PRODUCT QUALITY IMPROVEMENT USING MULTIVARIATE DATA ANALYSIS

Manabu Kano * **Koichi Fujiwara** * **Shinji Hasebe** *
Hiromu Ohno **

Kyoto University, Kyoto 615-8510, Japan
***Kobe University, Kobe 657-0013, Japan*

Abstract: A data-based methodology for improving product quality is proposed. Referred to as Data-Driven Quality Improvement (DDQI), the proposed method can cope with qualitative as well as quantitative variables, determine the operating conditions that can achieve the desired product quality, optimize operating condition under various constraints, and thus can provide useful information to improve product quality. This paper aims to formulate DDQI and demonstrate its usefulness with an case study of an industrial steel process. In addition, possible extensions and remaining problems are discussed based on the authors' experience of succeeding in improving product quality by applying DDQI to several industrial processes. *Copyright ©2005 IFAC*

Keywords: Quality Control, Statistical Process Control, Optimization, Principal Component Analysis, Linear Discriminant Analysis

1. INTRODUCTION

How can we improve product quality and yield? More than ever, the answer to this question is vital as product life cycles are getting shorter and international competition is getting keener. Since this question arises repeatedly when a new product is developed, quality improvement should be achieved faster and in a more systematic way.

Although the most reliable approach to quality improvement will be the use of precise first-principle models, such models are not available in most newly developed processes and modeling of a complex industrial process is very difficult and time-consuming. Another option is the use of operation data. Because of the recent advances in information technology, a huge amount of data is kept in data bases and available for analysis. In the last decade or so, data-based approaches have been widely accepted for process control and monitoring in many industries. In the chemical industry, Multivariate Statistical Process Control (MSPC) based on Principal Component Analysis (PCA) has been a dominant technique for fault detection and identification (Kourti and MacGregor, 1995). MSPC is profitable not only to the chemical industry but to the steel industry (Dudzic and Zhang, 2004).

In the present work, a data-based methodology for improving product quality is proposed. The proposed method, referred to as Data-Driven Quality Improvement (DDQI), is based on a statistical model. DDQI can cope with qualitative as well as quantitative variables, determine the operating conditions that can achieve the desired product quality, optimize operating condition under various constraints, and thus can provide useful information to improve product quality. This paper aims to formulate DDQI, show its usefulness via an industrial case study, and

also point out possible extensions and remaining problems.

2. FORMULATION OF DATA-DRIVEN QUALITY IMPROVEMENT

In this section, Data-Driven Quality Improvement (DDQI) is formulated. Jaeckle and MacGregor (1998) proposed a product design method based on linear/nonlinear multivariate analysis. Their method can derive the operating conditions that can achieve the desired product quality, but it does not account for qualitative variables. DDQI can handle qualitative as well as quantitative variables in a unified framework. This characteristic expands the application area and the usefulness of DDQI.

2.1 Preprocessing Data

A quality data matrix $Y \in \Re^{N \times Q}$ and an operating condition data matrix $X \in \Re^{N \times P}$ are observed. N, Q, and P are the numbers of samples, quality variables, and operating condition variables, respectively. The nth measurements of the qth quality variable and the pth operating condition variable are denoted by y_{nq} and x_{np}, respectively. For simplicity, it is assumed that quality data $y_q \in \Re^{N \times 1}$ and operating condition data $x_p \in \Re^{N \times 1}$ are mean-centered. In addition, they are appropriately scaled if necessary.

2.2 Modeling Quality and Operating Conditions

DDQI is based on a statistical quality model that relates operating conditions with quality. Multiple regression analysis (MRA) is the simplest method for building a quality model, but it cannot be used if a colinearity problem occurs. Principal component regression (PCR) can cope with colinearity, and it includes MRA as its special case. On the other hand, linear discriminant analysis (LDA) can be used for classification if a quality variable is qualitative, but it cannot cope with colinearity. The colinearity problem can be solved by applying principal component analysis (PCA) to reduce the dimensionality before LDA is used. This method is referred to as PCA-LDA. Since LDA for two classes is the same as MRA when qualitative variables are quantified, PCA-LDA is essentially the same as PCR. Therefore, the following formulation of DDQI is based on PCR.

The singular value decomposition of an operating condition data matrix X is written as

$$
\begin{aligned}
X &= USV^T \\
&= \begin{bmatrix} U_R & U_0 \end{bmatrix} \begin{bmatrix} S_R & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} V_R & V_0 \end{bmatrix}^T
\end{aligned} \quad (1)
$$

where U and V are orthogonal matrices. R denotes the number of principal components to be retained in a PCA model. The diagonal matrix S has singular values s_r in its diagonal elements in decreasing order. The rth principal component is given as the rth column v_r of the loading matrix V_R, and the rth score t_r is given by

$$
t_r = X v_r = s_r u_r . \quad (2)
$$

The score matrix is given by

$$
T_R = X V_R = U_R S_R . \quad (3)
$$

Since the score t_r are uncorrelated to each other, the covariance matrix of T_R is diagonal.

$$
\frac{1}{N-1} T_R^T T_R = \frac{1}{N-1} S_R^2 \quad (4)
$$

The dimensionality can be reduced from P to R via PCA, and T_R is obtained. Then, X can be reconstructed as $\hat{X}$ by projecting T_R onto the original P dimensional space.

$$
\hat{X} = T_R V_R^T = X V_R V_R^T \quad (5)
$$

The size of $\hat{X}$ is the same as that of X, but the rank of $\hat{X}$ is only R.

In PCR, the scores are used as input variables of MRA. The colinearity problem does not occur any more, because the scores are uncorrelated with each other. The PCR model can be written as

$$
Y = T_R K + E \quad (6)
$$

where K is a regression coefficient matrix and E is an error matrix. By least squares, K is determined:

$$
K = (T_R^T T_R)^{-1} T_R^T Y . \quad (7)
$$

The prediction of quality can be given by

$$
\hat{Y} = T_R K = X V_R K = X K_{PCR} \quad (8)
$$

$$
K_{PCR} \equiv V_R K . \quad (9)
$$

The coefficient matrix K_{PCR} shows the influence of operating conditions on quality. Basically, the operating condition variable with the larger coefficient has greater influence on the quality. It should be noted here, however, that the estimated coefficients are biased due to correlation among input variables.

2.3 Searching Operating Conditions That Can Achieve Desired Product Quality

A method to determine the operating conditions that can achieve desired product quality is described here. The method solves an inverse problem of the PCR model. In general, the number of quality variables Q is less than that of principal components R, and thus, the operating condition cannot be determined uniquely. In DDQI, a space where operating conditions can achieve the desired quality is searched within R dimensional space spanned by principal components.

Based on Eq. (8), the operating conditions $\tilde{x} \in \Re^{P \times 1}$ that can achieve the desired quality $\tilde{y} \in \Re^{Q \times 1}$ are determined. The operating conditions must exist in the space spanned by R principal components, because correlation among operating condition variables, which is theoretically defined by physical or chemical laws, is extracted from operation data by using PCA in DDQI. Therefore, at the first step to solving the inverse problem, the scores $\tilde{t} \in \Re^{R \times 1}$ that can achieve the desired quality $\tilde{y}$ are determined. The scores $\tilde{t}$ are related to the desired quality $\tilde{y}$ via

$$\tilde{y} = K^T \tilde{t}. \tag{10}$$

If $\tilde{t}$ is found, then the operating conditions $\tilde{x}$ can be determined.

$$\tilde{x} = V_R \tilde{t} \tag{11}$$

When the number of quality variables Q is larger than that of principal components R, there is no score $\tilde{t}$ that can achieve the desired quality. In such a case, the score that achieves as desired quality as possible should be determined.

$$\tilde{t} = (KK^T)^{-1} K \tilde{y} \tag{12}$$

From Eq. (11), the operating condition recommended for quality improvement is

$$\tilde{x} = V_R (KK^T)^{-1} K \tilde{y}. \tag{13}$$

On the other hand, there is a unique score vector $\tilde{t}$ that can achieve the desired quality when $Q = R$. In this case, the operating condition recommended for quality improvement is also determined by Eq. (13) or

$$\tilde{x} = V_R (K^T)^{-1} \tilde{y}. \tag{14}$$

In many cases, however, Q is less than R. Therefore, the scores $\tilde{t}$ that can achieve the desired quality cannot be determined uniquely. In such a case, a solution set $\mathcal{S}_{\tilde{t}}$ where any score

can achieve the desired quality is searched. By using a matrix $\tilde{T}_0 \in \Re^{R \times (R-Q)}$ whose columns are bases of the null space (kernel) of K^T, $\tilde{t}$ can be described as

$$\tilde{t} = (K^T)^+ \tilde{y} + \tilde{T}_0 \phi \quad . \tag{15}$$

where $(K^T)^+$ denotes the generalized inverse of K^T and $\phi \in \Re^{(R-Q) \times 1}$ is any coefficient vector. Finally the operating conditions that can achieve the desired quality $\tilde{y}$ are given by

$$\tilde{x} = V_R \left\{ (K^T)^+ \tilde{y} + \tilde{T}_0 \phi \right\}. \tag{16}$$

Any general solution $\tilde{x}$ is an element of the column space of the loading matrix V_R, and thus it is clear that $\tilde{x}$ exists in the space spanned by R principal components.

2.4 Handling Qualitative Variables

In previous subsections, the method to search the operating conditions $\tilde{x}$ that can achieve the desired product quality $\tilde{y}$ is explained. However, it is applicable only to cases where all quality variables are quantitative. The desired product quality cannot be specified quantitatively when the quality variables are qualitative, e.g., good and bad. In this subsection, a novel quantification method is proposed.

In order to build a quality model by using PCR, qualitative quality variables should be quantified. As well-known, for example, qualitative information such as good and bad can be quantified and denoted by 1 and 0, respectively.

$$y_{nq} = \begin{cases} 1, & \text{for good product} \\ 0, & \text{for bad product} \end{cases} \tag{17}$$

The quantified variable is then mean-centered, and the quality model can be built via PCR.

$$\hat{y}_q = X k_{PCR} \tag{18}$$

It should be noted here that the derived coefficient vector k_{PCR} is proportional to the vector describing the discriminant function of PCA-LDA. In other words, PCR and PCA-LDA is essentially the same when there are two classes. A quality variable can be quantified in a similar way even when there are more than two classes.

This conventional quantification method is useful for building a quality model, but not for solving its inverse problem. The serious problem is that the physical meaning of the quantified variable is not clear at all. For example, what 0.6 means when a qualitative quality variable is quantified

through Eq. (17)? Is it acceptable or not? Nobody can answer to this question. In the present work, therefore, a novel quantification method is proposed to give physical meaning to quantified variables.

For a qualitative quality variable, the yield, i.e., the percentage of good products to all products, can be specified instead of the quality itself on the basis of the histogram for each category. Each histogram can be obtained from operation data, and it can be drawn as the frequency distribution of good or bad samples against the axis defined by $\boldsymbol{k}_{PCR}$. Then, the yield against $\boldsymbol{k}_{PCR}$ can be defined by

$$\text{yield}(\Delta_i) = \frac{N_{good}(\Delta_i)}{N_{good}(\Delta_i) + N_{bad}(\Delta_i)} \quad (19)$$

where Δ_i is the ith bin of the histgram. $N_{good}(\Delta_i)$ and $N_{bad}(\Delta_i)$ are the numbers of good or bad samples contained in Δ_i , respectively. To calculate the yield through Eq. (19), the position of each bin must be the same between the histogram for good products and that for bad products. Once the desired yield is specified, operating conditions that can achieve the desired yield can be found by following the above-mentioned approach.

On the other hand, if an operator needs to decide whether to use particular equipment or to decide which equipment to use, such operating conditions should be quantified. For example, when one of K pieces of equipment is used for production, the information on which equipment is used can be quantified by introducing K 0-1 variables. The kth 0-1 variable e_{nk} is defined as:

$$e_{nk} = \begin{cases} 1, & \text{when } k\text{th equipment is used.} \\ 0, & \text{when } k\text{th equip. is not used.} \end{cases} (20)$$

By simple extension, combinations of equipment to use can be quantified in a similar way. It should be noted here that only $K-1$ variables are necessary for distinguishing K operating status because the Kth equipment must be used if the first to the $K-1$th equipment is not used. However, if only $K-1$ variables are used, the Kth variable does not have any coefficient in the quality model developed, and thus it becomes difficult to understand the influence of the Kth equipment on the quality variables and the relationship between the Kth equipment and other operating conditions. In DDQI, therefore, K 0-1 variables are introduced for distinguishing K operating status and then they are mean-centered.

The above-mentioned searching algorithm cannot cope with 0-1 variables e_k. To derive a feasible operating condition, 0-1 variables should be determined in advance. Then, operating

conditions that can achieve the desired product quality can be calculated. It should be noted here that only a few combinations should be checked because the quality model helps one to judge which equipment is to be used.

2.5 Optimizing Operating Condition

The operating conditions that can achieve the desired product quality cannot be uniquely determined in general, but it can be optimized if an objective function is provided.

An objective function is assumed to be provided as a quadratic function. The objective function is optimized under the following four constraints: 1) the desired product quality is achieved, 2) the operating condition exists in the space spanned by R principal components, 3) all operating condition variables exist within their upper and lower bounds, and 4) T^2 statistic of scores is below its upper control limit or approximately 4') all R scores exist within their upper and lower bounds. Here, the T^2 statistic is defined by

$$T^2 = \|\boldsymbol{S}_R^{-1}\boldsymbol{t}\|^2 . \quad (21)$$

The last constraint is necessary for finding a new optimal operating condition within the region where the PCR or PCA-LDA model is valid. In other words, extrapolation should be avoided by using the last constraint. Since the objective function is quadratic and the constraints are linear if 4' is used, the optimization problem can be solved by quadratic programming.

The operating conditions $\tilde{\boldsymbol{x}}$ that satisfy the constraints 1 and 2 are given by Eq. (16) as a function of the coefficient vector $\boldsymbol{\phi}$. Therefore, the first two constraints are automatically satisfied if Eq. (16) is substituted for the objective function $J(\tilde{\boldsymbol{x}})$. As a result, the objective function $J'(\boldsymbol{\phi})$ needs to be optimized under the third and fourth constraints. The upper and lower bound constraints on $\tilde{\boldsymbol{x}}$ can be transformed into the linear constraints on $\boldsymbol{\phi}$ via Eq. (16). Therefore, the optimization problem can be formulated as follows:

$$\max_{\boldsymbol{\phi}} \ J'(\boldsymbol{\phi}) \quad (22)$$

$$\text{s.t.} \ \boldsymbol{V}_R\tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \geq \boldsymbol{x}_l - \boldsymbol{V}_R(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} \quad (23)$$

$$\boldsymbol{V}_R\tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \leq \boldsymbol{x}_u - \boldsymbol{V}_R(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} \quad (24)$$

$$\|\boldsymbol{S}_R^{-1}(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{T}}_0\boldsymbol{\phi}\|^2 \leq T_u^2 \quad (25)$$

where $\boldsymbol{x}_u$ and $\boldsymbol{x}_l$ are the upper and lower bounds of $\tilde{\boldsymbol{x}}$ and T_{ucl}^2 is the upper bound of T^2. For example, the objective function is given as

$$J(\tilde{\boldsymbol{x}}) = \|\tilde{\boldsymbol{x}} - \boldsymbol{x}_0\|_{\boldsymbol{W}}^2 + \tilde{\boldsymbol{x}}^T \boldsymbol{\theta} \qquad (26)$$

where $\boldsymbol{x}_0$, $\boldsymbol{W}$, and $\boldsymbol{\theta}$ denote the reference operating condition, a weighting matrix, and a coefficient vector describing the operation cost.

If there is no solution that satisfies all constraints, i.e., the imposed specifications on quality are too severe, the operating condition that achieves as desired quality as possible should be determined under the second, third, and fourth constraints. This optimization problem can be solved by quadratic programming if its objective function is defined as a weighted squared norm of the difference between the desired quality and the achieved quality. The relationship between the quality and the scores is modeled by Eq. (10). Therefore, the second constraint is automatically satisfied if Eq. (10) is substituted for the objective function $J_{sub}(\boldsymbol{y})$. As a result, the objective function $J'_{sub}(\boldsymbol{t})$ needs to be optimized under the third and fourth constraints. The constraints on $\boldsymbol{x}$ can be transformed into the linear constraints on $\boldsymbol{t}$ via Eq. (11). Therefore, the optimization problem can be formulated as follows:

$$\max_{\boldsymbol{t}} \; J'_{sub}(\boldsymbol{t}) \qquad (27)$$

$$\text{s.t.} \; \boldsymbol{V}_R \boldsymbol{t} \geq \boldsymbol{x}_l \qquad (28)$$

$$\boldsymbol{V}_R \boldsymbol{t} \leq \boldsymbol{x}_u \qquad (29)$$

$$\|\boldsymbol{S}_R^{-1} \boldsymbol{t}\|^2 \leq T_u^2 \,. \qquad (30)$$

3. CASE STUDY: STEEL PROCESS

In this section, a case study of applying DDQI to an industrial steel process in Japan is described. Successful application results of PCA-based MSPC to steel processes were reported by Dudzic and Zhang (2004). Their system can detect undesirable operation and reduce the occurrence of caster strand breakouts. In general, however, monitoring each facility independently is not sufficient for improving final product quality because steel products are produced through several facilities including a blast furnace, a converter, a continuous caster, a reheat furnace, hot or cold rolling mills, and so on. Therefore, the scope must be extended to cover a whole production system.

A typical problem encountered in the steel process is various defects on the surface of sheet steel products. Since the product quality can be affected by many factors, it is very difficult to find a root cause of defects and take an appropriate action to reduce them. The objective in this case study is to minimize defects on the steel surface and improve product yield. A flow diagram of the steel making and finishing process, which is
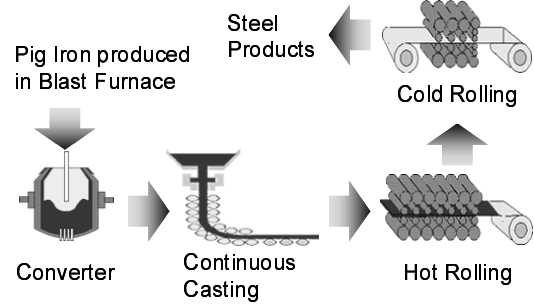

Fig. 1. Flow diagram of steel making and finishing process

investigated in this case study, is shown in Fig. 1.

On the basis of process knowledge, operation data measured in the converter, the hot rolling mill, and the cold rolling mill were analyzed. In addition, the number of variables was reduced from several hundreds to less than one hundred for efficient analysis. Since the remaining operating condition variables were highly correlated to each other, PCA was used for dimensionality reduction. The product quality, which is the surface condition of sheet steel, is inspected by human operators and evaluated qualitatively. Therefore, it must be quantified. In this study, the quality was quantified and denoted by 1 (good/no defect) or 0 (bad/defect). To model the relationship between the product quality and operating condition variables, PCA-LDA was used.

In one steel process, various types of steel are produced. Since the characteristics of products and process operating conditions are different depending on steel types, a PCA-LDA model must be built for each steel type. In this section, the results of applying DDQI to two types of sheet steel, high-carbon steel and alloy steel, are described.

In the high-carbon steel case, more than 70 operating condition variables with 50 samples classified into bad and 100 samples classified into good were used for modeling. PCA-LDA is used for discriminating between 50 bad samples and 100 good samples and also for building a model relating 70 operating condition variables with the product quality. On the basis of the model developed, the operating condition variables that significantly affect the product quality can be identified. In this study, the PCA-LDA model was compared with the LDA model. The classification power of LDA is higher than PCA-LDA when they are applied to data used for modeling. However, the LDA model provides misleading results because operating condition variables are highly correlated. For example, temperatures at two different measurement points had significant

coefficients with different sign in the LDA model, although these two temperatures have very strong positive correlation. In practice, it is impossible to change these two temperatures in different directions. This result suggests that PCA or another method must be used together with LDA to reduce the dimensionarity and get a reliable result. In fact, the PCA-LDA model with 10 principal components concluded that temperatures identified by the LDA model was not key variables to reduce defects, and it suggested that sheet steel should be cooled down rapidly after the hot rolling process. The results of the PCA-LDA model was confirmed to be correct by changing the operating condition.

In the alloy steel case, more than 50 operating condition variables with 120 bad samples and 20 good samples were used for modeling. The PCA-LDA model with six principal components suggested that the composition of additives fed to the converter should be changed. The results of the PCA-LDA model was confirmed to be correct by the laboratory tests.

4. EXTENSIONS AND PROBLEMS

In this section, possible extensions and remaining problems are discussed briefly.

Since DDQI formulated here is based on a linear model, its applicability to nonlinear processes is limited. A simple extension can be made by using a nonlinear model. However, a serious problem to consider is the limited number of samples as pointed out by Jaeckle and MacGregor (1998). In many processes, where product life cycles are short, product quality needs to be improved by using a small number of samples. In such a case, nonlinear modeling is not practical. From the authors' experience of succeeding in improving yield and product quality by applying DDQI to the steel and the semiconductor industries, linear models are sufficiently practical and useful for real problems. In practice, an operating condition cannot be changed drastically at a time, it should be changed carefully and step by step. Therefore, a linear model is useful because it can model the process with sufficient acceracy near the current operating condition.

In practice, only a few manipulated variables can be changed simultaneously. It would be usually unacceptable to change all variables at the same time. Therefore, it is necessary to derive a new operating condition that can achieve the desired product quality by changing only a few operating condition variables. From this viewpoint, a systematic methodology for selecting a few important manipulated variables needs to be developed. Although details are not shown in this paper, the authors have developed a novel method for selecting manipulated variables, which can reduce the variability of product quality and also take into account various performance measures including operation cost.

To date, DDQI has been tested in the steel and the semiconductor industries, and it has succeeded in finding new operating conditions to achieve higher product quality. However, since model building and validation are time-consuming and require expertise, an efficient and robust procedure needs to be developed. In addition, to realize an online quality improvement, DDQI is integrated with Run-to-Run control (Castillo and Huriwitz, 1997). The hierarchical quality improvement syswtem (HiQIS), consisting of DDQI, Run-to-Run control, local process control, and statistical process monitoring, is now tested.

5. CONCLUSION

In the present work, a data-based methodology for improving product quality – Data-Driven Quality Improvement (DDQI) – is proposed. DDQI can cope with qualitative as well as quantitative variables and provide useful information to improve product quality. The case study in the sheet steel process shows the usefulness of PCA-LDA modeling and DDQI.

REFERENCES

Castillo, E.D. and A.M. Huriwitz (1997). Run-to-Run process control: Literature review and extensions. *J. Qual. Technol.*, **29**, 184–196.

Dudzic, M. and Y. Zhang (2004). On-line Industrial Implementation of Process Monitoring/Control Applications Using Multivariate Statistical Technologies: Challenges and Opportunities. *IFAC Symp. on Dynamics and Control of Process Systems (DYCOPS)*, CD-ROM, Cambridge, July 5–7.

Jaeckle, C.M. and J.F. MacGregor (1998). Product Design through Multivariate Statistical Analysis of Process Data. *AIChE J.*, **44**, 1105–1118.

Kourti, T. and J.F. MacGregor (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemometrics and Intelligent Laboratory Systems*, **28**, 3–21.