INTERNATIONAL JOURNAL OF BUSINESS, 9(4), 2004 ISSN: 1083-4346

CASH: A New SPC Scheme Based on Cluster Analysis in Excess Variety

Bin Nie[a] and Er Shi Qi[b]
[a]*Department of Industrial Engineering, University of Tianjin, China*
niewenwu@hotmail.com
[b]*Department of Industrial Engineering, University of Tianjin, China*
esqi@tju.edu.cn

ABSTRACT

Precise control of process has become increasingly important in today's high-tech industries owing to the shrinking feature sizes and complex process technologies. In particular, there are excess varieties involved in the process. In this paper, we proposed a methodology called CASH that is a new SPC scheme based on cluster analysis and Shewhart control charts to deal with excess variety process control. This CASH scheme is implemented in a LED factory and the result is satisfactory.

JEL: G14, G15

Keywords: Statistical process control (SPC); Cluster analysis; Shewhart control chart; Analysis of variance (ANOVA); Semiconductor manufacturing

I. INTRODUCTION

Various product series is one of typical features of modern LED industry. A great deal of varieties are developed and produced in order to meet various demands from users. In a factory we investigated, every facility produces more than 20 different products. Even in one duty the products of different varieties are shifted several times. Because products are shifted frequently, process parameters are often regulated. As a result, the data collected from process are non-stationary in such manufacturing environment. As we known, Shewhart control chart is based on the supposition that the process is in stationary situation (Montgomery, 1991). So traditional SPC method could not be used directly to monitor the process. We propose a new methodology based on cluster analysis and analysis of variance (ANOVA) to solve this problem.

Seldom study was performed on process control of excess variety manufacturing environment. The object of the method is to make effective process control possible. This paper is organized as follows. In section II, the manufacturing process used in this study is described and problem of process control is also introduced. Section III provides detailed description of procedures of CASH methodology. In Section Ⅳ we apply the method in a real process and describe principle and operation method in detail. In the last section, a simple evaluation is made and we present further investigate way.

II. PROCESS DISCRIPTION

We focus on a main process named auto bonding (A/B) and make some progress of process monitoring on this procedure. Ball shearing strength is a critical quality character of the A/B. When a batch is completed, some samples are collected and test the ball shearing strength of them. The test results called Ball Shearing Testing (BST) data are used for making process monitoring.

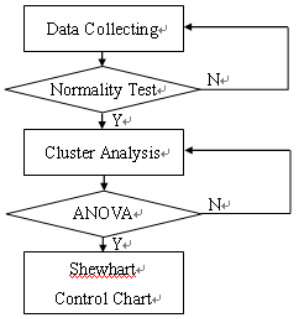
As usual, quality control people test all samples and a mean of a batch is calculated. Then they draw this point on a Shewhart control chart, and make decision on it. The control limit of control chart is calculated based on all test result of the last month.

In fact, systemic shift of process often occur accompany with regulation of variety. The reason is that different products have different process parameters. When such regulation occurs, the process should shifts to another situation. So the fundamental statistical hypothesis of Shewhart control chart is not appropriate any more. Some analysis is needed before apply Shewhart control chart to monitor process.

III. DESIGN OF CASH SCHEME

The object of CASH is to reduce varieties dimensions to a reasonable level and make SPC possible. This study proposed a new scheme to categorize hundreds of varieties in several clusters and then use Shewhart control chart to monitor the process, as shown in Figure 1.

Figure 1
Sequential CASH approach for quality control



The testing data are raw data of every product from calibration. According to prior knowledge to the process, if the process is in control, the BST from one variety is normality distribution. Before we construction control chart, we do not know whether the data is collected from a stationary manufacturing process exactly. In other word, we could not ensure that the data distribution is normal. So a normality test step is implemented after getting testing data. The object is to remove data from non-stationary process and make sure all data preserved are available. We arrange every variety testing data from different time together. Through normality testing some abnormal point are removed until the whole data of one variety pass the testing. If data is not enough, some adjustment of process maybe necessary and some new data are added into the data set. When each group of BST data of different varieties pass normality test, we will begin to implement the next step.

Cluster analysis is a powerful tool in data mining. Intuitively, the clustering problem can be described as follows: Let W be a set of n data points in a multi-dimensional space. Find a partition of W into classes such that the points within each class are similar to each other. The clustering problem has been studied in machine learning (Smyth, P., 1999), databases (Guha et al., 1998), and statistics (Brito et al., 1997) from various perspectives and with various methods.

The cluster result is a hiberarchy, which divided by similarity of different cluster result. In the procedure of cluster analysis, we need to put testing data of the same variety together firstly. Then a mean of every variety is calculated. In the end we use an algorithm to cluster all varieties in a hierarchy by mean of different varieties. Now we know how to divide varieties into several similar clusters, and the characteristic of each group could be regarded as the representation of all varieties in this cluster.

Because different similarity has different cluster result, we do not know which cluster result on a similarity is appropriate. So a hypothesis test is needed to test whether varieties in the same cluster has the same mean. Analysis of variance (ANOVA) is a good method to do it. We start from the top level of similarity that has the smallest cluster number. If all clusters pass ANOVA, procedure is ended. Otherwise

rising similarity to the next cluster result level, and make ANOVA again. We will not stop until every cluster pass the hypothesis test. As a result, the final cluster result tells us each variety in the same cluster come from the same population and has the same mean. If we suppose variance is constant, we can construct a serial of standard Shewhart control charts base on known mean and variance of each cluster. Then all statistical process control methods can be used to monitor process.

There are many key technologies in CASH methodology named normality test, cluster analysis, and ANOVA. All of them could be performed in Minitab software. In the next section we will introduce a real application and describe details of every procedure.

IV. APPLICATION STUDY

We apply CASH in a LED company to test its effect. The manufacturing process is described in section Ⅱ. Each procedure of CASH is described as follows:

A. Data Collection

BST data are collected at the end of A/B process. The object is to find systemic shift of this process through analysis of testing data and keep a high yield and reliability. During one time BST, one piece of PCB is chosen in each lot. Thus, BST data are labeled by an index $i = 1,\ldots,I$ according to their varieties. There are j samples are picked in each PCB. and k facilities are sampled in a A/B process. Let X_{ijk} be a BST measurement taken from the kth facility and it is the jth sample in the ith variety, where $k = 1,\ldots,K$, and $j = 1,\ldots,J$.

In the application study, we only focus on one facility. So let $k = 1$. A set of BST data from QC department is used to support our study.

Table 1
BST Data

DATE		12-1	12-1	12-2	12-2	…
TIME		0:35	10:51	3:53	21:45	…
PART		330A	230S	230B	550Q	…
MACHINE		W12	W12	W12	W12	…
LOT NO.		0833	0833	0833	1964	…
SHIFT		B	C	C	A	…
BALL SHEAR DATA	1	73.40	78.10	79.20	57.40	…
	2	71.50	73.20	78.70	56.90	…
	3	64.10	71.20	77.30	59.20	…
	4	78.90	78.90	71.30	53.30	…
	5	76.70	71.90	82.00	51.20	…
	6	77.40	73.30	85.80	56.50	…

B. Normality Test

When we attain original BST data set, we need to make some filtering on it. Part of the data is shown in Table 1.

W test designed by Shapiro and Wilk was selected for normality test because it can be use for small sample size, usually $3 \le n \le 50$. Furthermore, it is not necessary to know that the direction of drift is leaning or kurtosis (Shapiro and Wilk, 1965). Of course other methods can be choice depend on your real condition.

The procedure is as follows: At first, we put forward the null hypothesis H_0 and the alternative hypothesis H_1.

H_0: $X_{i1k}, X_{i2k}, \ldots, X_{ijk}$ come from a normal distribution, where $i = 1,2,\ldots,28$, $k = 1$.
H_1: $X_{i1k}, X_{i2k}, \ldots, X_{ijk}$ do not come from a normal distribution, where $i = 1,2,\ldots,28$, $k = 1$.
Ranking samples in a no decrease sequence: $y_1 \le y_2 \le \ldots \le y_n$
Calculating static W:

$$W = \frac{\left[\sum_{i=1}^{n} a_i (y_{n+1-i} - y_i)\right]^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1}$$

When $n=2M$, then $[n/2]=M$; When $n=2M+1$, then $[n/2]=M$. According to sample size n, check table and find $\alpha_l (1 \le l \le [n/2])$.

According to confidential interval $1-\alpha$ and sample size n, check the table and find W_α. If $W<W_\alpha$ then reject H_0.

One of the normality test results is shown as Figure 2. Every group of BST data form each variety should pass the normality test. Otherwise this group data is removed and cannot be used in the next step.

C. Cluster

Each variety BST group has a sample mean. We calculate all means of varieties, and prepare to cluster analysis. The object is to reduce quantity of varieties into an acceptable level.

Ward (1963) explored a cluster method called hierarchical grouping. We use this method to cluster sample mean. Every lot feature is described by $\{X_{i\cdot 1}, i = 1,2,\ldots,28\}$ which is the means of each variety sample from BST. Cluster procedure can be implement by Minitab software. Then we get a hierarchy cluster result as shown in Figure 3.

Figure 2
Normality test result of group No.9

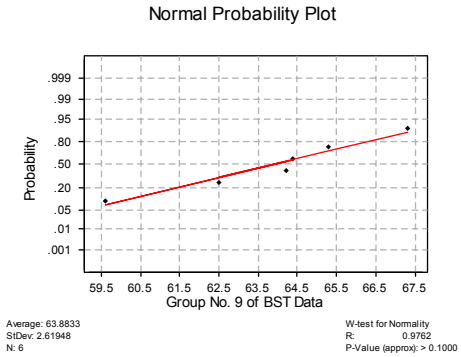
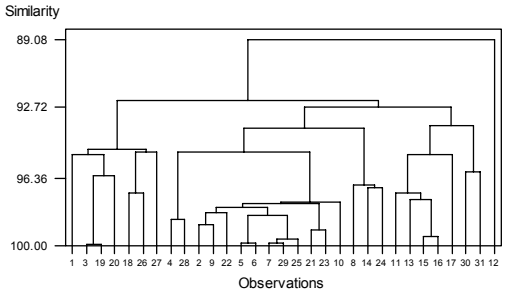


Figure 3
Cluster result



D. ANOVA

Analysis of variance is a procedure to test the hypothesis that several populations have the same mean (Watson et al., 1993). ANOVA can be employed to test whether the cluster result is convinced.

We now have a formal test of the null hypothesis of no difference in variety means in the same cluster (H_0: $\mu_1=\mu_2=\ldots=\mu_k$). Following statistical testing procedure is performed to test if the null hypothesis can be accepted.

Firstly, we calculate SB2, called the between-groups variance estimate that is based on the variation from one sample mean to the next.

$$S_B^2 = \frac{\sum_{j=1}^{k} n_j (\overline{X}_j - \overline{\overline{X}})^2}{k-1} \tag{2}$$

Secondly, within-group estimate of variance SW2 is calculated.

$$S_W^2 = \frac{\sum_{j=1}^{k} (n_j - 1) S_j^2}{n_T - k} \tag{3}$$

where j is the group no, X_j is the mean of group j. The grand mean X is the overall sample mean for all of the observations, k is the number of groups, n_T is the total number of items sampled, n_j is the number of the group j, S_{j2} is the variance of group j, as follows:

$$S_j^2 = \frac{\sum_{i=1}^{n_i} (X_{ij} - \overline{X}_j)^2}{n_j - 1} \tag{4}$$

where X_{ij} is the data i in group j. The SB2 figure has k-1 degree of freedom. And the S_W^2 figure has n_T-k degree of freedom.

Then the ratio of the between-groups estimates of variance to the within-groups estimates of variance is F value that follows a probability distribution:

$$F = \frac{S_B^2}{S_W^2} \tag{5}$$

To determine the reject rule, we must consult the F probability distribution table. In that table, we can find the F value that is exceeded by chance with only 5% for various combinations of degrees of freedom. If $F > F_{1-\alpha,k-1,nT-k}$, then we reject null hypothesis and conclude that the means of varieties are different. It means that there is no confidence to believe these varieties come from the same one population. So we conclude this cluster need to be separated further. Then we choose a higher similarity level, and make ANOVA again. At last we get 5 clusters, and every cluster pass ANOVA. That means there are 5 populations that have their own mean and variance.

E. Cluster Shewhart Control Chart

When we finish all procedures above, final data set could be used to design a Shewhart control chart. Every cluster has its own mean (μ) and variance (σ^2). So we could use

$\mu \pm 3\sigma$ principle to confirm upper/lower control limit of control chart. So we construct 5 cluster control charts. Then we mark $\{X_{i\cdot}\}$ point of ith variety on its own cluster for process control [1]. One of the cluster Shewhart control chart of No.5 cluster is shown in Figure 4.

Figure 4
Cluster Shewhart control chart of No.5 cluster

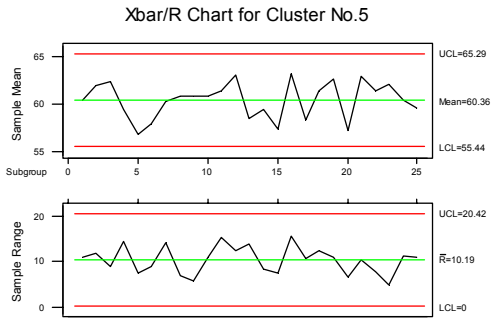


Table 2
Cluster result

Cluster No.	Quantity of Variety
1	4
2	8
3	1
4	3
5	12
Total Quantity	28

V. RESULTS

From the result (as shown in Table 2), there are 28 varieties need to be monitored during manufacturing process. After the whole procedure is performed, we get 5 clusters in the end. The biggest cluster has 12 varieties. In every cluster we have 95% confidential to convince that all varieties in the same cluster have the same statistical feature, such as mean and variance. So 5 control charts are enough to monitor all 28 varieties. Then Shewhart control chart is possible to be used in excess variety manufacturing process monitoring.

In this paper, cluster analysis and Shewhart control scheme (CASH) is proposed to deal with excess variety process control problem involved in highly automated manufacturing facility. Based on the case study, complexity of management is reduced. This method makes it possible to use traditional SPC control chart in excess variety manufacturing environment. Further research may be integration of CASH with fault diagnosis, and make a completed process control system.

REFERENCES

Brito, M., E. Chavez, A. Quiroz, and J. Yukich, 1997, "Connectivity of the Mutual K-Nearest-Neighbor Ggraph for Clustering and Outlier Detection," *Statistics and Probability Letters*, 35, 33-42.

Guha, S., R. Rastogi, and K. Shim, 1998, "CURE: An Efficient Clustering Algorithm for Large Databases," In Proceedings of ACM SIGMOD, 73-84.

Montgomery, D. C., 1991, *Introduction to Statistical Quality Control*, New York: Wiley.

Shapiro, S. S. and M. B. Wilk, 1965, "An Analysis of Variance Test for Normality," *Biometrika*, 52, 591.

Smyth, P., 1999, "Probabilistic Model-Based Clustering of Multivariate and Sequential Data," *In Proceedings of Artificial Intelligence and Statistics*.

Watson, C. L., P. Billingsley, D. J. Croft, and D. V. Huntsberger, 1993, *Statistics for Management and Economics*, 5th edition, Allyn and Bacon.

Ward. J. H. Jr., 1963, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236-244.