Performance Map Construction for a Centrifugal Diffuser with Data Mining Techniques*

Koji SHIMOYAMA**, Kazuyuki SUGIMURA***, Shinkyu JEONG** and Shigeru OBayashi**

** Institute of Fluid Science, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan
E-mail: shimoyama@edge ifs tohoku.ac.jp
*** Hitachi Plant Technologies, Ltd.
603 Kandatsu-machi, Tsuchiura, Ibaraki 300-0013, Japan

Abstract
Performance maps, which represent relations between performance and geometry parameters, are essential for engineers to make a first decision on preliminary specification of a product to be designed. However, actual design often needs to consider various performance and geometry parameters simultaneously. Therefore, the resulting performance maps must be constructed in a high-dimensional form. Based on these requirements, this paper proposes and demonstrates performance map construction with the aid of data mining techniques. Data mining can reveal characteristic patterns in high-dimensional data with performance and geometry parameters. Therefore, the data mining results make it easy to interpret complex features of performance vs. geometry relations, and help engineers to discover new knowledge for engineering design through interpretation. The present demonstration of a centrifugal diffuser demonstrated that the data mining techniques are suitable and applicable to high-dimensional performance map construction, together with actual acquisition of new knowledge for diffuser design that was unknown from conventional quasi-one-dimensional nozzle theory.

Key words: Data Mining, Computational Fluid Dynamics, Performance Map, Centrifugal Diffuser, Nozzle Theory

1. Introduction
Recent progress in computer technology has brought about great developments in the field of simulation engineering. Simulation algorithms (formulation, modeling, schemes, etc.) are constantly improving in terms of efficiency, accuracy, and applicability. With such progress, numerical simulations can output various data, help engineers to discover new knowledge for actual engineering design, and promote further progress of simulation technologies themselves.

For knowledge acquisition, numerical simulations are expected to output as much data as possible. In actual situations, however, huge amounts of simulation data often lead to difficulties at the data analysis stage. A major issue in data analysis is high-dimensionality. For example, computational fluid dynamics (CFD) simulations output the data of various physical properties (density, pressure, velocity, etc.) calculated at different spatial locations and different time iterations, and then evaluate various performances regarding fluid dynamics (output power, energy loss, noise level, etc.). Especially for fluid machinery design, CFD simulations are performed for different body geometries (shape, scale, topology, etc.). Thus, the output data from numerical simulations contain combination sets of various performance and geometry parameter values in a high-dimensional form.
After the simulations, it is necessary for engineers to extract specific relations between performance and geometries from the simulation data. For two-dimensional or three-dimensional data, such relations can be visualized simply by plotting the data points. However, conventional visualization becomes virtually impossible when the data have more than three dimensions.

Recently, a solution for high-dimensional data analysis, commonly called data mining\(^{(1)}\), has attracted attention in various fields of engineering. Data mining has its origins in the field of information science (pattern recognition, artificial intelligence, etc.). This can rewrite characteristic features in complex data into different formats, which are easy to understand (numerals, images, texts, etc.). Therefore, data mining is expected to be useful to acquire new engineering knowledge from high-dimensional simulation data.

There have been a number of studies utilizing data mining techniques in engineering research topics as follows. Jeong \textit{et al.} have applied them to the data for aerospace design optimization\(^{(2)}\), and Sugimura used these techniques for turbomachinery design optimization\(^{(3)}\). Both of these studies demonstrated capabilities of the data mining techniques for design knowledge acquisition. In addition to engineering design topics, data mining techniques have been applied in the field of physics. Using data mining techniques, Obayashi \textit{et al.} analyzed unsteady CFD simulation data and interpreted complex unsteady phenomena in flow fields (interactions between shock waves and shear layer oscillations)\(^{(4)}\). In a similar way, Morizawa \textit{et al.} also examined the relations between blood flow pulsation and circulatory diseases\(^{(5)}\).

This paper discusses further applicability of data mining techniques to a fundamental topic in engineering research, i.e., the construction of performance maps that represent relations between performance and geometry parameters. Performance maps are often used to make a first decision on preliminary specification of a product to be designed. Therefore, performance map construction is an essential area in the field of engineering.

This study was performed with a centrifugal diffuser as the target product for performance map construction. The centrifugal diffuser is one of the components commonly used in various household appliances (air cleaners, vacuum cleaners, etc.) as well as heavy industrial machineries (aircraft engines, marine engines, etc.). Conventionally, diffuser performance has been evaluated based on the quasi-one-dimensional nozzle theory. Figure 1 shows an illustration of the performance map for a linear nozzle with rectangular cross-section, which mimics the original map in Ref. (6). This map represents the contour lines of the pressure recovery coefficient \(C_{pr}\) in the nozzle throat to exit section, which are plotted on the plane regarding two nozzle geometry parameters (aspect ratio \(\lambda\) and expansion ratio \(\epsilon\)). It reveals the \(C_{pr}\) vs. \(\lambda\) and \(C_{pr}\) vs. \(\epsilon\) relations, which are suitable for diffusers with flat blades. However, actual diffusers mostly consist of cambered blades, and actual diffuser performance seems to be affected by the blade geometries ignored in the quasi-one-dimensional nozzle theory (blade camber, blade attack angle to flow, etc.). In addition, diffuser hub and case geometries may also affect diffuser performance (e.g., Ref. (7) discusses the case geometry effects based on experiments). A further consideration is that actual diffusers should ensure good overall performance, i.e., air pressure must be recovered not only in the nozzle section but also upstream and downstream of the nozzle. Thus, the performance map should be constructed in a higher-dimensional form, which allows comprehension of the various relations among many performance parameters and many geometry parameters. Therefore, the centrifugal diffuser is an appropriate target product to validate the data mining techniques for high-dimensional performance map construction, as well as to provide useful knowledge about the relations between diffuser performance and geometries.

The performance studies for centrifugal diffusers have also been reported by other researchers. Reference (8) made experimental measurements of the internal flow field development within an impeller-diffuser-interacted stage by means of laser velocimeters. Reference (9) experimentally investigated simultaneous adjustments of inlet guide blades and dif-
fuser blades in centrifugal compressors for the improvements in both performance and operating range. Reference (10) proposed and demonstrated a wave augmented diffuser to reduce the loss caused by the discharge flow turning from radial/tangential to axial direction by numerical simulations. However, diffuser geometries considered in those studies were limited to flat blades (8), (10) or cambered blades parameterized simply by the angle of attack (9). Although Ref. (11) compared and discussed the performance among three different diffusers (wedge, symmetric airfoil, and cambered airfoil) by numerical simulations, it still had lack in the varieties of diffuser geometries to be considered for performance map construction. Recently in Ref. (12), a three-dimensional aerofoil diffuser geometry, which allows spanwise variations in solidity, stagger, and lean angles, has been developed for industrial centrifugal compressor stages based on both numerical and experimental analyses, but its performance tendency has not been explained in a high-dimensional form of geometry parameters.

2. Centrifugal Diffuser

2.1. Whole Geometry Definition

Figure 2 shows the whole geometries of the centrifugal diffuser considered in this study. This diffuser has 13 similar cambered blades of constant thickness. The leading and trailing edges of these blades are linear and parallel to the diffuser center axis. The size of the blade is determined by $D_3$ and $D_4$, while that of the diffuser case is determined by $D_5$ and $D_{ex}$. In addition, the blade camber angle has a linear profile between the leading and trailing edges ($\beta_3$ and $\beta_4$). For simplicity, $D_3$ was fixed in this study, and the values of $D_4/D_3$, $D_5/D_4$, $D_{ex}/D_5$, $\beta_3$, and $\beta_4$ were treated as five independent variables for whole geometry definition of the diffuser.

2.2. Performance Quantification and Evaluation

The centrifugal diffuser should work so efficiently that it can to decelerate internal air flow without pressure loss. In general, such performance can be quantified by a pressure recovery coefficient, such that a larger value of the coefficient leads to more efficient air deceleration. For general discussions on diffuser performance, this study focuses on three pressure recovery coefficients in different sections as performance functions: $C_{prin-3}$ for inlet to blade entrance section, $C_{prin-4}$ for inlet to blade exit section, $C_{prin-5}$ for inlet to corner section, which are defined as

![Fig. 1 Performance map for a linear nozzle with rectangular cross-section](image)
In the above equations, \( p, \rho, \) and \( C \) are static pressure, density, and absolute velocity of internal air flow, respectively. The subscripts \( \text{in}, 3, 4, \) and \( 5 \) denote inlet, blade entrance, blade exit, and corner, respectively.

This study was performed to evaluate the values of \( C_{p_{\text{in}}^3}, C_{p_{\text{in}}^4}, \) and \( C_{p_{\text{in}}^5}, \) which were obtained at a constant mass flow rate, from computational fluid dynamics (CFD) simulations using the commercial software STAR-CD \(^{13}\). The simulation domain is set in one pitch, as shown in Fig. 2, and employs cyclic boundary conditions. Figure 3 shows an example of the present CFD simulation. The mesh size is \( 133 \times 37 \times 29 = 142,709 \), when mesh convergence has been confirmed successfully. The governing equations are the Reynolds-averaged Navier-Stokes (RANS) equations for turbulent compressible air. The turbulence model is the high-Reynolds-number \( k-\epsilon \) model. The semi-implicit method for pressure-linked equations (SIMPLE method) solves the governing equations, and the monotone advection and reconstruction scheme (MARS scheme) realizes second-order spatial discretization.

Although this solver does not have the validation data corresponding to the sole diffuser, this has already been validated for a turbomachinery consisting of centrifugal impeller and diffuser in Ref. (3); the CFD simulations using this solver showed a qualitative agreements with overall efficiency vs. flow rate curves, which were obtained from the experiments. In addition, the present CFD simulations assumed the inlet mass flow rate to be constant and the inlet velocity distribution to be uniform. Such settings are unrealistic but ideal as to construct performance maps used for preliminary design of a centrifugal diffuser.

3. Performance Map Construction Procedure

3.1. Performance Sampling

Three diffuser performance function values \( (C_{p_{\text{in}}^3}, C_{p_{\text{in}}^4}, \) and \( C_{p_{\text{in}}^5}) \) are sampled in a whole space regarding five geometry variables \( (D_4/D_3, D_5/D_4, D_{ex}/D_3, \beta_3, \) and \( \beta_4) \). This study generated 100 sample points through Latin hypercube sampling (LHS) \(^{14}\) in the five-variable space, and evaluated the three performance function values from the CFD simulations.
for 100 diffusers with different shapes, each of which corresponds to one of the sample points. Consequently, in the present implementation, the performance values were successfully obtained at 85 sample points, while the CFD simulations have illegally fallen into divergence at the rest.

As shown in Fig. 4, the LHS divides the whole \( n \)-variable space into \( N^n \) subspaces, where the range of each variable is divided into \( N \) equally spaced subranges. Then, it selects \( N \) subspaces at random, each of which satisfies the orthogonality condition. This condition indicates that the selected subspaces do not have overlapping subranges regarding all the variables, i.e., each subrange of one variable is combined with only one subrange of each other variable. The LHS can comprehend the whole variable space even with a smaller sample size, compared to simple sampling without considering the orthogonality condition.

3.2. Data Mining

The previous stage described in Section 3.1 obtains 85-point sample data regarding three performance functions of five geometry variables. To construct performance maps in a comprehensive manner, the next stage analyzes these data using several data mining techniques. To ensure reliability and variety of the information discovered from the results, this study employed two techniques, analysis of variance (ANOVA)(15)(16) and self-organizing maps (SOM)(17) for quantitative and qualitative analyses, respectively.

3.2.1. Analysis of Variance (ANOVA)  
ANOVA is a technique that is commonly used in statistical analysis. ANOVA can quantify the contributions of each variable and the contribution of combinations of variables to functions.

Consider a set of sample data regarding a function \( f(x) \) of \( n \) variables \( x = [x_1, x_2, \cdots, x_n]^T \)
Another data mining technique, SOM, described in this section can overcome such difficulties. Consider a set of sample data with \( m \) functions \([f_1(x), f_2(x), \ldots, f_m(x)]\) of \( n \) variables \( x = [x_1, x_2, \ldots, x_n]^T \) as the target vector dataset to be analyzed. For \( f(x) \), a total variance \( S \) is defined in an integral form as

\[
S = \sigma^2 = \int \cdots \int [f(x_1, \ldots, x_n) - \mu]^2 \, dx_1 \cdots dx_n
\]  

(2)

where \( \mu \) is a total mean of \( f(x) \):

\[
\mu = \int \cdots \int f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n
\]  

(3)

Based on \( \mu \), the contribution of one variable \( x_i \) (\( i = 1, 2, \ldots, n \)) to \( f(x) \) is formulated. Such an effect is called the main effect \( S_i \), which is defined as

\[
S_i = \int |\sigma_i(x_i)|^2 \, dx_i
\]  

(4a)

\( \sigma_i(x_i) \) is a functional form of the main effect defined as

\[
\sigma_i(x_i) = \mu_i(x_i) - \mu
\]  

(4b)

where \( \mu_i(x_i) \) is the mean of \( f(x) \) considering the variables except for \( x_i \):

\[
\mu_i(x_i) = \int \cdots \int f(x_1, \ldots, x_n) \, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n
\]  

(5)

Next, the contribution of a combination of two variables \( x_i \) and \( x_j \) (\( i, j = 1, 2, \ldots, n \) and \( i \neq j \)) to \( f(x) \) is formulated: the (two-way) interaction effect \( S_{ij} \), which is defined as

\[
S_{ij} = \iint |\sigma_{ij}(x_i, x_j)|^2 \, dx_i dx_j
\]  

(6a)

and its functional form \( \sigma_{ij}(x_i, x_j) \) as

\[
\sigma_{ij}(x_i, x_j) = \mu_{ij}(x_i, x_j) - \mu(x_i) - \mu(x_j) + \mu
\]  

(6b)

where

\[
\mu_{ij}(x_i, x_j) = \int \cdots \int f(x_1, \ldots, x_n) \, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_{j-1} dx_{j+1} \cdots dx_n
\]  

(7)

Furthermore, higher-order interaction effects can be also formulated in a similar way. Note that the following equation is satisfied in ANOVA:

\[
S = \sum_{i=1}^n S_i + \sum_{i, j=1}^n S_{ij} + \sum \text{(higher-order interaction effects)}
\]  

(8)

The results of ANOVA can reveal quantitative features in the sample data as follows. Larger \( S_i \) (\( i = 1, 2, \ldots, n \)) indicates that the value of a function \( f(x) \) seems to change more drastically when the value of a variable \( x_i \) changes, and larger \( S_{ij} \) (\( i, j = 1, 2, \ldots, n \) and \( i \neq j \)) does so when both \( x_i \) and \( x_j \) values change simultaneously. In addition, functionally-formed results represent the spatial distributions of positive and/or negative contributions. A function value \( f(x) \) seems to increase/decrease at a value of a variable \( x_i \) with larger/smaller \( \sigma_i(x_i) \) (\( i = 1, 2, \ldots, n \)), and at a combination of \( x_i \) and \( x_j \) values with larger/smaller \( \sigma_{ij}(x_i, x_j) \) (\( i, j = 1, 2, \ldots, n \) and \( i \neq j \)).

3.2.2. Self-Organizing Map (SOM)  

The ANOVA described in Section 3.2.1 examines the variable contributions to a single function. Therefore, in cases with multiple functions, the ANOVA must be implemented in a separate way for each of the functions. Thus, the ANOVA has difficulty in examining the relations among multiple functions (e.g., trade-offs). Another data mining technique, SOM, described in this section can overcome such difficulties.

Figure 5 shows an illustration of a SOM model. The SOM is a feedforward-type neural network model with unsupervised machine learning. Consider a set of sample data with \( m \) functions \([f_1(x), f_2(x), \ldots, f_m(x)]\) of \( n \) variables \( x = [x_1, x_2, \ldots, x_n]^T \) as the target vector dataset to be analyzed. First, the SOM assigns the dataset to each neuron in the input layer. On the other hand, the output layer has neurons with a two-dimensional hexagonal mesh topology, and \( m \)-dimensional weight vectors \([w_1, w_2, \ldots, w_m]\), which are randomly assigned to
each neuron. Note that the weight vectors have the same $m$ dimensions as the input data. Then, the output layer searches for the best-matching unit, which is the neuron with a closer weight vector to each data point in the input layer. During the learning process of the network, the weight vectors of the best-matching unit and its neighboring neurons are adjusted to become much closer to the input data, and as a result, sequences of closely related data in the original $m$-dimensional function space are projected onto a sequence of neighboring neurons in the two-dimensional hexagonal mesh plane.

![Self-organizing map](image)

Then, the mesh plane is colored by each function value $f_i(x)$ ($i = 1, 2, \cdots, m$). Comparison of the color patterns of $f_1(x), f_2(x), \cdots, f_m(x)$ on the mesh planes reveals various correlations among the functions. In addition, comparison with the color patterns of each variable value $x_i$ ($i = 1, 2, \cdots, n$) on the mesh plane reveals the relations between the functions and variables (a detailed description of SOM is given in Ref. (2)).

Thus, the SOM enables originally high-dimensional function/variable-combined data to be visualized in a low-dimensional form while preserving their own features. Note that SOM is different from conventional geographical maps because it cannot define any geographical features, such as coordinates and direction, except for closeness. This study used the commercial SOM software Viscovery®SOMine in the actual implementation.

4. Data Mining Results based on Whole Geometries

This section presents the ANOVA-based and SOM-based data mining results for the 85-point LHS data regarding three performance functions ($C_{prin-3}$, $C_{prin-4}$, and $C_{prin-5}$) of five whole geometry variables ($D_4/D_3$, $D_5/D_4$, $D_{ex}/D_{5}$, $\beta_3$, and $\beta_4$).

4.1. ANOVA Results

Figure 6 shows the breakdowns of main and interaction effects ($S_i$ and $S_{ij}$ defined by Eqs. 4a and 6a, respectively) for each performance function. Note that these results exclude subtle effects with less than 1% of the total variance ($S$ defined by Eq. 2). For $C_{prin-3}$, only the variable $\beta_3$ has a major contribution corresponding to more than 90% of the total. On the other hand, the breakdowns seem similar for $C_{prin-4}$ and $C_{prin-5}$; $\beta_3$ has the largest contribution, with $\beta_3$ and $\beta_4$ also making slight contributions in an interactive manner. These results suggest that the three performance functions can be categorized into two types with different contribution features.

4.2. SOM Results

Figure 7 shows the SOM color images, each of which is colored according to each performance function or geometry variable value. Here, note that a given LHS point maintains the same position in all SOMs. Figure 7(a) is red (good $C_{prin-3}$ to be maximized) on the lower-right, while both Figs. 7(b) and 7(c) are blue (bad $C_{prin-4}$ and $C_{prin-5}$ to be maximized).
at this position. These observations indicate trade-off relations between $C_{prin-3}$ and $C_{prin-4}$ improvements, and between $C_{prin-3}$ and $C_{prin-5}$ improvements. $C_{prin-4}$ and $C_{prin-5}$ are improved simultaneously because they have maximum values at the same position on the SOMs, as shown in Figs. 7(b) and 7(c).

For the geometry variables, Fig. 7(g) shows almost the reverse color pattern of Fig. 7(a), i.e., smaller $\beta_3$ leads to larger $C_{prin-3}$. Trends of the other variables are indistinct because these SOMs have spotted color patterns. These results were consistent with the ANOVA results described in Section 4.1.

As described above, the data mining results can show only a single rule between $C_{prin-3}$ and $\beta_3$ at present. This is still insufficient and insignificant as performance vs. geometry information. Therefore, an additional implementation of data mining focusing on the details of the geometries is needed to construct useful performance maps for a centrifugal diffuser.

5. Retrial of Data Mining

5.1. Nozzle Geometry Definition

Next, data mining considers the geometries in the nozzle part, as illustrated in Fig. 8. $h_t$ and $h_e$ are the nozzle widths at throat and exit, respectively. $l_p$ and $l_s$ are the nozzle lengths at pressure and suction sides, respectively, and the average of these values $l_{avg} = (l_p + l_s)/2$ is also considered as the midway nozzle length. $\theta$ is the deflection angle through the nozzle. $\alpha$ is the attack angle of the flow at the inlet. In addition to the geometries indicated in Fig. 8, nozzle curvatures $\kappa_p = \theta/l_p$, $\kappa_s = \theta/l_s$, and $\kappa_{avg} = \theta/l_{avg}$ are considered at the pressure side, suction side, and midway, respectively. Furthermore, two scaling ratios for the nozzle shape are considered: expansion ratio $\epsilon = h_e/h_t$ and aspect ratio $\lambda = l_{avg}/h_t$. Consequently, the number of nozzle geometry variables sums to 12 ($h_t$, $h_e$, $l_p$, $l_s$, $l_{avg}$, $\theta$, $\alpha$, $\kappa_p$, $\kappa_s$, $\kappa_{avg}$, $\epsilon$, and $\lambda$).

5.2. Performance Resampling

The values of twelve nozzle geometry variables were calculated for each of the 85-point...
Fig. 7  SOM results for whole geometries

Fig. 8  Nozzle geometries in a centrifugal diffuser
LHS data, which were sampled in the space regarding five whole geometry variables \((D_4/D_3, D_5/D_4, D_{ex}/D_5, \beta_3, \text{ and } \beta_4)\) as described in Section 3.1. With interpolation of these data, two performance functions \((C_{prin-3} \text{ and } C_{prin-4})\) were converted from the original five-variable space to the new twelve-variable space. For simplicity, \(C_{prin-5}\) was not considered because this function has the same trend to be maximized as \(C_{prin-4}\), which was described in Section 4.2. In other words, the performance map to be constructed for \(C_{prin-4}\) can be used as a substitute for \(C_{prin-5}\) performance map. Moreover, this matter also applies to the overall pressure recovery coefficient \(C_{prin-ex} = \frac{(p_{ex} - p_i)}{(\frac{1}{2}p_i C_{in}^2)}\), which is defined for inlet to outlet section.

Note that the converted functions are just the approximations of diffuser performance functions due to data interpolation, i.e., the function values may not agree with the real values exactly at the points where sample data are not given. This study adopted the Kriging model\(^{(16), (19)}\) for function interpolation because this model can adapt well to nonlinear functions even with a small sample size (a detailed description of the Kriging model is given in Ref. \((20))\). The present implementation resulted in accurate approximation of the performance functions with discrepancies within 0.04, which are far smaller than the maximum \(C_{prin-4}\) value of 0.61 (see the color range in Fig. 7(b)). On the approximated functions, 1,000 data points were resampled through the LHS in the twelve-variable space, and then analyzed using ANOVA and SOM.

6. Data Mining Results based on Nozzle Geometries

This section presents the ANOVA-based and SOM-based data mining results for the 1,000-point LHS data regarding two performance functions \((C_{prin-3} \text{ and } C_{prin-4})\) of twelve nozzle geometry variables \((h_t, h_e, l_p, l_e, l_{avg}, \theta, \alpha, \kappa_p, \kappa_s, \kappa_{avg}, \epsilon, \text{ and } \lambda)\).

6.1. ANOVA Results

Figure 9 shows the breakdowns of main and interaction effects \((S_i \text{ and } S_{ij})\) defined by Eqs. 4a and 6a, respectively) for each performance function. For \(C_{prin-3}\), only the variable \(h_t\) has a major contribution. This is equivalent to the \(C_{prin-3}\) vs. \(\beta_3\) contribution shown in Fig. 6(a), because larger \(\beta_3\) leads directly to smaller \(h_t\) at the throat. For \(C_{prin-4}\), various contributions of \(h_t, \kappa_p, \epsilon, \text{ and } \lambda\) were revealed. A significant difference was seen between the contributions of the whole geometries and those of the nozzle geometries; Figure 9(b) indicates the largest contribution of a combination of two variables, \(h_t \text{ and } \kappa_p\), while Fig. 6(b) shows a major contribution of the variable \(\beta_3\). As mentioned above, \(h_t \text{ and } \beta_3\) are associated with each other because both represent local geometries at the throat. On the other hand, \(\kappa_p\) represents the geometry of the blade, and thus \(\kappa_p\) is less relative to \(\beta_3\).

![Fig. 9 ANOVA results (breakdowns of main and interaction effects) based on nozzle geometries](image)

Figure 10 shows the functional-formed main and interaction effects \((\sigma_i(x_i) \text{ and } \sigma_{ij}(x_i, x_j))\), defined by Eqs. 4b and 6b, respectively) of the variables that showed large contributions in Fig. 9. Figure 10(a) shows the main effect distribution where \(h_t\) affects \(C_{prin-3}\) and indicates
that smaller $h_t$ leads to larger $C_{pr_{in-3}}$. Figure 10(b) shows the interaction effect contour plot where $h_t$ and $\kappa_p$ affect $C_{pr_{in-4}}$, and indicates that a combination of smaller $h_t$ and larger $\kappa_p$, and a combination of larger $h_t$ and smaller $\kappa_p$ lead to larger $C_{pr_{in-4}}$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig10}
\caption{(a) $h_t$ main effect on $C_{pr_{in-3}}$ (b) $h_t$ vs. $\kappa_p$ interaction effect on $C_{pr_{in-4}}$}
\end{figure}

The interaction effect of $h_t$ and $\kappa_p$ (Fig. 10(b)) is a remarkable issue to discuss diffuser performance because $\kappa_p$ is not considered in the quasi-one-dimensional nozzle theory (Fig. 1). Figure 10(b) specifies a large effect of the blade curvature on the diffuser performance. Therefore, the present results are useful as fundamental information to understand pressure recovery performance in a curved nozzle as well as a centrifugal diffuser.

6.2. SOM Results

Figure 11 shows the SOM color images, each of which is colored according to performance function or nozzle geometry variable value (only four nozzle geometries with large contributions are considered here). $C_{pr_{in-3}}$ becomes large in the left area on the SOM (denoted as "A" in Fig. 11(a)), while $h_t$ becomes small in this area. This relation is consistent with the main effect results from the ANOVA described in Section 6.1. On the other hand, $C_{pr_{in-4}}$ becomes large in the upper area on the SOM, which is denoted as a combination of "B" and "C" in Fig. 11(b). Area B has small $h_t$ and large $\kappa_p$, while area C has large $h_t$ and small $\kappa_p$. Therefore, each of areas B and C is also consistent with the interaction effect results from the ANOVA in Section 6.1. Furthermore, other higher-order interaction effects can also be determined by comparing the color patterns of all the SOM images.

The SOM can reveal various relations between many performance functions and many geometry variables by combining them into a single map. The SOM color images can serve as the high-dimensional performance maps themselves. Consequently, new performance maps with high dimensions can be constructed successfully. The present study confirmed the suitability and usefulness of the data mining techniques for performance map construction.

6.3. Discussions on the Data Mining Results

This section gives discussions on the data mining results, which were obtained by ANOVA and SOM, from the viewpoint of fluid dynamics. Now, four samples points (Samples I, II, III, and IV) were selected for the following comparisons and discussions. Table 1 compared the nozzle geometry variables ($h_t$ and $\kappa_p$) and pressure recovery coefficients ($C_{pr_{in-3}}$ and $C_{pr_{in-4}}$) among these four samples. Figure 12 compared the flow field structures among these samples; contour colors represent static pressure and white lines correspond to the locations where radial velocity $= 0$, i.e., the regions closed by the white lines involve reversed flows.

First, $h_t$ main effect on $C_{pr_{in-3}}$, as seen in Fig. 10(a), appears due to the flow blockage at
Fig. 11  SOM results for nozzle geometries


Table 1  A comparison of nozzle geometry variables and pressure recovery coefficients among four samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Nozzle geometry variables</th>
<th>Pressure recovery coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_t$ [mm]</td>
<td>$\kappa_p$ [rad/m]</td>
</tr>
<tr>
<td>I</td>
<td>3.9</td>
<td>17</td>
</tr>
<tr>
<td>II</td>
<td>4.1</td>
<td>20</td>
</tr>
<tr>
<td>III</td>
<td>2.6</td>
<td>17</td>
</tr>
<tr>
<td>IV</td>
<td>2.5</td>
<td>20</td>
</tr>
</tbody>
</table>

Fig. 12  A comparison of flow field structures among four samples (contour colors: static pressure, white lines: radial velocity $= 0$)

blade entrance. A comparison of the pressure contour plots for Samples I and III (Figs. 12(a) and 12(c)) indicates that, as $h_t$ becomes smaller, a nozzle throat is apt to dam up and compress the air flow coming from inlet more greatly. Eventually, a smaller value of $h_t$ surely leads to a larger value of $C_{pr_{in-3}}$.

Next, $h_t$ vs. $\kappa_p$ interaction effect on $C_{pr_{in-4}}$, as seen in Fig. 10(b), is mainly associated with flow separation in a nozzle. A nozzle with larger $\kappa_p$ is apt to cause flow separation and secondary flows in the downstream of the throat. Such a flow field involves the loss of air flow energy in blade entrance to exit section, i.e., smaller $C_{pr_{in-4}}$. This tendency applies to the nozzles with large $h_t$; e.g., Sample I without flow separation (Fig. 12(a)) has larger $C_{pr_{in-4}}$ than Sample II with flow separation (Fig. 12(b)).

In addition, note that $h_t$ vs. $\kappa_p$ interaction effect on $C_{pr_{in-4}}$ is also associated with an exception as stated below. The nozzles with smaller $h_t$ (e.g. Samples III and IV as shown in Figs. 12(c) and 12(d), respectively) show an opposite tendency, such that larger $\kappa_p$ leads to larger $C_{pr_{in-4}}$. Sample III has the blades, each of which stands very close, and thus the cross-sectional area in the nozzle part becomes almost constant, while Sample IV represents

a divergent nozzle. Therefore, Sample III cannot realize sufficient pressure recovery in blade entrance to exit section. In other words, Sample IV works better as a diffuser than Sample III although Sample IV involves huge flow separation.

7. Conclusions

This paper demonstrated the applicability of data mining techniques to high-dimensional performance map construction for a centrifugal diffuser. Two data mining techniques, analysis of variance (ANOVA) and self-organizing map (SOM), were employed to analyze the sample datasets obtained from computational fluid dynamics (CFD) simulations. ANOVA can be used for quantitative determination of the relations between performance functions and geometry variables, while SOM can do so qualitatively. Although the first trial of data mining based on whole diffuser geometries lacked relations to be obtained, the next trial based on local nozzle geometries successfully revealed various performance vs. geometry relations. The most remarkable observation was the blade curvature effect on pressure recovery, which cannot be derived from conventional quasi-one-dimensional nozzle theory. Therefore, this study confirmed that data mining techniques can be used to discover new engineering knowledge, and is suitable and applicable to performance map construction with high dimensions.

Finally, remember that the diffuser performance maps constructed in this study correspond to ideal conditions, which assumed the inlet mass flow rate to be constant and the inlet velocity distribution to be uniform. These maps seem usable for preliminary design, but still not usable for real-world design. Therefore, a future work should improve the performance maps by adding new parameters, which represent the nonuniformity of inlet flow conditions, to the map dimensions.

Acknowledgments

The first author was partially supported through a Grant-in-Aid for Young Scientists (B), No. 19760098, 2007, from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

(3) Sugimura, K., Design Optimization and Knowledge Mining for Turbomachinery, Ph.D. Dissertation, Tohoku University, Sendai, 2009.
(9) Simon, H., Wallmann, T., and Mönk, T., Improvements in Performance Characteristics


