International Journal of Precision Engineering and Manufacturing (2019) 20:973–982
https://doi.org/10.1007/s12541-019-00136-7

Online ISSN 2005-4602
Print ISSN 2234-7593

REGULAR PAPER

Big Data Analysis Methodology for Smart Manufacturing Systems

Hyun Sik Sim[1]

Received: 12 February 2019 / Revised: 22 April 2019 / Accepted: 23 April 2019 / Published online: 7 May 2019

Abstract

The goal of smart factories is to improve productivity and reduce production costs, but it is more important to attain manufacturing competitiveness through improvements to product quality and yield. As product functions become more advanced and processing becomes increasingly miniaturized, the yields of micro-manufacturing processes have become an important management factor, determining the production cost and quality of a product. Micro-manufacturing processes generally pass through many stages to produce a product; therefore, it is difficult to find the process or piece of equipment where a fault has occurred. As such, it is difficult to realistically ensure high yields. This paper presents an S-EES (smart-equipment engineering system) construction and big data analysis methodology for manufacturing to increase product yield and quality in a smart factory environment. It also presents plans for acquiring the data needed for big data analysis of a manufacturing site and for constructing the system. To improve product yield, it is necessary to analyze the fault factors causing low yield; similarly, the critical processes and equipment that affect these fault factors must be identified and managed. However, interrelations exist between pieces of equipment, and complex faults are caused by the downstream as well as upstream in the processing sequence that a certain lot passes through. Because of this, yield management is important but also difficult. This study finds the fault-responsible processes and machines that affect yields by using a method that utilizes PLS-VIP (partial least squares with variable importance of projection) and association rules in micro-manufacturing line processes, and it classifies these processes and machines as single factors or cumulative factors. In addition, it applies the specific methodology to an actual site, extracts the fault-responsible processes and machines, and confirms the effects of important processes and equipment on yields.

Keywords Nano-scale manufacturing process · Yield analysis · Smart-EES · Big data analysis · Association rule · Partial least squares regression · Single factor(SF) · Cumulative factor(CF)

1 Introduction

As technological environments rapidly evolve and technology development periods gradually shorten, technology gaps in micro-manufacturing processes are gradually shrinking. In addition, there is intense competition between leading companies to capture high market share, and the time period for new products is gradually becoming shorter because businesses are devoting a great deal of time and effort to process improvements that increase yields. Examples of micro-manufacturing processing include PBGA-PCB (plastic ball grid array-printed circuit board, hereafter referred to as PBGA), semiconductor, and LCD processing, and these are widely used not only in common electronic products, such as TVs, but also in precision devices such as mobile phones and tablet computers. As the circuit structures in recently produced smartphones, and the like. Are more complex due to their high functionality and miniaturization, the parts that compose them have multilayer, high-function board structures, and this development has increased processing complexity. These complex board structures and processes not only increase product manufacturing costs, but also reduce product yield and make it difficult for businesses to ensure competitiveness. To achieve high yields, businesses have long since introduced statistical process management techniques into the manufacturing process in order to perform quality management. These techniques determine whether faults exist by checking board circuits and measuring plated

✉ Hyun Sik Sim
hssim@kgu.ac.kr

[1] Department of Industrial and Management Engineering, Kyonggi University, Gyeonggi, Suwon-city 154-42, Republic of Korea

thickness and line width, and so on after product processing has been completed [1]. However, production lot measurements require time and money, and they cannot be performed for every process. In reality, they are performed only for important processes in the processing line or at the final stage. In the PBGA manufacturing process, which is the target of this study, around 30 fault types are examined during the inspection after the etching process. Faults that are discovered during the inspection process are important factors in the creation of high production costs when the process progresses downstream, and they increase overall product production costs. Activities that minimize faults and maximize yields are absolutely necessary. Therefore, it is important to analyze which fault types are major causes of low yields and accurately find and manage the equipment and processes where faults are occurring. Each process in a PBGA production line is complex, consisting of many pieces of the same equipment. As such, it is very difficult to find which process and which piece of equipment experienced the major fault that is causing low yield. Furthermore, PBGA processing does not have just a single process that is responsible for causing faults; rather, faults are caused in a complex way through multiple processes and equipment.

This study uses data on the equipment paths that each production lot goes through to discover fault-responsible machines that affect the yields of micro-manufacturing processes. Fault-responsible machines are not just a single piece of equipment that affects faults. Instead, they comprise complex groups of equipment that cause faults at a higher level because they include participation from the downstream as well as the upstream. This is due to a phenomenon in which the possibility of faults increases due to the chemical and physical interrelationships between processes.

In order to perform big data analysis on this kind of manufacturing site, it is necessary to construct an environment that gathers lot histories and process/equipment parameters for major processes in a manufacturing site in real time and provides feedback. Among smart factory features, priority must be given to an environment that can connect the site's equipment and receive and control equipment information in real time. The key to implementing smart factories is to connect internal and external management resources of factories based on manufacturing IoT technology, and to form a platform for optimizing manufacturing and services [2]. The composition of the platform is based on real-time collection of production data, analysis and application of production big data [3]. Therefore, this study also presents an architecture for creating a manufacturing site equipment engineering system (EES) and data gathering, which are needed to create a smart factory.

The rest of this paper is organized as follows. Section 2 examines related research. Section 3 presents an Smart-EES implementation plan and analysis model for discovering fault-responsible processes and machines. Section 4 describes the experiments and data analysis process, as well as the results obtained from the proposed model. Section 5 presents the paper's conclusions and research for future development.

2 Related Works

PBGA-PCB processing and semiconductor processing are typical fields in micro-manufacturing processing, and they consist of the circuitry, plating, and etching processes. They are considered some of the most demanding processes in the manufacturing industry. To analyze processes that have such characteristics, various studies have, for a long time, relied on methodologies related to fault detection, diagnostics, and prognostication at manufacturing sites. Montgomery and Douglas [4] reported that traditional univariate quality management methods, such as Shewhart and CUSUM charts, have often been used in the past, but they suffer from a problem of frequent false alarms due to increases in interrelated management variables.

In the case of multivariate quality management, methods [5–7] have been proposed that use PCA (principal component analysis) on the many variables that occur during processing to reduce their dimensions and monitor product quality through multivariate statistics, such as Hotelling's T^2. In addition, methods that perform monitoring using various data mining techniques (artificial neural networks) [8], decision making trees [9] Support vector machines [10], and K-nearest neighbors algorithm [11] have also been proposed.

Looking at studies that find equipment and equipment parameters that influence yields during multi-stage manufacturing processes, Ma et al. [12]. applied statistical methods to the CVD (chemical vapor deposition) process, among several processes related to production, and they increased yields based on important variables that influence quality. First, they used clustering and stepwise regression on a chamber in which the CVD process is performed, and they found 31 important variables. Then, they used the important variables and the concept of Mahalanobis distance to set a new specific limit line, and they achieved high yields by doing so. Additionally, they provided a framework that uses process analysis by expressing important variables in a form that is similar to a DNA microarray. However, these methods are unable to consider phenomena in which multiple pieces of equipment, in several processes, simultaneously affect yields when many processes are performed. They have a limitation in that they only analyze a single process. The approaches mentioned up to now are all analysis methods that can be used to manage a single process or equipment parameter. By contrast, this paper is the first to propose a methodology that discovers equipment by analyzing the cumulative effects of multiple pieces of equipment in multiple processes, rather than just a single piece of equipment in a complex micro-manufacturing process.

In order to perform this type of big data analysis at a manufacturing site, all the devices and equipment in the factory must be connected to one another, and it must be possible to collect and analyze data based on this connectivity. Therefore, the functions that connect all the equipment at a site as well as collect and analyze the required data can be considered the most basic functions of a smart factory [13]. Methodologies have been proposed for constructing various smart factories for manufacturing businesses and EES [14]. However, these studies have focused on how to implement smart factories using information systems. There have been no studies published on methodologies for connecting to sites' big data and implementing actual smart factories. Therefore, this paper proposes a method for finding fault-responsible processes and machines that affect yields through a big data analysis of manufacturing, and it proposes a new Smart-EES construction plan, which is needed for this method.

3 Proposed Methodology

3.1 Smart-Equipment Engineering System (S-EES)

EES were created to find and remove the causes of equipment faults and preemptively prevent faults by managing equipment parameters in real time in order to maximize the operating capacity of equipment, which accounts for a large portion of production costs in the equipment industry [14]. In order to implement a smart factory, generally, all equipment at the site must be connected by Internet of Things sensors, and centralized control must be possible. To do this, the factory must be upgraded so that it is possible to connect the major equipment at the production site through the internet, import the required information from the equipment in real time, and control the equipment at the manufacturing execution system (MES) level. Therefore, in order to perform the functions required by a smart factory, EES construction must be given priority at a fundamental level. The functions required by this kind of EES, which is capable of centralized control, include process control, task condition management, equipment efficiency management, and big data. Process control consists of fault detection and classification (FDC), which detects equipment abnormalities, and Run to Run (R2R) systems, which control the products' processing conditions. Task conditions management includes recipe management systems (RMS), which automatically manage recipes. Equipment performance tracking (EPT) is a system that monitors equipment status in real time and analyzes and manages the status information.

Big data applications receive required data from the EES framework (data layer), execute analysis algorithms, and carry out the required control. EES at a manufacturing site perform the initial informationization of the real-time data generated by the equipment, and they support user and manager decision-making. At the higher levels, EES help with accurate ordering and production plan management by connecting with MES, enterprise resource planning (ERP), and supply chain management (SCM). At the lower levels, EES control equipment, such as transport and programmable logic controller (PLC)/post office protocol (POP) equipment, and connect with automation systems, which makes automation possible. They perform the role of integrating controls from enterprise-wide resource management to the lowest-level production equipment. Here, big data functions have been added to implement systems that can diagnose and predict processing faults (Fig. 1).

As shown in Fig. 1 above, the S-EES consists of the RMS, FDC, R2R, EPT, and big data modules. As for the analysis stages, Stage 1 collects required information from the MES and EES, such as the yield information for each product and lot, equipment task history, and equipment information, and creates a data set. Stage 2 uses a data mining algorithm to find fault-responsible processes and machines that affect yields, and it finds important process factors and equipment parameters that can actually be managed. Stage 3 uses the FDC to perform real-time monitoring of the critical parameters found in Stage 2, and it sets up the system is paused if an abnormality is detected.

In this study, we found fault-responsible processes and machines that affect yields through the analysis of the first and second stage, and additional functions that are controlled in connection with the management of critical equipment parameters and FDC will be carried out in the next study.

In EES, the functions needed for equipment management in many IT businesses are being developed and commercialized, and most manufacturing businesses are using these functions individually to manage equipment. However, some businesses have developed the required functions by

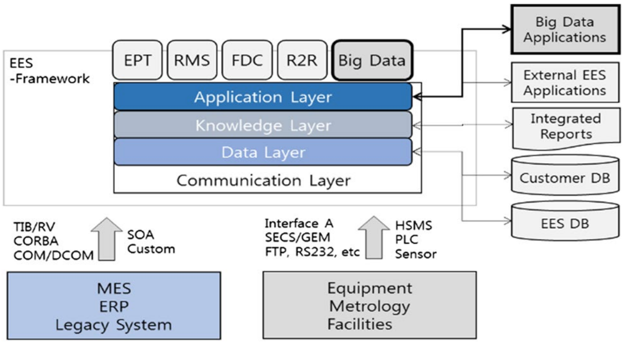

Fig. 1 Smart-EES configuration

themselves to safeguard their production method know-how. In this study, Oracle/Unix was used for the database management system (DBMS). For communication between site equipment, RS-232C communication was used, and for communication between the equipment and server, TCP/IP communication was used. All the manufacturing site data were created such that labels are semi-automatically attached to all equipment and lots, and bar code readers (BCR) are used to input the data. Figure 1 shows the final complete S-EES application architecture, and its layout is broadly divided into four areas: the application layer, knowledge layer, data layer, and communication layer. In the application functions that are configured here, the product, processes, and equipment are configured differently according to the characteristics of the tasks of the products being produced. As such, it is desirable to configure the required functions so that they are suitable to the task concept. Therefore, in order to implement an S-EES that is suitable to the product being produced, it is desirable to use a common framework and set up the features of the S-EES on top of the framework so that it is customized for the product, process, and equipment. The communication layer defines the functions needed according to the smart factory's overall system organization, and it performs the role of connecting to related higher and lower systems.

3.2 Data Mining Approach

3.2.1 Methodology Procedures

In this study, data pre-processing was performed on the equipment trace data using association rules before finding the fault-responsible machines. Equipment trace data are also called the process history, and refer to the sequence of process equipment that a single lot goes through. The trace can be considered a sequence of 0 and 1, in which a 1 indicates that a lot went through a certain piece of equipment, while a 0 indicates it did not. The PLS-VIP (partial least squares with variable importance of projection) method is used for equipment trace data to resolve the multicollinearity that exists between pieces of equipment. In addition, if association rules are applied directly, a large number of them are created when there is a large amount of equipment. Therefore, PLS-VIP is used first to select the important equipment causing the faults, and then association rules are used to find equipment that affect yields. Because the algorithm proposed in this study is based on association rules, this paper first explains association rules and then describes PLS-VIP.

3.2.2 Association Rules in our Application

Association rules are a method that helps to find hidden rules that are worth focusing on within vast amounts of data. The rules divide the relationships between items into LHS (left hand side) and RHS (right hand side), and they are expressed in a {LHS ⇒ RHS} format. In this paper, the LHS refers to the process and equipment sequence, and the RHS refers to the class (normal or fault). In this study, RHS is limited to lots that show faults. By using association rules, this study can identify a correlation between two or more pieces of equipment and quality variables. This compensates for the disadvantages of correlation coefficients, which can only show the relationship between a single piece of equipment and a quality variable. Of course, one alternative to this is to use multiple regression analysis to discover the fault-responsible machine; however, the results of multiple regression analysis are not reliable in cases where multicollinearity exists between pieces of equipment. Therefore, as explained in 3.2.1, a step for resolving multicollinearity is necessary. PLS was used to overcome the multicollinearity problem. To make simultaneous analysis easy, the PLS-VIP method was used to select only equipment that contribute greatly to faults, and then association rules were applied. Before looking at the association rule-based methodology, which will be introduced in the next chapter, this paper will provide a simple description of the indices that are related to association rules.

3.2.2.1 Support, Confidence and Lift Association rules for mutually separated items X and Y are expressed in the format {X ⇒ Y}. X refers to a process and equipment sequence, and Y refers to a class. Association rule strength measures the rule's support and confidence values [15]. Support generally refers to the rate at which rules occur in the total data. In this paper, it can be considered the ratio of total lots to the faults that occur as the lots that go through the equipment specified in the rules. N refers to the total number of lots. As for u, the support is shown in Eq. (1), when there is a quantity that includes both X and Y.

$$Support\{X \Rightarrow Y\} = \frac{\sigma(X \cup Y)}{N} \tag{1}$$

Confidence is the ratio at which Y is also included when X is included. It refers to the frequency of faults occurring after the lots having passed through certain equipment. Confidence is shown in Eq. (2).

$$Confidence\{X \Rightarrow Y\} = \frac{Support\ (X \cup Y)}{Support\ (X)} \tag{2}$$

Finally, lift refers to the ratio at which Y occurs in association with X compared to the ratio at which Y occurs by itself, and it can be understood as an index that is similar to a correlation coefficient. That is, it refers to the degree of correlation between the equipment that is passed through and the faults that occur. If they are independent of each other, the lift value is 1 [15]. This means that the equipment that is passed through and the values occur independently of each other. Therefore, a value of more than 1 implies that there is a positive correlation between the equipment that is passed through and a fault. A value of less than 1 means a negative correlation. Because of this, analyses that use lift normally focus on rules with a value of more than 1. Lift is shown in Eq. (3).

$$Lift\{X \Rightarrow Y\} = \frac{Support(X \cup Y)}{Support(X) \cdot Support(Y)} \tag{3}$$

3.2.3 Single & Cumulative Factor Detection Based on Association Rules

This paper proposes a new algorithm that considers correlation analysis and process complexity. The key point of this paper is not just discovering a single fault-responsible machine but understanding the degree to which the downstream and upstream affect the fault while simultaneously managing the fault-responsible machines and increasing yield. In this paper, one fault-responsible machine is defined as a single factor (SF), and multiple fault-responsible machines, in which the downstream and upstream contribute to the fault, are defined as cumulative factors (CF). Here, the following is clearly critical. Because there are cases where multiple fault-responsible machines contribute to a fault to a smaller degree than a single piece of equipment, CF must be able to judge fault prediction at a higher level than SF. This is determined by accuracy, which refers to the ratio of the number of faults, which have been distinguished by the rules, compared to all the lots that have passed through the equipment specified in the rules. A comparison is made between the upstream accuracy and the accuracy when the downstream is included. If the ratio of increased accuracy with downstream participation in the process, compared to the accuracy when it does not participate (i.e., accuracy of the upstream only), is above a fixed level, this means that the rule is a cumulative factor (CF). In this paper, this ratio is called the cumulative effect, and the cumulative effect is shown in Eq. (4).

$$Cumulative\ effect(\%) = \frac{Incremental\ Accuracy\ by\ downstream\ process}{Accuracy\ of\ upstream\ process} \times 100\% \tag{4}$$

Figure 2 shows a flowchart of the algorithm proposed in this study.

```
Algorithm: Cumulative factor detection based on association rules
Input:
D                       // Process data
min-sup                 // Minimum support
min-conf                // Minimum confidence
min-lift                // Minimum lift
min-cum                 // Minimum cumulative effect
Output:
Sin_Fac, Cum_Fac        // Single, Cumulative factor
Procedure:
  Rule set R ← Apriori algorithm (D, min-sup, min-conf)
  Construct tree-shaped rule structures using R
  Find single factor and cumulative factor using the tree structures and
  min-cum
```

Fig. 2 Cumulative factor detection algorithm based on association rules

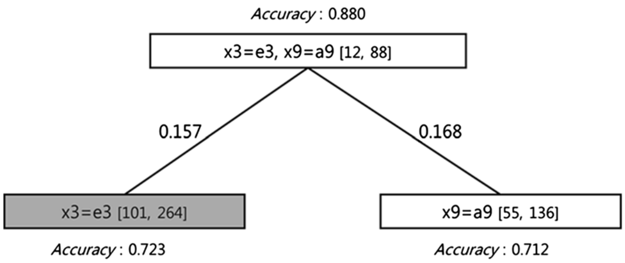

Fig. 3 Tree-shaped rule expression between upstream and downstream processes

This algorithm can be broadly divided into two stages. In the first stage, the Apriori algorithm [16] is used to create a rule set (R) that satisfies minimum support (min-supp), minimum confidence (min-conf), and minimum lift (min-lift). The second stage finds the SF and CF from the created rules, and it sets the minimum cumulative effect threshold value (min-cum). If a rule composed of multiple pieces of equipment does not show a cumulative effect, it is removed from the rule set (R). Specifically, rules with a length of 1, from among the rule set (R) created in the first stage, belong to the SF. The cumulative effect is measured in the rules with a length of 2 or more. In this process, rules are depicted in tree form to easily understand the inclusion relationships between the upstream and downstream. In the tree, which shows the inclusion relationships between rules, as a rule's length increases, it is placed on the tree's upper layer. In this study, such rules are called upper layer rules. Also, the partial sets which compose upper layer rules are called lower layer rules. Lower layer rules naturally have a shorter length than upper layer rules.

Figure 3 shows one relationship tree that depicts rules created by the Apriori algorithm when the minimum support (min-sup) and minimum confidence (min-conf) are 0.05 and the minimum lift (min-lift) is greater than 1. In Fig. 3, the nodes show the rules, and the shaded part shows the lower layer rules that depict the upstream of the upper layer rules. In the relationship tree, rules with a length of 1 are SF. Therefore, the SF in Fig. 3 is rule {x3 = e3} and rule {x9 = a9}. The figure represented by the right side of all the rules that make up the tree is the number of normal and fault states discovered when passing through the equipment that the rule represents.

In Fig. 3, {x3 = e3, x9 = a9} [12,88] means that 12 normal states and 88 fault states occurred when the lot passed through equipment e3 in process x3 and through equipment a9 in process x9. The accuracy can be calculated based on the number of times normal states and fault states occurred. The numbers above the lines, which connect rules to other rules, refer to the accuracy differences between upper layer rules and lower layer rules. If this number has a positive value, then accuracy increases when moving from a lower layer rule to the upper layer rule directly above it. From this, it can be seen that a cumulative effect occurred due to the downstream.

In Fig. 3, in order to examine the cumulative effect of downstream {x9 = a9} on upstream {x3 = e3}, the accuracy of rule {x3 = e3, x9 = a9} and rule {x3 = e3} is used. However, rule {x9 = a9} is not upstream of another process on the tree; therefore, it is an unnecessary rule in the process of examining cumulative effects. Cumulative effect is the ratio of upstream accuracy to the accuracy increased by the downstream. The cumulative effect between the two rules connected by the line on the left side of Fig. 3 is 21.7% (= 0.157/0.723 × 100). If this value is larger than the minimum cumulative effect threshold value (min-cum), it means that the rule {x3 = e3, x9 = a9} is a cumulative factor. However, if the difference in accuracy between an upper layer rule and a lower layer rule is a negative value, it means that the accuracy was actually reduced by downstream participation. In this case, not only did a cumulative effect clearly not occur, but it shows that a single fault-responsible machine (SF) displayed excellent performance in terms of determining faults compared to the cumulative factor (CF).

3.2.4 Partial Least Squares (PLS)

The reason that multicollinearity is a problem is because high correlation between independent variables can lead to bad judgments. For example, if multicollinearity exists, the size of the estimated regression coefficient can vary within a wide range if just one or two variables are added or omitted [17, 18]. To resolve this, a variable selection method is generally used to exclude some independent variables with high correlation, or a principal component analysis (PCA) is used to extract mutually independent principal components, and then a regression analysis is performed. Unlike PCA, PLS can find latent variables that can simultaneously describe independent and dependent variables in order to perform a more meaningful analysis. Below is a PLS model for a data ix matrix X ($n \times k$) made of k number of independent variables, n number of observations, and a dependent variable vector y (n × 1).

$$X = TP^t + E \tag{5}$$

$$y = Tb + f \tag{6}$$

In Eq. (5), X can be decomposed into T, p^t, and E. T is the X-score that refers to the position where the original variable exists in the space made of latent variables. p^t is the loading that responds to X. E is the error matrix, which cannot be completely described by latent variables. In Eq. (6), b is the coefficient that describes the relationship between the X-score and y, and it refers to y-loading [19]. PLS is a model that is robust against noise and missing values, and it can be used even with a small amount of data. It also has the benefit of being able to deal with a variety of variable types, such as nominal, ordinal, and continuous [20].

3.2.4.1 Partial Least Squares with Variable Importance of Projection (PLS-VIP) Unlike multiple regression analysis, which selects key variables based on the estimated value of the regression coefficient, PLS regression analysis sets latent variables to configure the model. Therefore, even though it is also called a regression analysis, it must use a different variable selection standard. PLS regression analysis simultaneously considers the degree to which independent variables influence latent variables and the influence of latent variables on dependent variables, and it measures the importance of independent variables through the index show below [21].

$$VIP_j = \sqrt{\frac{k\sum_{a=1}^{a^*}\left[\left(b_a^2 t'_a t_a\right)\left(w_{aj}/\,||w_a||\right)^2\right]}{\sum_{a=1}^{a^*}\left(b_a^2 t'_a t_a\right)}}, \quad j = 1, \ldots, k, \qquad a = 1, \ldots, a^* \tag{7}$$

In Eq. (7) above, j is the original independent variable, a is the latent variable, and a* is the number of latent variables created by PLS. In Eq. (7), w_{aj} is the loading weight of variable j when latent variable a is used. This can be considered variable j's level of contribution to latent variable t_a b_a^2 t'_a t_a consists of the variance that indicates latent variable a and the y-loading value. It can be understood as the level of contribution by latent variable t_a to y [22]. $(w_{aj}/\|\, w_{al}|)^2$ is the importance of variable j in latent variable a. Finally, VIP_j can be considered an index that evaluates the importance of variable j based on the importance of the variance, which describes the latent variables and the independent variables that make up the latent variables.

Table 1 PBGA processes and machines

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
c1		c3	c4	c5	c6	c7	c8	c9	
		d3	d4			d7			
		e3				e7			

4 Case Study

4.1 Setup

A single lot observed by this study goes through 10 processes sequentially. Then, it can be depicted in trace form, which shows the equipment history. The trace has a structure like $x_{11}, x_{12}, \ldots, x_{ij}$. Here, x_{ij} refers to equipment j in process i. It can be considered a path variable that is a 1 if the lot passed through it, and 0 if the lot did not. Because the 10 processes are performed by several pieces of equipment, if all possible combinations of trace types are calculated from Table 1, there are 97,200 results. This number is rather difficult to analyze. To resolve this, it is necessary to sort the traces that are discovered in the processes by type, and 312 different trace types were discovered. In all the lots, at least 1, and at most several dozen, of each trace type were discovered; therefore, it is necessary to define the typical number of faults shown by the traces according to type. As such, this study calculated the typical number of faults to be the average number of all faults that occur when passing through the equipment indicated in a trace. After preprocessing was completed, x_{ij}, which makes up the trace, was defined as the independent variable, and the typical number of faults was defined as the dependent variable for the 312 trace types. From this, the experiment used a PLS regression analysis to describe faults based on 33 pieces of equipment that make up 10 processes.

4.2 PLS-VIP

This stage finds the degree of importance that the equipment, which comprises the trace $x_{11}, x_{12}, \ldots, x_{ij}$, has to the quality variable Y's fault, based on the VIP score value. The number of latent variables in the PLS regression analysis was selected through a five-fold cross validation, which is widely used in estimating prediction error [23], and the VIP score values of the quality variables can be seen in the graph below. Normally, the square of the VIP value's mean is 1, and independent variables larger than 1 are selected as significant variables. However, in accordance with studies that have reported good results when the VIP value was between 0.83 and 1.21, this experiment selected variables with values of 0.83 and above [24].

Figure 4 shows the VIP score values of 33 independent variables for quality variable Y. The x3c variable had the lowest VIP score of 0.1016, and x10b had the highest value at 2.1488. Table 2 shows the fault-responsible machine candidates selected for quality variable Y. Here, Y is the major item among the inspection process fault types.

4.3 Association Rule Analysis Results

Association rules are applied to the fault-responsible machines and machine groups that affect the yield of quality variables in the previous stage. First, in order to apply the association rules, the minimum support (min-sup) and minimum confidence (min-conf) parameters must be set. In this paper, the min-sup and min-conf were set at 0.05, and the experiment was performed. Low level values were selected because, even though they may only represent a single fault in an environment with similar levels of production technology between competing firms (due to the previously mentioned technological changes), they still cause significant production disruption from a company's perspective. Also, even though a low frequency is currently seen, low level values were selected to take into account the possibility of a fault-responsible machine, or machine group, that causes latent faults that exceed the limit values as data accumulates. Figure 5 is a scatter plot that shows the set of rules in which the min-lift value is greater than 1 under the support and confidence conditions that were set.

The horizontal axis in Fig. 5 shows the support values, and the vertical axis on the left side shows the lift values. The shaded area on the right side shows the confidence values. Darker shading in the shaded area means a higher value. In the case of Y, 19 rule sets were discovered. A point worth mentioning is that the support values were low, but most of the rules had high confidence values. If they are examined, 15 out of the 19 rules that were discovered had confidence values of 0.78 or more, and, furthermore, three rules had confidence values of 1. This means that even though they are not frequent rules, these rules lead to a high level of faults. Because of this, these rules must be managed, and this supports the fact that setting low levels of support values is rational. When rule creation is finished, the next step is to use the previously proposed algorithm to find the fault-responsible machines (SF), which independently cause

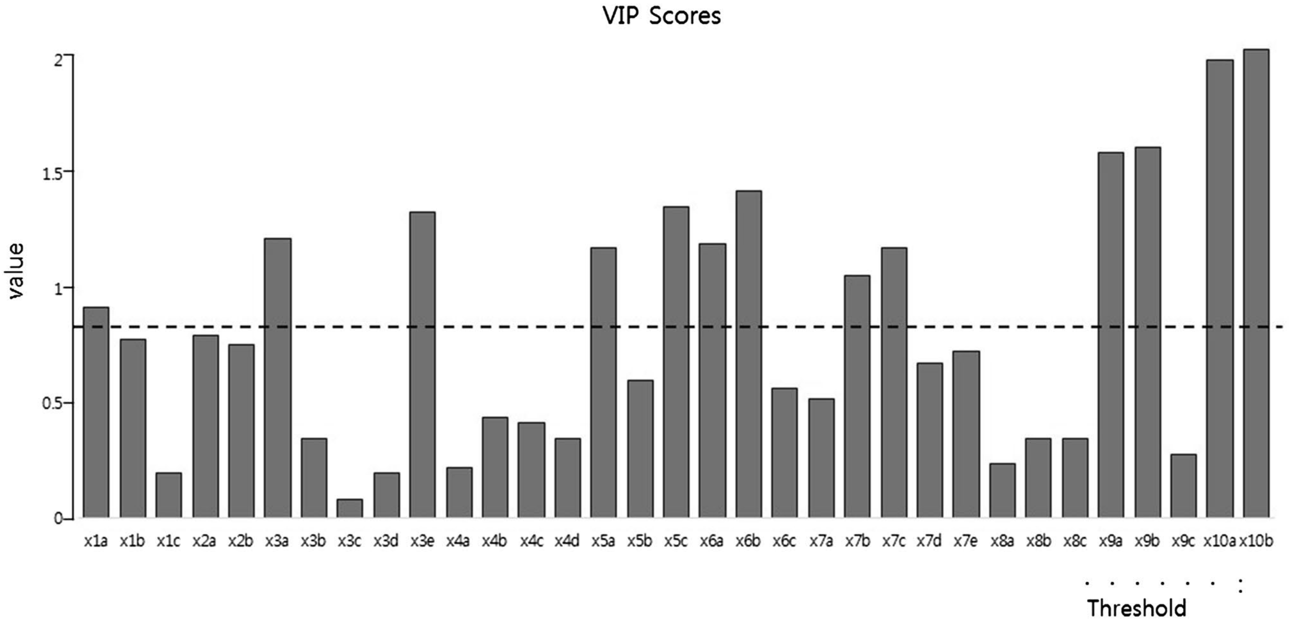

Fig. 4 VIP scores for quality parameter

Table 2 Results from the VIP scores

Quality parameter	Selected machines (independent variable)
Y	x_{1a}, x_{3a}, x_{3e}, x_{5a}, x_{5c}, x_{6a}, x_{6b}, x_{7b}, x_{7c}, x_{9a}, x_{9b}, x_{10a}, x_{10b}

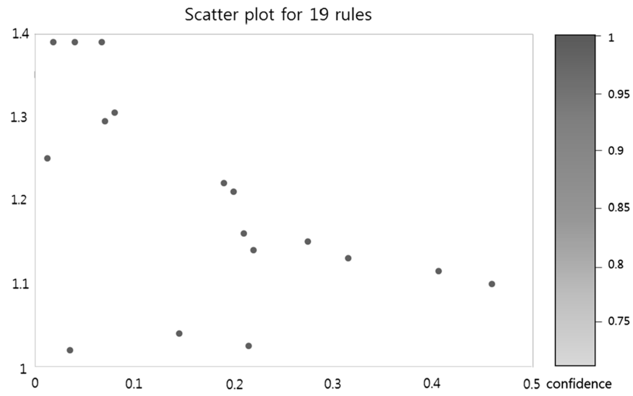

Fig. 5 Scatter plot of rules generated by association rules

faults, and the fault responsible machine groups (CF), in which the downstream causes faults together with the upstream. Figures 6 and 7 show trees composed of upper layer rules and lower layer rules, according to the rule length, in order to find the SF and CF in quality variable Y.

In order to discover the cumulative factors (CF) in the relationship tree, this paper set the minimum cumulative effect threshold (min-cum) at 5%. That is, the cumulative

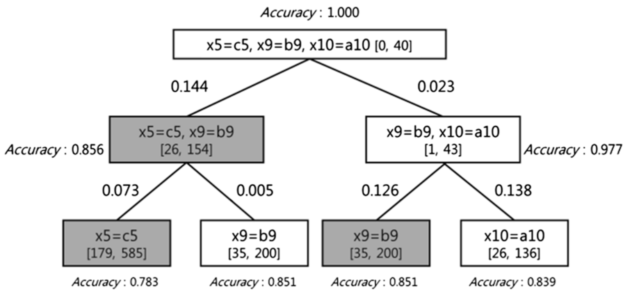

Fig. 6 The rules for quality parameter y relationship tree (rule length = 3)

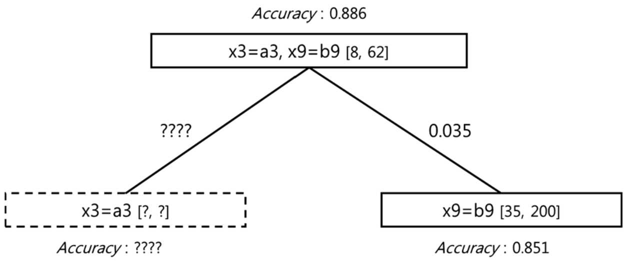

Fig. 7 The rules for quality parameter y relationship tree (rule length = 2)

factors were selected by selecting rules that showed a cumulative effect of 5% or more based on the accuracy before and after the downstream participated in the upstream. Looking at the results in Fig. 6, the shaded rule {x5 = c5, x9 = b9} refers to the upstream of the upper layer rule {x5 = c5, x9 = b9, x10 = a10}. The cumulative effect caused by

Table 3 Cumulative effects on quality parameter *y*

Cumulative factor	Accuracy(%)	Cumulative effect (%)
$(x7 : c7,\ x9 : b9)$	79.4	11.6
$(x3 : e3,\ x5 : c5)$	82.3	5.1
$(x3 : e3,\ x7 : b7)$	84.8	8.3
$(x3 : e3,\ x10 : a10)$	92.0	17.5
$(x9 : b9,\ x10 : a10)$	97.7	14.8
$(x3 : e3,\ x5 : c5,\ x7 : b7)$	87.0	5.7
$(x3 : e3,\ x5 : c5,\ x10 : a10)$	92.3	12.1
$(x5 : c5,\ x9 : b9,\ x10 : a10)$	100	16.8

downstream {x10 = a10} is 16.8% (= 0.144/0.856 × 100%). Because this figure is larger than the minimum cumulative effect threshold (min-cum), the rule {x5 = c5, x9 = b9, x10 = a10} is a cumulative factor. However, there were cases in which the cumulative effects for rules consisting of multiple pieces of equipment were unknowable. In order to identify the cumulative effect of downstream {x9 = b9} for the rule {x3 = a3, x9 = b9} in Fig. 7, the accuracy of upstream {x3 = a3} must be known. However, rule {x3 = a3} is a rule that cannot be discovered from the algorithm; therefore, certain information for finding the cumulative effect of rule {x3 = a3, x9 = b9} cannot be provided. For such cases, a relationship tree was created, as in Fig. 7, which shows undiscoverable rules as dotted lines and unknowable information as question marks. Table 3 below shows the cumulative factors for quality variables, accuracies that indicate cumulative factors, and cumulative effect figures.

In summary, eight cumulative factors exist that cause faults in quality variable Y, and the cumulative effect of these is distributed from 5% to 17%. An accuracy that indicates a cumulative factor is relatively high can be seen, and the cumulative factors discovered in this experiment have an average accuracy of 89%. When considering the study's goal of managing faults caused by cumulative factors, rather than just single factors, in a process, the above results are notable. Specifically, the cumulative factor {x9 = b9, x10 = a10} in Table 3 shows that faults are found in 97.7% of all lots that go through equipment b9 in process x9 and then go through equipment a10 in process x10, and it can be known that the cumulative factor shows a 14.8% higher performance than the fault detection performance of single factor {x9 = b9}. From the above results, a new fact is discovered: the upstream of cumulative factors {x3 = e3, x5 = c5, x7 = b7} and {x3 = e3, x5 = c5, x10 = a10}, which have lengths of 3, includes the cumulative factor {x3 = e3, x5 = c5}, which has a length of 2. That is, one cumulative factor can be composed of a different cumulative factor's partial factor. The results of these analyses can be considered a result that clearly shows the cumulative effects caused by the downstream in the manufacturing process.

5 Conclusion

The goal of this study is to ensure the competitiveness of businesses by improving manufacturing line yields and productivity through an analysis of the processes and equipment that affect micro-manufacturing process yields. To do this, an analysis technique was proposed that analyzes manufacturing-line fault data and equipment parameters. It is used to determine which processes affect yields and to examine the manufacturing equipment in each process. This information is used to find the fault-responsible machines that have the most influence on faults. The experiment results confirmed that the factors that cause faults are not just single process variables but also cumulative factors, in which the downstream causes faults together with the upstream. Specifically, the cumulative factor {x9 = b9, x10 = a10} shows that faults are found in 97.7% of all lots that go through equipment b9 in process x9 and then go through equipment a10 in process x10, and it can be known that the cumulative factor shows a 14.8% higher performance than the fault detection performance of single factor {x9 = b9}.

The use of these methodologies is expected to greatly contribute to manufacturing companies finding causes for product defect, increasing yields and improving quality.

The processes and equipment that are classified as important factors must be managed intensively by collecting opinions from site engineers. Also, this kind of big data analysis of manufacturing is only possible if an environment is constructed in which lot histories and process and equipment data for important processes can be gathered at the manufacturing site and feedback can be provided. That is, priority among smart factory functions must be given to an environment in which the site's equipment is connected and the equipment information can be received and controlled. Therefore, in this study, construction of a PBGA-line S-EES for big data analysis and an architecture for data gathering were proposed together.

In the future, we will find a critical parameter that affects yield through an analysis in conjunction with the equipment parameter, and we will conduct further research on methodology that automatically controls the equipment parameter in connection with the FDC.

References

1. Spanos, C. J. 1992. Statistical process control in semiconductor manufacturing. In *Proceedings of the IEEE*, vol. 80, No. 6, (pp. 819–830).
2. Lee, J. Y., Yoon, J. S., & Kim, B. H. (2017). A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. *International Journal of Precision Engineering and Manufacturing, 18*(10), 1353–1361.
3. Park, J. M. (2015). Technology and Issue on Embodiment of Smart Factory in Small-Medium Manufacturing Business. *The Journal of Korean Institute of Communications and Information Sciences, 40*(12), 2491–2502.
4. Montgomery, D. C. (2009). *Introduction to statistical quality control* (pp. 288–506). Hoboken: Wiley.
5. Cherry, G. A., & Qin, S. J. (2006). Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis. *IEEE Transactions on Semiconductor Manufacturing, 19*(2), 159–172.
6. Yan, L. (2006). A PCA-based PCM data analyzing method for diagnosing process failures. *IEEE Transactions on Semiconductor Manufacturing, 19*(4), 404–410.
7. Spitzlsperger, G., Schmidt, C., Ernst, G., Strasser, H., & Speil, M. (2005). Fault detection for a via etch process using adaptive multivariate methods. *IEEE Transactions on Semiconductor Manufacturing, 18*(4), 528–533.
8. Chen, F. L., & Liu, S. F. (2000). A neural-network approach to recognize defect spatial pattern in semiconductor fabrication. *IEEE Transactions on Semiconductor Manufacturing, 13*(3), 366–373.
9. Chien, C. F., Wang, W. C., & Cheng, J. C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications, 33*(1), 192–198.
10. Sarmiento, T., Hong, S. J. & May, G. S. (2005). Fault detection in reactive ion etching systems using one-class support vector machines. In *Conference and workshop on advanced semiconductor manufacturing*, *IEEE/SEMI*, (pp. 139–142).
11. He, Q. P., & Wang, J. (2007). Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing, 20*(4), 345–354.
12. Ma, M. D., Wong, D. H., Jang, S. S., & Tseng, S. T. (2010). Fault detection based on statistical multivariate analysis and microarray visualization. *IEEE Transactions on Industrial Informatics, 6*(1), 18–24.
13. Special Report. (2017). Smart factory. Dong-A business review, no. 227, (pp. 67–68).
14. International SEMATECH. (2002). Equipment engineering capabilities (EEC) guidelines, Version 2.5.
15. Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques* (2nd ed.). Burlington: Morgan Kaufmann.
16. Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of 20th international conference on very large data bases* (VLDB), (pp. 487–499).
17. Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (2013). *Applied regression analysis and other multivariable methods*. Duxbury: Duxbury Press.
18. Myers, R. H. (1986). *Classical and modern regression with applications*. Duxbury: Duxbury Press.
19. Boulesteix, A. L., & Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics, 8*(1), 32–44.
20. Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics, 3*(4), 283–297.
21. Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). *Multi-and megavariate data analysis: Principles and applications*. Umeå: Umetrics Academy.
22. Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A partial least squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology, 6*(1), 1–12.
23. Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning* (pp. 214–217). New York: Springer.
24. Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*(1), 103–112.

Hyun Sik Sim Professor in the Department of Industrial & Management Engineering at Kyonggi University. His research interest is Smart Factory and Smart Manufacturing System, Manufacturing Big Data Analysis.