Shewhart. Nonetheless, they all have an impact on the injection molding industry, each contributing to defect prevention.

Description of the Problem

"Fire-fighting" or a reaction-oriented approach is defensive in dealing with defects. It emphasizes action and how fast the problem reaction wheel can be circled (see Figure 8). It may be effective for short-term fixes, but unfortunately it is incomplete,

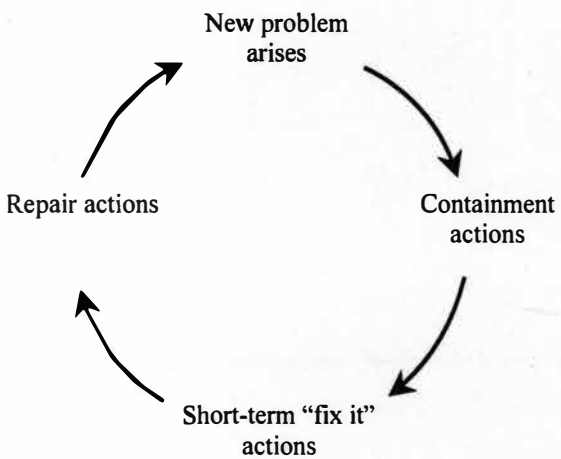


Figure 8. Problem Reaction Wheel (Kane, 1989)

neglecting problem analysis and prevention (Kane, 1989). Defect prevention is an offensive approach that strives for goals such as "zero defects" and supports philosophies such as "doing it right the first time" (Crosby, 1967, p.32).

The goal of this work is to develop a technique for classifying and predicting attribute defects in conventional injection molding. The chosen methodology bridges several approaches used in defect prevention. Categorized as an exploratory observational study, this work can be divided into two parts. First, a causal approach is used to cluster and reduce the types of attribute defects. This is necessary for the second part, the use of classification tree analysis, which presents the design and construction variables that lead to certain defects in a tree like structure. The reason for selecting these methods will be discussed further in the methodology. The combination and application of both is what distinguishes work in this thesis. In setting the stage for this work, it is necessary to describe two methods of defect prevention.

Root Cause Analysis

Most people will maintain that problems are inevitable. And if always "doing the right thing right the first time" is impossible, then "doing the right thing right the second time" is the next best thing (Wilson, Dell, & Anderson, 1993, p.8). Root cause analysis by definition is a reactive method, uncovering the reasons to a problem that has already occurred. However, when effective root cause analysis is continuously performed in a proactive or forward-looking sense, future problems can be prevented. This can only be done by properly distinguishing the root cause from symptoms and apparent causes (see

Figure 9). "Like weeds, problems may reappear if not properly removed or treated. Perhaps more ominously, they also can spread to other areas" (Wilson et al., 1993, p.7). The root cause is the most basic reason for an undesirable condition, which if eliminated would have prevented it from occurring. Symptoms are the tangible evidence or manifestation(s) indicating the existence of something wrong. The "smoking gun" often

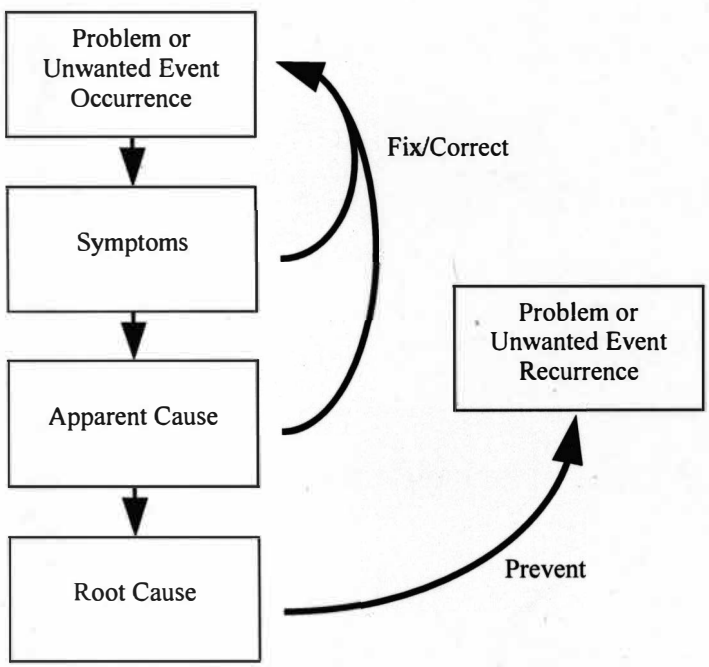


Figure 9. Identification of Root Causes Can Prevent Defects (Wilson et al., 1993)

referred to as the apparent cause, represents the immediate or obvious reason for a problem (Wilson et al., 1993). Of course, the apparent cause may be the root cause, but only when confirmed by analysis. It should be noted that causes are also categorized (in a statistical sense) by their variation in a process. Special or assignable causes of variation

arise because of special circumstances not due to chance. Common causes of variation are inherent in a process over time and affect every outcome of the process (ASQ, 2004).

Root cause analysis is a high level of problem solving, encompassing knowledge, ability and experience while utilizing tools and techniques. It is a backwards approach that makes use of cause-effect diagrams, logic trees and Pareto analysis.

Cause-effect diagram- is a tool for analyzing process dispersion. It is also referred to as the "Ishikawa diagram" because Kaoru Ishikawa developed it, and the "fishbone diagram" because the complete diagram resembles a fish skeleton. The diagram illustrates the main causes and sub-causes leading to an effect (symptom). The cause-effect diagram is one of the "seven tools of quality" (ASQ, 2004).

Logic trees- are a means of organizing data into an understandable and logical format. In the purest sense it is a graphical representation of how experts think. In a tree-like format, it displays the process of deductive reasoning of an event or undesirable outcome from a macroview to a microview. A logic tree differs from logic diagrams and fault trees in that it is factual versus hypothetical. The later methods often deal with the probability of some hypothetical event occurring from a list of potential causes, such as in FMEA (Latino & Latino, 1999).

Pareto analysis- involves ranking problems by some criterion and then focusing on the more significant. Frequently the results of the analysis follow the 80-20 rule, where 80 percent of the troubles are caused by 20 percent of the problems. This is also known as the significant few and trivial many. Pareto analysis helps identify the causes of problems and aids the overall classification process. "In addition, the codes used in

root cause analysis may be helpful in problem characterization" (Wilson et al., 1993, p.33).

Once a problem has been fully defined, the next step is to determine the best course of action to remove it. Classifying the type of problem does this by matching the problem to the most logical causal approach (Mobley, 1999). Of course, these clusters of problems and approaches should be the result of trained professionals. In the field of injection molding, most problems have been well documented.

Classification Tree Analysis

The intentions and applications of classification analysis tend to place it in the classical/correlation research or exploratory observational study. Breiman, Friedman, Olshen, and Stone (1984) frame two purposes of classification analysis: 1) to produce an accurate classifier or 2) to uncover the predictive structure of the problem. One may supercede the other, but most often the two are inseparable. Understanding which variables and interactions are needed to build a robust algorithm that characterizes inputs of an unknown is an important criterion for good classification practice.

Concerns with Data Sets

With large data sets of many variables comes more structure. The number of techniques available to mine the relationships also rises. However, size does not necessarily imply information richness. The components of complexity are what make a data set interesting. These include: high dimensionality, a mixture of data types, nonstandard data structure and nonhomogeneity.

High Dimensionality

A data set with a fixed number of observations (data points) can succumb to "the curse of dimensionality" (Bellman, 1961). Consider 100 points distributed with a uniform random distribution on the interval [0, 1]. If the interval is divided into 10 equal cells, then there is a high probability that all cells will contain some points. However, if the same points are distributed in 2-D over a unit square and a cell size of 0.1 is maintained for each dimension, it is likely that many of the 100 cells will be empty (Steinbach, Ertoz, & Kumar, 2003). For three dimensions, the 100 points will be "worlds apart".

Although enormous efforts have been undertaken to develop methods for reducing high dimensionality in multivariate analyses, shortcomings are present. For example, multiple regression includes a stepwise procedure for selecting variables, yet it has limitations. Nevertheless, in order to analyze and understand complex data sets, approaches are needed to separate useful information from noise (Breiman et al., 1984).

Types of Data

Most literature suggests two general types of variables: categorical and numerical. Categorical variables, also known as qualitative variables, contain a finite set of values. These values must be mutually exclusive and exhaustive. Categorical variables include nominal and ordered variables. A nominal variable represents a set of categories that have no natural order, while an ordered variable specifies a natural order or ranking. For example, male and female are categories of a nominal variable and good, average and poor are levels of an ordered categorical variable.

Numerical variables, also called quantitative variables, take on measurable values from real numbers. Note that assigning classes to real numbers is not a numerical variable. For example, expressing a level of pain on a scale of 1 to 10 is an ordered categorical variable. Furthermore, the difference between the numbers in ordered data may not be equal; there is no exact relationship of "how much difference" (StatSoft, Inc., 2004). Numerical variables can either be continuous or discrete. Continuous variables can assume an infinite number of real values. The values can be ordered, counted and measured (Statistics Canada, 2003). These may later be grouped into intervals. For example, height could be distinguished as less than 4 feet, 4 to 6 feet, and over 6 feet. Discrete variables utilize a finite set of real numbers. The values are separate and countable (Statistics Canada, 2003). A four-digit password and the number of defects in a lot of 100 parts are examples of discrete variables.

Data Structure

A nonstandard data structure adds complexity to a data set because there is no fixed set of variables for each case. A nonstandard data structure occurs when measurement depends on the observation. For example, a survey of banks may inquire about ATM service. Only banks offering such a service are able to respond to those particular questions.

Homogeneity

Breiman et al. (1984) warns that nonhomogeneity may add the greatest complexity to a data set. That is, different relationships held between variables in different parts of the measurement space.

Classification Trees Defined

To maintain uniformity, the notation of the founding authors of classification and regression trees will be followed. Define a classifier or classification rule, $d(\mathbf{x})$ such that for every measurement or variable, $x_1, x_2, x_3, \ldots, x_N$, of the measurement vector $\mathbf{x}$ corresponding to any case M, the classifier $d(\mathbf{x})$ is assigned to only one of the J classes in C, where C is the set of classes ($C = \{1, 2, \ldots, J\}$) and X is the measurement space containing all measurement vectors. In short, a classifier can be expressed as $d(\mathbf{x}) = j, j \in C$.

In terms of classification, data can serve as one of two functions to the analyst: the learning set or the test set. The learning set is used to construct the classifier. The learning set, L, can be defined as a set of data $(x_1, j_1), (x_2, j_2), (x_3, j_3), \ldots, (x_N, j_N)$ on M cases, where $x_i \in X$ and $j_i \in C$, $C = \{1, 2, \ldots, J\}$, and $i = 1, 2, \ldots, N$.

The test set is used to estimate the accuracy of the classifier. Given a classifier, $d(\mathbf{x})$ defined on X, taking values in C, the proportion of new samples incorrectly classified by $d(\mathbf{x})$ is called the misclassification rate, $R^*(d(\mathbf{x}))$. If $P(\mathbf{x}, j)$ is the probability that a case drawn at random from the same distribution as L, has its measurement vector $\mathbf{x}$ in X and its class in $j, j \in C$, then $R^*(d(\mathbf{x})) = P(d(\mathbf{x}) \neq j \mid L)$.

Construction of Tree Classifiers

Classification trees strive to find binary splits, $s \in X$, in the learning sample, L, that create subsets, $t_i \in X$, which contain "purer" data than in the preceding, larger subset, also called a node. Breiman et al. (1984, p. 28) outlined the following four elements in the tree construction:

"1. A set Q of binary questions…

2. A goodness of split criterion Φ(s, t) then can be evaluated for any split s of any node t

3. A stop-splitting rule

4. A rule for assigning every terminal node to a class."

Assume a set Q of binary questions yields a set S of splits s ($s \in S$) for every node t. The affirmative cases in parent node t descend to the left child node t_L, while the negative respondents move to the right child node t_R (see Figure 10). Note that the root node contains all measurement vectors of *X*, that is $t_1 = X$, intermediate nodes contain a subset of *X*, and terminal nodes are also subsets of *X*, but are assigned a class label. The split label, s* is given to intermediate nodes.

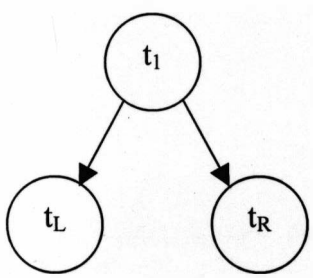


Figure 10. Sample of Splits to Left and Right Child Nodes (Breiman et al., 1984)

There are a limited number of splits for the various types of variables. Nominal variables have $2^{k-1} - 1$ admissible splits, while ordered variables have k – 1 splits, where k represents the number of categories. Numerical variables have r – 1 splits for r distinct values (Aluja-Banet & Nafria, 1998).

Define an impurity function Φ for the proportions of all classes p_j, such that $p_j \geq 0$; $j = 1, 2, \ldots J$; $\Sigma_j p_j = 1$ and with the properties:

"1. Φ is a maximum only at the point (1/J, 1/J, …, 1/J),

2. Φ achieves its minimum only at the points (1, 0, 0, …, 0), (0, 1, 0, …, 0), (0, 0, 0, …, 1).

3. Φ is a symmetric function of $p_1, p_2, \ldots, p_j$" (Breiman et al., 1984, p. 32).

Based on the impurity function Φ, a measure of impurity can be obtained at any node t as $i(t) = \Phi[p(1|t), p(2|t), \ldots, p(J|t)]$. And when a split s of node t divides the cases into the proportions p_L and p_R for nodes t_L and t_R, respectively, the decrease in impurity can be defined as $\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$. The goodness of split $\Phi(s, t)$ becomes $\Delta i(s, t)$.

Two of the most common splitting rules are the Gini index and the Twoing rule. The Gini index attempts to minimize node impurity using Equation 1 (Breiman et al., 1984; Aluja-Banet & Nafria, 1998):

$$i(t) = \sum_{j} \sum_{i \neq j} p(j|t)p(i|t) \text{ or } i(t) = 1 - \max \sum_{j \in J} p(j|t) \qquad (1)$$

Simply put, Gini recursively focuses on the largest or most important (in terms of cost or weight) class and looks to separate it from the other classes (Salford Systems, n.d.). In contrast to Gini, Twoing strives to combine and then separate groups of classes that account for 50% of the data (Salford Systems, n.d.). Equation 2 defines the split as

$$i(t) = \left[(p_L p_R) / 4 \right] \left(\sum_{j} |p(j|t_L) - p(j|t_R)| \right) \text{ (Breiman et al., 1984).} \qquad (2)$$

Splitting can terminate in a number of ways. A simple threshold value, β can be set so that $\max_{s \in S} \Delta i(s, t) < \beta$, where $\beta > 0$. A Fact-style stopping rule (Loh & Vanichestakul, 1988) splits the tree until all nodes are pure or have the minimum number of cases for a class that has been specified beforehand. Another method grows the tree until terminal nodes contain single cases or only one class, then prunes upward. Pruning cuts off terminal nodes and replaces them with the parent node until a specified minimum number of cases or standard error rule (cross-validation method) is met.

Each terminal node is assigned the class j(t) for which p(j|t) is greatest. If a tie exists, one of the largest classes is randomly assigned. Class assignments may be adjusted based on priors or misclassification costs.

Benefits and Drawbacks

There are many advantages to the classification tree approach. First, this approach is a nonparametric technique that requires minimal specification (Breiman et al., 1984; Salford Systems, n.d.). Classification tree analysis eliminates the need for advance selection of variables by using a stepwise procedure. It chooses the best variables in the sample space in which it is working. However, "performance can be much enhanced by a judicious selection and creation of predictor variables" (Salford Systems, n.d.). Another merit is its ability to handle categorical and numerical variables. Also, transformations of variables have no effect. For example, if the split on a nominal variable is $x_1 < 4$, then the respective splits for logarithm, square root, and cube would be $x_1 < 0.6$, $x_1 < 2$, and $x_1 < 64$. Another advantage is that classification trees treat cases as one among the total observations, rather than using the case's intrinsic value. Therefore, outliers have no

effect on splits. It also has the ability to handle missing values through the use of surrogate splits. If a case x_m is missing a value, it searches for a split on the most similar variable in **x**, the surrogate (Breiman et al., 1984). Finally, the simplicity of the classification tree structure makes it easy to understand and interpret by the statistical novice. Hand (1997, p. 64) reiterates this in the context of machine learning in that such a collection of simple if-then rules provides a model "of the way human brains hold procedural knowledge". Its outputs give an estimate of misclassification and an index of variable importance.

In terms of drawbacks, classification trees have their share. Nonparametrics aside, some knowledge of how to use classification trees is imperative. It is unlikely a nonstatistically oriented person will develop a model without some critical error. Should the type of variable, splitting rule or method of pruning be incorrectly specified for the data structure, erroneous and/or nonoptimal results are inevitable. For instance, if variable combinations are used, the standard tree program will perform poorly at separating the linear relationship in comparison to linear discriminate analysis (Loh & Vanichesetakul, 1988; Loh & Shih, 1997) (see Figure 11). Many perpendicular splits are needed to partition the structure, resulting in large trees. If a linear structure is suspected, it is important that splits be extended to use a linear combination of the form $\sum_{m} a_m x_m \leq c$, where **a** is a set of coefficients and c represents all possible values. Boolean combinations provide similar challenges.

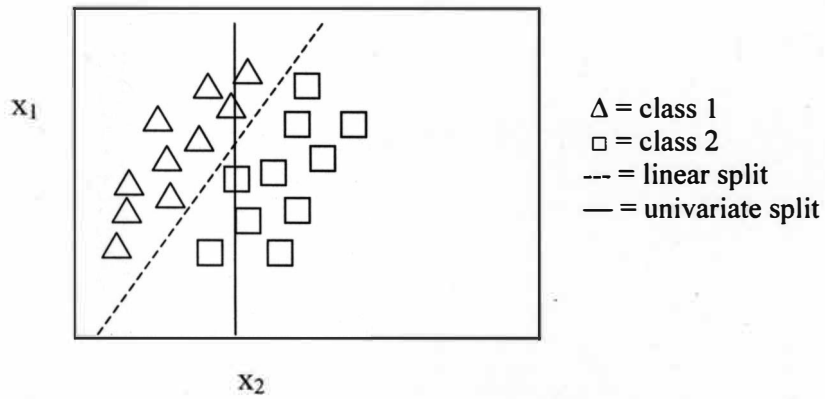


Figure 11. Difference in Effectiveness Between a Linear and Univariate Split (Breiman et al., 1984)

Some have charged the recursive nature of classification trees to be a near-sighted deficiency. The search for the best split among all variables at each node does not consider the impact on the future growth of the tree. The concept is best represented by Breiman et al.'s (1984, p. 97) analogy to bridge: "...the team goal in every deal is to take as many tricks as possible. But a team that attempts to take the trick every time a card is led will almost invariably lose against experienced players."

Another difficulty is the instability of the tree structure. The difference in goodness of split values can be negligible. Considering the inherent noise, the choice between competing splits is almost random (Breiman et al., 1984). Small fluctuations in data can change which variables are split and which are not split. The impact can affect the growth from that node downward. Furthermore, the condition can be deceptive, leading to misinterpretation. The "unused" variables could be considered unrelated to the dependent variable; in actuality, their predictive power is quite high. To realize this masking effect a user can review cross-validation trees, compare surrogate splits with

optimal splits, and examine the variable rankings in terms of their potential effect on the classification.

Other deficiencies come from the tendencies for certain splitting rules to perform better with the size of data sets, types of variables chosen, and priors.

Summary

Liang, Ou and Tang (2003) combined decision trees and survival analysis to diagnose causes behind failure rates from defective hardware warranty claims for automobiles. Although results were confidential, they claimed the ability to identify where, when and how failures occurred. The limited publication suggested the independent variables were attributes of cars, the dependent variable was failure rate of product defects and the observations were warranty claims of hardware defects. It was not clear if multiple defects could occur from the same automobile or set of attributes.

In the search of related literature, the aforementioned project was the only case similar to this thesis. Although root cause analysis and classification trees have been widely used to solve a variety of issues, there is limited application for assessing defects.

CHAPTER III

EXPERIMENTAL PROCEDURE

Background

This study is an outgrowth of questions raised during proprietary research at Western Michigan University. The current project has been a long-term investigation into forecasting the performance of tools prior to production. Recently, the attention has been directed toward modeling the types of problems a tool may encounter based upon design and construction variables. Figure 12 outlines a generalized methodology for developing a classification tree and analyzing manufacturing defects. Although this methodology was developed specifically for this project, it is intended to be transferable among most manufacturing industries, especially those dealing with attribute defects. Injection molding was used to demonstrate the proposed methodology. Specific elements of the flow chart will be discussed in the remainder of this chapter.

Data Collection

As is the case with most exploratory observational studies, the intuition of numerous experts and the collection of data for many variables were utilized. Six Michigan companies were involved in this project. Expertise within these companies and relevant literature established which independent variables could potentially predict tool performance. Microsoft Access was the media used to enter and store the data. Companies were responsible for entering their own data. A series of extensive checks

were programmed to minimize errors and missing entries. For example an error message would appear for an extremely large value or non-numeric entry, or the user may be prompted to complete a form upon closing it. Unfortunately, some information was not

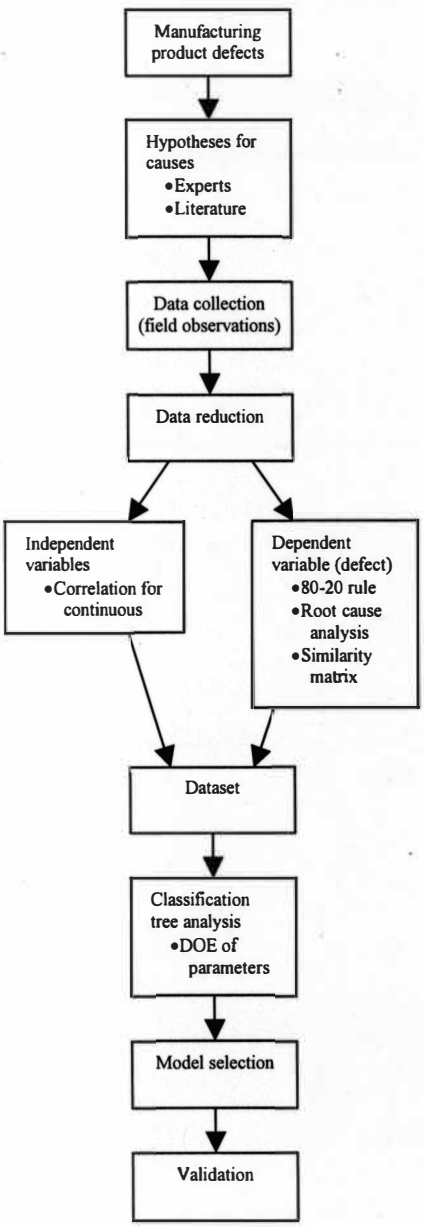


Figure 12. Generalized Methodology for Developing a Classification Tree and Analyzing Manufacturing Defects

available and entries were intentionally left blank. As an extra precaution various data cleansing techniques were applied to insure all data were appropriate. For example, variable minimums, maximums, ranges, standard deviations and sample size (valid N's) were compared against expected values. Graphical techniques were also used to aid in detecting groups of outliers. Finally, frequency tables were run to examine the distribution (variance) within each variable.

Analysis Selection

As a classical/correlation study might suggest, the selection of the type of analysis to be used was dictated by several issues in the data set. For the following reasons, classification tree analysis appeared most suitable:

- Dependent variable was categorical (its nominal nature eliminated most linear techniques)
- Large number of independent variables (no need for transformations)
- Independent variables include both categorical and numerical data
- High dimensionality (many variables, m have numerous levels, k yielding dimensionality, k^M)
- Missing data (ability to be handled through surrogate functions)
- Nonstandard data structure
- Nonhomogeneity (complex relationships can be easily interpreted in a simple tree-like layout)
- Nonparametric nature of the analysis

The outputs of classification trees also served to meet two demands of the companies' goal, which were to produce an accurate classifier and reveal the predictive structure of the problem.

Observations

The following excerpt from Rosato and Rosato (1995) summarizes the complex function of an injection mold (or tool).

> Optimizing the injection molding process to reach higher productivity requires careful examination of individual components. Compromises in the performance of any one of these can adversely affect productivity. Specifically, overall performance is related to designing the mold for maximum productivity and specifying the machine to obtain maximum output.
>
> A mold is a highly sophisticated piece of machining. It comprises many parts requiring high-quality steels. It also includes cooling channels and possibly hot runner channels for the hot feed of molten plastics. In many cases, it will also contain a number of moving parts, such as ejector pins and moving cores.
>
> To capitalize on the advantages of injection molding, the mold tool may incorporate many cavities, adding further to its complexity. All these parts must function efficiently and smoothly, at high temperature and very high pressure, in a reciprocating machine that may well cycle several times a minute or even parts of a minute for long production runs (p.204).

The passage justly describes how tools ultimately share the same functions and objectives, yet can be so different in terms of complexity and structure. The 508 tools of

this thesis's data set reflected the same distribution. Some were built for presses over 4000 tons, while others for 100 tons and below. The number of simultaneous parts produced spanned from one to sixteen. Some geometries were simple with few projections, while others had many contours with holes, depressions, and projections. Although most parts served the automotive industry, a significant number of them belonged to furniture, appliance, and container markets. A wide variety of production materials were present as well. However, as much as each company's philosophy, skill, resources and product lines differed, tools were designed and constructed to maximize productivity, meaning a large process window free of defects with an efficient cycle time. Each company's representation in the data set is shown in Figure 12. It should be noted that this study only considered tools processed by conventional methods.

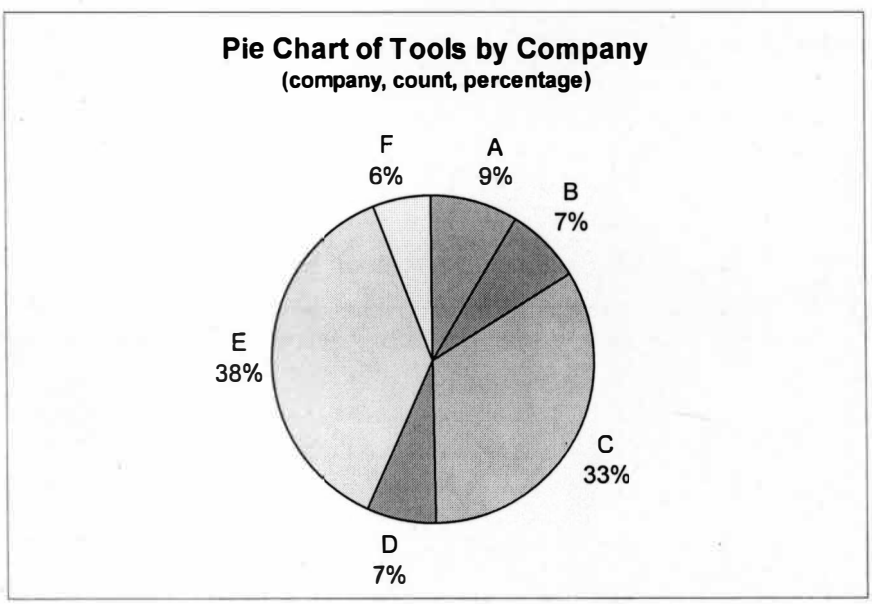


Figure 13. Breakdown of Company Representation in the Data Set

Independent Variables

As mentioned earlier, the initial selection of independent or predictor variables was dictated by the knowledge that represented a wide variety of published sources. The reasons for selection were based on experiences with similar tools. Furthermore, it was agreed that nonhomogeneity, dimensionality and company differences could play havoc in formulating relationships. The predictive importance of variables would be left to interpretation after classification models were developed. Another reason was there was no good method to test relationships between independent and dependent variables prior to modeling when the dependent variable is nominal. In other words, a shotgun was used for initial variable selection rather than a sharp-shooting rifle. This issue of nonparametric analysis will be discussed further as a topic for future work.

The first issue handled was missing data. Many of the independent variables were missing data from more than 40 percent of the tools. Although the use of surrogates was an option, these relatively incomplete variables would in a sense be phony and introduce high levels of masking. As a quick screening, variables with a sample size of less than 300 were removed from initial analyses. There were 151 variables, which was comprised of 109 categorical and 42 numerical variables used in this study.

These remaining variables caused an issue with dimensionality. In other words, the search algorithm used to generate a tree would be computationally challenged. To reduce the magnitude of the data set, a capacity of twenty-five independent variables was set for the classification tree analysis. The variables were initially selected using three criteria.

- The degree of missing data. Variables roughly 80 percent or more complete (N = 400) were used as "starter variables". The remaining predictors were set aside from opening analyses, but could be introduced later. This condition reduced the data set by 64 variables.
- The distribution of observations within categorical variables. Any variable with less than seven observations in any given category or level was put on "statistical probation". Seventeen more variables were placed on reserve.
- Highly correlated variables. Although correlations against nominal variables with more than three levels could not be achieved (unless symmetrical), many of the binary and ordered variables were scrutinized for high correlations. Searching for correlated variables was not left to chance. A judicious selection based on design and construction was used to seek out likely correlations. For example, if a tool's primary cavity is inserted, then there is a strong possibility the primary core is also inserted (see Table 1). Other variables exhibiting a significant correlation coefficient (r) greater than or equal to 0.7 (with $p < .05$) were chosen on two premises: 1) the ability to represent the most variables and 2) expert opinion. These variables could be swapped for those not used at any time (see Table 2). This step allowed 14 variables to be temporarily discarded. Note: variables exhibiting a high correlation ($r > 0.7$) were highlighted and variable names were coded.

Table 1. Correlation Matrix Between Similar and Paired Tooling Variables

Variables on Cover Side of Tool	Variables on Core Side of Tool						
	A	B	C	D	E	F	G
A	0.87	0.14	-0.04	0.23	0.07	0.06	-0.14
B	0.17	0.83	-0.01	0.01	-0.09	-0.02	0.03
C	-0.04	-0.01	1.00	0.02	0.05	-0.01	0.02
D	0.27	-0.02	0.02	0.89	0.21	0.11	-0.20
E	0.14	-0.09	0.05	0.25	0.89	-0.23	0.35
F	0.05	-0.02	-0.01	0.11	-0.23	1.00	0.04
G	-0.14	0.04	0.02	-0.14	0.29	0.04	0.92

In several instances, independent variables were reduced by merging them into one entity. For example, three "yes/no" questions about the quality requirements of a product were visual, structural and dimensional. Since these were binary and served the same purpose of explaining the types of quality requirements, these variables were good candidates for aggregation. This turned simple binary variables into a nominal variable with 2^k combinations or levels, where k is the number of variables (see Figure 13). Unfortunately, the combinations caused one or more of the newly formed categories to have less than seven observations, violating the aforementioned selection criterion.

Table 2. Correlation Matrix of Ordered Predictor Variables

	a																	
1	0.03	b																
2	0.00	-0.07	c															
3	0.11	0.02	-0.03	d														
4	-0.06	-0.01	0.12	0.00	e													
5	-0.11	0.11	0.18	-0.02	0.43	f												
6	0.08	0.01	0.14	0.04	0.37	0.40	g											
7	-0.02	-0.12	0.40	-0.01	0.25	0.34	0.24	h										
8	0.05	0.11	-0.14	-0.13	-0.03	-0.01	0.00	-0.13	i									
9	0.09	-0.04	0.44	0.26	0.06	0.07	0.19	0.65	-0.31	j								
10	-0.04	0.03	0.40	0.25	0.12	0.17	0.16	0.76	-0.23	0.82	k							
11	0.06	-0.04	0.08	0.42	0.05	0.12	0.12	0.25	0.08	0.36	0.21	l						
12	-0.01	0.08	-0.10	-0.04	-0.09	-0.03	-0.03	-0.28	0.19	-0.19	-0.14	0.01	m					
13	0.13	0.04	0.38	0.35	0.10	0.13	0.22	0.70	-0.22	0.92	0.72	0.43	-0.19	n				
14	0.04	0.14	0.22	-0.04	-0.03	-0.05	0.04	0.24	0.04	0.41	0.18	0.15	0.02	0.41	o			
15	0.36	0.34	0.26	0.22	0.00	0.07	0.05	0.21	-0.02	0.32	0.21	0.12	-0.15	0.34	0.15	p		
16	0.75	0.61	0.22	0.13	-0.04	0.03	0.12	0.02	0.07	0.19	0.12	0.08	0.13	0.24	0.18	0.53	q	
17	0.17	0.02	0.38	0.31	0.07	0.11	0.22	0.65	-0.19	[redacted]	0.75	0.44	-0.17	[redacted]	0.42	0.31	0.26	r
18	0.01	0.05	0.00	-0.02	0.00	0.07	0.16	0.21	0.09	0.07	0.02	0.14	-0.06	0.23	0.15	0.14	0.02	0.38

However, the merged and original variables could always be analyzed separately. This step was also taken for the part design rules and styles of venting.

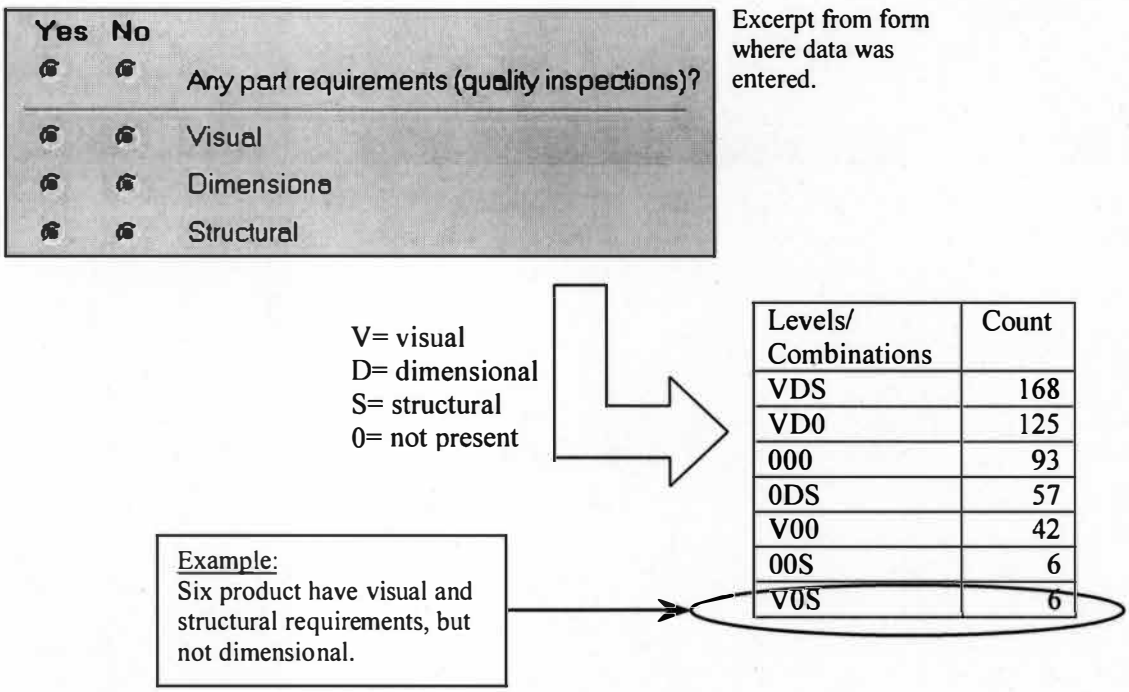


Figure 14. Conversion of Similar Independent Binary Variables into One Multilevel Variable

The versatility of classification trees implied there was no need to alter the data, such as transforming, so they remained "as collected". Also, an independent to dependent variable ratio of 10:1 was considered ideal to avoid overparameterizing the model(s). Of course, preliminary classification tree analysis indicated that roughly 30 to 40 independent variables were being used to classify attribute defects.

The Dependent Variable (Defect)

Companies were permitted to select as many as four different tool related problems from a possible twenty-three (and "other") for each tool. A tool could have no problems. Along with the frequency of the problem, a brief description, cause, and solution were entered. Each problem was carefully reviewed to insure only mold related problems were recorded. For example, certain defects caused from running the tool in an insufficient press were not true tool related problems; thus, they were discarded.

The focus of this study was on attribute defects. Not surprisingly, this was in line with the frequencies of the problems observed in the data set. Ten of the eleven attribute defects were also the most observed. In order to distinguish the "other" non-attribute, tool deficiencies from tools with no problems, these deficiencies were tagged as "OtherA". Figure 14 shows the Pareto analysis of the various defects. Those bolded and italicized were the attribute defects. Weld line was the only problem occurring as infrequently as the Other A group.

Since a tool could have no problems or up to four different problems simultaneously, many unique combinations (or classes in terms of classification trees) for the nominal dependent variable existed. In fact, a total of 1186 combinations were possible. This was obtained from the following.

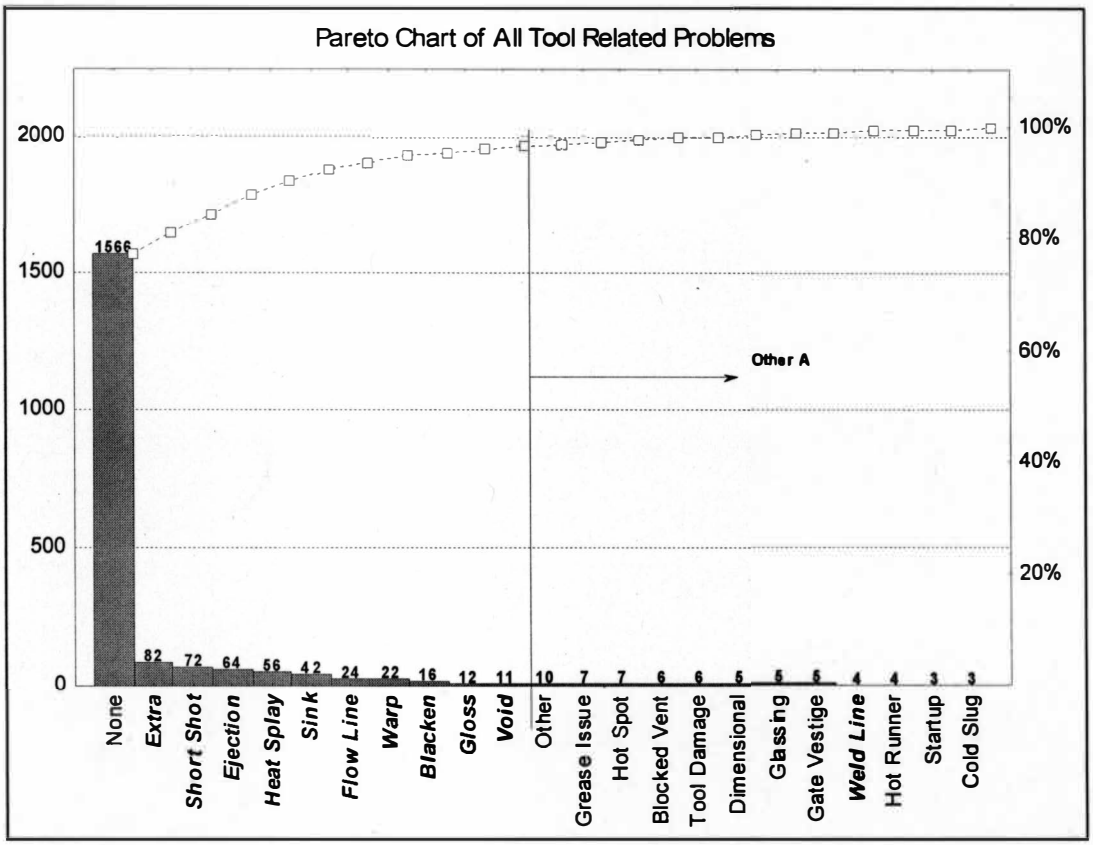


Figure 15. Pareto Chart of Tool Related Problems

Define **x** as a vector of defects, where $\mathbf{x} = \{x_1, x_2, x_3 \ldots x_n\}$ and n represents the entire set of defects. Let Ø and A symbolize no problem and a problem defined as "Other A", respectively. Then a 4-tuple, $P = (p_1, p_2, p_3, p_4)$ can be defined as the set of possible outcomes for a given tool m, where p_i (i = 1, 2, 3, 4) takes on values in $(x_1 \ldots x_n, Ø, A)$. The following conditions were:

- If a tool had no problem, then no other outcomes were possible. That is for any tool m, $P = (Ø, Ø, Ø, Ø \mid p_1 = Ø)$.

- A tool could have multiple occurrences of OtherA. That is $p_1 = p_2 = p_3 = p_4 = A$ was possible for any m.
- A tool could not possess multiple occurrences of the same attribute defect. That is $p_1 \neq p_2 \neq p_3 \neq p_4$ for any measurement of **x** for any tool m.

The probability density function for the 4-tuple of defects was expressed as a series of conditional probabilities as in Equation 3.

$$f(p_1, p_2, p_3, p_4) = f(p_4 \mid p_1, p_2, p_3)\, f(p_3 \mid p_1, p_2)\, f(p_2 \mid p_1)\, f(p_1) \qquad (3)$$

The total number of possibilities was obtained as Equation 4.

$$= \sum_{i=0}^{4} \binom{n}{i} + \sum_{i=0}^{3} \binom{n}{i} + \sum_{i=0}^{2} \binom{n}{i} + 1 \qquad (4)$$

where n is the total number of defects (attributes and others). In this situation, n =12 (11 attribute defect plus OtherA). For this application, the equation can be reduced to form Equation 5.

$$= 5\binom{n}{0} + 4\binom{n}{1} + 3\binom{n}{2} + 2\binom{n}{3} + \binom{n}{4} \quad = \sum_{i=0}^{5} (5-i)\binom{n}{i} \qquad (5)$$

For all practical purposes, the arrangement of defects within the 4-tuple did not matter. Any arrangement of the same four values was considered the same. However, software required the combinations to be in order; otherwise, each arrangement of problems was treated as unique.

Even though it was unlikely that all possible arrangements were actually observed, it was unthinkable to attempt modeling even a portion of such a large number of classes. Furthermore, the chosen classification tree software limited the number of classes to twenty-two. Intuitively, if n could be reduced, the resulting combinations would decrease. A clustering technique was employed to reduce the number of combinations and will be discussed in the next section.

Causal Analysis

Although only 107 problem combinations (of 508 observations) of the possible 1186 were present, high dimensionality and the sizeable amount of single occurrences made classification tree analysis virtually impossible. Logic suggested that if the initial number of attribute defects (n) could be reduced, the number of combinations or classes would drastically decrease per Equation 3. Product defects being a well-documented topic provided a foundation for clustering the attribute defects via a causal approach.

The first step was to develop a uniform definition of the symptoms for the defects. Several troubleshooting guides (Rosato & Rosato, 1995; Advanced Process Engineering, 1996; Nova Chemicals, 2000) were used to compile the symptom-like definitions listed in Table 3. The next and most difficult step was to define the most basic cause for the symptoms. The cause needed to be the backbone of the fishbone diagram or the root that all other branching causes would feed into. The most logical approach for this task was to characterize the problem in terms of the plastic (see Table 3). This standardized the definitions, so that clusters could easily be identified. The third and final step was to group similar attribute defects. In other words, which root causes were synonymous?

Again, this was accomplished by referencing troubleshooting guides (Rosato & Rosato, 1995; Advanced Process Engineering, 1996; Nova Chemicals, 2000).

Table 3. Definitions of Symptoms and Root Causes Compiled from Troubleshooting Guides

Defect	Symptom	Root Cause
Extra	thin layer of excess material	penetration of material into mating surfaces
Short Shot	an incomplete part	insufficient material in terminal path of flow
Ejection/ Pulling	deformation of the part (whitening or drag mark)	improper shrinkage or ejection design
Gloss	local area of excessive or deficient gloss	thermal increase during material flow
Heat Splay (Blush)	hazy or discolored surface	thermal increase during material flow
Sink	local depression not following tool's surface	insufficient material flow
Flow Line	circular ripples or wavelets on part's surface	thermal decrease in material flow
Warp	part's shape does not conform to tool	nonuniform shrinkage
Blacken	dark or black spot of charred material	trapped air in terminal path of material flow
Void	vacuum inside the part	localized nonuniform shrinkage
Weld Line	line or streak on part's surface	thermal decrease in joining flow fronts

The most visible cluster was short shot and sink. Although a short shot typically occurs in the terminal path of material flow during the injection phase, while sink appears in relatively thick sections of a part during the pack/hold and cooling phases, both stem from a lack of material flow. Possibly the most arguable decision was to group void with sink and short shot. Although this defect deviates more from short shot, literature often considered it an extension of a sink mark. Where sink can form, so can a void. And ultimately, a void is caused by a lack of material flow compensating the shrinkage in the

material. These three attribute defects formed a new group called *insufficient material* (abbreviated InsuffMatl).

Another evident grouping was flow line and weld line. Although different in how and where they occur, their effects are rooted in a thermal decrease in the material during flow. There is also a lot of overlap in the methods used to correct them. Flow line and weld line were fittingly united as a cluster called *lines*.

The third and final cluster of defects consisted of heat splay and gloss. Heat splay is a severe version of gloss. The thermal increase in the material is so great that the material fractures or degrades leaving hazy or discolored streaks on the surface of the part. Gloss on the other hand results from a thermal difference (usually an increase) in the material as it contacts the tool's surface. These defects were paired and named *thermal*. Note: heat splay and blush were grouped prior to this analysis. Blush is the term for heat splay that occurs near the gate(s) or entrance into the cavity.

As an added measure to ensure appropriate groupings were made, a more mathematical clustering technique, also based on causal analysis, was performed (see Table 4). Apparent causes of tool design and construction attributes were compiled and summarized from the troubleshooting guides (Rosato & Rosato, 1995; Advanced Process Engineering, 1996; Nova Chemicals, 2000). In Table 4, a relationship between a cause and a defect is identified by a "1".

Table 4. Matrix of Apparent Causes on Attribute Defects

	Attribute Defects										
Apparent Causes	Extra	Short Shot	Ejection/ Pulling	Heat Splay (Blush)	Sink	Flow Line	Warp	Blacken	Void	Weld Line	Gloss
Poor parting line or mating surfaces	1	1									
Insufficient clamp	1				1				1		
Misalignment of tool halves	1		1								
Multiple cavities (imbalance)	1	1			1				1		
Insufficient support	1										
Insufficient venting	1	1		1	1	1		1		1	1
Vents too large	1										
Poor design of land area	1	1			1				1		
Wear- parting line/shutoffs	1		1	1							1
Nonuniform tool temp		1		1	1	1	1		1	1	1
Sizing of gates/runners/sprue		1		1	1	1	1	1	1	1	1
Insufficient number of gates		1			1				1	1	
Improper gate location		1		1	1	1	1	1	1	1	1
Texture			1			1				1	1
Insufficient taper/draft			1								
Improper sprue puller design			1	1							
Nonuniform section thickness	1	1	1		1	1	1		1	1	1
Insufficient ejection			1				1				
Variation in contour							1				
Unbalanced multiple gates							1				
Improper projection design								1			
Tool temperature (layout)	1	1	1	1	1	1	1		1	1	1

A simple clustering algorithm was then used to generate a matrix of pairwise comparisons (see Table 5). The measure of similarity between defects was defined as:

$S_{ij} = \frac{\sum i \cap j}{\sum i \cup j}$, where i = the i^{th} defect and j = the j^{th} defect (Gupta & Seifoddini, 1990).

Large similarities ($S_{ij} > 0.6$) are highlighted in Table 5.

Table 5. Similarity Matrix for Attribute Defects Based on Apparent Causes

					Attribute Defects						
Attribute Defects	Extra	Short Shot	Ejection/ Pulling	Heat Splay (Blush)	Sink	Flow Line	Warp	Blacken	Void	Weld Line	Gloss
Extra	-										
Short Shot	0.40	-									
Ejection/ Pulling	0.27	0.13	-								
Heat Splay (Blush)	0.20	0.42	0.25	-							
Sink	0.40	0.82	0.13	0.42	-						
Flow Line	0.20	0.55	0.25	0.56	0.55	-					
Warp	0.12	0.38	0.23	0.36	0.38	0.50	-				
Blacken	0.07	0.27	0.00	0.38	0.27	0.38	0.20	-			
Void	0.33	0.73	0.13	0.33		0.45	0.42	0.18	-		
Weld Line	0.19	0.64	0.23	0.50	0.64		0.45	0.33	0.55	-	
Gloss	0.27	0.50	0.33	0.67	**0.50**	**0.88**	0.45	0.33	0.42	0.78	-

The results supported the groupings from the first causal analysis. The insufficient material cluster shared many of the apparent causes. Sink and void had the highest relationship of 0.9, short shot and sink exhibited a similarity of 0.82, and as expected a bit more distant was short shot and void at 0.73. Flow line and weld line boasted a similarity measure of 0.88, while the thermal group of heat splay and gloss had a modest relationship of 0.67. This exercise highlighted a couple other potential

arrangements. Gloss showed a tendency to be categorized with flow line and weld line at 0.88 and 0.78 respectively. And weld line had a relatively weak similarity to insufficient material group, posting values of 0.64 with both short shot and sink.

By decreasing the number of attribute defects from eleven to seven, the theoretical total possibilities for the dependent variable were reduced from 1186 to 308 (n = 8). The actual number of combinations was shrunk from 107 to 59, of which 27 were single occurrences (see Figure 15). Remember that besides the high dimensionality that would occur, the maximum number of levels or classes for the dependent variable was limited to twenty-one. To remedy the situation, a 70-30 rule (an ideal 80-20 rule exceeded the threshold by 10 levels) was adopted to remove relatively low occurring problem classes. These were labeled "*Other B*", which became the second largest class of problems (see Figure 15).

Classification Tree Analysis

Classification tree analysis was performed with Insightful's S-PLUS 6 statistical software. As discussed earlier, fifty independent variables were selected on simple statistical guidelines. A limitation in the software forced the selection to be halved. These twenty-five variables used in the model construction were established on a trial and error basis and expert opinion. The binary recursive partitioning procedure used to determine splits was S-PLUS's deviance splitting criteria. This rule attempts to minimize impurity through Equation 6:

$$\min\ i(t) = -2\sum_{t}\sum_{j} n_{tj}\, log(\, p_{tj}\,), \tag{6}$$

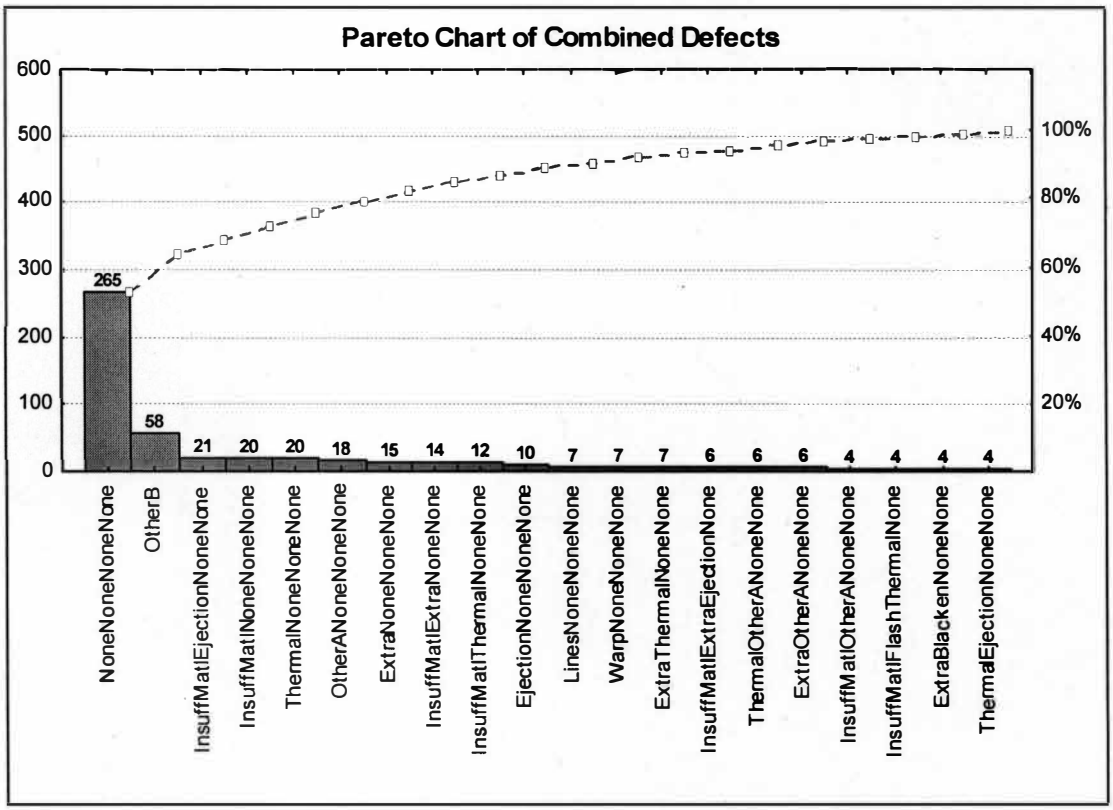


Figure 16. Pareto Chart of Defect Combinations with Combinations Less Than Three as "Other"

where t is the number of nodes, j is the number of classes, n_{tj} is the number of observations of class j in node t, p_{tj} is the proportion of observations of class j in node t. Two fitting options were evaluated in a DOE to establish the most suitable tree size and accuracy. The first was a combination of the minimum number of observations before a split and the minimum node size. Logically, the number of observations in a parent node had to be greater than or equal to twice the setting for the minimum node size. The levels for the split and size were 4 and 2, 10, and 5, 14 and 7, and 20 and 10, respectively. The second factor was the minimum node deviance, which is the measure of node

heterogeneity (a pure node has adeviance of zero). Five levels were used: 0.01, 0.05, 0.1, 0.15, and 0.2. The response variables were the misclassification error rate and the number of terminal nodes. The misclassification error rate measured how accurately the model was explaining the learning set. The rate was obtained by counting the total number of misclassified objects and dividing by the total number of observations. The number of terminal nodes provided a good indication of over-parameterization or a tree that may be more complex than necessary to describe the data. A full-factorial design was used. Tools with missing values in the predictor variables were omitted from the analysis. Pruning, a backward analysis to tree optimization (as opposed to a forward analysis with the fitting options), was not used for this study.

CHAPTER IV

FINDINGS

The runs from the DOE for evaluating tree size and performance provide several ancillary results. Of the twenty-five variables selected for analyses, only seventeen were actually used in tree construction. These also remained constant throughout the DOE trials. This was important because a change may have caused the number of observations, which was 308 to also change due to missing data. The results of the DOE are also listed in Table 6. The factorial plots are displayed in Figures 16 through 19.

Table 6. Design of Experiment for Assessing Tree Parameters on Responses

Model	Minimum Node Deviance	Min Obs for Node Split (Size)	Misclassification Error Rate	No. of Terminal Nodes
1	0.01	4(2)	0.289	59
2	0.05	4(2)	0.442	21
3	0.10	4(2)	0.510	13
4	0.15	4(2)	0.539	9
5	0.20	4(2)	0.558	6
6	0.01	10(5)	0.386	42
7	0.05	10(5)	0.445	21
8	0.10	10(5)	0.516	12
9	0.15	10(5)	0.539	9
10	0.20	10(5)	0.558	6
11	0.01	14(7)	0.419	36
12	0.05	14(7)	0.451	21
13	0.10	14(7)	0.516	12
14	0.15	14(7)	0.539	9
15	0.20	14(7)	0.558	6
16	0.01	20(10)	0.474	23
17	0.05	20(10)	0.474	19
18	0.10	20(10)	0.523	12
19	0.15	20(10)	0.542	9
20	0.20	20(10)	0.558	6

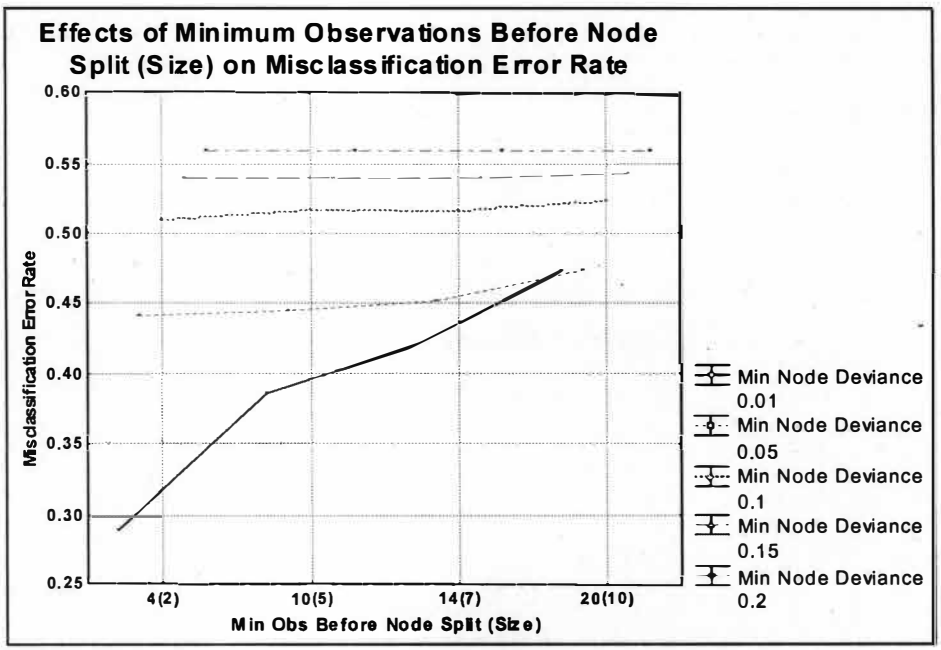


Figure 17. Effects of Minimum Observations Before Node Split (Size) on Misclassification

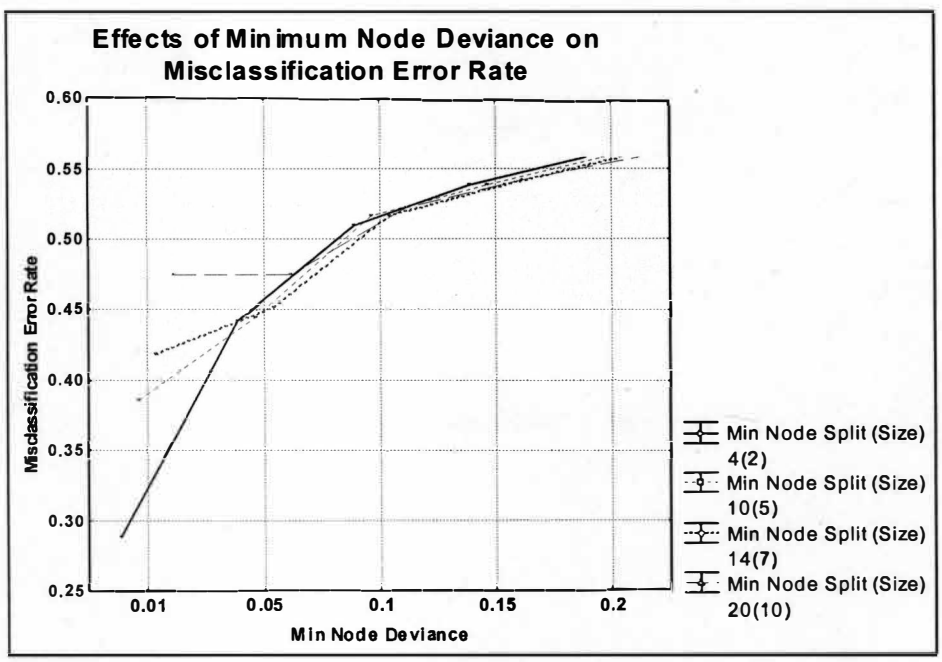


Figure 18. Effects of Minimum Node Deviance on Misclassification Error Rate

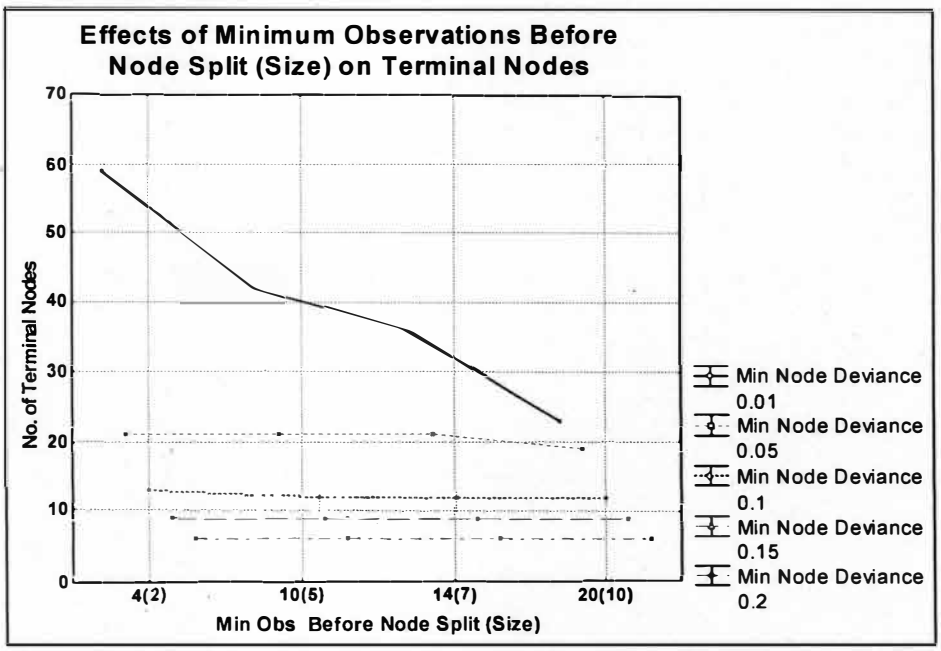


Figure 19. Effects of Minimum Observations Before Node Split (Size) on Nodes

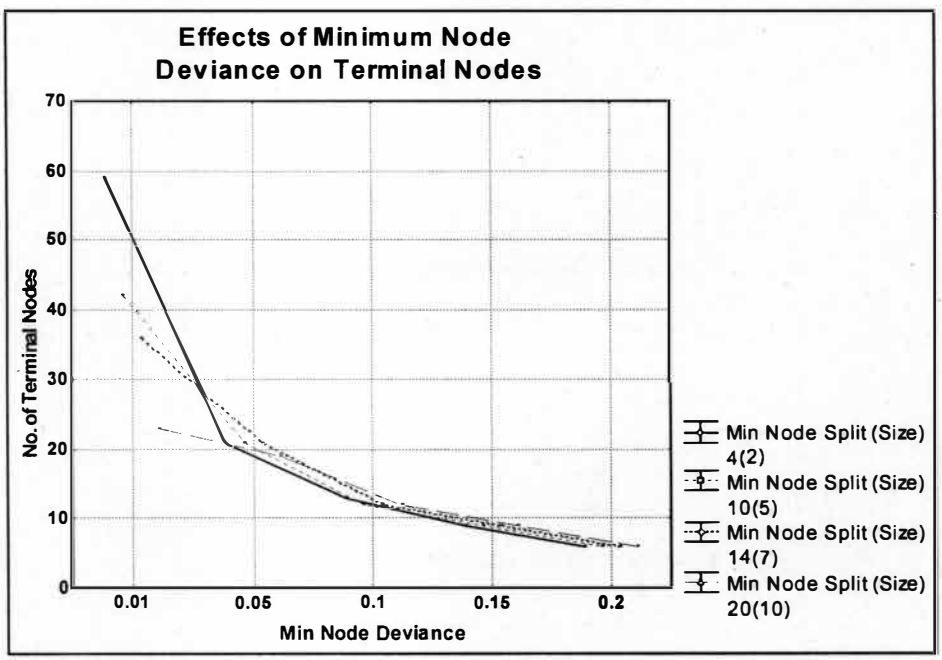


Figure 20. Effects of Minimum Node Deviance on the Number of Terminal Nodes

For sake of discussion the minimum number of observations before splits and minimum node size will simply be referred to as the minimum node size. In Figure 16, it is clear from the flatness of the lines that the minimum node size had little effect on the misclassification error rate, except at the 0.01 level of node deviance. Of course, the separation between the higher levels indicated the deviance was affecting the error rate. This effect is exacerbated in Figure 17, where the levels of minimum node size converge after a 0.01 node deviance. The plots suggest that the interaction of node size and deviance at the 4(2) and 0.01 levels, respectively had the greatest impact on the misclassification error rate.

An inverse of the same relationships was experienced with the number of terminal nodes, which implies the response variables were related (see Figure 20). In fact, the correlation between the two was -0.9792 (p –value 0.0000). The variation of the

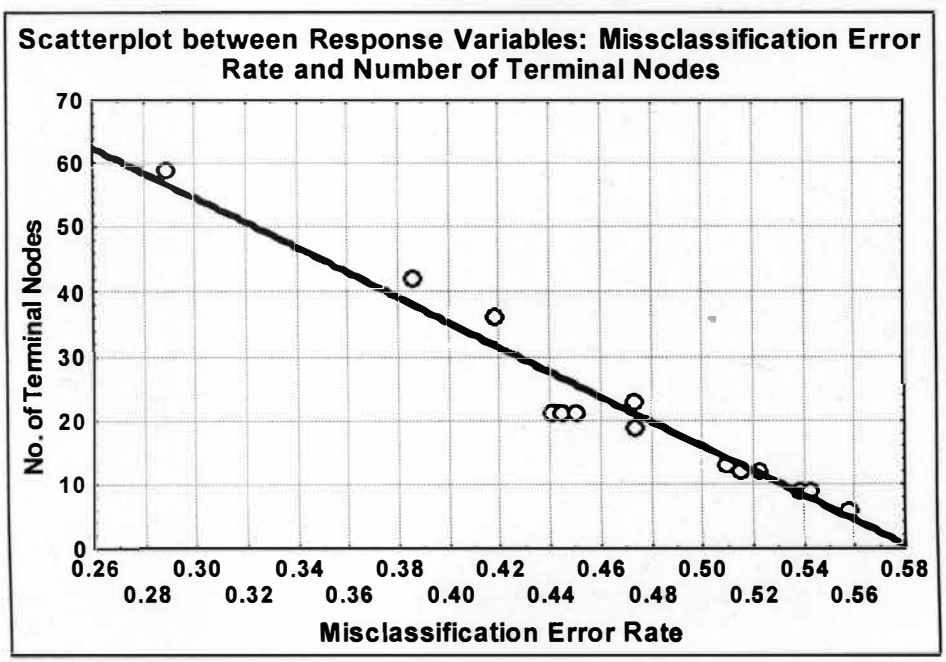


Figure 21. Relationship Between Response Variables

misclassification error rate variables was most noticeable in Figure 21. As the minimum node size and deviance increased the variation decreased in the response variables.

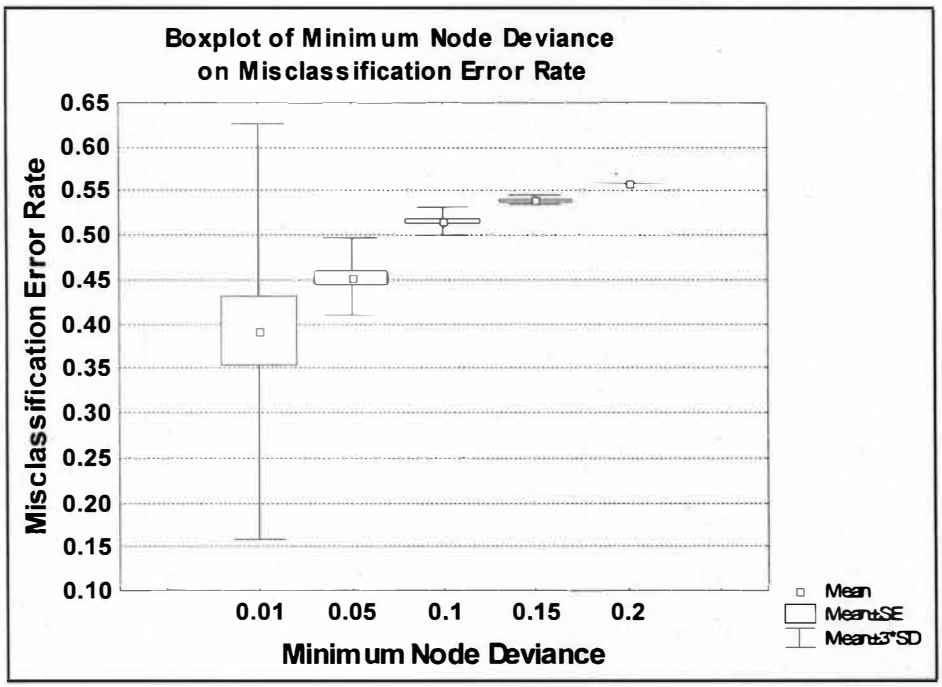


Figure 22. Boxplot Illustrating the Variation in Levels of Node Deviance on Error Rate

From these findings, it was decided that for this application the most appropriate fitting options were 14(7) and 0.05 for the node size and deviance, respectively. Although the misclassification error rate of 0.4513 was only slightly better than a coin toss, the 21 terminal nodes were appealing. It was believed that the tree model was not overparameterized, more suited to correctly classify new tools and possibly more robust. Of course, this decision and these assumptions were subjective and warranted testing through a series of validation methods. Furthermore, a series of variable substitutions could improve model performance. It was also believed that a detailed investigation of

node distributions and placement of misclassified defects may be more important than the overall tree performance. These and other recommendations for future work are discussed in the Chapter 5.

As expected, the node deviances were higher in nodes with more observations (see Figure 22). The correlation between the two was 0.7941and the p-value was 0.00002. This also supports the relationship between the minimum number of observations before splitting (and node size) and the misclassification error rate. The structure of the chosen classification tree is displayed in Figure 23. The split conditions are given at the parent nodes. The number of tools, node deviances and predicted class are listed respectively, at each terminal node. Note: the variables are coded for confidentiality.

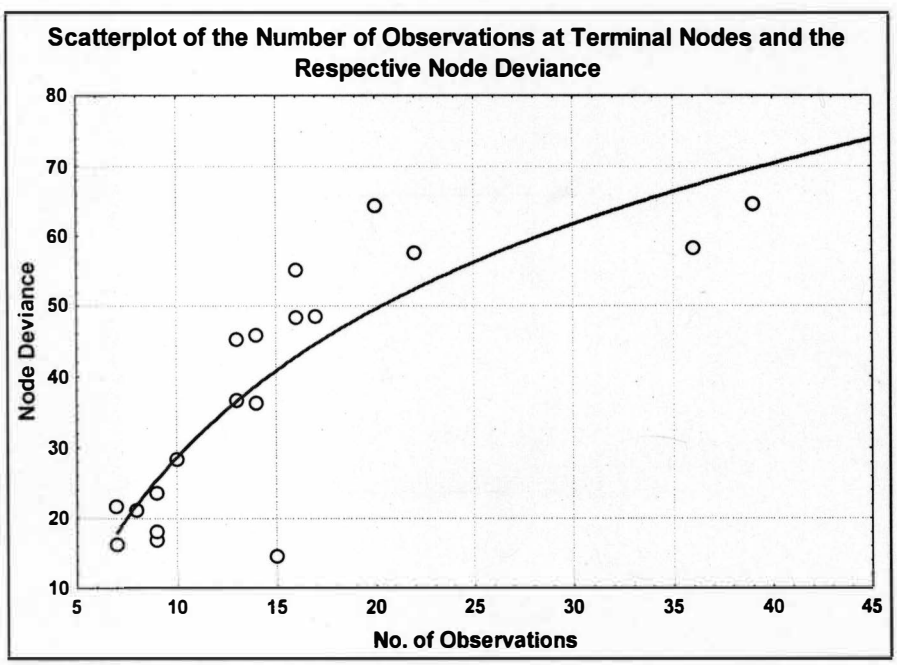
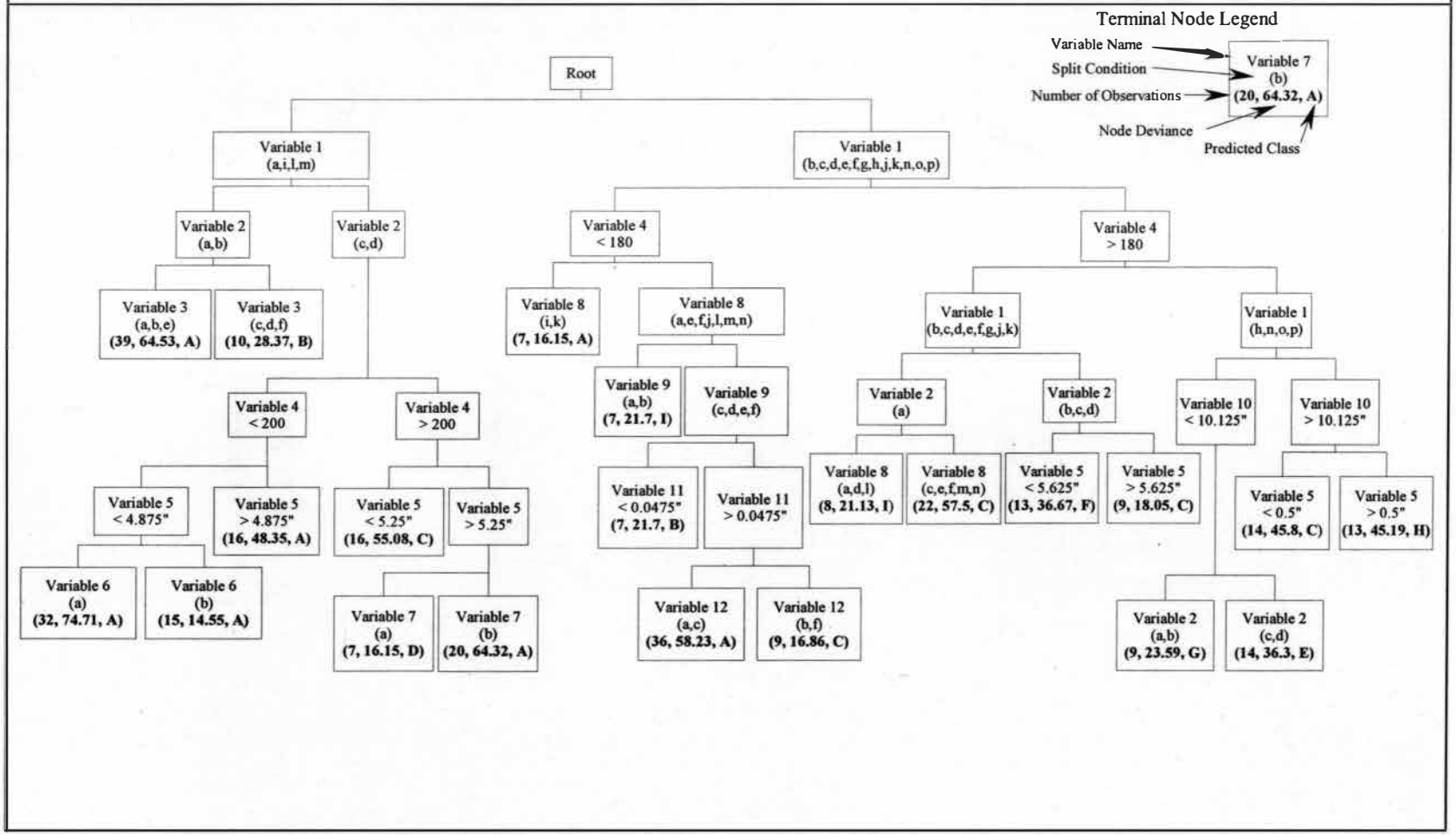


Figure 23. More Observations in Terminal Nodes Resulted in Higher Node Deviances

Figure 24. Classification Tree for Model 12



In regard to the types of defect combinations, Figure 24 illustrates the number of observations classified. Although predictor variables that were roughly 80 percent or more complete (N = 400) were used, the alignment of missing data between variables caused some classes to be drastically reduced. The class or tuple of defects (None, None, None, None) dropped from 265 observations to 120, OtherB 58 to 51, (Insufficient Material, Ejection, None, None) 21 to 15, (Insufficient Material, None, None, None) 20 to 12, (Thermal, None, None, None) 20 to 14, (OtherA, None, None, None) 18 to 15 and (Extra, None, None, None) from 15 to 12 observations.

Figure 25 points out the performance of the tree in terms of each class of defects. For the most part, the proportionally larger class sizes were classified more accurately. In three situations a class with more than ten tools was entirely misclassified. They were (Insufficient Material, None, None, None), (Insufficient Material, Thermal, None, None) and (Thermal, None, None, None) with 12, 11, and 14 observations respectively.

Most important for users may be the information found in Table 7. Since many of the defect combinations overlap, some misclassified objects may not be as "incorrect" as others. For example, the observed category (Insufficient Material, Extra, None, None) had 12 observations of which eight were predicted correctly, three misclassified as (Extra, None, None, None) and one misplaced as *OtherB*. The three misclassified as (Extra, None, None, None) only differ from the observed defect combination by *insufficient material*, while *OtherB* may be considered much more different. This evaluation implies a weighting scheme may have application at the time of tree construction.

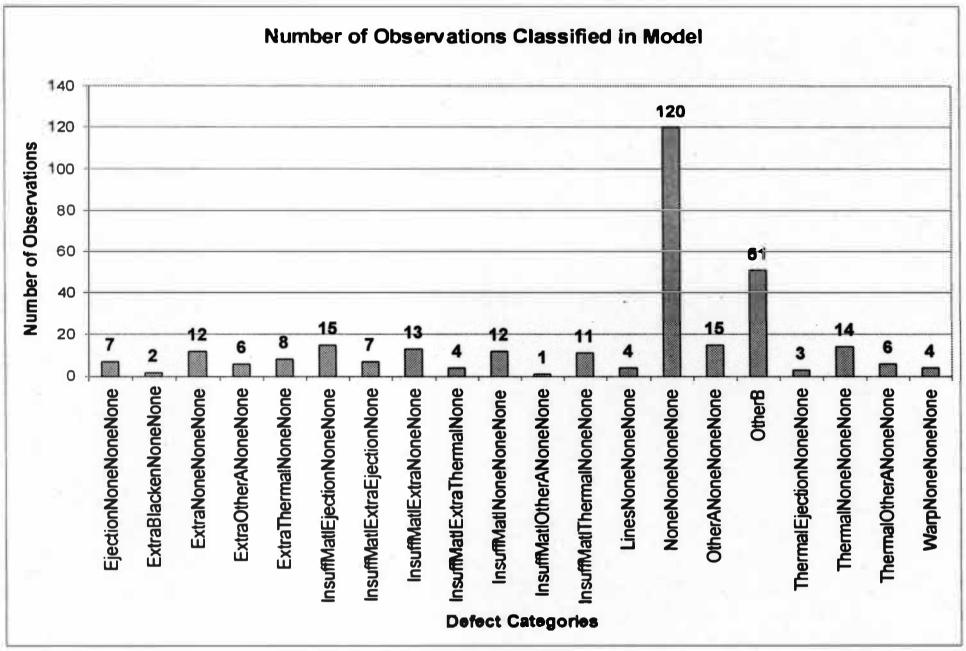


Figure 25. Number of Observations Per Defect Class in Model 12

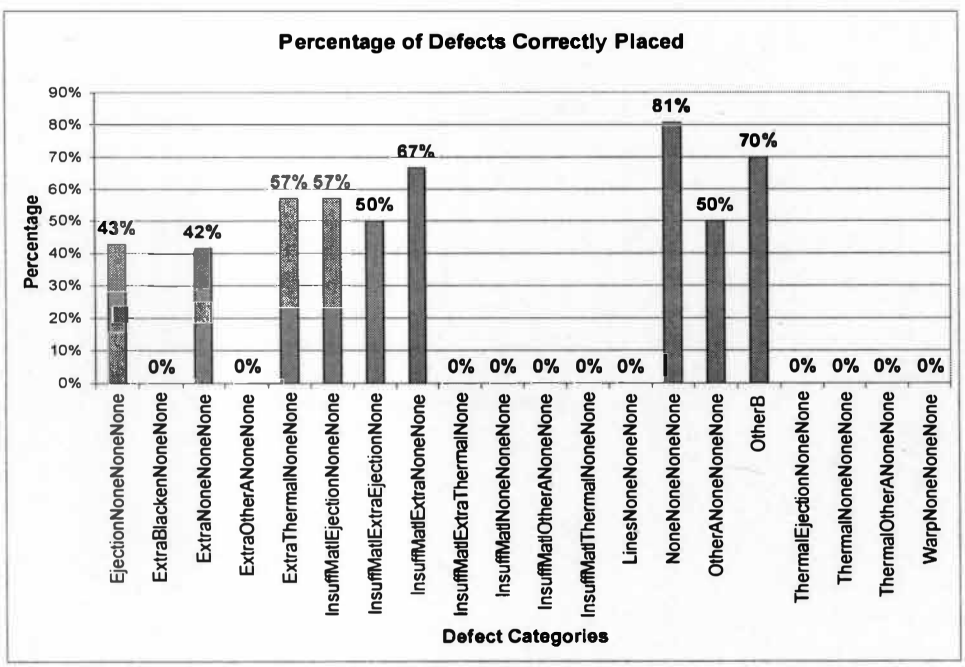


Figure 26. Percentage of Defect Classes Predicted Correctly in Model 12

Table 7. Matrix of Observed Versus Predicted Defects from the Classification Tree

	Observed Defect																			
Predicted Defect	EjectionNoneNoneNone	ExtraBlackenNoneNone	ExtraNoneNoneNone	ExtraOtherANoneNone	ExtraThermalNoneNone	InsuffMatlEjectionNoneNone	InsuffMatlExtraEjectionNone	InsuffMatlExtraNoneNone	InsuffMatlExtraThermalNone	InsuffMatlNoneNoneNone	InsuffMatlOtherANoneNone	InsuffMatlThermalNoneNone	LinesNoneNoneNone	NoneNoneNoneNone	OtherANoneNoneNone	OtherB	ThermalEjectionNoneNone	ThermalNoneNoneNone	ThermalOtherANoneNone	WarpNoneNoneNone
EjectionNoneNoneNone	3	1	0	0	0	2	0	0	0	0	0	0	0	0	0	3	0	0	0	0
ExtraBlackenNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ExtraNoneNoneNone	0	0	5	2	0	1	1	3	0	2	0	0	0	1	0	0	0	0	0	0
ExtraOtherANoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ExtraThermalNoneNone	0	0	0	0	4	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0
InsuffMatlEjectionNoneNone	0	0	0	1	0	8	0	0	0	1	0	1	0	4	1	0	0	0	1	0
InsuffMatlExtraEjectionNone	0	0	0	0	0	0	3	0	0	2	0	0	3	1	0	2	0	2	0	0
InsuffMatlExtraNoneNone	1	0	0	0	0	0	1	8	0	1	0	0	0	7	1	1	0	1	0	0
InsuffMatlExtraThermalNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
InsuffMatlNoneNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
InsuffMatlOtherANoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
InsuffMatlThermalNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LinesNoneNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NoneNoneNoneNone	2	0	3	0	3	0	0	0	3	4	0	5	1	96	4	9	1	8	3	1
OtherANoneNoneNone	1	0	0	0	0	0	0	0	0	1	0	0	0	2	7	0	0	0	1	1
OtherB	0	1	4	3	0	3	1	1	0	1	1	4	0	8	1	35	1	3	1	2
ThermalEjectionNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ThermalNoneNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ThermalOtherANoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WarpNoneNoneNone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

CHAPTER V

CONCLUSIONS

Implications for Industry

The first part of the methodology, root cause analysis was intended to reduce the number of classes in the dependent variable. By tracing defects back to their basic causes, it was found that many could be grouped without compromising the integrity of the individual defects. The first causal approach standardized the defects in terms of the plastic. This technique required a high level of expertise and left room for error. The second approach was more grounded in literature. Troubleshooting guides provided a commonality to draw comparisons between the similarities in attribute defects. The results of both methods were near equivalents and served to compliment each other quite well. Besides successfully preparing the data for classification tree analysis, two other outcomes were realized from the causal analysis. First, it maintained the focus of the research on predicting attribute defects. It provided insight for the subsequent statistical stages. The second outcome was more tangible. The compiled definitions and causes from the various documents resulted in a thorough troubleshooting guide centered about tool design and construction attributes. For many industries battling non-conformities, performing an exercise like this could bring about solutions never imagined.

The second part of the methodology, classification tree analysis generated a model capable of classifying simultaneous product defects. It is important to remember, this

model is one tool of many defect prevention methods. It most likely fits as part of some larger system or pairs well with a similar method. Checking a tool design against the tree model could be written as a task of ISO-9000 documentation. It could be strategically sequenced before a design review, so that engineers know what elements of the design to pay special attention. It should not be a replacement for sound design practice, but rather a flag for potential problems or reassurance that defects are unlikely. The classification tree provides the opportunity for managers to apply additional resources, such as veteran personnel, flow and cooling analyses and sophisticated tooling. Or more importantly, the tree can identify when and where not to allocate these resources in a program of tools. Providing the deviances and distribution of predicted classes at terminal nodes gives users a level of confidence in the prediction.

A majority of industrial experts would agree that the most reliable products are a result of well-practiced personnel. The classification tree model is most suited to assist inexperienced designers and engineers in identifying potential problem areas. It provides an added measure of making sure that items that should be obvious are not overlooked in the design phase. However, designers and engineers at all levels can benefit from the model's ability to provide information from the production floor in terms of design variables. In a sense, it "closes the loop" between design and manufacturing. Designers can be aware of the battles fought on the factory floor. Listing each tool at the terminal nodes allows a designer to reference the person(s) responsible for those previous objects and learn from their success or failure. If the tool is still in production, its location could be traced to witness its performance firsthand.

If utilized early in the design phase, say at the time of quote, a problematic product design identified by the model enables a company to adjust the quote or possibly reject the product all together. Using the model as a historical representation of past projects can help in leveraging the request for additional money from customers to compensate for potential problems. For example, it may be worth asking for a guided ejection system for a tool likely to experience ejection problems.

Problems in tooling are often related. It is not uncommon for one problem to be caused when another problem is solved. This is usually a sign of a narrow processing window. For example, increasing the pack pressure to fix sink can cause warpage in another area of the part. Having the ability to classify combinations of problems could be another method of troubleshooting for process technicians. The predicted defect combination may suggest a different problem exists along with a new list of remedies. Furthermore, some of the subtle differences in the clustered defects may offer unlikely solutions. For example, short shot and void were grouped together as Insufficient Material based upon their similarity value. The causes that did not overlap could be a solution to one or the other.

Maintenance

Several things should be considered in building and implementing a classification model. Were the predictor variables readily accessible? How efficient and effective was the process of classifying a new object? Did technical and non-technical users easily interpret the results? As business climates change and new product lines are introduced,

is the model capable of being updated? Answering questions like these will help prepare and maintain successful analyses.

Extended Uses

The methodology and results presented in this work provide a framework for the classification of simultaneous responses. It appears this methodology has application in many fields beyond defect prevention. Causal analysis could be an effective technique to support sophisticated clustering software. Practitioners in medical diagnosis may be interested in modeling and forecasting patient costs, where multiple medical conditions exist. It might be beneficial for financial consultants to assess the probability of clients opening and maintaining different accounts. Of course, the data, prior knowledge of the problem and goals dictate the efforts and direction needed to build a successful model.

Recommendations for Future Work

It is obvious that many variations of a classification tree could be generated through a variety of statistical practices and parameters. The accuracy of these models can be quickly assessed by their misclassification matrix of the learning sample and/or a test sample. However, as stated earlier there is more to classification analysis than producing an accurate classifier; the second purpose entails uncovering the predictive structure of the problem (Breiman et al., 1984). This especially rings true for such a complex process as in injection molding. Understanding which variables and interactions are needed to build a robust algorithm that characterizes inputs of an unknown is an important criterion. In cases like this, where real industrial customers exist, a model must be logical. Failure to do so will only lead to skepticism on the part of the expert and

ultimately rejection. The following approach is suggested for gaining acceptance and evaluating the rational behind the structure. The approach is centered on matching the expectations of the end user.

The first objective is the classification analysis should meet a desired level of accuracy in predicting the learning set. The specified performance measures could be adjusted for each class in the dependent variable. The next step should be an in-depth look at the misclassified objects. Identifying trends that continuously misplace many objects into a certain class could uncover something potentially important in the way the analysis was conducted or more importantly a solution to a problem. On the other hand, the misclassified objects may make sense. For example, misclassifying a handful of tools as the tuple (Extra, None, None, None) instead of (Extra, Ejection, None, None) is not terribly unrealistic. The problem combination only differs by one and the joint problem of extra and ejection may be related; meaning, if the extra was corrected the ejection issue would also cease (i.e. extra inside a core pin that causes the part to stick or break). Researching individual misclassifications may also point out outliers (of inputs) or an error in the initial classification of the object.

The second objective is to test the robustness and prove the model is reliable. Of course this can be performed by a variety of statistical techniques through test samples. However, it has been observed that the most effective way to meet expectations is to have the user validate the model first hand on a new object. Even if the outcome was not what was anticipated, it allows the analyst (and often expert) to figure out why. Of course, repeated misclassifications in the hands of the customer could be detrimental to its

acceptance, thus this step should be handled cautiously. Validation could also be very time consuming, which is why the sample size should be predetermined.

Concurrent to model validation is an intuitive approach that evaluates the structure of the tree. Regardless of the accuracy, the branches consisting of variable interactions should do one of two things: 1) verify a series of relationships that were already known or 2) reveal an unfamiliar arrangement that makes sense. If the variables leading to a terminal node cause some confusion, it may be beneficial to work backward. Remember, this is expected. The power of classification analysis lies in the ability to recursively analyze a multitude of relationships in a highly dimensioned sample space, otherwise impossible to the human brain. By asking a series of questions that rebuilds the partitions used to classify an object, the logic of split conditions can be examined. It also enables an expert to follow the paths of objects and whether or not it was reasonable.

Finally, the classification tree model can be compared against an expert model. This differs from the previous approach in that it is performed independent of the analysis. The expert model can take the form of a logic tree or fault tree diagram. The model needs to map the expert's knowledge of expected outcomes. This is probably the most time consuming suggestion, but it may be the most useful in matching expectations. It begins a transition from a nonparametric analysis to a parametric one, where the analyst has some insight as to how the structure should act. Weights, misclassification costs and strategic positioning of independent variables could be imposed according to the expectations.

In terms of building upon this classification tree analysis, several additional steps could be taken. First, missing data can be imputed or the surrogate option used to reduce the drastic loss of observations. Secondly, pruning and shrinking options could be used to eliminate the weakest splits and control the size of the tree. Misclassification costs and weights could be added to influence the classification of certain defects. This may be especially useful for tuples that differ by only one value. Other split criterion, such as Twoing and Entropy are also warranted since class sizes can fall below the set node size. Establishing a new measurement based on what the model is capable of predicting may be more useful. Finally, the idea of a four-tuple and the levels within should give insight as to how difficult it can be to manage so many possible arrangements. It may be necessary to prioritize the defects beforehand, which allows for smaller dimensions in the tuple.

Some final suggestions for future work include working backwards to see if certain types of variables, such as cooling, ejection, delivery and venting relate accordingly to the type of defect. For example it would make sense to see the flow length, nominal wall thickness, material and design requirements as a set of split conditions for *insufficient material*. Another suggestion is looking at the products at the terminal nodes to see if non-predictor type variables are clustered at these nodes. It could be that some outside influences, such as company tendencies or certain product lines are being classified inherently.

Limitations

An exploratory observational study is not a controlled study. The results are only as good as the data of those who collected it. The nonparametric nature of data collection

is not an exact science and relies heavily on hypotheses of experts. The extent of detail often has to be compromised because of the degree of difficulty in gathering certain data or due to poor record keeping. For example, recording water line dimensions on a tool is very tedious and a tool drawing, if available, can be outdated leading to incorrect values.

There are several issues that raise the need for a variable selection method. The analyst must rely on classification tree outputs to evaluate a predictor's performance. Literature in the field indicates that there is no good method to select the best variables prior to the analysis. The nominal nature of the dependent variable eliminates linear techniques such as correlation. Furthermore, graphical methods are difficult to handle past three dimensions and can be very labor intensive for many variables. A large number of variables also increases the risk of missing important relationships due to masking. Choosing to analyze variables by way of best subsets can be very time consuming on the part of the analyst and expert. Moreover, variables of a subset that are closely competing for splits can be very sensitive to missing data. This can create confusion on the part of the analyst (especially if surrogates are used). A nonstandard data structure can complicate matters further. For example, a tool built with a cold runner possesses data on delivery measurements, while a hot runner tool contains different information, such as the number of drops and valve gates.

Another limitation is in the class sizes of the dependent variable. For classes containing observations less than half of the fixed minimum node size, there is no opportunity to be classified correctly. Mathematically, these small classes cannot get majority at a terminal node and be predicted correctly. For example, in this study (Extra,

Blacken, None, None) began with four tools, but due to missing data was reduced to two. The minimum node size for model 12 was seven. Even if tools were placed into the same terminal node, they fall two cases short of majority. This supports the need to investigate other split criterion.

BIBLIOGRAPHY

Advanced Process Engineering. (1996). *Injection molding troubleshooting guide.* Seminole, FL: Author

Aluja-Banet, T., & Nafria, E. (1998). Generalized impurity measures and data diagnostics in decision trees. In Blasius, J., Greenacre, M. (Eds.) *Visualization of categorical data.* (Chapter 5). San Diego: Academic Press

American Society for Quality. (2004). Glossary. Retrieved July 15, 2004, from http://www.asq.org/info/glossary.

Anna-Reddy, M. (2003, February). Introducing Moldflow Plastics Advisers 6.0. [Electronic version]. *Flowfront.*

Badiru, A. B. (1992). *Expert systems applications in engineering and manufacturing.* Englewood Cliffs: Prentice-Hall.

Bellman, R. (1961*). Adaptive control processes: a guided tour.* Princeton, NJ: Princeton University Press.

Berins, M. L., (Ed.). (1991). *Plastics engineering handbook of the Society of the Plastics Industry* (5th ed.). New York: Chapman & Hall.

Boothroyd, G., Dewhurst, P., & Knight, W. (2002). *Product design for manufacturer and Assembly* (2nd ed.). New York: Marcel Decker.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsorth International Group.

Burke, L. (2001). Case study: neural network applications. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 11.6). New York: McGraw-Hill.

Crosby, P. B. (1967). *Cutting the cost of quality (the defect prevention workbook for managers).* Boston: Farnsworth Publishing.

Dewhurst, P. (2001). Design for manufacture and assembly. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 13.2). New York: McGraw-Hill.

Groover, M. P. (2001). *Automation, production systems, and computer integrated Manufacturing* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Gupta, T. & Seifoddini, H. (1990). *Clustering algorithms for the design of a cellular manufacturing system - an analysis for their performance.* Computers & Industrial Engineering, 19, 432-436.

Gyrna, F. M. (1988a). Quality costs. In Juran, J. M., Gyrna, F.M., (Eds.), *Juran's quality control handbook* (4th ed.) (pp. 4.2-4.29). New York: McGraw-Hill.

Gyrna, F. M. (1988b). Quality improvement. In Juran, J. M., Gyrna, F.M., (Eds.), *Juran's quality control handbook* (4th ed.) (pp. 22.2-22.72). New York: McGraw-Hill.

Gyrna, F. M. (1988c). Product development. In Juran, J. M., Gyrna, F.M., (Eds.), *Juran's quality control handbook* (4th ed.) (pp. 13.2-13.74). New York: McGraw-Hill.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Grablowsky, B. J. (1979). *Multivariate data analysis with readings.* Tulsa, OK: Petroleum Publishing

Hand, D. J. (1997). *Construction and assessment of classification rules.* Chichester, England: John Wiley & Sons Ltd.

Hartley, J. R. (1992). *Concurrent engineering.* Portland OR: Productivity Press

Ichida, T., & Voight, E. C., (Ed.). (1996). *Product design review: a method for error-free product development.* Portland, OR: Productivity Press.

Insightful Corporation. (2001). S-PLUS 6 (Version 6.2) [Data analysis software and manual]. Seattle, WA: Insightful Corporation.

Kane, V. E. (1989). *Defect prevention: use of simple statistical tools.* New York: Marcel Decker.

Lamberson, L. R. (2000). *Reliability engineering.* Kalamazoo, MI: Western Michigan University.

Latino, R. J., & Latino, K. C. (1999). *Root cause analysis: improving performance for bottom line results.* Boca Raton, FL: CRC Press.

Liang, Y., Ou, W., & Tang, G. (2003). *Product defect diagnosis.* [Unpublished manuscript, Ford Motor Company and Michigan State University]. Retrieved July 15, 2004, from http://www.mth.msu.edu/Graduate/msim/MSIMProjectReports/Spring2003/Ford.pdf.

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica, 97*, 815-840.

Loh, W. Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association, 83*, 715-728.

Mobley, R. K. (1999). *Root cause failure analysis.* Boston: Butterworth-Heinemann.

Moldflow Corporation. (2000). Internet Enabled Part Advisor (Release 5.0) [Plastic part manufacturability analysis software and manual]. www.moldflow.com.

National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods.* Retrieved July 15, 2004, from http://www.itl.nist.gov/div898/handbook.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman , W. (1996). *Applied linear statistical models (4th ed.).* Boston, MA: WCB/McGraw Hill.

Nova Chemicals. (2000). *Injection molding trouble-shooting guide.* Retrieved July 15, 2004, from http://www.novachem.com/04_products/PS/documents/processing_guides/Injection%20Troubleshooting%20Guide.pdf.

Rosato, D. V., & Rosato, D. V. (1995). *Injection molding handbook* (2nd ed.). New York: Chapman & Hall.

Routh, R. L. (2001). Artificial intelligence and knowledge management systems. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 12.5). New York: McGraw-Hill.

Salford Systems. (n.d.). *Technical note for statisticians.* [White Papers]. Retrieved August 15, 2004 from http://www.salford-systems.com/422.php

Sorensson, P. A. (2001). Quality management. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 13.6). New York: McGraw-Hill.

Statistics Canada. (2003). *Glossary.* Retrieved August 15, 2004, from http://www.statcan.ca/english/edu/power/ch8/variable.htm.

StatSoft, Inc. (2004). STATISTICA (Version 6) [Data analysis software and manual]. www.statsoft.com.

Steinbach, M., Ertoz, L., & Kumar, V. (2003). *The challenges of clustering high dimensional data.* New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition. Retrieved August 15, 2004 from http://www-users.cs.umn.edu/~kumar/papers/high_dim_ clustering_19.pdf

Tsuyuki, T. (2001). The role of statistical process control in improving quality. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 13.6). New York: McGraw-Hill.

Walker, J. M. (2001). Product development. In Zandin, K. B. (Ed.), *Maynard's industrial engineering handbook* (5th ed.) (Chapter 13.1). New York: McGraw-Hill.

Wilson, D. F., Larry, D. D., & Anderson, G. F. (1993). *Root cause analysis: at tool for Total Quality Management.* Milwaukee, WI: ASQC Quality Press.