DATA-DRIVEN QUALITY IMPROVEMENT: HANDLING QUALITATIVE VARIABLES

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

[*] *Kyoto University, Kyoto 606-8501, 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 constraints, and also evaluate the validity of the results. The desired yield is specified instead of the quality for a qualitative quality variable. This paper aims to formulate DDQI and demonstrate its usefulness with an illustrative example. In addition, possible extensions and remaining problems are discussed based on the authors' experience of suceeding in improving product quality by applying DDQI to several industrial processes.

Keywords: Quality Improvement, Statistical Quality Control, Statistical Process Control, Principal Component Regression, Optimization

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.

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 constraints, and also evaluate the validity of the results. This paper aims to formulate DDQI, show its usefulness via a case study, and also point out possible extensions and remaining problems.

The most important contribution of this work is to provide a new method that can handle qualitative quality variables. By building a model that can relate operating condition to yield, the desired product quality can be specified, the operating condition can be optimized, and also the achievable quality at a certain operating condition can be estimated, even if the product quality is not measured quantitatively.

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 $\boldsymbol{Y} \in \Re^{N\times Q}$ and an operating condition data matrix $\boldsymbol{X} \in \Re^{N\times P}$ 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 $\boldsymbol{y}_q \in \Re^{N\times 1}$ and operating condition data $\boldsymbol{x}_p \in \Re^{N\times 1}$ are mean-centered. In addition, they are appropriately scaled if necessary.

If quality is given as qualitative information, such as good and bad, the qualitative variable should be quantified to build a model. For example, good and bad can be quantified and denoted by 1 and 0, respectively.

$$y_{nq} = \begin{cases} 1, & \text{for } y_{nq} \in C_1 \\ 0, & \text{for } y_{nq} \in C_2 \end{cases} \tag{1}$$

where C_1 and C_2 are classes consisting of good and bad samples, respectively. A quality variable can be quantified in a similar way even if there are more than two classes. This quantified variable is then mean-centered. This quantification is standard and widely used.

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:

$$e_{nk} = \begin{cases} 1, & \text{when } k\text{th equipment is used.} \\ 0, & \text{when } k\text{th equipment is not used.} \end{cases} \tag{2}$$

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–1th 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.

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. That is, PCR is more general than MRA. 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 formulation of DDQI is based on PCR.

The singular value decomposition of an operating condition data matrix $\boldsymbol{X}$ is written as

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \\ &= \begin{bmatrix} \boldsymbol{U}_R & \boldsymbol{U}_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_R & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_R & \boldsymbol{V}_0 \end{bmatrix}^T \end{aligned} \tag{3}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices. R denotes the number of principal components to be retained in a PCA model. The diagonal matrix $\boldsymbol{S}$ has singular values s_r in its diagonal elements in decreasing order. The rth principal component is given as the rth column $\boldsymbol{v}_r$ of the loading matrix $\boldsymbol{V}_R$, and the rth score $\boldsymbol{t}_r$ is given by

$$\boldsymbol{t}_r = \boldsymbol{X}\boldsymbol{v}_r = s_r\boldsymbol{u}_r\,. \tag{4}$$

The score matrix is given by

$$\boldsymbol{T}_R = \boldsymbol{X}\boldsymbol{V}_R = \boldsymbol{U}_R\boldsymbol{S}_R\,. \tag{5}$$

The score $\boldsymbol{t}_r$ are uncorrelated to each other. In fact, the covariance matrix of $\boldsymbol{T}_R$ becomes diagonal.

$$\frac{1}{N-1}\boldsymbol{T}_R^T\boldsymbol{T}_R = \frac{1}{N-1}\boldsymbol{S}_R^2 \tag{6}$$

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

$$\hat{\boldsymbol{X}} = \boldsymbol{T}_R\boldsymbol{V}_R^T = \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R^T \tag{7}$$

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

$$\boldsymbol{Y} = \boldsymbol{T}_R\boldsymbol{K} + \boldsymbol{E} \tag{8}$$

where $\boldsymbol{K}$ is a regression coefficient matrix and $\boldsymbol{E}$ is an error matrix. By least squares, $\boldsymbol{K}$ is determined:

$$\boldsymbol{K} = (\boldsymbol{T}_R^T\boldsymbol{T}_R)^{-1}\boldsymbol{T}_R^T\boldsymbol{Y} \quad . \tag{9}$$

The prediction of quality can be given by

$$\hat{\boldsymbol{Y}} = \boldsymbol{T}_R\boldsymbol{K} = \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{K} = \boldsymbol{X}\boldsymbol{K}_{PCR} \tag{10}$$

where

$$\boldsymbol{K}_{PCR} \equiv \boldsymbol{V}_R\boldsymbol{K}\,. \tag{11}$$

The coefficient matrix $\boldsymbol{K}_{PCR}$ 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.

Desired product quality cannot be specified when the quality variable is qualitative. For the qualitative quality variable, the yield, i.e., the percentage of good products to all products, or a similar quantitative measure can be specified instead of the quality itself on the basis of a histogram for each category.

Based on Eq. (10), the operating conditions $\tilde{\boldsymbol{x}} \in \Re^{P\times 1}$ that can achieve the desired quality $\tilde{\boldsymbol{y}} \in \Re^{Q\times 1}$ 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 $\tilde{\boldsymbol{t}} \in \Re^{R\times 1}$ that can achieve the desired quality $\tilde{\boldsymbol{y}}$ are determined. The scores $\tilde{\boldsymbol{t}}$ are related to the desired quality $\tilde{\boldsymbol{y}}$ via

$$\tilde{\boldsymbol{y}} = \boldsymbol{K}^T\tilde{\boldsymbol{t}}\,. \tag{12}$$

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

$$\tilde{\boldsymbol{x}} = \boldsymbol{V}_R\tilde{\boldsymbol{t}} \tag{13}$$

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

$$\tilde{\boldsymbol{t}} = (\boldsymbol{K}\boldsymbol{K}^T)^{-1}\boldsymbol{K}\tilde{\boldsymbol{y}} \tag{14}$$

From Eq. (13), the operating condition recommended for quality improvement is

$$\tilde{\boldsymbol{x}} = \boldsymbol{V}_R(\boldsymbol{K}\boldsymbol{K}^T)^{-1}\boldsymbol{K}\tilde{\boldsymbol{y}}\,. \tag{15}$$

On the other hand, there is a unique score vector $\tilde{\boldsymbol{t}}$ that can achieve the desired quality when $Q = R$. In this case, the operating condition recommended for quality improvement is also determined by Eq. (15) or

$$\tilde{\boldsymbol{x}} = \boldsymbol{V}_R(\boldsymbol{K}^T)^{-1}\tilde{\boldsymbol{y}}\,. \tag{16}$$

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

In general, a column space $\mathcal{S}_c(\boldsymbol{A})$, a row space $\mathcal{S}_r(\boldsymbol{A})$, and a null space (kernel) $\mathcal{S}_0(\boldsymbol{A})$ of a matrix $\boldsymbol{A}$ have the following relationships:

$$\mathcal{S}_r(\boldsymbol{A}) = \mathcal{S}_0(\boldsymbol{A})^{\perp} \tag{17}$$

$$\mathcal{S}_r(\boldsymbol{A}) = \mathcal{S}_c(\boldsymbol{A}^T) \tag{18}$$

where $^{\perp}$ denotes orthogonal complement. Therefore, any element $\tilde{\boldsymbol{t}}$ in the solution set $\mathcal{S}_{\tilde{\boldsymbol{t}}}$ is uniquely decomposed into the orthogonal projection, $\tilde{\boldsymbol{t}}_{ps}$, of $\boldsymbol{K}^T$ onto its row space $\mathcal{S}_r(\boldsymbol{K}^T)$ and the other orthogonal projection, $\tilde{\boldsymbol{t}}_0$, of $\boldsymbol{K}^T$ onto its null space $\mathcal{S}_0(\boldsymbol{K}^T)$.

$$\tilde{\boldsymbol{t}} = \tilde{\boldsymbol{t}}_{ps} + \tilde{\boldsymbol{t}}_0 \tag{19}$$

Since $\tilde{\boldsymbol{t}}_0$ is the orthogonal projection onto the null space $\mathcal{S}_0(\boldsymbol{K}^T)$,

$$\boldsymbol{K}^T\tilde{\boldsymbol{t}}_0 = \boldsymbol{0}\,. \tag{20}$$

Substituting these equations for Eq. (12) gives

$$\tilde{\boldsymbol{y}} = \boldsymbol{K}^T(\tilde{\boldsymbol{t}}_{ps} + \tilde{\boldsymbol{t}}_0) = \boldsymbol{K}^T\tilde{\boldsymbol{t}}_{ps}\,. \tag{21}$$

That is, $\tilde{\boldsymbol{t}}_{ps}$ is a particular solution of the inverse problem, and it can be determined by using the minimum norm generalized inverse $(\boldsymbol{K}^T)^+$ of $\boldsymbol{K}^T$.

$$\tilde{\boldsymbol{t}}_{ps} = (\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} \tag{22}$$

Since a general solution $\tilde{\boldsymbol{t}}$ is given by Eq. (19), it is necessary to determine $\tilde{\boldsymbol{t}}_0$ that is an element of the null space $\mathcal{S}_0(\boldsymbol{K}^T)$. By defining a matrix $\tilde{\boldsymbol{T}}_0 \in \Re^{R\times(R-Q)}$ which has all bases of $\mathcal{S}_0(\boldsymbol{K}^T)$ as its columns, $\tilde{\boldsymbol{t}}_0$ can be described as

$$\tilde{\boldsymbol{t}}_0 = \tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \tag{23}$$

by using any coefficient vector $\boldsymbol{\phi} \in \Re^{(R-Q)\times 1}$. Therefore,

$$\tilde{\boldsymbol{t}} = (\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \quad . \tag{24}$$

Finally the operating conditions that can achieve the desired quality $\tilde{\boldsymbol{y}}$ are given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{V}_R\left\{(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{T}}_0\boldsymbol{\phi}\right\}\,. \tag{25}$$

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

Unfortunately, desired product quality cannot be specified when the quality variable is qualitative. Even in such a case, the quality variable can be quantified as described before, and the quality model can be built via PCR.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{k}_{PCR} \tag{26}$$

For the qualitative quality variable, the yield, i.e., the percentage of good products to all products, or a similar quantitative measure can be specified instead of the quality itself on the basis of a histogram for each category. The histograms can be obtained from operation data, and they can be drawn against the axis defined by $\boldsymbol{k}_{PCR}$. For example, histograms of both good products and bad products can be calculated and drawn against the axis defined by $\boldsymbol{k}_{PCR}$, and then the desired yield can be specified. Once the desired yield is specified, operating conditions that can achieve the desired yield can be found by following the above-mentioned approach. An illustrative example is shown in the next section.

The searching algorithm cannot cope with 0-1 variables e_k, which is used for quantifying operating status when an operator selects particular equipment to use. 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.

2.4 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 three constraints: 1) the desired product quality is achieved, 2) the operating condition exists in the space spanned by principal components, and 3) all operating condition variables exist within their upper and lower bounds. Since the objective function is quadratic and the constraints are linear, 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. (25) as a function of the coefficient vector $\boldsymbol{\phi}$. Therefore, the first two constraints are automatically satisfied if Eq. (25) 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 constraint. The upper and lower bound constraints on $\tilde{\boldsymbol{x}}$ can be transformed into the linear constraints on $\boldsymbol{\phi}$ via Eq. (25). Therefore, the optimization problem can be formulated as follows:

$$\max_{\boldsymbol{\phi}}\ J'(\boldsymbol{\phi}) \tag{27}$$

$$\text{s.t. } \boldsymbol{V}_R\tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \geq \boldsymbol{x}_l - \boldsymbol{V}_R(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} \tag{28}$$

$$\boldsymbol{V}_R\tilde{\boldsymbol{T}}_0\boldsymbol{\phi} \leq \boldsymbol{x}_u - \boldsymbol{V}_R(\boldsymbol{K}^T)^+\tilde{\boldsymbol{y}} \tag{29}$$

where $\boldsymbol{x}_u$ and $\boldsymbol{x}_l$ are the upper and lower constraints on $\tilde{\boldsymbol{x}}$. For example, the objective function is given as

$$J(\tilde{\boldsymbol{x}}) = \|\tilde{\boldsymbol{x}} - \boldsymbol{x}_0\|^2_{\boldsymbol{W}} + \tilde{\boldsymbol{x}}^T\boldsymbol{\theta} \tag{30}$$

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 and the third 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. Or, the problem can be solved by linear programming if its objective function is defined as a linear function of the achieved quality. The relationship between the quality and the scores is modeled by Eq. (12). Therefore, the second constraint is automatically satisfied if Eq. (12) 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 constraint. The constraints on $\boldsymbol{x}$ can be transformed into the linear constraints on $\boldsymbol{t}$ via Eq. (13). Therefore, the optimization problem can be formulated as follows:

$$\max_{\boldsymbol{t}} \; J'_{sub}(\boldsymbol{t}) \tag{31}$$

$$\text{s.t. } \boldsymbol{V}_R\boldsymbol{t} \geq \boldsymbol{x}_l \tag{32}$$

$$\boldsymbol{V}_R\boldsymbol{t} \leq \boldsymbol{x}_u \,. \tag{33}$$

2.5 Evaluating Validity of Quality Model

DDQI provides a unified framework for improving product quality. An operating condition can be optimized so that the desired quality is achieved under constraints. Since DDQI is based on PCR, the results are valid inside the region where the PCR model can give a close approximation to the real process. The validity will deteriorate outside the operating conditions where process data used for modelling are obtained.

To check the validity of the optimization results, that is, to check the validity of the PCR model, SPC is integrated with DDQI. In PCA-based SPC, two statistics, T^2 and Q, are monitored (Jackson and Mudholkar, 1979).

$$T^2 = \sum_{r=1}^{R} \frac{t_r^2}{\sigma_{t_r}^2} \tag{34}$$

$$Q = \sum_{p=1}^{P} (x_p - \hat{x}_p)^2 \tag{35}$$

where t_r is the rth score and $\sigma_{t_r}^2$ is the variance of t_r. x_p and $\hat{x}_p$ are a measurement of the pth variable and its predicted (reconstructed) value, respectively. The T^2 statistic is a measure of the variation within the PCA model, and the Q statistic is a measure of the amount of variation not captured by the PCA model. The optimization results as well as the PCR model are judged to be valid when both statistics are below the predetermined thresholds.

Furthermore, the operating condition can be optimized within the space where the quality

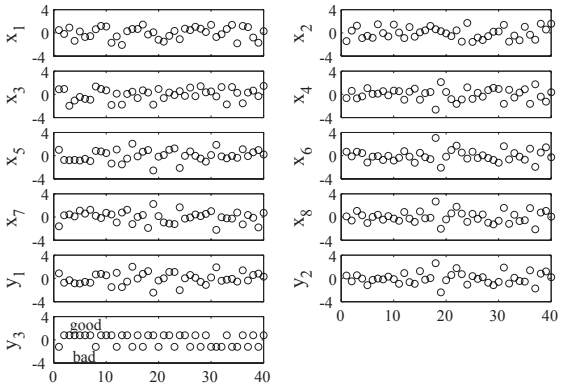

Fig. 1. Operating condition and quality data

model is valid by imposing the constraints on scores

$$\boldsymbol{t}_l \leq \boldsymbol{t} \leq \boldsymbol{t}_u \,. \tag{36}$$

By using constraints on scores instead of T^2 statistic, the optimization problem can be solved by quadratic programming.

3. NUMERICAL EXAMPLE

In this section, a case study shows how DDQI works, in particular, how it handles a qualitative quality variable. Simulation data of eight operating condition variables $\boldsymbol{x}$ and three quality variables $\boldsymbol{y}$ are generated. In this case study, $\boldsymbol{x}$ is linear combinations of random variables. The quality variable y_3 is qualitative, the judgment of which is given as good or bad, while y_1 and y_2 are quantitative variables. All variables are normalized and shown in Fig. 1.

PCR is used for modeling, and four principal components are retained. The operating condition $\tilde{\boldsymbol{x}}$ that can achieve the desired product quality $\tilde{\boldsymbol{y}}$ must exist in one dimensional space because there are three quality variables ($Q = 3$) and four principal components ($R = 4$). In this case study, four optimization problems are solved. The objective function is given as Eq. (30), where

$$\boldsymbol{x}_0 = \boldsymbol{0},\, \boldsymbol{W} = \boldsymbol{I},\, \boldsymbol{\theta} = [5\,1\,4\,0\,0\,0\,0\,0]^T \quad . \tag{37}$$

The desired product quality and the constraints are summarized in Table 1. The desired quality of y_3 cannot be specified because y_3 is a qualitative variable. Therefore, the desired yield is specified by using the expected yield function, which is determined on the basis of histograms of good products and bad products as shown in Fig. 2. The horizontal axis y_3 represents the dummy quality variable estimated via Eq. (26). The optimal operating conditions are summarized in Table 1 and Fig. 2 (right-bottom). Each solid line in Fig. 2 shows the projection of the solution set

Table 1. Constraints and optimal operating conditions

	y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
upper bound				2.0	2.0	2.0	3.0	3.0	4.0	3.0	3.0
lower bound				-4.0	-2.0	-2.0	-3.0	-3.0	-3.0	-3.0	-3.0
optimization 1	1.0	0.5	70%	-1.59	-1.78	-1.03	-0.09	-0.77	0.22	-0.36	-0.35
optimization 2	1.5	0.7	75%	-2.68	-1.89	-1.16	0.02	-1.15	0.30	-0.37	-0.52
optimization 3	2.0	0.9	80%	-3.77	-2.00	-1.27	0.12	-1.53	0.38	-0.38	-0.69
optimization 4	3.0	2.0	100%	The desired quality cannot be achieved							
(achieved)	2.5	2.4	95%	2.00	2.00	0.42	-1.84	2.08	1.52	-1.12	2.29

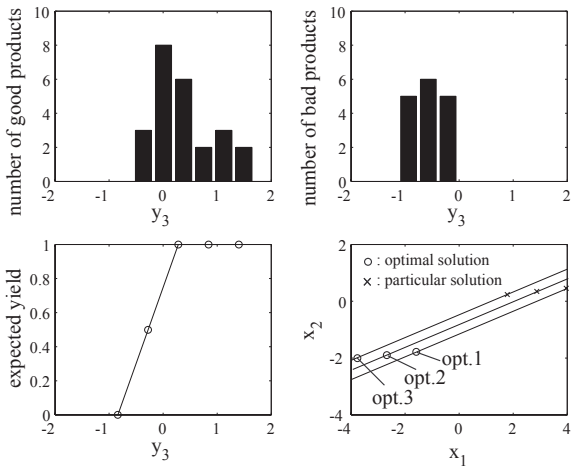

Fig. 2. Histogram of good products (left-top), histogram of bad products (right-top), expected yield (left-bottom), and derived operating conditions (right-bottom)

onto x_1-x_2 space. The derived optimal operating conditions exist inside the space spanned by the principal components and limited by the constraints, and they can achieve the desired product quality. In the last case (optimization 4), however, the desired quality cannot be achieved under constraints. Therefore, operating condition is optimized to realize as desired quality as possible.

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 and should be avoided. 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.

In practice, only a few manipulated variables can be changed simultaneously. It would be usually unacceptable to change all variables. 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. In DDQI, new operating conditions must exist in the space spanned by principal components. This constraint might need to be relaxed for searching the optimal operating condition in which only a few variables need to be changed. For this purpose, constraints on T^2 and Q may be used.

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, model building and validation are time-consuming and require expertise. To realize an online DDQI system as automatic process control, an efficient and robust procedure needs to be developed.

5. CONCLUSIONS

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, determine the operating conditions that can achieve the desired product quality, optimize operating condition under constraints, and evaluate the validity of the results. In addition, possible extensions and remaining problems are discussed.

REFERENCES

Jackson, J.E. and G.S. Mudholkar (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, **21**, 341–349.

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