Industrial Management & Data Systems

Prognostic analysis of defects in manufacturing
Jongsawas Chongwatpol

Received 25 May 2014
Revised 29 August 2014
Accepted 20 October 2014

Prognostic analysis of defects in manufacturing

Jongsawas Chongwatpol
NIDA Business School, National Institute of Development Administration, Bangkok, Thailand

Abstract

Purpose – Since works-in-process (WIPs) are highly vulnerable to defects because of the variety and complexity of manufacturing processes, the purpose of this paper is to describe how to utilize existing analytics techniques to reduce defects, improve production processes, and reduce the cost of operations.

Design/methodology/approach – Three alternatives for diagnosing causes of defects and variations in the production process are presented in order to answer the following research question: "What are the most important factors to be included in prognostic analysis to prevent defects?"

Findings – The key findings for the proposed alternatives help explain the characteristics of defects that have a great impact on manufacturing yield and the quality of products. Consequently, any corrective action and preventive maintenance addressing the common causes of defects and variations in the process can be regularly evaluated and monitored.

Research limitations/implications – Although the focus of this study is on improving shop-floor operations by reducing defects, further experimentation with business analytics in other areas such as machine utilization and maintenance, process control, and safety evaluation remains to be done.

Practical implications – This study has been validated with several scenarios in a manufacturing company, and the results demonstrate the practical validity of the approach, which is equally applicable to other manufacturing sub-sectors.

Originality/value – This study is different from the others by providing alternatives for diagnosing the root causes of defects. Control charts, costs of defects, and clustering-based defect prediction scores are utilized to reduce defects. Additionally, the key contribution of this study is to demonstrate different methods for understanding WIP behaviors and identifying any irregularities in the production process.

Keywords Manufacturing, Analytics, Defects, Data mining

Paper type Research paper

Introduction

The manufacturing industry has become more and more complex, involving a variety of sub-sectors across the supply chain. Drastic changes in consumer markets over the last decades have increased the pressure on production businesses. The main challenges are to speed up the introduction of new products to the market and to meet customer demand; thus inefficient production processes have an adverse impact on maintaining competitive advantages. The key performance measures are to increase yield, to maintain the quality of products, and to reduce the cost of operations.

The number of defects is a critical factor because defects not only lower manufacturing yields and the quality of products, but also cause potential reliability problems. Especially in high-tech manufacturing industries, works-in-process (WIPs) are highly vulnerable to defects because of the variety and complexity of manufacturing processes. In fact, a huge amount of data on WIPs, yield, defects, and process and equipment from shop-floor operations is, to some extent, automatically or semi-automatically recorded and accumulated in a database. Inevitably, these data can be used to signal whether the process is out of control. The important task is to enable

Industrial Management & Data Systems
Vol. 115 No. 1, 2015
pp. 64-87

0263-5577
DOI 10.1108/IMDS-05-2014-0158

organizations to understand and utilize the ample data so that they can find ways to leverage the data collected at each function in their manufacturing enterprises into actionable strategies to tackle their challenges and pressures.

Although, numerous studies (Chi *et al.*, 2009; Choudhary *et al.*, 2009; Ciflikli *et al.*, 2010; Harding *et al.*, 2006; Hsu and Chien, 2007; Kusiak, 2006) have been done over the past decades on ways to detect defects and why they occur, many organizations still face difficulties in diagnosing the root causes of defects, resulting in high costs for reworks and repairs. Specifically, these studies focus on building models to predict defects, considering merely "defects vs non-defects" scenarios to identify the critical causes of defects. Consequently, such analysis may not systematically and proactively recognize the patterns of defects and the results may not include all the factors that can cause defects since factors such as process variation, costs of defects, defect clusters, and signs of potential defects may have been ignored in the analysis.

This study seeks to fill this gap and answer the following research question "What are the most important factors to be included in prognostic analysis to prevent defects?" A framework is proposed for how analytics can help manufacturing industry leaders make strategic decisions to improve shop-floor operations. This study differs from others by providing several alternatives (control charts, cost bucketing, and clustering-based defect prediction scores) for diagnosing the characteristics of defects and causes of variations in the process so that corrective and preventive actions are properly promoted. The main purpose is to describe how to get the most value from enterprise data and utilize analytics techniques such as decision trees, neural networks, regression models, and cluster analysis to reduce defects, improve production processes, reduce costs of operations, and stay competitive.

The rest of this paper is organized as follows. First, related literature on analytics in defect prediction, yield improvement, and quality control in manufacturing is briefly reviewed in the next section. In third section, the research methodology along with the problem domain, data description, and predictive models are presented. A case study from a manufacturing enterprise utilizing analytics is then described. Fourth section outlines the proposed alternatives for diagnosing the causes of defects. After the results and discussion in Fifth section, the last two sections contain the managerial implications and the conclusions and directions for future research.

Brief literature review

Advanced analytics through data mining approaches has been used successfully to identify intrinsic patterns in data and interpret them into useful information within a particular context. Data mining incorporates both statistical and analytical techniques to effectively and efficiently understand and use data (Jackson, 2002; Turban *et al.*, 2011). Business analytics has been applied in many fields across all industries, from customer relationship management, behavioral profiling, healthcare, and genome analysis to supply chains (Davenport, 2006; Harding *et al.*, 2006). Many high-technology companies such as Toshiba and Dell take an integrated approach to analytics in their business to improve yield management and to ensure that the right products are being manufactured at the right time (Davenport, 2006). Numerous studies have examined the implementation of data mining in different areas in manufacturing such as production processes, quality improvement, fault detection, optimization of manufacturing yield, material requirement planning, and preventive machine maintenance (Chi *et al.*, 2009; Choudhary *et al.*, 2009; Harding *et al.*, 2006; Kusiak, 2006). Particularly in fault detection, data mining is used to identify patterns of defects, factors influencing the failure of processes, types of defects, and error rates in manufacturing (Harding *et al.*, 2006).

The common approach to detecting the root causes of defects is to focus on analyzing the process parameters. Machine breakdowns in the production process are reported as one of the most important factors in defect diagnosis. For instance, Ciflikli and Kahya-Özyirmidokuz (2010) develop a data mining solution for enhancing carpet manufacturing productivity that employs attribute relevance analysis, decision trees, and rule-based induction. The results indicate that the isolation of machine breakdowns in the production process and the proposed decision tree model produce a significant improvement (72.8 percent) in the accuracy ratio. Further investigation is toward identifying problems in the procedure that lead to defects. The historical data of environmental (humidity or temperature) and machine conditions (pressure, voltage, or current) in the production process is then assessed to ensure that the quality standards have been met. Recently, a data mining approach has been applied toward a zero-defect manufacturing (ZDM) system to ensure non-failures in the manufactured products. The key to achieving ZDM is to concentrate not only on the quality of the products and the product conditions, but also on the condition of the equipment and the degradation of performance as well (Wang, 2013). Additionally, the raw materials for each product and any working condition-related factors including duration, shift, and experience level of the assigned workers should also be considered in identifying the potential sources of the problems that influence the success or failure of the production process (Chi *et al.*, 2009; Choudhary *et al.*, 2009; Dean, 2014; Harding *et al.*, 2006; Kusiak, 2006).

In analytics specifically, many examples of using different data mining techniques in defect prediction can be found in the literature. Decision trees, logistics regressions, and neural network models are often used to identify the root causes of defects or the specific failure patterns. Applying such techniques usually results in increasing the accuracy rate of recognition of defects, which helps support the decision making in defect classification (Tan *et al.*, 2014; Yuen *et al.*, 2009). Cluster analysis is also applied to aid in examining the relationships between manufacturing practices and plant performance, including new product development, flexibility, efficiency, and market-based performance (Narasimhan *et al.*, 2005). In semiconductor manufacturing, data mining and knowledge discovery techniques such as the Kruskal-Wallis test, *k*-means clustering, and variance reduction splitting criteria are used in defect diagnosis to identify the possible causes of faults and manufacturing process variation in order to improve the yield of fabricated wafers (Chien *et al.*, 2007). Turhan *et al.* (2009) propose a defect prediction approach by applying principles of analogy-based learning such as nearest neighbor filtering. The main purpose is to utilize cross-company data to initiate the defect prediction process while developing a local repository of within company data to construct similar defect predictors (Turhan *et al.*, 2009). Control charts have also been used to monitor defects in most integrated circuit manufacturing. Hsieh *et al.* (2007) apply fuzzy theory with a control chart to monitor defects, with the goal of defect clustering. The results from using the proposed control chart help identify whether the process is significantly out of control and eliminate non-normal causes of defects (Hsieh *et al.*, 2007). Hessinger *et al.* (2014) apply data mining methods to determine and quantify root causes of yield loss from defects in the semiconductor industry. The main goal is to reduce defect risks by considering different defect sizes and types and analyzing the non-random distributions of failure and defect data. Recently, data mining is being applied to diagnose the anomalies in wind turbine bearings. The key method is to employ classification techniques such as anomaly detection and support vector machines to discriminate between defect examples in order to provide advanced failure warning and precise fault detection in early stages (Purarjomandlangrudi *et al.*, 2014). The comparison of various data mining techniques used to process fault diagnosis for manufacturing process parameters is summarized in Perzyk *et al.* (2014). These techniques are evaluated based on accuracy, robustness of results, and applicability. Interestingly, for the case presented in their study, the results indicate that simple statistical methods appear to be more robust than advanced techniques such as neural networks and support vector machines.

These studies are just a few examples of efforts to realize the benefits of analytics and data mining approaches in the manufacturing setting. This study is different from the others by providing alternatives for diagnosing the root causes of defects. The main focus is not just on the incidence of defects, but rather on the potential for defects due to variations in the operations caused by shop-floor operators, parts and components, machines, and production processes. The research methodology and problem scenarios are presented in the next section.

Methodology

We follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology as a guide in structuring the data mining project for diagnosing defects in the semiconductor industry. The CRISP-DM methodology breaks down data mining projects into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Shearer, 2000; Wirth and Hipp, 2000).

Phase I: business understanding

The case study presented here comes from one of the optical receiver-transceiver (ORT) production lines from a company in Thailand. This production line manufactures several ORT models, which can be grouped into four product families. Operating on a quarterly timeframe, a production planner releases work orders, based on both demand forecasts and actual purchase orders, to the shop-floor operators at the beginning of each week. After preparing all parts and components such as fiber optics, the printed circuit board (PCBA), heater cup, laser, and photodiode, an operator manually assembles the unit, called a WIP, based on the instructions for each model. Product families may share the same major parts and components, and common components are available for all models. Additionally, different models may go through similar or different mechanical sub-assembly processes. After assembly, the WIP is transferred to the testing workstation for the uploading of several software packages to test its performance. If the results from the performance testing are satisfactory, the WIP is then moved to the next workstation (final assembly, inspection, and packaging). Otherwise, the WIP is returned to the mechanical assembly line so that problematic WIPs can be inspected and changed accordingly. This WIP is now called a failed-test unit, or defect. Traditionally, when a defect occurs, the testing engineer collects data from the defective unit and identifies possible causes of the problem.

Currently, data mining in the form of decision trees and stepwise regressions is employed by this manufacturer to predict defects for each product family. Since beginning to use these techniques, the rework rate has decreased. However, the current data mining process focusses on determining the leading causes of defects at the product family level and considers only "defects vs non-defects" scenarios. Other factors such as costs associated with defects or source, pattern, and characteristics of defects have not been included in the analysis. Thus, some useful information is being ignored and the results produced by the current data mining models may not systematically and proactively represent the factors and recognize the patterns of defects. In this study, we offer several alternatives for diagnosing causes of defects and include the costs of defects in our analysis. Figure 1 presents the research framework of this study.

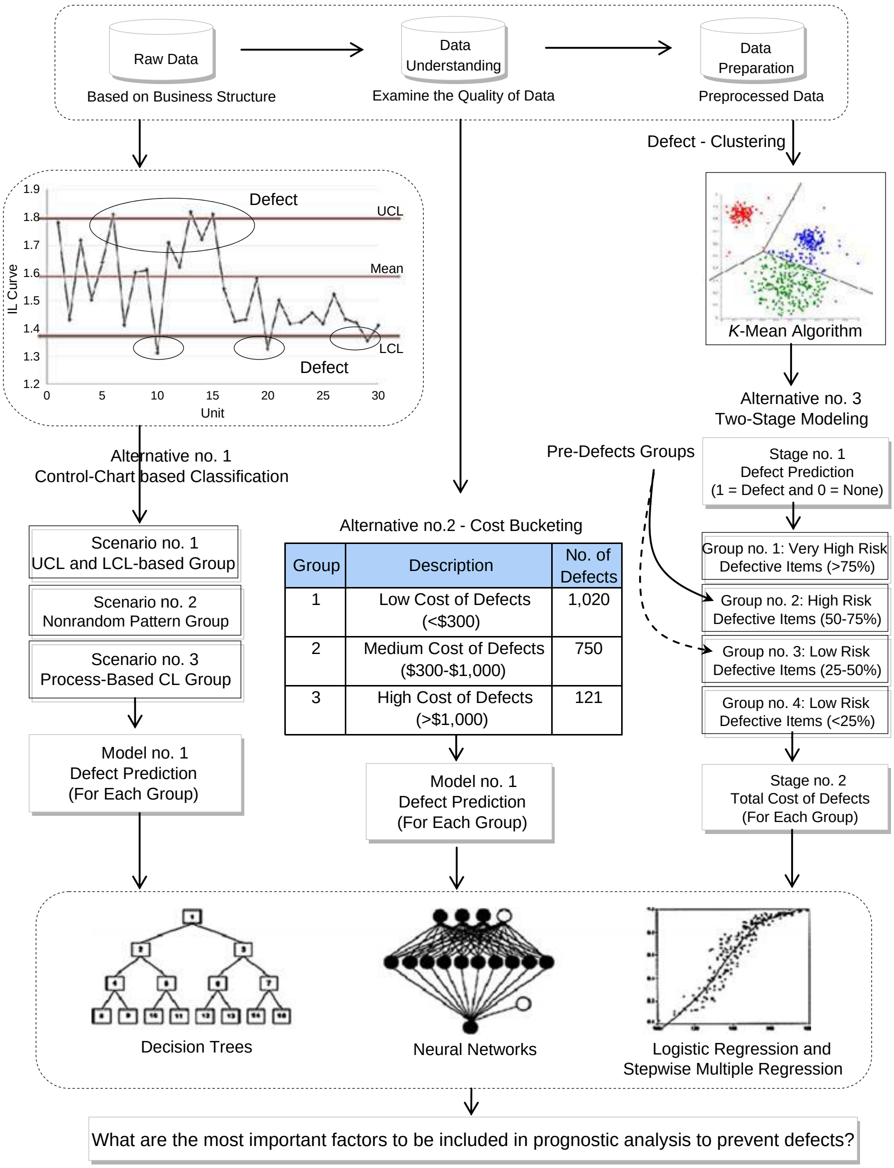


Figure 1.
Research framework

Phases II and III: data description and data preparation

After understanding the domain of shop-floor operations, the objective of this study is to achieve prognostic analysis of defects through the data mining approach. The first step is to understand the relevant data sources, assess data quality, and look at the data for preliminary insights. The next step is to preprocess the data from its initial raw state to the final data set, ready for model development. This preprocessing step - cleaning, transforming, constructing, and formatting the relevant data - takes about 60 percent of the project time. The original data set consists of over 10,000 records, with 26 variables related to staff, processes, machines, and raw materials, from January to December, 2013. The first step is to remove any inconsistencies, errors, and outliers. The next step is to assess whether the data are complete and which variables are to be included in the model. Because including variables with a high number of missing values can lower the quality of the findings, even after applying a missing value imputation method, the assumption is that the model excludes variables with over 50 percent of the information missing. Additionally, the data set is stratified to construct a model set with approximately equal numbers of each target variable. With a 50 percent adjustment for oversampling (for instance, 1,891 defects and 1,891 non-defects), the contrast between the two values is minimized, which makes pattern recognition in the data set easier and more reliable. For each alternative, the data set is partitioned into 60 percent for training and 40 percent for testing the data set before applying analytical data mining techniques to predict and explain the leading causes of defects.

Phase IV: modeling

A variety of data mining techniques is available in predictive analytics (Wang, 2007). However, only three popular models are used in this study to analyze data sets with multiple predictor variables: logistic regression, a decision tree, and an artificial neural network (ANN):

- Logistic regression is often used to predict a binary outcome variable or multi-class dependent variables. It builds the model to predict the odds of discrete variables (dependent variables) by a mix of continuous and discrete predictors, instead of by point estimate events as in the traditional linear regression model, as the relationship between dependent variables and independent variables is non-linear.
- A decision tree is another data classification and prediction method commonly used because of its intuitive explainability. A decision tree divides the data set into multiple groups by evaluating individual data records which can be described by their attributes. It is also simple and easy to visualize the process of classification where the predicates return discrete values and can be explained by a series of nested if-then-else statements.
- An ANN is a mathematical and computational model for pattern recognition and data classification through a learning process. It is a biologically inspired analytical technique, simulating a biological system, where a learning algorithm indicates how learning takes place and involves adjustments to the synaptic connections between neurons. Data input can be discrete or real-valued; the output is in the form of a vector of values and can be discrete or real-valued as well.

For a technical summary, including both algorithms and their applications for each method, see Delen (2010), Delen *et al.* (2005, 2010), Jackson (2002), and Turban *et al.* (2011).

Phase V: evaluation

To ensure the accuracy and validity of the decision tree, logistic regression, and neural network models built in the previous step, their performance is measured based on the misclassification rate and the average squared error. Using the proportion of incidents of misclassified data is very common in predictive modeling when the target variable is binary (1 = defect and 0 = non-defect). The observed misclassification rate should be relatively low for model justification. For the interval target variable (total costs of defects, $), the average squared error is evaluated among the three models built on the testing data set. The lower the average squared error, the better the model.

Phase VI: deployment

The last phase of the CRISP-DM methodology is the creation of a strategy for deployment. Identifying the sources of defects is generally not the end of the data mining project. The key important task is not only to understand why and how defects occur but also to utilize such key findings to prevent major and costly failures. Thus, any monitoring and maintenance strategies or recommendations to improve the performance of production process are outlined. The details on ways to improve shop-floor operations in terms of reducing defects or variations in the process are presented in section five "Analysis of Results and Discussion".

Defect prediction-based data mining approach

This section explains three alternatives for diagnosing defects. The first alterative utilizes a control chart to identify the sources of variation and the patterns of defects. The second alternative considers the costs associated with defects in the predictive models. The third alternative focusses on partitioning WIPs into four groups based on defect prediction scores (very low, low, high, and very high risk) so that priority can be given to WIPs in the very high risk of defect group.

Alternative no. 1: control charts and the characteristics of defects

In the performance testing workstation, each WIP is tested to ensure that certain indexes such as output power, operating voltage, frequency range, and responding time meet the threshold criteria. When the testing results are not satisfactory (greater or lower than threshold values), the testing engineer records the results and marks those units as failed-test units or defects. Table I presents an example of defects (1 = defect and 0 = non-defect) when the voltage between transistor gate and source (VGS) measured during the IL curve performance testing process is greater than 1.80 or lower than 1.40 VGS. The shop-floor analyst then builds predictive models to understand and explain the causes of defects. Attributes related to staff, processes, machines, and raw materials are included in the predictive models. Once the major causes of defects are identified, shop-floor operators conduct corrective maintenance and "quality at the source" strategies to reduce defects in the operations.

Although the current analysis of failed-test units seems to work very well, the shop-floor analyst focusses merely on the "defect count" or "number of defects" when

Staff Related Attributes

Process/Machine Related Attributes

Work In Process (WIP) – Performance Testing Measurement Criteria

Raw Materials Related Attributes

Dependent (Target) Variable

No.	ID	Lot Number	Staff_ID	Shift	....	Freq_MHZ	SubModel1	SubModel2	Line	....	Frame Stock #	Compound Stock #	Fiber Stock #	....	IL Curve	Defect
1	1002101	BZZ5058700	1002	1	....	27	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.78	0
2	1002102	BZU0298100	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.43	0
3	1002103	BZZ5061200	1002	1	....	27	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.72	0
4	1002104	BZU0238000	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.50	0
5	1002105	BZU0513400	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.64	0
6	1002106	BZZ8984500	1002	1	....	25	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.81	1
7	1002107	BZU0642800	1002	1	....	16	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.41	0
8	1002108	BZU1241500	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.60	0
9	1002109	BZU0205400	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.61	0
10	1002110	BZZ7741000	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.31	1
11	1002111	BZZ7080300	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.71	0
12	1002112	BZU0356100	1002	1	....	26	B015G213	A3316	CA0023	....	F0018	C2003	FB2100	....	1.62	0
13	1002113	BZU0314000	1002	1	....	26	B015G213	A3316	CA0023	....	F0018	C2003	FB2100	....	1.82	1
14	1002114	BZU0295600	1002	1	....	26	B015G213	A3316	CA0023	....	FD1002	C2003	FB2100	....	1.72	0
15	1002115	BZU0453500	1002	1	....	26	B015G213	A3316	CA0023	....	FD1002	C2003	FB2100	....	1.81	1
16	1002116	BZU0432500	1002	1	....	26	B015G213	A3316	CA0023	....	FD1002	C2003	FB2100	....	1.54	0

Table I. An example of defects at the performance testing workstation

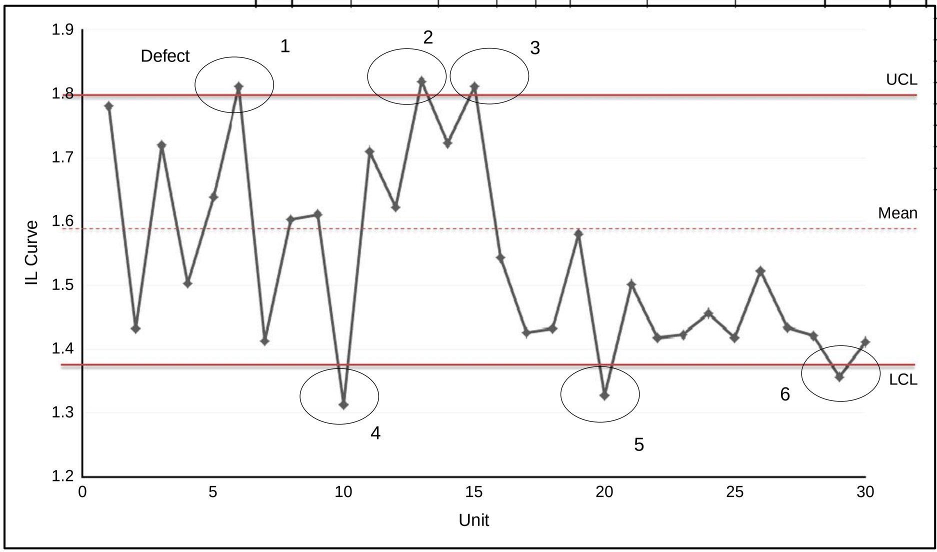
defect = 1, as presented in Table I, and ignores the assumption of randomness for defects or defect clustering.

In this study, a control chart is employed to initially monitor whether the current production process is unstable, out of control, or not predictable. The data from the process, gathered in the same manner as in the current operations, are presented in Table I, but using the control chart helps to determine the sources of variation and the patterns of defects. Any operating (IL curve) measures that are outside the control limits are recorded as defects. In other words, any plots that are above or below the predefined control limits indicate that assignable causes are present and the process is out of control. As presented in Figure 2, the target threshold value of the IL curve is 1.6 VGS. However, the actual IL curve varies appreciably for each unit. If the IL curve is too far from 1.6 VGS or outside the range of control limits, the performance of the WIP is degraded and unacceptable in terms of quality aspects. In Figure 2, there are six failed-test WIPs: three above the upper control limit and three below the lower control limit.

Upper and lower control limit groups of defects. Without looking at the control chart, these six failed-test units (Figure 2) are considered defects with similar characteristics, a so-called "homogeneous group of defects," in the eyes of the shop-floor analyst, so only two groups (defect vs non-defect) are explored in the current data mining project. However, one might argue that defects may be caused by variation in raw materials, operating machines, environmental changes, or the way shop-floor operators assemble the product; therefore, classifying defects based on their characteristics before building predictive models might help the shop-floor analyst better understand the causes of defects. Figure 3 shows two groups of defective WIPs to be analyzed separately: those whose IL curve values are above the upper control limit and those below the lower control limit.

Non-random pattern group. Although the production process is under control, meaning that the values of the IL curve for each unit are within the control limits so that all WIPs pass through the performance testing workstation favorably, any sort of pattern in the control chart can still suggest a non-random process. Stevenson (2009)

No.	ID	Lot Number	Staff_ID	Shift	....	Freq_MHZ	SubModel1	SubModel2	Line	....	Frame Stock no.	Compound Stock no.	Fiber Stock no.	....	IL Curve	Defect
1	1002101	BZZ5058700	1002	1	....	27	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.78	0
2	1002102	BZU0298100	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.43	0
3	1002103	BZZ5061200	1002	1	....	27	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.72	0
4	1002104	BZU0238000	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.50	0
5	1002105	BZU0513400	1002	1	....	26	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.64	0
6	1002106	BZZ8984500	1002	1	....	25	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.81	1
7	1002107	BZU0642800	1002	1	....	16	B0150426	A3305	CA0023	....	F0018	C2003	FB2100	....	1.41	0
											F0018	C2003	FB2100	....	1.60	0
											F0018	C2003	FB2100	....	1.61	0
											F0018	C2003	FB2100	....	1.31	1
											F0018	C2003	FB2100	....	1.71	0
											F0018	C2003	FB2100	....	1.62	0
											F0018	C2003	FB2100	....	1.82	1
											FD1002	C2003	FB2100	....	1.72	0
											FD1002	C2003	FB2100	....	1.81	1
											FD1002	C2003	FB2100	....	1.54	0



<u>Baseline Model no.1</u>
Dependent Variables
- (Defect vs Non-Defect): 0 and 1

Data Mining Method:
- Classification Tree
- Logistic Regression
- Neural Network

Figure 2.
The control chart of defects for the IL curve process

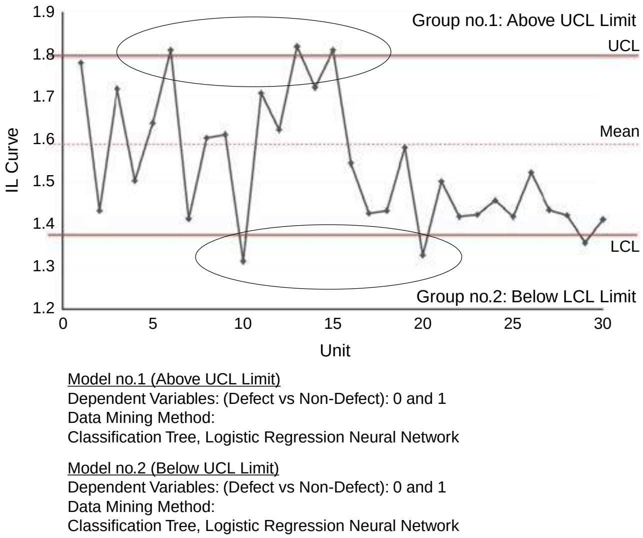
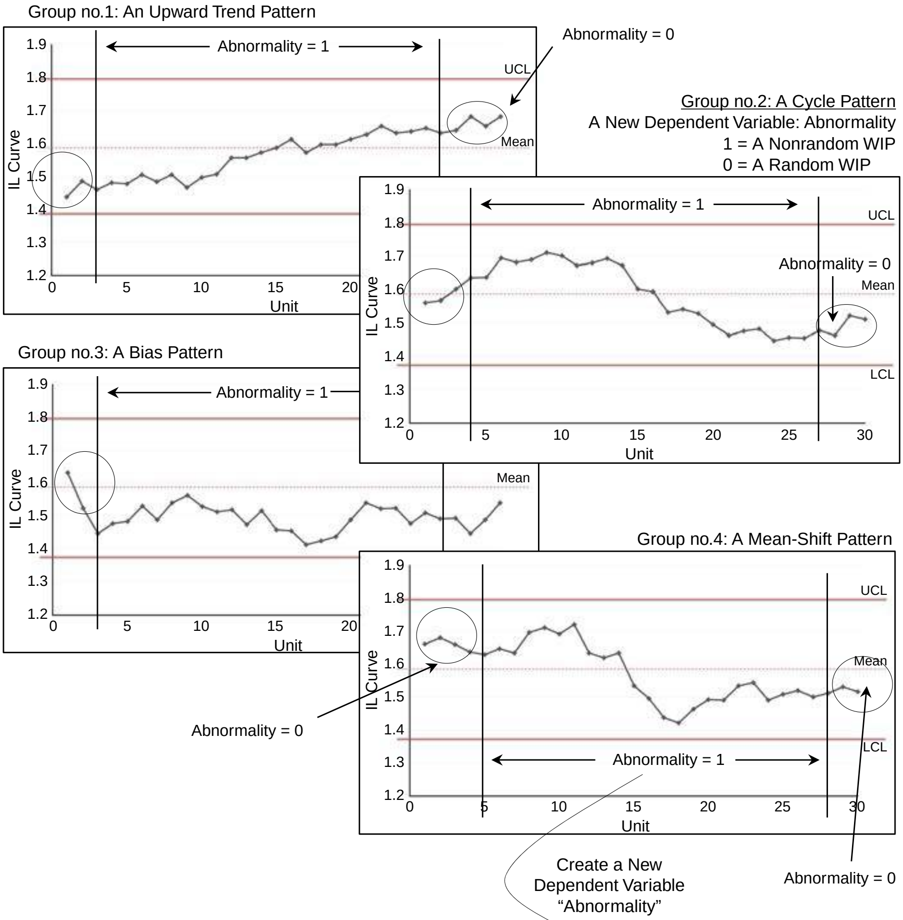


Figure 3. Defect clustering (upper and lower control limit groups)

provides some examples of non-random patterns in control chart plots that can be present in the operations and can signal whether the process has a tendency to go out of control. Figure 4 presents four different patterns that can be found in the production line based on the IL curve testing results. Essentially, WIPs can be categorized into sub-groups based on the presence of patterns such as trend (an upward/downward movement), cycle (a wave pattern), bias (on one side of the center line pattern), and mean shift (a shift in the average). For instance, as presented in Figure 4, a mean shift pattern occurs when there is an abrupt shift in the series of observations (IL curve), a sudden change in level within the series from above the target value to below the target value. For each product line, a new dependent variable "Abnormality" is created so that any WIPs that show signs of a non-random pattern are recoded as "1" (0 = random pattern). Thus, the shop-floor analyst can figure out the leading causes of abnormalities in the process.

Process-based control chart group. Because of the limited capacity in the testing stations, the testing engineer can test a maximum of 30 units of the same product line at a time. The testing results in Figure 5 show only a few defects in product model no. 1 when the target threshold value of the IL curve is 1.6 VGS. Thus, it seems that the process is stable and fairly in control. However, the data in the control chart actually indicate three distinct patterns and further analysis confirms that these WIPs are from different sources:

- The first group is the pre-built WIPs, called the "WIP Inventories," that were released to the shop-floor based on the forecast from previous quarters. Usually, WIP inventories are processed at the beginning of the quarter, not only to avoid a tight schedule due to the upcoming demand at the end of the quarter, but to utilize idle resources and capacity in each operating week as well. Once the actual purchase orders are received, these prebuilt WIPs are processed to the testing stations and final assembly.



A Mean-Shift Pattern

No.	ID	Lot Number	Fiber_Stock	....	IL Curve	Defect	Abnormality
1	1002301	BZZ5058700	FB2100	....	1.66	0	0
2	1002302	BZU0298100	FB2100	....	1.68	0	0
3	1002303	BZZ5061200	FB2100	....	1.66	0	0
4	1002304	BZU0238000	FB2100	....	1.64	0	0
5	1002305	BZU0513400	FB2100	....	1.63	0	1
6	1002306	BZZ8984500	FB2100	....	1.65	0	1
7	1002307	BZU0642800	FB2100	....	1.64	0	1
8	1002308	BZU1241500	FB2100	....	1.70	0	1
9	1002309	BZU0205400	FB2100	....	1.71	0	1

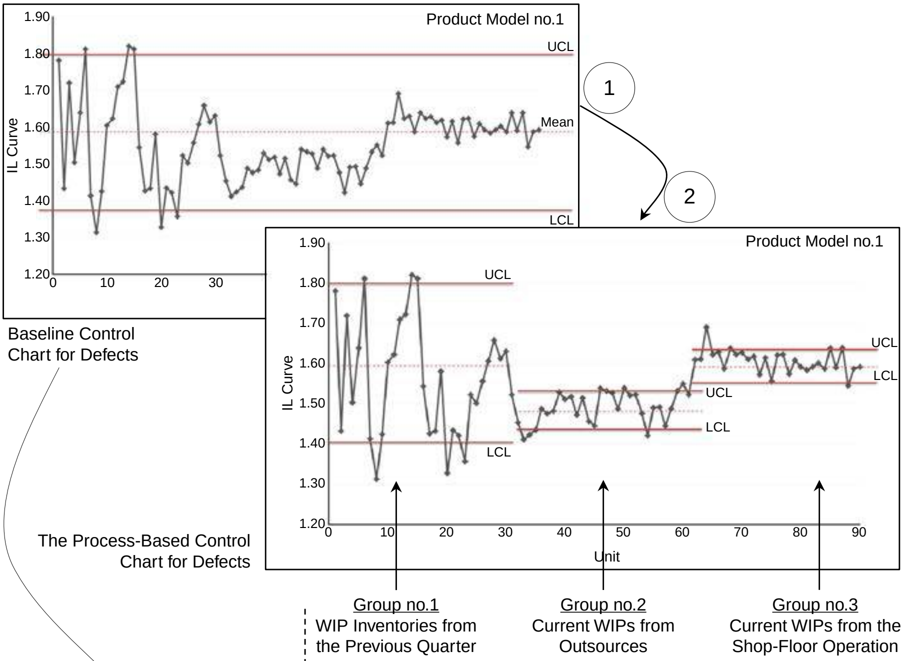
Model no.1 (Mean-Shift Pattern)
Dependent Variables:
- (Abnormality): 0 and 1

Data Mining Method:
- Classification Tree,
- Logistic Regression,
- Neural Network

Figure 4.
Defect clustering (non-random patterns)

- The second group is current WIPs outsourced to a local plant. When rush orders occur, the company sometimes decides to outsource partially assembled units to satisfy those demands.
- The third group is the WIPs from the current shop-floor operation itself.

As presented in Figure 5, the testing results for the first group deviate greatly from the target value of 1.6, compared to the other groups. The outsourced WIPs (the second group) show a non-random pattern since all IL curve values are below the target value; only the WIPs from the current shop-floor operations (the third group) seem to be stable



Only 3 Defects Shown

Unit	ID	.....	IL Curve	Defect	Abnormality (Within Process)	Abnormality (Among Process)	Cluster
1	1004101	.....	1.78	0	0	0	Group1
.....	.....	.....	.....	.....	.....	.....	.....
5	1004105	.....	1.64	0	0	0	Group1
6	1004106	.....	1.81	1	1	1	Group1
7	1004107	.....	1.41	0	0	0	Group1
.....	.....	.....	.....	.....	.....	.....	.....
13	1004113	.....	1.72	1	1	1	Group1
14	1004114	.....	1.82	0	0	0	Group1
15	1004115	.....	1.81	1	1	1	Group1
.....	.....	.....	.....	.....	.....	.....	.....
53	1004153	.....	1.48	0	0	1	Group2
54	1004154	.....	1.42	0	1	1	Group2
55	1004155	.....	1.49	0	0	1	Group2
56	1004156	.....	1.49	0	0	1	Group2
57	1004157	.....	1.45	0	0	1	Group2
58	1004158	.....	1.49	0	0	1	Group2
59	1004159	.....	1.53	0	1	0	Group2
60	1004160	.....	1.55	0	1	0	Group2
61	1004161	.....	1.52	0	0	0	Group2
62	1004162	.....	1.61	0	0	0	Group3
63	1004163	.....	1.61	0	0	0	Group3
64	1004164	.....	1.69	0	1	0	Group3

Abnormality among groups

Abnormality within a group

Figure 5.
Process-based control chart

and under control. Clearly, there are variations among these groups (1 = abnormality in the process and 0 = none). The shop-floor analyst should analyze and treat them separately so that the causes of variations or any signs of abnormality can be better captured. Additionally, the shop floor analyst can further analyze the causes of variation within each group (1 = abnormality within process and 0 = none) for signals of the potential cases of variation based on each process.

Alternative no. 2: cost bucketing

When the incidents of failed-test WIPs occur, the consequence is an amassing of costs associated with those WIPs, for example, the material costs of repairs, labor costs of rework, and any penalty costs due to delayed shipments. The impact of the costs of defects is demonstrated in Figure 6, which depicts the quantity and cost of defects for the first-quarter (13-week period) of product families nos 1 and 2. The number of defects in both product families increases throughout the period, which is quite normal since the company usually receives more purchase orders at the end of the quarter. The total costs of defects for product families nos 1 and 2 are \$5,250 and \$6,912, respectively. Although the total costs are not significantly different, product family no. 2 has a relatively high number of defects throughout the quarter, a total of 188 units, whereas product family no. 1 has only a few defects (25 units). This difference indicates that the average cost of defects per unit for product family no. 1 is significantly higher than that for product family no. 2.

Since cost information can provide an overview of shop-floor performance, including both the quantity and cost of defects as part of a data mining project is another way to help the shop-floor analyst understand the characteristics of defects and prioritize the emphasis on those defects with higher costs. The range of costs associated with defects for all models in the learning sample data set is from \$10 to \$300 per unit. As presented in Table II, we categorize our analysis into three different focusses based on the cost of defects. Our small sample size limits us to three groups partitioned in such a way that the sum of all costs of defects in each group is approximately the same at approximately \$33,900. We realize that in the high-cost group, approximately 33 percent of the overall cost of defects originates from 6.4 percent of defects. For the low-cost and medium-cost groups, the number of defects is 53.94 and 39.66 percent, respectively. A decision tree model is built on the training data set and is validated on the testing data set to predict the groups of defects (1 = low-cost group, 2 = medium-cost

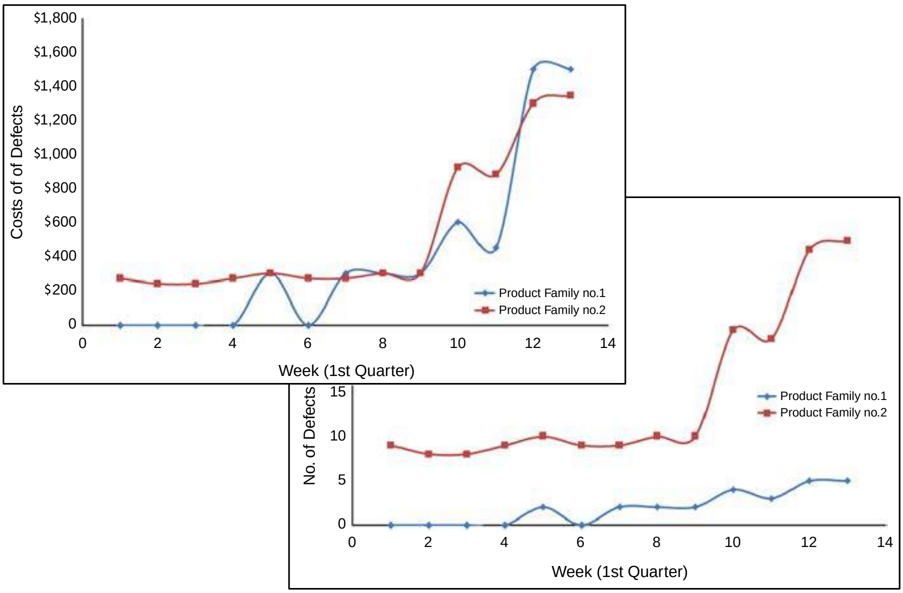


Figure 6.
Number of defects and costs of defects comparison (first quarter)

Group	Description	No. of Defects	Sample (%)	Total Cost ($)
1	Low Cost of Defects (<$40)	1,020	53.94%	$33,900
2	Medium Cost of Defects ($40-$100)	750	39.66%	$33,756
3	High Cost of Defects (>$100)	121	6.40%	$33,970

Dependent (Target) Variable

No.	ID	Lot Number	Model	....	IL Curve	Defect	Total Cost	Defect Group
1	1002105	BZU0513400	1	....	1.82	1	$300	3
2	1002212	BZU0432500	1	....	1.39	1	$300	3
3	1002351	BZZ5067500	1	....	1.39	1	$300	3
4	2002354	BZZ0236300	2	....	1.39	1	$270	3
5	2002409	BZU0514500	2	....	1.81	1	$270	3
6	2002453	BZZ0374500	2	....	1.81	1	$270	3
....	....	....	....	....	....	....	....	....
1889	2002354	BZZ0236300	2	....	1.37	1	$10	1
1890	2002409	BZU0514500	2	....	1.37	1	$10	1
1891	2002453	BZZ0374500	2	....	1.85	1	$10	1

Model no. 1
Dependent Variables (Defect Group):
1 = Low Cost Group,
2 = Medium Cost Group, and
3 = High Cost Group
Data Mining Method: Classification Tree

Model no. 2
Dependent Variables (Defect Cost): $$$$
Data Mining Method: Stepwise Regression

Table II.
Defect-cost bucket information

group, and 3 = high-cost group). A stepwise regression model, which provides additional information for analyzing defective WIPs, can also be applied to predict the cost of defects. The key assumption is that the shop-floor analyst can first focus primarily on any WIPs that are likely to fall into the high-cost group to reduce substantial rework or repair expenses.

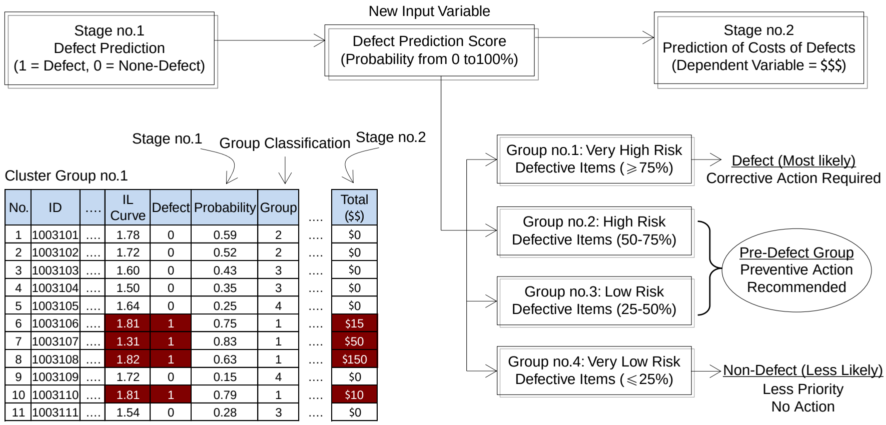
Alternative no. 3: cluster analysis and two-stage modeling

The first task is to develop a process for defect prediction that is based on the WIP segmentation profile rather than considering a "one equation fits all" scenario. Cluster analysis is used, as opposed to building a predictive model for each product family, because some product models share resources such as parts and components, machines, or processes. Thus, WIPs are segmented into sub-clusters that share similar characteristics. The *k*-means algorithm is deployed in this clustering process (see Collica (2011) for details on clustering techniques using SAS® Enterprise Miner™.

Instead of just predicting defects and non-defects (0 and 1) within each cluster group, WIPs are partitioned into four groups based on defect prediction scores (the probability of being defective; see Figure 7): very high risk (⩾75 percent), high risk (50-75 percent), low risk (25-50 percent), and very low risk (⩽25 percent). For Group nos 1 and 2, the shop-floor analyst can also further investigate the characteristics of defects by applying the stepwise regression or neural network models to predict the total costs associated with defects. The key interest is in the pre-defect group (Group nos 2 and 3), which are defined as follows:

> A pre-defect group is a group of WIPs that have not been certainly diagnosed as defective but show signs of potential defects, especially when the current production process has not been changed or controlled.

The very high risk group of defects (⩾75 percent) indicates that immediate attention and usually corrective action is needed. The group with a very low risk of defects (⩽25 percent) can be monitored periodically. The pre-defect group is further investigated to determine the potential causes of defects. This pre-defect issue is relevant to this study, as the main focus is on preventing defects and reducing defect prediction scores by



No.	ID	....	IL Curve	Defect	Probability	Group	....	Total ($$)
1	1003101	....	1.78	0	0.59	2	....	$0
2	1003102	....	1.72	0	0.52	2	....	$0
3	1003103	....	1.60	0	0.43	3	....	$0
4	1003104	....	1.50	0	0.35	3	....	$0
5	1003105	....	1.64	0	0.25	4	....	$0
6	1003106	....	1.81	1	0.75	1	....	$15
7	1003107	....	1.31	1	0.83	1	....	$50
8	1003108	....	1.82	1	0.63	1	....	$150
9	1003109	....	1.72	0	0.15	4	....	$0
10	1003110	....	1.81	1	0.79	1	....	$10
11	1003111	....	1.54	0	0.28	3	....	$0

Figure 7.
Two-stage modeling

monitoring key risk factors for defects helps improve the shop-floor performance. With defect prediction scores, the shop-floor analyst understands which groups are the main priority and can properly promote preventive action.

Analysis of results and discussion

Alternative no. 1: control chart and defect prediction

When defects occur, shop-floor engineers follow trouble-shooting instructions step-by-step to diagnose the causes of the problem, carefully consider the historical data of specific types of defects, and then perform corrective action. However, this traditional process takes time and does not solve the problem of defects. Using advanced analytics and a control chart to classify WIPs into sub-groups, causes of defects or non-random patterns are identified so that preventive actions can be taken.

After excluding variables with outliers and high numbers of missing values, the first step is to build a predictive model from the original sample data set focussing only on the defects and non-defects (targets = 1 and 0). This predictive model serves as the baseline scenario for comparing subsequent models in three scenarios. Scenario no. 1 considers the upper and lower control limit groups of WIPs. Scenario no. 2 focusses both on defects and on non-random patterns of WIPs. Scenario no. 3 includes the abnormalities when WIPs are derived from various sources of production. Table III presents the fit statistics for Scenario no. 2, where stepwise regression produces the best results with the lowest misclassification rate, 0.34086. Similar results can be seen for Scenario no. 1 as well. Although the neural network model shows better results in predicting the target variable in the baseline scenario and in Scenario no. 3, the

Table III.
Fit statistics for scenario no. 2 (non-random pattern group)

No.	Model description	Misclassification rate Testing	Misclassification rate Training	Average squared error Testing	Average squared error Training
1	Stepwise regression	0.34086	0.33383	0.21466	0.21051
2	Stepwise + neural network	0.34128	0.3594	0.21448	0.21256
3	Decision tree	0.3436	0.3805	0.2255	0.22127
4	Maximum tree	0.34365	0.3815	0.226	0.2213
5	Neural Network	0.35015	0.31377	0.2519	0.20072
6	Neural Network (6 Hidden Units)	0.35184	0.3296	0.2244	0.20601

misclassification rate is just slightly lower than that from the stepwise regression model. Thus, the stepwise regression model is selected because of its better explanation.

Table IV presents a summary of the final variable selection from the stepwise regression model for all scenarios. Most of the variables, such as SH Part Number, SH Line SA Lot, Spindle Motor Stock, or Freq_MHZ, are important causes of defects among all scenarios. For Scenario no. 1, the final variables selected in both upper and lower control limit groups remain exactly the same as in the baseline scenario. Only the odds ratio estimates are different among these variables. For instance, the top three variables, those with the highest odds ratio estimates for the upper control limit group, are Freq_MHZ (Odds = 1.495), Fiber_Stock (FB3100 with Odds = 1.132), and SH Line SA Lot (SH2100 with Odds = 1.110). For the lower control limit group, Test_PGMVER_3600, Freq_MHZ, and Fiber_Stock (FB3100) are the most important variables, with odds ratio estimates of 1.321, 1.307, and 1.110, respectively.

Interestingly, the results show that Wafer_MD, Main_AssemLine_003, Shift, and Operators are also leading causes of non-random patterns in the process (Scenario no. 2), while Rec_Time, PF_Code, and Compound_Stock are additional causes of defects in the process-based group (Scenario no. 3). These key findings signal the shop-floor analyst to focus not only on the common causes of defects but also on factors that indicate potential causes of problems in the production process. For instance, the shop-floor analyst starts monitoring the following three variables: Compound_Stock (Scenario no. 3), Operators (Scenario no. 2), and Shifts (Scenario no. 2). The key objective is to perform *post hoc* analysis to promote preventive action by tracking the performance of these variables. The findings from the results of this data mining project are as follows:

- Figure 8 shows a significant variation in the IL curve testing results of WIPs Model no. 1 between Operators nos 1 and 2. Since the production process relies on the manual assembly of some parts and components (fiber optics, PCBA, Chassis, etc.), the skill of operators is an important factor in determining the quality of WIPs. Clearly, the shop-floor engineers can periodically examine the WIPs from Operator no. 1 to reduce the chance of defective WIPs.

Baseline scenario Overall model	Scenario no. 1 UCL/LCL group	Scenario no. 2 Non-random pattern	Scenario no. 3 Process-based group
SH Part Number	SH Part Number	SH Part Number	SH Part Number
SH Line SA Lot	SH Line SA Lot	SH Line SA Lot	SH Line SA Lot
Spindle Motor_Stock	Spindle Motor_Stock	Spindle Motor_Stock	Spindle Motor_Stock
Fiber Stock	Fiber Stock	Fiber Stock	Fiber Stock
Test_PGMVER_3600	Test_PGMVER_3600	Test_PGMVER_3600	Test_PGMVER_3600
Test_PGMVER_1800	Test_PGMVER_1800	Test_PGMVER_1800	Test_PGMVER_1800
SODLot	SODLot	SODLot	SODLot
Freq_MHZ	Freq_MHZ	Freq_MHZ	Freq_MHZ
Line	Line	Line	Line
		Wafer_MD	Rec_Time
		Main_AssemLine_003	PF_Code
		Shift	Compound_Stock
		Operators	

Table IV. Summary of variable importance from stepwise regression

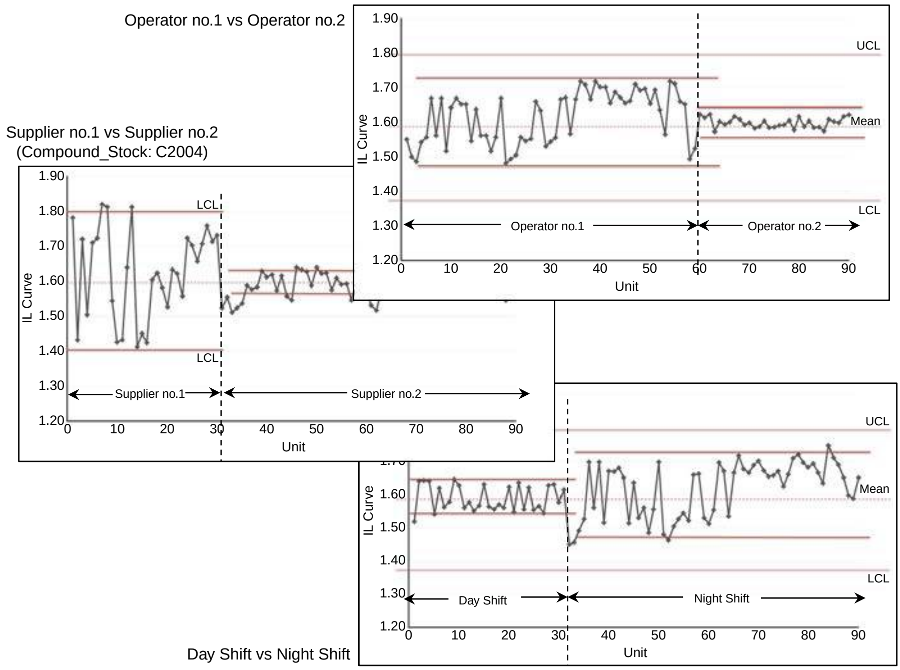


Figure 8.
WIP monitoring for product model no. 1

- The production process is very sensitive to the quality of raw materials such as Fiber Optics, Compound_Stocks, or Frame_stocks. A low quality in these parts and components usually results in reworks and repairs. Figure 8 shows that the IL curve testing results from WIPs with Compound_Stock (CS2004) from Supplier no. 1 tend to deviate more from the target IL curve value of 1.6 VGS than those with Compound_Stock (CS2004) from Supplier no. 2. This result implies that the quality of Compound_Stock no. CS2004 from Supplier no. 1 is questionable and needs further quality inspection. Thus, a "Quality at the source" strategy can be implemented once the shop-floor analyst knows which parts and components are potentially causes of defects.
- Figure 8 also shows that WIPs from the night shift are at a higher risk of being defective than those from the day shift. The results of testing WIPs from the night shift confirm that the plots of IL curve values are dispersed from the target threshold value and several points are outside the range of control limits. The shop-floor analyst can follow up on whether operators in the night shift receive proper training or follow the instructions in assembling parts for the released work orders.

Alternative no. 2: cost bucketing. A decision tree model is used to predict groups of defects based on cost information (1 = low-cost group, 2 = medium-cost group, and 3 = high-cost group). The following examples illustrate rule-based predictions from the decision tree model:

> ***IF*** a WIP is in [SubModel1 = "B0150426"] process ***AND*** gets through [LINE_ID = "BA06"] ***AND*** is operated by [Staff_ID = "1002"] ***THEN*** the probability of "Low-Cost Group" = 63.53%.

IF a WIP is in [SubModel1 = "B015G213"] process ***AND*** gets through [LINE_ID = "BA06"] ***AND*** is assembled with [Fiber_Stock # = C2004] ***THEN*** the probability of "Medium-Cost Group" = 78.1%.

IF a WIP is assembled with [Compound_Stock = CS2004] with the Freq_MHZ of > 100 ***AND*** is processed in the night shift ***THEN*** the probability of "High-Cost Group" = 82.1%.

A total of 24 rule-based algorithms are used classify the groups of defects. However, implementing all of these rules may indicate an overfit scenario, especially when the model performs well on the training data set but poorly on another independent data set. In this study, the original data set is partitioned into training and testing data sets. The decision tree model is developed on the training data set and is evaluated on the testing data set. Figure 9 presents the misclassification rate comparison when the decision tree algorithm is applied to both data sets. As the plot indicates, the misclassification rate decreases significantly when the number of leaves in the decision tree model increases, representing how well the developed classification tree fits the sample training data set. However, after applying the same model to the testing data set, when the number of leaves increases the plot indicates a big gap in the misclassification rate. Thus, for model simplicity, the decision tree model with nine leaves is selected to predict the defects in each cost bucket.

Practically, the shop-floor analyst is now able to first pay more attention to those WIPs that are likely to fall in the high-cost-of-defects group. Sub_Model 1 (B015G213), Line_ID (BA06), Fiber_Stock (C2004), or Compound_Stock (CS2004), for example, can be inspected routinely once work orders are released to the shop-floor operations. When preventive maintenance on each machine and the quality of raw materials are properly managed, the chance of defect incidents can be reduced.

Alternative no. 3: two-stage modeling. After WIPs are clustered into sub-groups based on similar characteristics, predictive analytics is applied to develop the defect prediction score for each cluster. The main focus is not on the predicted outcome of defect or non-defect but rather the probability of WIPs that are likely to be defects. Figure 10 presents examples of prediction results for four different groups based on the defect prediction scores.

A total of 171 WIP units are diagnosed with a very high risk of defects (⩾75 percent), while 304 units are diagnosed with a very low risk of defects (⩽25 percent). The pre-defect group, which includes both high risk (50-75 percent) and low risk

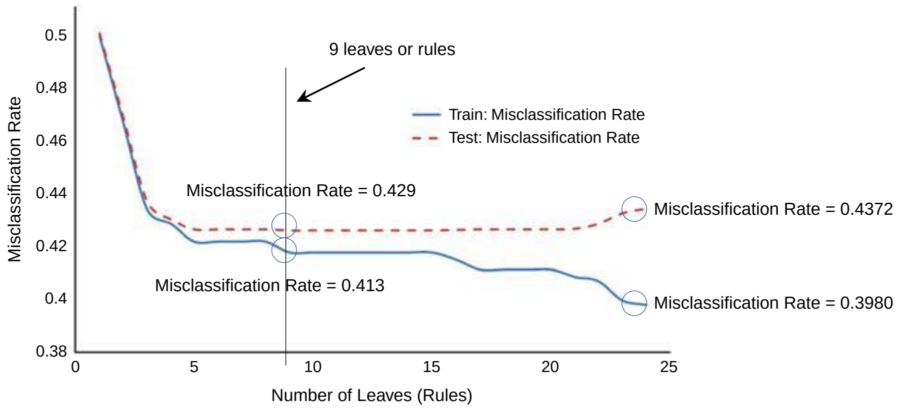
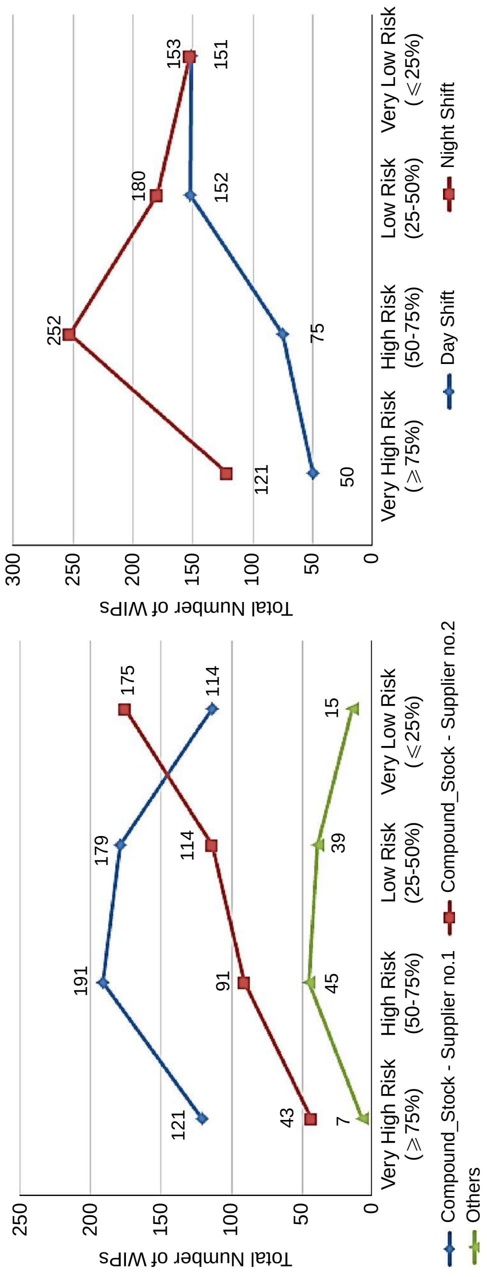


Figure 9. Misclassification-rate comparison



Very High Risk (≥75%)	High Risk (50-75%)	Low Risk (50-75%)	Very Low Risk (≤25%)	Total
171	327	332	304	1,134
15.08%	28.84%	29.28%	26.81%	100.00%

Figure 10.
Defect prediction by supplier and shift from cluster no. 1

(25-50 percent) groups, is accounted for 659 units, or approximately 58 percent of the total WIPs in the sample data set. Reducing the number of WIPs in this pre-defect group helps improve the overall shop-floor performance. Since the shop-floor analyst is interested in identifying the potential causes of defects, the details of defect prediction by Compound_Stock and Shift, which are also presented in "Alternative no. 1: control chart and defect prediction", are illustrated in Figure 10. The key findings are as follows:

- Compound_Stock from Supplier no. 1 has a higher possibility of defects than that from Supplier no. 2 in each defect risk category. For instance, 121 units of WIPs with Compound_Stock from Supplier no. 1 are at very high risk, while only 43 units of those from Supplier no. 2 are in the very high risk category. The results imply that raw materials from Supplier no. 1 are problematic; even the number of WIPs in the pre-defect group is almost twice the number using materials from Supplier no. 2.
- The ratio of WIPs with a very high risk and a high risk of defects from the day shift is much lower than that from the night shift. For instance, of the 327 WIP units in the high risk group (50-75 percent), approximately 77 percent are from the night shift. This indicates that "Shift" is one of the most important variables for defect prediction.

After getting some sense of defect prediction from the sample data set, the next focus is on understanding the difference in the total costs associated with defects. The average cost of defects in Group nos 1 (⩾75 percent) and 2 (50-75 percent) are \$162 and \$73, respectively. The predicted results are reasonable in that the higher the risk of defects, the higher the cost associated with defects. Similarly, the average cost of defects for WIPs with Compound_Stock from Supplier no. 1 and WIPs from the night shift is higher than the cost for those from Supplier no. 2 and from the day shift.

Managerial implications

Identification of sources of defects is always critical in manufacturing industries. The proposed predictive models utilize data mining techniques to detect both defects and abnormalities in operations. The first task is to collect relevant data from the production process so that shop-floor analysts can examine what has happened from the historical data to better detect any problems in the process and then encourage corrective action and preventive care for the factors that are possible root causes of defects and abnormalities. The input data should include attributes related to staff, processes and machines, and parts and components. The proposed model is neither complex nor time consuming but rather very efficient and cost effective in helping analysts understand the behavior of the process.

The key findings from the defect prediction-based data mining approach are used for knowledge acquisition to improve future system performance. Such knowledge can be integrated with the existing manufacturing enterprise system to make better strategic decisions to enhance process performance and product improvement. For instance, let us consider the case of variable "Freq_MHZ," which refers to the frequency range of the ORT products, measured during the performance testing process. Since the results of the stepwise regression in Alternative no. 1 and the decision tree in Alternative no. 2 indicate that Freq_MHZ is a key risk factor for defects, the shop-floor analyst is hypothetically interested in implementing a technology in the assembly line which can reduce the Freq_MHZ of WIPs by 10 percent. The idea is to study the impact of preventive measures on WIPs which have a Freq_MHZ higher than 100 MHz. Figure 11 provides the results of defect prediction before and after the preventive measure program starts. The number of WIPs in the very high risk group (⩾75 percent) decreases from 171 to 134 units (approximately 21.63 percent). Similar results can also be seen for the pre-defect group. As expected, the number of WIPs in the very low risk group (⩽25 percent) increases reasonably, to 22.04 percent (from 304 to 371 units). The expected cost saving from reducing defects only in the high risk-groups is estimated at $5,550 ((171-134 units) × (the average repair cost of $150 per units)). Thus, implementing the preventive measure program has a great impact on the quality of WIPs, especially in reducing the number of WIPs with a high risk of defects.

Conclusion

The need for predictive analytics to analyze ample data for timely decision making in manufacturing is apparent. This study offers three alternatives for diagnosing the problems of defective WIPs in shop-floor operations. The first alternative utilizes control charts to classify WIPs into sub-groups so that any causes of defects or any signs of non-random patterns can be further analyzed. The second alternative considers costs associated with defects as a part of the data mining project. The third alternative focusses on defect prediction scores and how preventive measures for the potential causes of defects reduce the predictive risk score, especially in the pre-defect groups. Defects and non-random patterns result from many factors, such as a change in operations, shop-floor operators, raw materials, machines, or production lines. The key findings from these three alternatives are noted so that any corrective action and preventive maintenance related to those factors can be regularly evaluated and monitored, helping reduce the defect rate in the shop-floor operations.

Many studies report that traditional data mining techniques such as neural networks, decision trees, and logistic regressions have been used successfully to predict defects. And, of course, some other advanced data mining techniques (fuzzy logic, genetic algorithms, support vector machines, and polynomial regressions) may be applicable in this study to improve the accuracy of defect prediction. However, reducing defects is not the only criterion to measure shop-floor performance. Thus, in addition to defect prediction, the contribution of this study is to demonstrate different methods for understanding WIP behaviors and identifying any irregularity in the production process. Although the

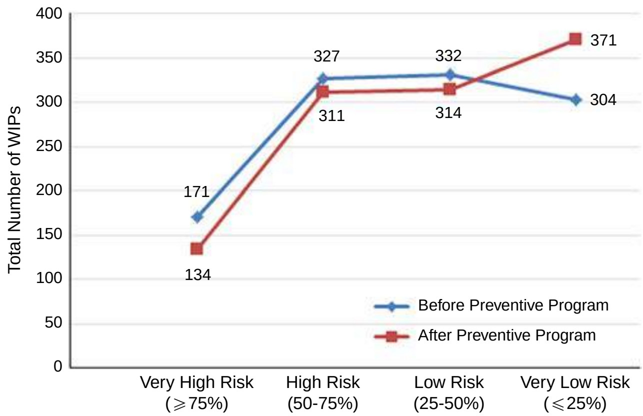


Figure 11. Hypothetical defect prediction after preventive program (reducing Freq_MHZ by 10 percent)

problem of high defect rates noted in this study is specific to this organization, the situation depicted in this company is quite typical of what is observed in most job-shop manufacturing environments where parts and components are assembled and tested for a quality standard. The proposed alternatives and the results for this company are applicable to a wide group of manufacturing operations, though not to all.

Another important task in successful data mining projects is in the data preparation process. In this study, about 60 percent of the project time was devoted to understanding and managing the data source from the production line and to cleaning and formatting data to use in the analysis. Major effort is needed in this process to ensure that the relevant factors related to machines, processes, materials, and staff are captured to efficiently and effectively diagnose defects. The higher the quality of data processed, the better the validity of the model generalized. Once analytics are integrated into each function in manufacturing, the next step is to ensure that the enterprise not only realizes the operational improvements but continues providing support and awareness of the value of applying data analytics in the long run.

It is also worth noting that emerging technologies play a vital role in collecting, analyzing, and turning data into useful information so that actionable plans for reducing defects and improving processes can be achieved. This study focusses only on applying analytics techniques in the manufacturing domain. Thus, future work should emphasize the integration of emerging identification technologies like Radio Frequency Identification (RFID) or Business Intelligence (BI) tools in the data mining projects. Additionally, further experimentation with business analytics in other areas such as machine utilization and maintenance, process control, and safety evaluation remains to be done.

References

Chi, H.-M., Moskowitz, H., Ersoy, O.K., Altinkemer, K., Gavin, P.F., Huff, B.E. and Olsen, B.A. (2009), "Machine learning and genetic algorithms in pharmaceutical development and manufacturing processes", *Decision Support Systems*, Vol. 48 No. 1, pp. 69-80.

Chien, C.-F., Wang, W.-C. and Cheng, J.-C. (2007), "Data mining for yield enhancement in semiconductor manufacturing and an empirical study", *Expert Systems with Applications*, Vol. 33 No. 1, pp. 192-198.

Choudhary, A.K., Harding, J.A. and Tiwari, M.K. (2009), "Data mining in manufacturing: a review based on the kind of knowledge", *Journal of Intelligent Manufacturing*, Vol. 20 No. 5, pp. 501-521.

Ciflikli, C. and Kahya-Özyirmidokuz, E. (2010), "Implementing a data mining solution for enhancing carpet manufacturing productivity", *Knowledge-Based Systems*, Vol. 23 No. 8, pp. 783-788.

Collica, R.S. (2011), *Customer Segmentation and Clustering using SAS Enterprise Miner*, SAS Institute, Cary, NC.

Davenport, T.H. (2006), "Competing on analytics", *Harvard Business Review*, Vol. 84 No. 1, pp. 150-151.

Dean, J. (2014), "Case study of a high-tech product manufacturer", *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*, John Wiley & Sons Inc., Hoboken, NJ, pp. 229-232.

Delen, D. (2010), "A comparative analysis of machine learning techniques for student retention management", *Decision Support Systems*, Vol. 49 No. 4, pp. 498-506.

Delen, D., Oztekin, A. and Kong, Z.J. (2010), "A machine learning-based approach to prognostic analysis of thoracic transplantations", *Artificial Intelligence in Medicine*, Vol. 49 No. 1, pp. 33-42.

Delen, D., Walker, G. and Kadam, A. (2005), "Predicting breast cancer survivability: a comparison of three data mining methods", *Artificial Intelligence in Medicine*, Vol. 34 No. 2, pp. 113-127.

Harding, J., Kusiak, A. and Shahbaz, M. (2006), "Data mining in manufacturing: a review", *Journal of Manufacturing Science and Engineering*, Vol. 128 No. 4, pp. 969-976.

Hessinger, U., Chan, W. and Schafman, B. (2014), "Data mining for significance in yield-defect correlation analysis", *IEEE Trans Semicond Manuf IEEE Transactions on Semiconductor Manufacturing*, Vol. 27 No. 3, pp. 347-356.

Hsieh, K.-L., Tong, L.-I. and Wang, M.-C. (2007), "The application of control chart for defects and defect clustering in IC manufacturing based on fuzzy theory", *Expert Systems with Applications*, Vol. 32 No. 3, pp. 765-776.

Hsu, S.-C. and Chien, C.-F. (2007), "Hybrid data mining approach for pattern extraction from Wafer Bin Map to improve yield in semiconductor manufacturing", *International Journal of Production Economics*, Vol. 107 No. 1, pp. 88-103.

Jackson, J. (2002), "Data mining: a conceptual overview", *Communications of the Association for Information Systems*, Vol. 8 No. 1, pp. 267-296.

Kusiak, A. (2006), "Data mining: manufacturing and service applications", *International Journal of Production Research*, Vol. 44 Nos 18/19, pp. 4175-4191.

Narasimhan, R., Swink, M. and Kim, S.W. (2005), "An exploratory study of manufacturing practice and performance interrelationships: implications for capability progression", *International Journal of Operations & Production Management*, Vol. 25 No. 10, pp. 1013-1033.

Perzyk, M., Kochanski, A., Kozlowski, J., Soroczynski, A. and Biernacki, R. (2014), "Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis", *Information Sciences*, Vol. 259 No. 1, pp. 380-392.

Purarjomandlangrudi, A., Ghapanchi, A.H. and Esmalifalak, M. (2014), "A data mining approach for fault diagnosis: an application of anomaly detection algorithm", *Measurement*, Vol. 55 No. 1, pp. 343-352.

Shearer, C. (2000), "The CRISP-DM model: the new blueprint for data mining", *Journal of Data Warehousing*, Vol. 5 No. 4, pp. 13-22.

Stevenson, W.J. (2009), *Operations Management*, 10th ed., McGraw-Hill Higher Education, New York, NY.

Tan, S.C., Watada, J., Ibrahim, Z. and Khalid, M. (2014), "Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects", *IEEE Transactions on Neural Networks and Learning Systems*, doi:10.1109/TNNLS.2014.2329097.

Turban, E., Sharda, R. and Delen, D. (2011), *Decision Support and Business Intelligence Systems*, Pearson Education, Upper Saddle River, NJ.

Turhan, B., Menzies, T., Bener, A.e.B. and Di Stefano, J. (2009), "On the relative value of cross-company and within-company data for defect prediction", *Empirical Software Engineering*, Vol. 14 No. 5, pp. 540-578.

Wang, K. (2007), "Applying data mining to manufacturing: the nature and implications", *Journal of Intelligent Manufacturing*, Vol. 18 No. 4, pp. 487-495.

Wang, K. (2013), "Towards zero-defect manufacturing (ZDM) – a data mining approach", *Advances in Manufacturing*, Vol. 1 No. 1, pp. 62-74.

Wirth, R. and Hipp, J. (2000), "CRISP-DM: towards a standard process model for data mining", paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, The Practical Application Company, Manchester, April 11-13.

Yuen, C., Wong, W., Qian, S., Chan, L. and Fung, E. (2009), "A hybrid model using genetic algorithm and neural network for classifying garment defects", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2037-2047.

About the author

Jongsawas Chongwatpol is a Lecturer in the NIDA Business School at National Institute of Development Administration. He received his BE in Industrial Engineering from Thammasat University, Bangkok, Thailand, and two MS Degrees (in Risk Control Management and Management Technology) from University of Wisconsin - Stout, and PhD in Management Science and Information Systems from Oklahoma State University. His research has recently been published in major journals such as *Decision Support Systems*, *Decision Sciences*, *European Journal of Operational Research*, *Energy - The International Journal*, and *Journal of Business Ethics*. His major research interests include decision support systems, RFID, manufacturing management, data mining, and supply chain management. Jongsawas Chongwatpol can be contacted at: jongsawas.c@ics.nida.ac.th

This article has been cited by:

1. ZhouBinghai Binghai Zhou LiuZilong Zilong Liu Tongji University, Shanghai, China . 2016. Optimizing preventive maintenance: a deteriorating system with buffers. *Industrial Management & Data Systems* **116**:8, 1719-1740. [Abstract] [Full Text] [PDF]
2. PistofidisPetros Petros Pistofidis M-2831-2013 EmmanouilidisChristos Christos Emmanouilidis PapadopoulosAggelos Aggelos Papadopoulos BotsarisPantelis N. Pantelis N. Botsaris ATHENA Research and Innovation Centre, Xanthi, Greece School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, UK Kleemann Lifts SA, Kilkis, Greece Department of Production Engineering and Management, School of Engineering, Democritus University of Thrace, Xanthi, Greece . 2016. Management of linked knowledge in industrial maintenance. *Industrial Management & Data Systems* **116**:8, 1741-1758. [Abstract] [Full Text] [PDF]