# Retrospective Analysis for Mining the Causes in Manufacturing Processes

Kwok-Pan PANG
*Gippsland School of IT,*
*Monash University*
*Email: benpang@netspace.net.au*

Shawkat ALI
*School of Information Systems*
*Central Queensland University*
*Email: Shawkat.Ali@ieee.org*

## Abstract

*There has been a considerable growth in the use of Statistical Process Control (SPC) for improving the quality in business, industries, or software development since the last decade. However, the processes are growing much more complex, and there is a tremendous increase of data size owning to the use of automated record machine. The conventional SPC tools become less effective in analyzing and identifying the cause of the process failures. This paper extends the idea of the Modified Centered CUSUMS, and proposes a new data selection procedure so that the associative discovery technique can be used in retrospective SPC analysis. Through our approach, the common data mining method (i.e. associative discovery) can be used to find the hidden knowledge from the data, and identify the causes of the process failure or success for the quality improvement. Besides, the hidden information that we extracted from the data can be used as supplement for the cause and effect diagram in the on-line process control.*

## 1. Introduction

The use of Statistics in quality management and quality improvement has a long history. Since the study of Deming (1986, 1993), the Statistical Process Control (SPC) has formed the basis for continuous quality improvement that is the key component of Total Quality Management (TQM). In TQM, all activities are considered as processes. The performance of the process may have some fluctuation or variation over time. TQM will use the statistical techniques and relevant tools to identify the cause of the problem and to reduce the variation. Statistical Process Control (SPC) has been developed as one of the important components of the quality control activities for detecting and identifying the process failure in manufacturing industry, service organization or software development industry.

In this paper, we mainly concentrate on identifying the persistent special causes for the improvement in the future. From the statistical point of view, the persistent special causes are a result of the structural change. The analyzed object changes from in-control distribution to out-of-control distribution. The persistent special causes will remain in out-of-control states until corrective actions are taken. As depicted in Figure 1, if the observed variability of the attributes of a process is within the range of variability from common causes, the process is said to be under statistical control. The practitioner of SPC tracks the variability of the process to be controlled. When that variability exceeds the range to be expected from the common causes, one then identifies assignable causes, and takes the corrective action and removes the persistent special causes.

In common practice, cause and effect diagram is used to help to identify the persistent special cause. The SPC provides the path for continuous improvement through learning from the mistakes.

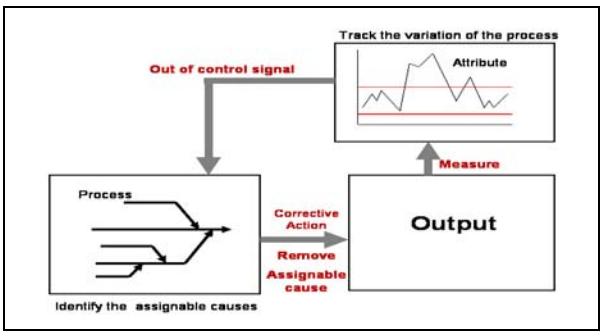



**Figure 1**: Statistical Process Control

### Cause and Effect Diagram

In the traditional approach of off-line SPC analysis, the company may identify the cause of the process failure with the aid of the cause and effects diagrams. The cause-and-effect diagram, which is also known as "Ishikawa Diagram" or "Fishbone Diagram", is designed to detect all possible contributing factors or all possible causes of effect. As indicated in Figure 2, the head of central 'spine' elicits the effect and the causes will be shown at the 'rib' ends. The diagram suggests that we should first work with the principal factors or causes and then reduce to sub-cause level, and even sub-sub-causes if needed. The process continues until all possible causes are extracted. (Smith 1998). However, it has become more difficult for the traditional approach to find out the cause of process failure or improvement because of the increasing complexity of the production process and product's composition. Besides, due to the new measurement device and modern database capabilities, a large amount of production data is yield. There is an increasing demand of using Data Mining to extract the relevant information for quality improvement. Data mining has

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

been proved effective and efficient in analyzing large data sets and it has already been applied to many different areas, such as business, economic or ecology etc. (Giudici 2003, Milne, Drummond and Renoux 1998 and Perner 2002). It is convinced that the great potential of data mining can be integrated effectively with SPC.

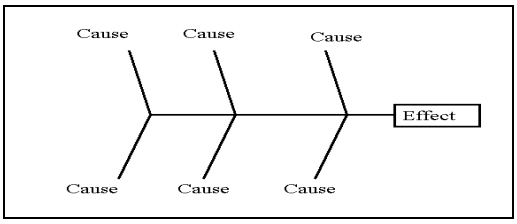



**Figure 2:** Basic form of cause and effect diagram

**Limitations of the conventional SPC**

1. The improvement is mainly made through minimizing the mistake. SPC normally ignores the data with the success experience, which may contain some hidden knowledge for further process improvement. e.g. the process may give a very good performance in a particular period. Mining the hidden knowledge about the good performance in the data will be invaluable for further process improvement.

2. The cause and effect diagram is constructed based on the previous experiences and knowledge. With the growing complexities of the process, and the tremendous increase of data size data size as well as the increasing number of input variables, it becomes more difficult to find out all possible causes, and construct an effective cause and effect diagram.

An increasing number and complexity of historical production data increase the difficulties to apply SPC for quality improvement. Lam (1996) conducted the survey and showed that the lack of ability of quality improvement tools to solve existing quality problem was cited by respondent as the major barrier to their use. As Guter (1998) mentioned, "The reality of modern production and service processes has simply transcended the relevant and utility of this honored but ancient tool." The manufacturing environment in which SPC is used is changing rapidly. In view of the great need to improve the SPC tools to cope with the changing manufacturing environment, we are motivated to propose a new approach to mine the cause for the process improvement.

In order to extract effectively and exhaustively the information hidden in the data, we propose a preprocessing process that convert the historical process data into the format that Data Mining technique (i.e. Association discovery) can be used to identify the cause.

Conventionally, the retrospective analysis for quality improvement normally focuses on a portion of most recent segment data for finding out the cause of the process failure. However, it is not appropriate in some situations. Let us use Figure 3 to explain. The conventional retrospective analysis for the quality improvement normally focuses on Segment $A_6$ only. The out of the control Segment $A_2$ is easily ignored as the people may consider the problem has been solved. It may not be easy to identify the cause if the size of segment $A_6$ is small, and the same problem may appear intermittently. For example, the supplier provides two synthetic resins to the company, everything of these synthetic resins are the same; the only difference is that they are produced from the different supplier's branches. If one of supplier's branches makes a mistake in particular production lot of the synthetic resins, and the company does not notice the problem, the problem will re-appear after a certain period. In this situation, it will be helpful for mining the cause of the process failure if both segments $A_2$ and $A_6$ are selected, as some hidden information is available in Segment $A_2$

In addition, the ignorance of the success experience in the Segment $A_4$ will make us lose some important information for further improvement.

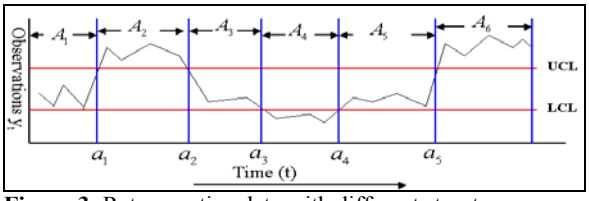



**Figure 3**: Retrospective data with different structures

In this paper, we extend the idea of the Modified Centered CUSUMS (Pang and Ting 2004) for estimating all change points of the process of the retrospective data, so the process data is separated into multiple segments. Take Figure 3 as an example, we try to estimate the change points $a_1$, $a_2$,.. $a_4$ and $a_5$, and find the segments $A_1$, $A_2$,.. $A_5$ and $A_6$. We try to find out the factors of the process failure from the segments with the abnormal performance (i.e. Segments $A_2$ and $A_6$), and the factor of achieving success in the segment $A_4$.

## 2. Description of the process data

The main purpose of this paper is to propose the new pre-processing process that can convert the original process data to the new data set with the format that Data Ming technique can be applied to find out the hidden reason of better performance or poor performance (i.e. what input variables with what parameter setting will affect the performance of the process.)

International Conference on Computational Intelligence for Modelling
Control and Automation,and International Conference on
Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

We assume the process are presented by $\{(U_t, y_t), t = 1,2,...,n\}$ where $U_t=(u_{1t},u_{2t},...,u_{kt})$ is a vector of k input variables at time t, $u_1$, $u_2$, ... $u_k$ are the numerical or categorical input variable. $y_t$ is the output numerical variable at time t. $y_t$ is assumed to be constant and stable over time if nothing is changed in the process. We suppose observation of $y_t$ will be used for tracking the performance of the process. Let us use the sample of the injection machine production to illustrate the process. Suppose that $y_t$ is the number of qualified product produced by injection machine per hour (QPR), GPR provides the index indicating the performance of the injection machine. The value of GPR can be affected by the set of the input variables that may include the type of synthetic resin, production lot number of the resin, types of dye, production lot no. of dye, name of vendor of that dye, name of machine operator, or the machine settings (e.g. temperature, pressure or cooling time setting) etc. In many situations, a factory normally has lots of alternative sources of material (e.g. several materials may have the identical specification, but from different manufacturers). When there is an increase of input variables and data size, the conventional SPC approach may not be capable to find out the possible cause.

## 3. The proposed approach

To improve the conventional SPC in finding the cause for process improvement, we propose a new preprocessing procedure that consists of 3 steps as described in Figure 4: (i) Split a whole process segment data into multiple process sub-segment data , (ii) group the segments with the optimum performance and (iii) transform as described in Figure 3, After completing all steps, two different new data sets in the format that data mining technique (i.e. association Discovery) can be used will be generated. One is used for mining the causes of the mean change of the process, another one is for mining the cause of the variance change of the process. Through our approach, the association discovery technique can be used to identify the possible cause.

Whole set of the historical process data is assumed to be A, $A=\{(U_t, y_t), t=1,2,3,...,n\}$. The descriptions of the process variable are provided in last section.

**Step 1: Split a whole process segment data into multiple segments**

The whole process segment data A will be split into multiple process sub-segment data $\{\{A_1\},\{A_2\}, \{A_3\},\ldots,\{A_m\}\}$ according to the structure of $y_t$, where m is the number of change points in the original process segment data. $y_t$ is expected to have the distribution $N(\mu_i,\sigma_i)$ in the $i^{th}$ segment. We apply the Modified Centered CUSUMS with the binary segmentation to locate all change points, and then split the whole process data into multiple process sub-segment data. The details about Modified Centered CUSUMS for change point detection will be described in the next section. Binary Segmentation is the algorithms for estimating the multiple break points. If a change is detected, then the data is divided at the most likely location for a single change, and the change-point procedure is applied to each new group of data. This process is repeated until no group shows evidence of a change.

**Modified Centered CUSUMS**

The process sub-segments will be generated based on the structure of $y_t$. Modified Centered CUSUMS is used to identify the change points of data $\{y_t, t = 1,2,...,n\}$

Each sample is interpreted as:

$y_t = \mu_t + \sigma_t \varepsilon_t$ for t = 0, 1,2,3,…,n

where $\varepsilon_t$ has the standard normal distribution.

We use the following notations to denote the observation matrices $Y_{m,g}$ and $X_{m,g}$ which consist of (g-m+1) samples in the process starting from the $m^{th}$ sample to the $g^{th}$ sample.

$$Y\text{m,g} = \begin{bmatrix} y_m \\ y_{m+1} \\ \ldots \\ \ldots \\ y_g \end{bmatrix}, \quad E_{m,g} = \begin{bmatrix} \sigma_m \varepsilon_m \\ \sigma_{m+1} \varepsilon_{m+1} \\ \ldots \\ \ldots \\ \sigma_g \varepsilon_g \end{bmatrix} \quad \text{and}$$

$$X_{m,g} = \begin{bmatrix} x_m \\ x_{m+1} \\ . \\ . \\ x_g \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ .. \\ .. \\ 1 \end{bmatrix}$$

where m < n, and $x_m, x_{m+1},..., x_n$ are the row vectors. If the process parameters $\sigma$ and $\mu$ are constant in all samples, the relationship between the variable expressed in Equation (1) can be rewritten as:

$Y_{0,n} = X_{0,n}\beta + E_{0,n}$, where $\beta = [\mu]$, and

$\sigma_0 = \sigma_1 = \sigma_2 = ...\sigma_n$

With the above notation of the matrices, we can construct Modified Centered CUSUMS for detecting all change points.

The standardized recursive prediction residual $w_r$ can be obtained by:

$w_r = (y_r - x_r\hat{\beta}_{r-1})/d_r = (y_r - \mu_{r-1})/d_r$, $r = 1,2\ldots,n\text{-}1,n$

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

where

$\hat{\beta}_r = (X_{0,r}^{/} X_{0,r})^{-1} X_{0,r}^{/} Y_{0,r}$,

$d_r = 1 + x_r (X_{0,r}^{/} X_{0,r})^{-1} x_r^{/}$, $r = 0, 1, 2, 3, \ldots, n$;

Modified Centered CUSUMS will be:

$$M(r, n_1) = \sqrt{\frac{n_1 + 1}{2}} \left| \frac{\sum_{i=1}^{r} w_i^2}{\sum_{i=1}^{n_1} w_i^2} - \frac{r}{n_1} \right| \quad , r = 1,2,3,..,n_1, \; n_1 \le n$$

where $n_1$ is the selected sample size that the maximum $T_{n1}$ is generated.

$T_{n_1} = \max_{0<r\le n_1} M(r, n_1)$ $\quad , r = 1,2,...,n_1,$ where $n_1 \le n$

The estimated change point $(\hat{r})$ will be:

$\hat{r} = \arg\max_{0<r\le n_1} M(r, n_1)$ $\quad$ , where $0 \le n_1 \le n$

When the maximum value $T_{n1}$ is over the control limit, the process will be deemed to be 'Out-of-Control'. The critical values (i.e. control limits) of $T_{n1}$ for the specified n1 are tabulated in Inclan and Tiao (1994).

**Step 2: Group the segments with the optimum performance together**

In this step, we will generate six different groups of segments, three groups are generated for analyzing the change of the process mean, and the other three groups are for the change of the process variance.

We assume m sub-segments are generated in step 1, these sub-segment are $A_1$, $A_2$, $A_3$,…,$A_m$, the distribution of $y_t$ of $i^{th}$ segment is assumed to be $N(\mu_i, \sigma_i^2)$. We calculate the mean and variance for each segment from $A_1$ to $A_m$, and find the segments which have the largest mean, the smallest mean, the largest variance and the smallest variance. We suppose that the maximum process sub-segment mean $\mu_p$ is found in $A_p$, the minimum process sub-segment mean $\mu_q$ is found in $A_q$, the maximum process sub-segment variance $\sigma_R$ is found in $A_R$, and the minimum process sub-segment variance $\sigma_g$ is found in $A_g$.

After the segments with the optimum performance are identified, we will group all segments which structure is the same as the segment with the optimum performance. This step composes of two independent procedures, and these procedures are almost identical, the only difference is: (a) One procedure is to group the segments for analysis of process mean, (b) another procedure is to group the segments for analysis of process variance.

**(a) group the segments for analysis of the change of the process mean**

Three subsets $G_1$, $G_2$, and $G_3$ will be generated. Whole data set will be $G$={$G_1$, $G_2$, $G_3$}

$G_1$: A group of segments in which $y_t$ have the same structure as the segment with the maximum process mean (i.e $G_1 = \{(U_t, y_t) \bullet mean(y_t) = \mu_p\}$);

$G_2$: A group of segments in which $y_t$ have the same structure as the segment with the minimum process mean (i.e. $G_2 = \{(U_t, y_t) \bullet mean(y_t) = \mu_q\}$);

$G_3$: A group formed by the rest of segments (i.e. $\{(U_t, y_t) \bullet (U_t, y_t) \notin (G_1 \cup G_2)\}$).

**(b) group the segments for analysis of the change of the process variance**

Three subsets $Q_1$, $Q_2$, and $Q_3$ will be generated. Whole data set will be $Q$={$Q_1$, $Q_2$, $Q_3$}

$Q_1$: A group of segments which $y_t$ have the same structure as the segment with the maximum process variance (i.e. $Q_1 = \{(U_t, y_t) \bullet \text{var}(y_t) = \sigma_R^2\}$).

$Q_2$: A group of segment {($U_t$, $y_t$)} which $y_t$ have the same structure as the segment with the minimum process variance (i.e. $Q_2 = \{(U_t, y_t) \bullet \text{var}(y_t) = \sigma_g^2\}$).

$Q_3$: A group formed by the rest of segments (i.e. $Q_3 = \{(U_t, y_t) \bullet (U_t, y_t) \notin (Q_1 \cup Q_2)\}$).

In order to group the segments with the same structure, we use Chow test to detect whether the specified two segments have the same structure.

Chow test = $\dfrac{(SSM - SSM_1 - SSM_2)}{(SSM_1 + SSM_2)/(e-2)}$

Chow test has $F_{1, n-2}$ distribution

Where

SSM = the sum of square deviation about the process mean in which process mean and process variance are assumed to be the same in two specified segments (i.e. Segment 1 and Segment 2);

$SSM_1$ = the sum of square deviation about the process mean estimated in Segment 1;

$SSM_2$ = the sum of square deviation about the process mean estimated in Segment 2;

e is total number of observations.

Given the sub-segment data sets $A_s$ and $A_v$, let us define Chow test as a function as follows:

$$\text{Chow}(A_s, A_v) = \begin{cases} 1 \text{ if } A_s \text{ and } A_v \text{ are found to have the different structure of } y_t; \\ 0 \text{ if } A_s \text{ and } A_v \text{ are found to have the same structure of } y_t; \end{cases}$$

If $A_s$ and $A_v$ are two data sets, $A_s \cup A_v$ denotes the union of sets $A_s$ and $A_v$.

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

The Pseudo Code for $G_1$, $G_2$ and $G_3$ is described as:

```
G1=Ap
G2=Aq
G3=[]
i = 1
Do while i <= m and  i!=p and i!=q
        If Chow(Ap, Ai) = 0
                G1= G1∪Ai
            elseif Chow(Aq, Ai) = 0
                G2= G2∪Ai
            else
                G3= G3∪Ai
        endif
        i = i +1
enddo
```

The Pseudo Code for $Q_1$, $Q_2$ and $Q_3$ is described as:

```
Q1=AR
Q2=Ag
Q3=[]
i = 1
Do while i <= m and  i!=R and i!=g
        If Chow(AR, Ai) = 0
                Q1= Q1∪Ai
            elseif Chow(Ag, Ai) = 0
                Q2= Q2∪Ai
            else
                Q3= Q3∪Ai
        endif
        i = i +1
enddo
```

**Step 3: Transform the data into the format for using associative discovery technique**

This step aims to transform the format of the data subset generated in step 2 to the format that can be used by associative discovery technique.

Two final data sets $G'$ and $Q'$ will be generated. $G'$ will be used for mining the causes or conditions for improving the process mean. $Q'$ will be used for mining the causes or conditions for improving the process variance.

The G' consists of three subsets, $G' = \{G_1', G_2', G_3'\}$, $G_1', G_2'$ and $G_3'$ are transformed from the subsets $G_1$, $G_2$ and $G_3$ generated in Step 2.

$$G_1' = \{(U_t, 'good') \bullet U_t, \in domG_1\}$$
$$G_2' = \{(U_t, 'poor') \bullet U_t \in domG_2\}$$
$$G_3' = \{(U_t, 'normal') \bullet U_t \in domG_3\}$$

The $Q'$ consists of three subsets, $Q' = \{Q_1', Q_2', Q_3'\}$, $Q_1', Q_2'$ and $Q_3'$ are transformed from the subsets $Q_1, Q_2$ and $Q_3$ generated in Step 2.

$$Q_1' = \{(U_t, 'good') \bullet U_t, \in domQ_1\}$$
$$Q_2' = \{(U_t, 'poor') \bullet U_t \in domQ_2\}$$
$$Q_3' = \{(U_t, 'normal') \bullet U_t \in domQ_3\}$$

## Association Discovery

After obtaining the transformed data sets $G'$ and $Q'$, we will apply the simple associative discovery technique for the data set $G'$ and $Q'$ separately. $G'$ will be used for identifying the cause of the change of the process mean, and $Q'$ will be used to identify the cause of the change of the process variance. Association Discovery is to find items what imply the presence of other items in the same record. The discovery process produces association rules in the form: "If C inputs are used then R will happen."

In this paper, we suggest using Leverage to indicate the validity and importance of the rule.

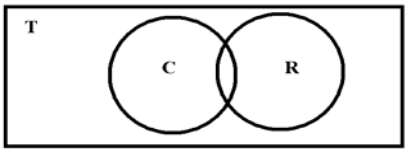



**Figure 4**:Venn Diagram

Let us assume we have an association rule indicated as LHS $\rightarrow$ RHS

|T| is the total number of records in the database.

|C| is the number of records covered by the LHS.

|R| is the number of record covered by RHS.

$|C \cap R|$ is the number of records covered by both the LHS and RHS, indicated by the overlapping area in Figure 4.

The measures Leverage can be expressed as follows.

$$Leverage = \frac{|C \cap R|}{|T|} - \frac{|C|}{|T|} \times \frac{|R|}{|T|}$$

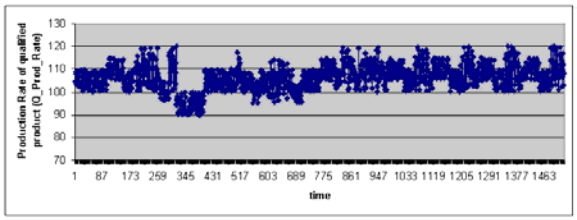
## 3. Example

Let us consider the following data set from the injection machine. The data set contains the attributes: Production_date (date of production), operator_no (no. of workers operating the machine), shift (period of operating time), PN1 (part number of synthetic resins), Prod_Lot1 (production lot of the synthetic resins), PN2 (part no. of dye), Prod_Lot2 (production lot of the dye), QPR (production rate of the qualified product). QPR is the attributes which observations can be used to present the performance of the injection machine. The attributes other than QPR will be used as input attributes.

This injection machine produces a single product using several alterative synthetic resins and alternative dyes. The quality of the product from the machine is unstable. The supplier of the injection machine checks the machine and reports that the problem is not related to the machine, so further investigation on the cause of process failure is needed.

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

From the data set, we plot the data of QPR as the following:




We split the data of QPR into multiple segments using Modified Centered CUSUMS according to its distribution, then we group the relevant segments with the transformation and form $G'$ and $Q'$. After that, we generate the following rules using measure Leverage.

R1: operator_no=00124 and shift=night → low process mean;
R2: PN1=R-0024 and PN2=D0054 → high process mean;
R3: Prod_Lot1=950415L → high process variance;
R4: Operator_no=00321 → low process variance.

After the rules are generated, we can investigate and verify the rule, e.g. we can see why the production lot 950415L increases the process variance.

## 4. Discussion and Conclusion

This paper contributes to propose a new preprocessing process which enables the association discovery technique to find out the condition for improving and the causes of deteriorating the process performance.

Our approach is not only designed as the supplement of the cause and effect diagram, but is also deployed to update the cause and effect diagram for the on-line and off-line process analysis.

The process can be improved through our proposed approach in two ways:

(1) Learning from the mistake;
(2) Learning from data with success experience.

The proposed idea enriches the information in the cause and effect diagram as shown in Figure 5. The proposed preprocess procedure with associative discovery technique provides an alternative method for mining the condition for improving the process performance. Besides, the knowledge generated from our approach can update the knowledge shown in the cause and effect diagram. The update knowledge increases the effectiveness of the cause and effect diagram for the future on-line or off-line process analysis.

The Product traceability plays the significant role for the success of our approach. Many causes of scrap and rework originate from poor raw material quality or the problems in an earlier process step. To successfully identify the root causes of the problems, it is often necessary to trace the product back to an earlier process. If we fail to fulfill this condition, hidden knowledge will not be found effectively.

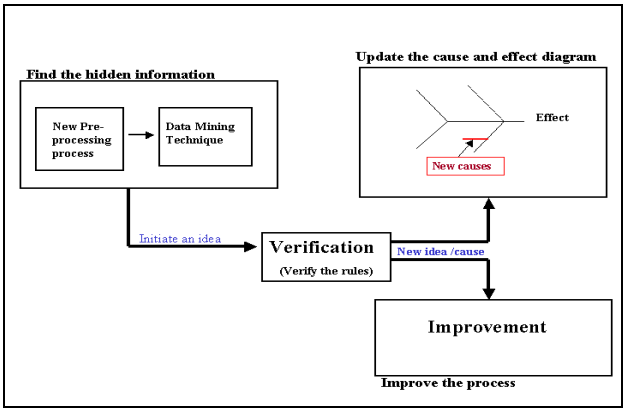



**Figure 5:** Updating the cause and effect diagram

## Reference


[1] Chow, G. (1960), "Test of equality between sets of coefficients in two linear regressions". Econometrica, Vol.28, No. 3, 591-605

[2] Deming, W.E. (1986), *Out of the Crisis*, Cambridge, MA: MIT Press.

[3] Deming, W.E. (1993), *The New Economics for Industry, Government, Education*, Cambridge, MA: MIT Press.

[4] Giudici P. (2003), *Applied Data Mining: Statistical Methods for Business and Industry*, Wiley: New York (in press).

[5] Guter, Bert (1998), " Column: Statistics Corner: Farwell Fusillade", Quality Progress, Vol. 31, No. 4, 111-114

[6] Inclan, C and Tao, G.C (1994) "Use of Cummulative Sums of Squares for Retrospective Detection of Changes of Variance", Journal of the American Statistical Association, Vol. 89, no. 427, 913-923.

[7] Lam, Simon(1996), "Applications of quality improvement tools in Hong Kong: An empirical analysis", Total Quality Management, Vol. 7, no . 6, 675-680

[8] Milne R, M. Drummond, P. Renoux (1998) "Predicting Paper making defect on-line using data mining", *Knowledge-Based Systems*, 11, 331-338.

[9] Pang, K.P. and Ting, K.M.(2004), "Improving the Centered CUSUMS Statistics for Structural Break Detection in Time Series", *Lecturer Notes in computer Science*, 3339, 402-413.

[10] Perner P. (2002), *Advance in Data Mining Applications in E-commerce, Medicine, and Knowledge Management*. Springer: Heidelberg.