14th AIAA/AHI Space Planes and Hypersonic Systems and Technologies Conference, 6–9 November 2006, Canberra, Australia

Knowledge Discovery in Aerodynamic Design Space for Flyback-Booster Wing Using Data Mining

Kazuhisa Chiba*
Japan Aerospace Exploration Agency, Tokyo 182-8522, Japan
Shinkyu Jeong† and Shigeru Obayashi‡
Tohoku University, Sendai 980-8577, Japan
and
Kazuomi Yamamoto§
Japan Aerospace Exploration Agency, Tokyo 182-8522, Japan

The data mining has been performed for the aerodynamic design optimization result of two-stage-to-orbit reusable launch vehicle flyback booster wing. Three data mining techniques were used such as self-organizing map, functional analysis of variance, and rough set theory. The optimization problem had four aerodynamic objective functions and 71 design variables regarding wing shape. The optimization obtained the result as the hypothetical design database with 302 all solutions including the 102 non-dominated solutions. Consequently, the knowledge in the design space was acquired regarding the correlation between objective functions, and the influence of the design variables to the objective function, for non-dominated and all evaluated solutions, respectively. The features of three data mining techniques were revealed. Although the combination among three techniques discovered detailed design knowledge, self-organizing map was especially a key technique for knowledge discovery. Moreover, design knowledge from all solutions conserved the information from non-dominated solutions. Data mining was essential to solve multi-objective optimization problem.

I. Introduction

Although design optimization problem is an important manner for engineering, the most significant point is the extraction of the knowledge in design space from optimization result. The result obtained by multi-objective optimization problem using evolutionary algorithm is not a sole solution but a set of optimum solutions. That is, as multi-objective optimization result is only figure enumeration, there is insufficient information regarding design. However, that set of optimum solutions can be considered hypothetical design database. Recently, data mining technique is applied for this hypothetical design database to obtain the fruitful design knowledge efficiently[1–3]. As data mining application is developing field, there is no effective manner.

In this study, three data mining techniques as self-organizing map (SOM), functional analysis of variance (ANOVA), and rough set theory were applied to the aerodynamic design optimization result regarding a two-stage-to-orbit (TSTO) reusable launch vehicle (RLV) flyback booster. A space transport system with a substantially reduced cost is needed so that space can be utilized in many more fields. One of the focused research is the RLV system, suggested as a replacement for the present expendable launch vehicle system. Because of the difficult assignments such as a higher performance propulsion system and greater reduction of its structure weight, current proposals for the introduction of reusable components in space transportation involve the TSTO configuration with winged flyback booster powered by liquid rocket engines for vertical takeoff and horizontal landing. As the wing geometry of flyback booster generates the

*Project Researcher, Aviation Program Group. Member AIAA.
†Research Associate, Institute of Fluid Science. Member AIAA.
‡Professor, Institute of Fluid Science. Associate Fellow AIAA.
§Section Leader, Aviation Program Group. Senior Member AIAA.

aerodynamic performance, it is the most important element. Therefore, the correlations among aerodynamic characteristics, such as lift, drag, and moment, are significant design information. Moreover, it is important to find design variables sensitive to the aerodynamic performance; that is, acquisition of knowledge in the design space is essential to improve the aerodynamic performance of winged flyback booster. The objective of this study is the acquisition of beneficial knowledge in the design space to apply data mining which is an emerging area of computational intelligence.

II. Design Optimization Problem

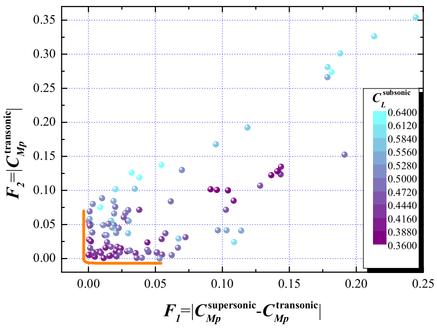
The reference mission of the TSTO RLV was to transport a 10-t payload into low earth orbit, similarly to the present H-IIA mission. The booster sizing was obtained by preliminary computation using the empirical equations developed by the Japan Aerospace Exploration Agency. The fuselage geometry was fixed to a given size and only the wing shape was allowed to be optimized in the present parameterization system.

Trajectory analysis around a typical TSTO configuration based on the present mission showed that the separation of the booster and orbiter took place at roughly Mach 3. Then the flyback booster turned over, slowed down, cruised at transonic speeds, and landed at subsonic speed. Note that the major part of its cross-range was in the transonic region. The considered four objective functions were to minimize the shift of aerodynamic center between supersonic and transonic conditions (F_1), transonic C_{Mp} (F_2) and transonic C_D (F_3), as well as to maximize subsonic C_L (F_4).

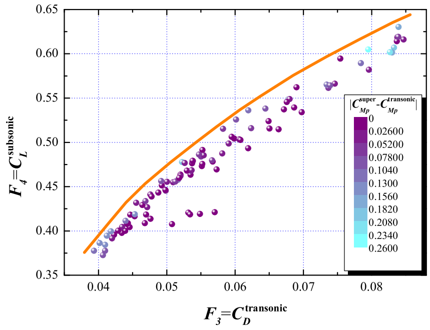
The three-dimensional Reynolds-averaged Navier-Stokes computation using the modified Spalart-Allmaras one-equation model on unstructured hybrid mesh was employed in aerodynamic evaluation. Adaptive range multi-objective genetic algorithm (ARMOGA) was used as an optimizer.

The design variables were related to planform, airfoil shape, wing twist, and position relative to the fuselage. A wing planform was determined by five design variables. A kink was place on the leading edge. Airfoil shapes were defined at the wing root, kink, and tip using thickness distributions and camber lines. The thickness distributions were described by Bézier curves using 11 control points and linearly interpolated in the spanwise direction. The camber line distributions were parameterized using Bézier curves with four control points and incorporated linearly in the spanwise direction. Wing twist was refined using B-splines with six control points. The position of the wing root relative to the fuselage was parameterized by x and z coordinates of the leading edge, angle of attack, and dihedral angle. The entire wing shape was thus defined using 71 design variables. Once a wing was defined, the junction line between the wing and fuselage was extracted and the final wing-fuselage geometry was derived by neglecting part of the wing inside the fuselage.

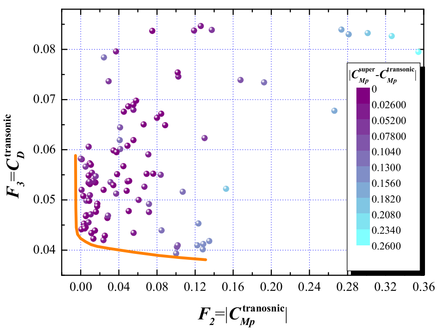
As the population size was set to eight. The total evolutionary computation of 40 generations was performed. Consequently, a total of 102 non-dominated solutions extracted from 302 all solutions of 40 generations were obtained. Figure 1 shows the two two-dimensional projections of the non-dominated solutions to better understand the tradeoffs among the four objective functions. The optimum values of F_1 and F_2 are zero, and the non-dominated solutions reached the origin, *i.e.*, the optimum values of F_1 and F_2 in Fig. 1a. As the plots in Fig. 1a are the non-dominated solutions for not two but four objective functions, there is a tradeoff surface spread of that the non-dominated solutions near the origin. Figure 1a shows that there is no tradeoff between the shift of aerodynamic center and the transonic C_{Mp}. The Pareto front can be seen clearly between F_3 and F_4 in Fig. 1b. Thus, Figure 1b indicates that there is a marked tradeoff between the transonic drag and the subsonic lift. This result indicated that a high lift device may be needed for an RLV booster for landing, similar to aircraft. The Pareto front for F_2 attained the optimum front. However, the Pareto front for F_3 did not reach C_D of zero. Thus, there is a slight tradeoff between F_2 and F_3, and the transonic drag can be improved while the transonic C_{Mp} increases. Moreover, the Pareto front for F_1 attained the optimum front. However, the Pareto front for F_4 did not have an apparent limit. Therefore, there was a slight tradeoff between F_1 and F_4. This indicated that shift and transonic C_{Mp} optimized simultaneously, while the subsonic lift was reduced slightly[4].



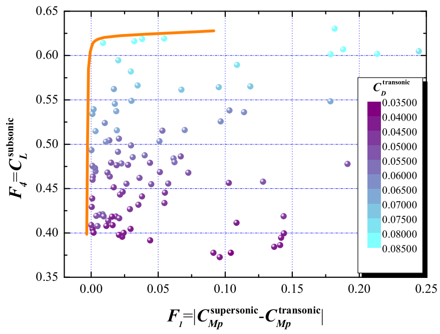
(a) between F_1 and F_2



(b) between F_3 and F_4



(c) between F_2 and F_3



(d) between F_1 and F_4

Figure 1. Derived non-dominated solutions on two-dimensional plane.

III. Data Mining Techniques

A. Self-Organizing Map

1. General SOM Algorithm

SOM[5] is an unsupervised learning, nonlinear projection algorithm from high to low-dimensional space. This projection is based on self-organization of a low-dimensional array of neurons. In the projection algorithm, the weights between the input vector and the array of neurons are adjusted to represent features of the high dimensional data on the low-dimensional map. The close two patterns are in the original space, the closer is the response of two neighboring neurons in the low-dimensional space. Thus, SOM reduces the dimension of input data while preserving their features.

2. Batch-SOM

In this study, SOMs were generated by using commercial software Viscovery® SOMine 4.0 Plus[6] produced by Eudaptics GmbH. Although SOMine is based on the general SOM concept and algorithm, it employs an advanced variant of unsupervised neural networks, *i.e.*, Kohonen's Batch SOM[7,8]. The algorithm consists of two steps that are iterated until no more significant changes occur: search of the best-matching unit c_i for all input data $\{\boldsymbol{x}_i\}$ and adjustment of weight vector $\{\boldsymbol{m}_j\}$ near the best-matching unit. The Batch-SOM algorithm can be formulated as follows:

$$c_i = \arg\min \|\boldsymbol{x}_i - \boldsymbol{m}_j\| \tag{1a}$$

$$\boldsymbol{m}_j^* = \frac{\sum_i h_{jc_i}\boldsymbol{x}_i}{\sum_i h_{jc_i}} \tag{1b}$$

where $\boldsymbol{m}_j^*$ is the adjusted weight vector. The neighborhood relationship between two neurons j and k is defined by the following Gaussian-like function:

$$h_{jk} = \exp\left(-\frac{d_{jk}^2}{r_t^2}\right) \tag{2}$$

where d_{jk} denotes the Euclidean distance between the neuron j and the neuron k on the map, and r_t denotes the neighborhood radius which is decreased with the iteration steps t.

The standard Kohonen algorithm adjusts the weight vector after all each record is read and matched. On the contrary, the Batch-SOM takes a 'batch' of data (typically all records), and performs a 'collected' adjustment of the weight vectors after all records have been matched. This is much like 'epoch' learning in supervised neural networks. The Batch-SOM is a more robust approach, since it mediated over a large number of learning steps. In the SOMine, the uniqueness of the map is ensured by the adoption of the Batch-SOM and the linear initialization for input data. Much like some other SOMs[9], SOMine creates a map in a two-dimensional hexagonal grid. Starting from numerical, multivariate data, the nodes on the grid gradually adapt to the intrinsic shape of the data distribution can be read off from the emerging map on the grid. The trained SOM is systematically converted into visual information[10, 11].

B. Analysis of Variance

ANOVA[12] is one of the data mining techniques showing the effect of each design variable to the objective and the constraint functions in a quantitative manner. ANOVA uses the variance of the model due to the design variables on the approximation function. By decomposing the total variance of model into the variance due to each design variable, the influence of each design variable on the objective function can be calculated. The decomposition is accomplished by integrating out the variables of model $\hat{y}$. The total mean ($\hat{\mu}_{\text{total}}$) and the variance ($\hat{\sigma}^2_{\text{total}}$) of model $\hat{y}$ are as follows:

$$\hat{\mu}_{\text{total}} \equiv \int \cdots \int \hat{y}\left(x_1, x_2, \cdots, x_n\right) dx_1 dx_2 \cdots dx_n \tag{3a}$$

$$\hat{\sigma}^2_{\text{total}} = \int \cdots \int \left[\hat{y}\left(x_1, x_2, \cdots, x_n\right) - \hat{\mu}_{\text{total}}\right]^2 dx_1 dx_2 \cdots dx_n \tag{3b}$$

The main effect of variable x_i and the two-way interaction effect of variable x_i and x_j are given as follows:

$$\begin{aligned} &\hat{\mu}\left(x_i\right) \\ &\equiv \int \cdots \int \hat{y}\left(x_1, x_2, \cdots, x_n\right) dx_1 dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \\ &- \hat{\mu}_{\text{total}} \end{aligned} \tag{4}$$

$$\begin{aligned} &\hat{\mu}_{i,j}\left(x_i, x_j\right) \\ &\equiv \int \cdots \int \hat{y}\left(x_1, x_2, \cdots, x_n\right) dx_1 dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_{j-1} dx_{j+1} \cdots dx_n \\ &- \hat{\mu}_i\left(x_i\right) - \hat{\mu}_j\left(x_j\right) - \hat{\mu}_{\text{total}} \end{aligned} \tag{5}$$

$\hat{\mu}(x_i)$ and $\hat{\mu}_{i,j}(x_i, x_j)$ quantify the effect of variable x_i and interaction effect of x_i and x_j on the objective function. The variance due to the design variable x_i is obtained as follows:

$$\hat{\sigma}^2_{x_i} = \int \left[\hat{\mu}_i\left(x_i\right)\right]^2 dx_i \tag{6}$$

The proportion of the variance P due to design variable x_i to total variance of model can be expressed by dividing Eq. (6) with Eq. (3b).

$$\begin{aligned} P &= \frac{\hat{\sigma}^2_{x_i}}{\hat{\sigma}^2_{\text{total}}} \\ &= \frac{\int \left[\hat{\mu}_i\left(x_i\right)\right]^2 dx_i}{\int \cdots \int \left[\hat{y}\left(x_1, x_2, \cdots, x_n\right) - \hat{\mu}_{\text{total}}\right]^2 dx_1 dx_2 \cdots dx_n} \end{aligned} \tag{7}$$

This value indicates[13] the effect of design variable x_i on the objective function $\hat{y}$.

C. Rough Set Theory

The rough set theory introduced by Pawlak[14,15] is based on the assumption that data and information is associated with every objects of the universe of discourse[16]. Objects described by the same properly selected information are indiscernible. The rough set theory can be used for a) reduction of data sets; b) finding hidden data patterns; c) generation of decision rules. The rough set theory algorithms fall into a broad area of machine learning such as a) neural networks; b) genetic algorithms; c) case-based learning; d) rule induction; e) analytical learning.

A *reduct* is a minimal sufficient subset of features RED $\subseteq A$ such that[16,17]:

a) $R(\text{RED}) = R(A)$, *i.e.*, RED produces the same classification of objects as the collection A of all features;

b) for any feature $f \in \text{RED}$, $R(\text{RED} - \{f\}) \neq R(A)$, *i.e.*, a reduct is a minimal subset with respect to the property a);

Core is the collection of features appearing in all reducts and is computed as the product of all reducts.

Pawlak introduced the concept of lower and upper approximations, which are useful for measuring of the quality and accuracy of classification. Denote U a finite set of objects, Q as a finite set of features, and let $P \subseteq Q$ and $Y \subseteq U$.

The *P-lower approximation* of Y, denoted as $\underline{P}Y$, is the set of all elements of U, which can be certainly classified as elements of Y based on the set of features P.

The *P-upper approximation* of Y, denoted as $\overline{P}Y$, is the set of elements of U, which can be possibly classified as elements of Y based on the set of features P. The two definitions are expressed formally as follows:

$$\underline{P}Y = \bigcup X\{X \in P^{*} \text{and} X \subseteq Y\} \tag{8a}$$

$$\overline{P}Y = \bigcup X\{X \in P^{*} \text{and} X \cap Y \neq \emptyset\} \tag{8b}$$

where P^{*} is the family of all equivalence classes of indiscernibility relation P^{r} on the set U. Two objects x and y are indiscernible on the set of features $P(xP^{r}y)$ if $r(x,q) = r(y,q)$ for every $q \in P$.

Equivalence classes of P^{r} are called P-elementary sets in the set of objects (data set).

Approximation accuracy (AA) of a data set is the ratio of the total lower approximation for all decision classes and the total upper approximation for all decision classes.

Boundary approximation is the difference between the upper and lower approximation.

Classification accuracy (CA) of a rule set is the ratio of the number of correctly classified objects from the test set and all objects in the test set[18].

Classification quality (CQ) of a feature set is the ratio of the number of objects in the lower approximation and the total number of objects in the data set.

In some areas, a broader definition of accuracy is used[19]. Accuracy is defined as the total number of true positives added to the total number of true negatives divided by the total number of patients studies[20], *i.e.*, accuracy= $(A + D)/(A + B + C + D)$). Based on the quadrant in Fig. 2, the following metrics are defined in addition to accuracy:

- *Sensitivity* (true positive rate)= $A/(A + C)$.
- *Specificity* (true negative rate)= $D/(B + D)$.
- *Positive predicted value*= $A/(A + B)$.
- *Negative predicted value*= $D/(C + D)$.

Exact rule= an outcome corresponds to one or more different conditions.

Approximate rule= the same condition corresponds to more than one outcome. Note that exact rules are generated for the set of objects in the lower approximation, while approximate rules are generated for the boundary.

The most basic definitions introduced above are illustrated with the data set in Fig. 3 containing six objects, four features, and the decision D. The classification quality of each single feature is as follows: CQ(F1)=.167, CQ(F2)=0, CQ(F3)=0, CQ(F4)=.333. For example, for feature F1 object 3 can be uniquely identified, therefore for F1=1, CQ(F1)=1/6=.167.

The classification quality of selected pairs of features is as follows: CQ(F1, F2)=.5, CQ(F2, F3)=.667. For example, for the feature set {F1, F2} three objects 1, 2, and 5 can be uniquely identified, therefore CQ(F1, F2)=3/6=.5.

The classification quality of selected triple features is as follows: CQ(F1, F2, F3)=.667, CQ(F2, F3, F4)=.667. For example, for the feature set {F1, F2, F3} four objects 1, 2, 3, and 5 can be uniquely identified, therefore CQ(F1, F2, F3)=4/6=.667.

The classification quality of all features is CQ(F1, F2, F3, F4)=.667.

Reducts: {F1, F4}, {F3, F4}, {F3, F4}

Core: { $\emptyset$ }

The classification quality of the core: 0

Number of decision classes: 3 (D=0, 1, 2 in Fig. 3)

Number of atoms: 5

Class $D = 0$
Number of objects: 3
Lower approximation: 2
Upper approximation: 4
Approximation accuracy: 0.5
Class $D = 1$
Number of objects: 1
Lower approximation: 1
Upper approximation: 1
Approximation accuracy: 1
Class $D = 2$
Number of objects: 2
Lower approximation: 1
Upper approximation: 3
Approximation accuracy: .333

Based on the above values for the data set in Fig. 3

- The classification quality of all features is (2+1+1)/6=.667, and
- The approximation accuracy is (2+1+1)/(4+1+3)=.5.

The knowledge extracted from a data set may follow different formats, with the most typical being decision tree, structured matrix, and decision rules. A typical format of a rule extracted from a data set is as follows:

> IF (Condition) THEN (Outcome) [Rule support, Rule coverage, Discrimination level] [List of supporting objects]

The following four exact and approximate rules have extracted from the set in Fig. 3.

In this study, ROSETTA[21,22] was used.

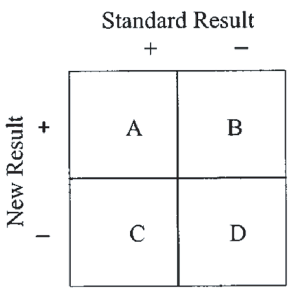


Figure 2. Classification quadrant.

	Features				
No.	F1	F2	F3	F4	D
1	0	1	Yes	2	0
2	0	0	Yes	3	1
3	1	1	No	2	2
4	0	0	No	1	0
5	0	1	Yes	0	0
6	0	0	No	1	2

Figure 3. Six object data set.

```
Exact rules
Rule 1. IF (F2 = 1) AND (F3 = Yes) THEN (D = 0); [2, 66.67%, 100.00%][1,5]
Rule 2. IF (F4 = 3) THEN (D = 1); [1, 100%, 100%][2]
Rule 3. IF (F1 = 1) THEN (D = 2); [1, 50%, 100%][3]

Approximate rule
Rule 4. IF (F4 = 1) THEN (D = 0) OR (D = 2); [2, 100%, 100%][4, 6]
```

Figure 4. Exact and approximate rules derived from the data set in Fig. 3[15].

IV. Data Mining Results

A. Design Space from Non-Dominated Solutions

The knowledge in the design space generated by the non-dominated solutions gives the correlation of tradeoff between the objective functions, and the influence of the design variables for tradeoffs. As tradeoff information is in the hypothetical design database formed by optimum solutions, data mining is the manner to give designers tradeoff and design knowledge directly.

1. Knowledge by SOM

The resulting 102 non-dominated solutions have been projected onto the two-dimensional map of SOM. Figure 5 shows the resulting SOM with 10 clusters taking the four objective functions into considering. Figure 6 shows the SOMs colored by the four objective values, respectively. This color figures show the SOM can be grouped as follows: Upper center area on SOM corresponds to the designs with the low shift of aerodynamic center. Upper right corner corresponds to the designs with the low shift of aerodynamic center, transonic C_{Mp}, and transonic C_D. Lower right corner corresponds to the designs with the low transonic C_D. Lower left corner corresponds to the high shift of aerodynamic center, transonic C_{Mp} and transonic C_D. Left center region corresponds to the high subsonic C_L.

The comparison between these colored SOMs reveals the tradeoffs and correlation between the objective functions. Figures 6a and 6b show high-value regions for the shift of aerodynamic center and the transonic C_D coincide with each other. That is, when one objective function is increased, another objective function is also increased strictly. Furthermore, because Fig. 6c is very similar to Fig. 6d, there is a severe tradeoff between transonic C_D and subsonic C_L. This knowledge from SOM corresponds to the results of Fig. 1b.

SOM can be also contoured by 71 design-variable values. Figure 7a shows the SOM colored by the design variable of the x coordinate of wing position to fuselage illustrated in Fig. 7b. Here, the x coordinate is held on the fuselage. Higher values are located in the lower left corner in Figs. 6a and 6b. High values of the shift of aerodynamic center, transonic C_{Mp}, and transonic C_D are clustered in this area. Thus, this means that the values of shift, transonic C_{Mp} and transonic C_D become poorer when the wing position is behind the fuselage.

The SOMs colored by the other characteristic design-variable values indicate the following knowledge. An individual with lower rearward camber height at the wing tip has lower transonic C_D. And, an individual with higher rearward camber height at the kink has higher subsonic C_L.

When the sweepback angle of the inboard wing becomes larger, the inboard wing acts as a strake. In general, as a strake generates a vortex, it may be effective to increase lift due to the leading-edge separation. However, the SOM colored by the design variable of the sweepback angle of the inboard wing shows that there are mixed values in the area in which high values of the subsonic lift clustered. That is, the sweepback angle of the inboard wing is not effective to increase C_L. The CFD visualization, which is one individual under subsonic flow condition where leading-edge separation is indicated by vortex centerlines, shows that the primary vortex occurs not from a strake but from a kink corner on the leading edge. Hence, the strake vortex is not essential to increase C_L. The knowledge for the CFD visualization is also obtained from SOM.

As the SOMs colored by several other design variables have jumbled coloring, there design variables had no effect in determining tradeoffs among the four objective functions. That is, it means that the sorting of the design variables can be also performed from SOM.

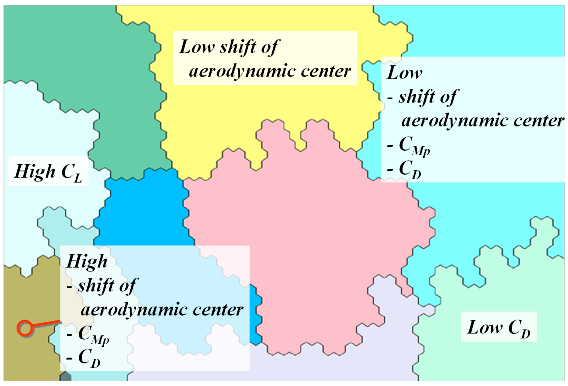
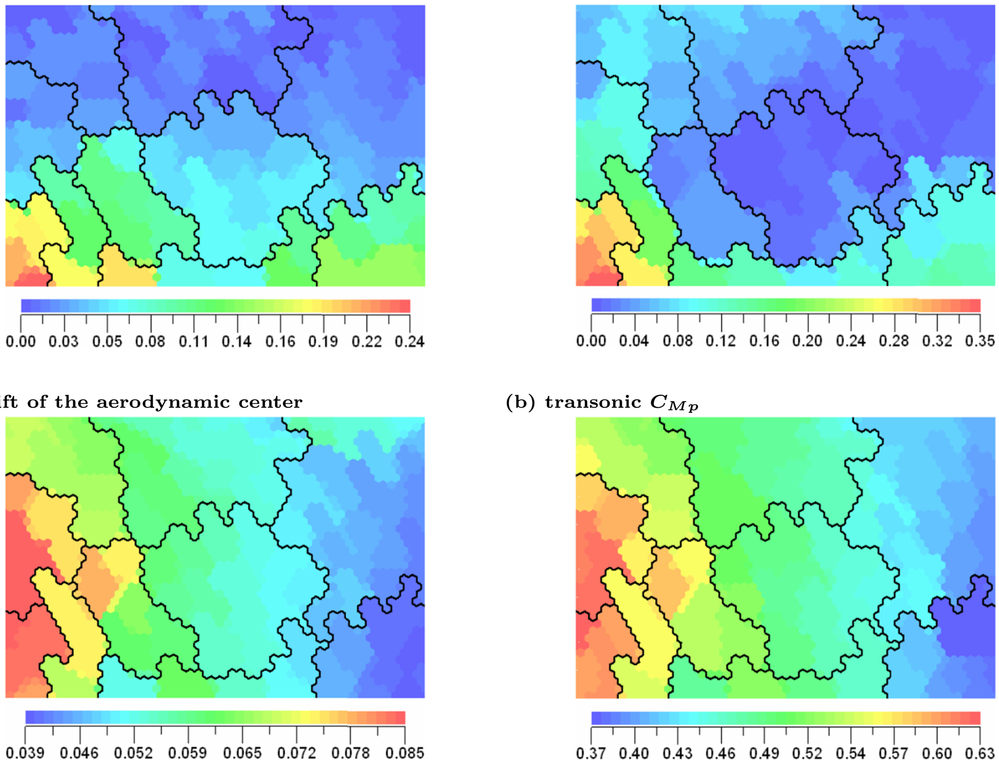


Figure 5. SOM of the non-dominated solutions in the four dimensional objective function space.



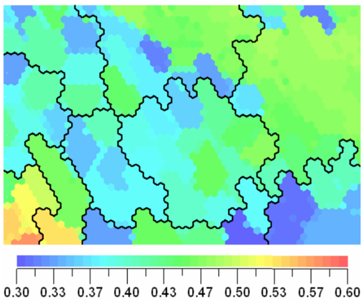
(a) shift of the aerodynamic center

(b) transonic C_{Mp}

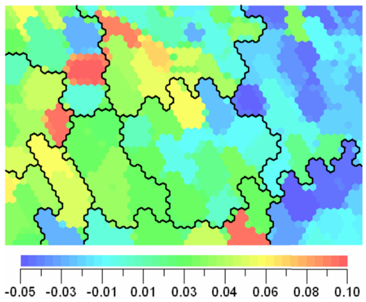
(c) transonic C_D

(d) subsonic C_L

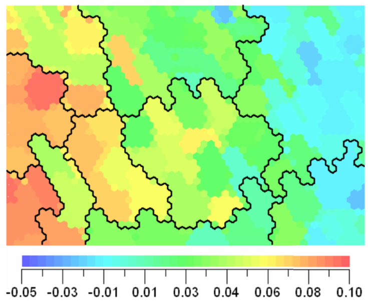
Figure 6. SOM colored by the four objective functions.



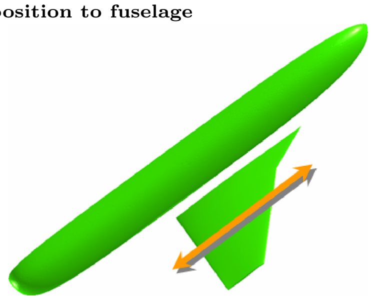
(a) dv7 as the x coordinate of wing position to fuselage



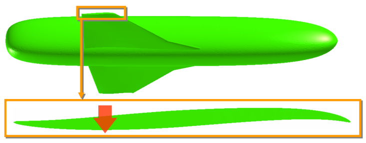
(b) dv22 as the camber height at tip rearward



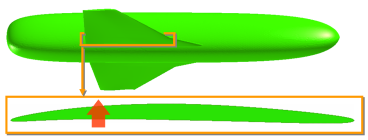
(c) dv18 as the camber height at kink rearward



(d) sketch of dv7



(e) sketch of dv22



(f) sketch of dv18

Figure 7. SOM colored by important design variables and their sketches.

2. Knowledge by ANOVA

The variance of the design variables and their interactions are shown in Fig. 8. The proportion of them are shown which is larger than 1% to the total variance. Note that 'dv' indicates design variable and '-' indicates interactions between two design variables.

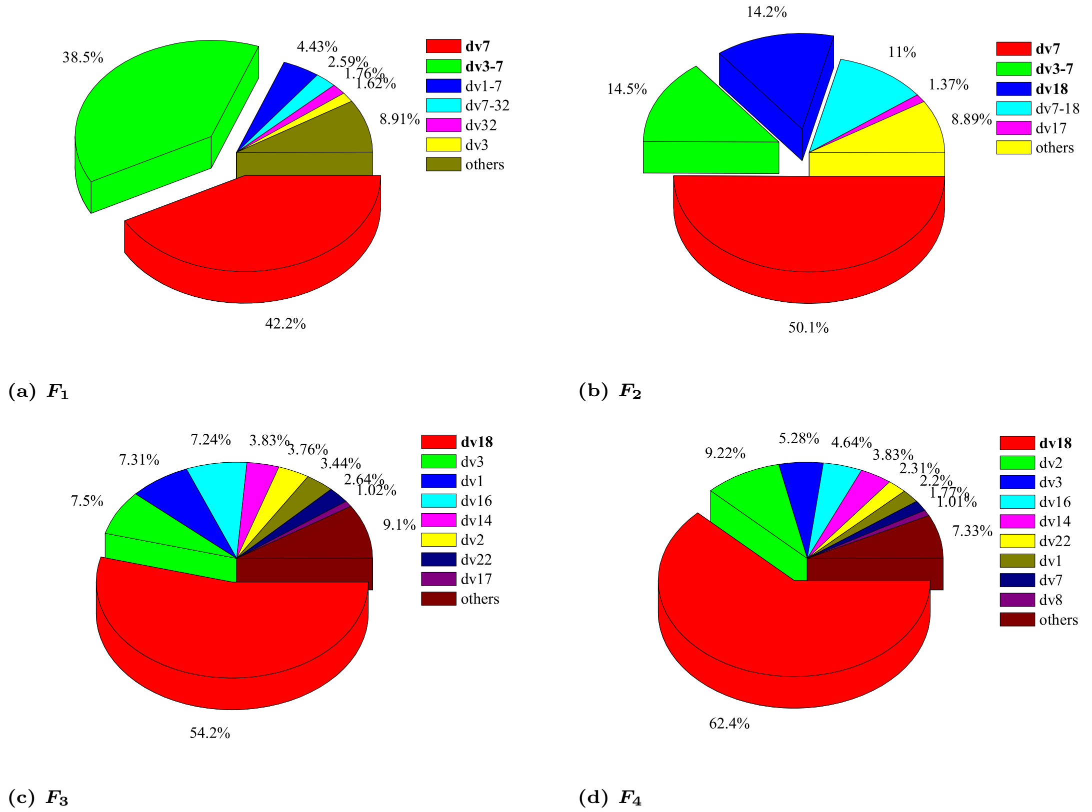
The results reveal that dv7, which is the x coordinate of relative wing position to fuselage, gives the largest effect on the objective function F_1 and F_2, and dv18, which is the rearward camber height at wing tip, gives the largest influence for F_3 and F_4. When the wing position relative to the fuselage is changed, the aerodynamic center is also changed, and this design variable varies the transonic C_{Mp}. In addition, it is known that the camber line has the influence to C_L and C_D. The knowledge obtained by ANOVA corresponds to general knowledge regarding aerodynamics.

When the results from ANOVA are compared with the results from SOM, the influence regarding dv7 and dv18 corresponds well. However, the results from ANOVA do not have the much influence regarding dv22. In this case, dv22 with specific smaller value gives the influence to reduce the transonic C_D. As the decrease of dv22 value reduces the induced drag at tip, the result from SOM is appropriate. Therefore, it is revealed that ANOVA cannot express the influence which only a design variable with a particular range gives, and this defect can be avoided by using SOM simultaneously.

3. Knowledge by Rough Set Theory

The flow of data mining using rough set theory is summarized as follows.

1. Preparation of data
2. Dispersion of data
3. Reduct
4. Generation of rules
5. Filtering
6. Construction of rules



(a) F_1 (b) F_2 (c) F_3 (d) F_4

Figure 8. Proportion of design-variable influence for the objective functions using ANOVA.

The present data was generated from 102 non-dominated solutions, and had four objective functions and 71 design variables. Thus, the object U denotes the non-dominated solutions, the condition attribute C is the design variables, and decision attribute D is the objective functions.

In the present study, each cluster classified by SOM was employed as decision attribute D. The name of each cluster is summarized in Fig. 9. Table 1 shows the cluster names which give the effects on the objective functions. Rough set theory made the rules regarding C1, C2, C5, and C10. Each rule is summarized in Tables 2 to 5. These results show that a high value of dv7 has influence to the shift of aerodynamic center, the characteristic value of dv7 has influence to transonic C_{Mp}, and dv18 has influence to transonic C_D and subsonic C_L. As this knowledge corresponds to the results obtained by SOM and ANOVA, these rule is generated appropriately. As ANOVA shows the total intensity in whole design space directly, it cannot show the intensity in particular design space. For example, as Table 2 shows dv18 has a strong intensity to F_1, local region of dv18 has influence to the shift of aerodynamic center. SOM and ANOVA do not reach this knowledge, and then these rules obtained by rough set theory are useful to narrow down to a detailed design space.

However, physical analyses is needed to the rules generated by rough set theory, and it is difficult to acquire the knowledge with flair. The order of rules does not correspond to the intensity of influence to the objective functions. Moreover, a rule becomes accurate as many individuals with a decision attribute as it generates. When a small number of individuals with characteristics satisfies a decision attribute, the rule makes insufficient result. Rough set theory is effective manner, when there are many individuals to satisfy a decision attribute, and also there are considerable design variables.

B. Design Space from All Solutions

The data mining for the design space generated by all evaluated solutions in optimization gives the knowledge regarding the sensitivity of design variable to objective function, *i.e.*, the direction for a better design. In addition, it can reveal the sweet spot in the design space. That is, it shows the design precept for the problem without severe tradeoffs.

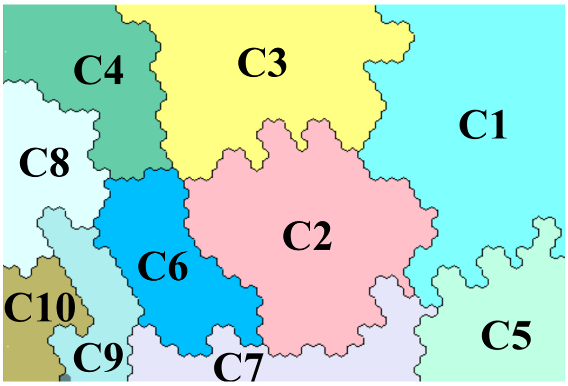


Figure 9. Cluster name on SOM from the non-dominated solutions in the four dimensional objective function space.

Table 1. Features of the clusters with the extreme objective function values on SOM from non-dominated solutions.

cluster	*particularity of performance improvement*
C1	shift of aerodynamic center
C2	transonic C_{Mp}
C5	transonic C_D
C10	subsonic C_L

Table 2. Rules for C1 generated by rough set theory using non-dominated solution data.

rule	*number of data*
dv7([0.462112,*)) AND dv61([*,0.978189)) → C1	17
dv7([0.462112,*)) AND dv69([*,-4.687520)) → C1	16
dv18([0.014534,*)) AND dv46([0.913306,*)) → C1	16
dv18([0.014534,*)) AND dv23([0.002630,*)) → C1	16
dv18([0.014534,*)) AND dv43([0.025468,*)) → C1	16
dv18([0.014534,*)) AND dv71([-6.687150,*)) → C1	16
dv22([*,-0.010476)) AND dv68([0.749406,0.799224)) → C1	15
dv15([0.172055,0.198975)) AND dv18([*,0.014534)) → C1	15
dv22([*,-0.010476)) AND dv46([0.913306,*)) → C1	14
dv22([*,-0.010476)) AND dv40([0.039847,0.047539)) → C1	14

Table 3. Rules for C2 generated by rough set theory using non-dominated solution data.

rule	*number of data*
dv11([0.259898,*)) AND dv20([0.070858,*)) AND dv23([0.001868,0.002630)) → C2	5
dv7([0.385122,0.462112)) AND dv10([*,2.145600)) AND dv46([0.889645,0.913306)) → C2	4
dv18([0.014534,0.044527)) AND dv43([0.023762,0.025468)) AND dv46([0.913306,*)) → C2	4
dv17([*,0.642730)) AND dv29([*,0.026988)) AND dv61([0.984426,*)) → C2	4
dv23([0.001868,0.002630)) AND dv24([*,0.001607)) AND dv65([1.291140,1.363690)) → C2	4
dv13([0.703171,0.732237)) AND dv35([*,0.994927)) AND dv65([1.291140,1.363690)) → C2	4
dv20([0.070858,*)) AND dv21([0.707026,*)) AND dv34([*,0.007870)) → C2	4
dv11([0.259898,*)) AND dv30([*,0.500564)) AND dv64([0.006496,0.006996)) → C2	4
dv12([*,0.015801)) AND dv25([0.012418,*)) AND dv66([0.327994,0.351801)) → C2	4
dv11([0.259898,*)) AND dv23([0.001868,0.002630)) AND dv30([*,0.500564)) → C2	4

Table 4. Rules for C5 generated by rough set theory using non-dominated solution data.

rule	*number of data*
dv11([*,0.192853)) AND dv18([*,0.014534)) AND dv53([*,0.010481)) → C5	4
dv5([3.205220,4.976520)) AND dv18([*,0.014534)) AND dv45([*,0.622205)) → C5	4
dv37([*,0.001320)) AND dv40([*,0.039847)) AND dv49([0.994318,0.995602)) → C5	4
dv40([*,0.039847)) AND dv49([0.994318,0.995602)) AND dv55([*,0.015348)) → C5	4
dv22([*,-0.010476)) AND dv31([*,0.588815)) AND dv35([*,0.994927)) → C5	3
dv21([*,0.601322)) AND dv29([0.026988,0.028812)) AND dv51([*,0.001391)) → C5	3
dv2([*,0.410756)) AND dv33([0.972706,0.984433)) AND dv38([*,0.001229)) → C5	3
dv5([3.205220,4.976520)) AND dv8([-0.050745,-0.044714)) AND dv54([*,0.038919)) → C5	3
dv40([*,0.039847)) AND dv43([0.023762,0.025468)) AND dv59([*,0.686297)) → C5	3
dv29([0.026988,0.028812)) AND dv40([*,0.039847)) AND dv43([0.023762,0.025468)) → C5	3

Table 5. Rules for C10 generated by rough set theory using non-dominated solution data.

rule	*number of data*
dv13([0.732237,*)) AND dv22([0.030320,*)) AND dv39([0.011883,*)) → C10	3
dv33([0.984433,*)) AND dv48([0.007403,*)) AND dv56([0.423662,*)) → C10	2
dv20([0.049659,0.070858)) AND dv44([0.496361,0.512868)) AND dv51([0.001885,*)) → C10	2
dv23([0.002630,*)) AND dv33([0.984433,*)) AND dv51([0.001885,*)) → C10	2
dv10([2.14560,2.967960)) AND dv32([0.875713,*)) AND dv59([0.721385,*)) → C10	2
dv13([0.732237,*)) AND dv18([0.044527,*)) AND dv51([0.001885,*)) → C10	2
dv13([0.732237,*)) AND dv34([0.008203,*)) AND dv51([0.001885,*)) → C10	2

1. *Knowledge by SOM*

The resulting 302 all evaluated solutions have been projected onto the two-dimensional map of SOM. Figure 10 shows the resulting SOM with nine clusters taking the four objective functions into considering. And Fig. 11 shows the SOMs colored by the four objective values, respectively. Figures 11a and 11b shows that the objectives of the shift of aerodynamic center and the transonic C_{Mp} can obtain the lower values simultaneously. These figures also reveal the tendency which both objectives can acquire the high values simultaneously. However, as there is an individual that has highest value of the shift of aerodynamic center and does not have highest value of the transonic C_{Mp}, that tendency is obscure on colored SOM. Especially, as ARMOGA generates a large number of better solution more than worse solutions due to the range adaptation, the obscurity is encouraged. Therefore, it is not good that the bias of solutions exists in the design space from all solutions to discuss the correlation for all evaluated solutions. The cleaning or disposition of solution is needed using response surface method. As Figs. 11c and 11d show that there is a severe tradeoff between the transonic C_D and the subsonic C_L in the general design space, a sweet spot does not exist for all objectives. However, this design space can have a sweet spot when the subsonic C_L is sacrificed for the other objectives. In the case of the subsonic C_L sacrifice for substantial flyback-booster design, high-lift devices must be considered for its landing.

In Figs. 12a, 12c, and 12d, the similar design knowledge is confirmed regarding the dv7, dv18, and dv22 which have the influence for the objective functions in the design space generated by the non-dominated solutions. However, the knowledge in the design space is not clear because of the diversity and bias of the evaluated solutions. Although the other design variables, such as dv12, dv40, dv47, dv54, dv55, and dv61, seems to have the influence for the transonic C_D in Figs. 12b, 12e, 12f, 12g, 12h, and 12i, there is no specific characteristics on colored SOM as a whole.

2. *Knowledge by ANOVA*

Figure 13 shows the ANOVA results for 302 all solutions. This reveals that the influence of design variables for all solutions is similar to one for non-dominated solutions shown in Fig 8. That is, the design knowledge corresponds to the information in the design space generated by the non-dominated solutions. F_1 and F_2 have the subordinate relation each other. Dv18 is effective to F_3 and F_4. Therefore, there is a severe tradeoff between them in the design space generated by all solutions. The notable information is to correspond between the knowledge from all and non-dominated solutions. That is, the knowledge from non-dominated solutions is conserved for design space generated by all solutions.

3. *Knowledge by Rough Set Theory*

The rules are generated by similar procedure which uses for non-dominated-solution data. The obtained rule is summarized in Tables 7 to 10. This rule is also similar to one from non-dominated solutions. Although SOM has perturbation because of a large number of data shown in Fig. 12, rough set theory reveals the similar characteristic design variables with the influence to objective functions compared with the rule from non-dominated solutions. But, notable design variables should find out to interpret rules. Generally, it becomes the problem difficulty to use rough set theory for a large number of design variables.

Table 6. Features of the clusters with the extreme objective function values on SOM of all solutions.

cluster	*particularity of performance improvement*
C1	shift of aerodynamic center
C3	transonic C_{Mp}
C4	transonic C_D
C8	subsonic C_L

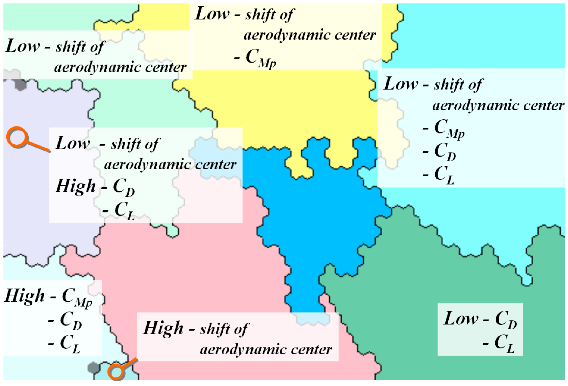
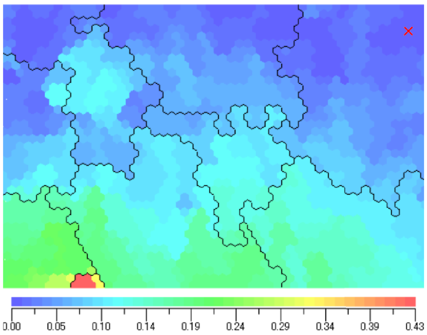
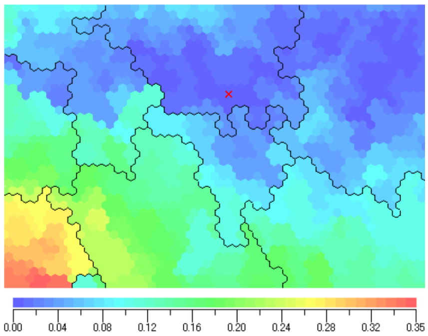


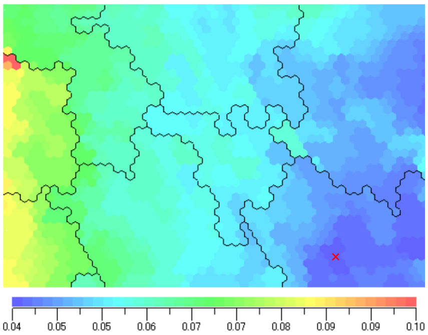
Figure 10. SOM of the evaluated all solutions in the four dimensional objective function space.



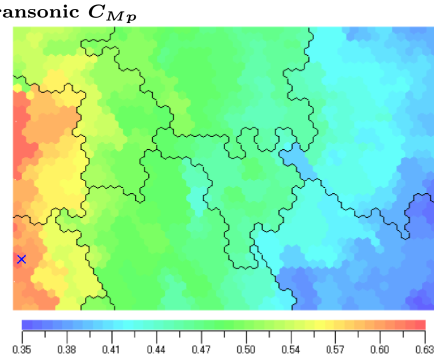
(a) shift of the aerodynamic center



(b) transonic C_{Mp}

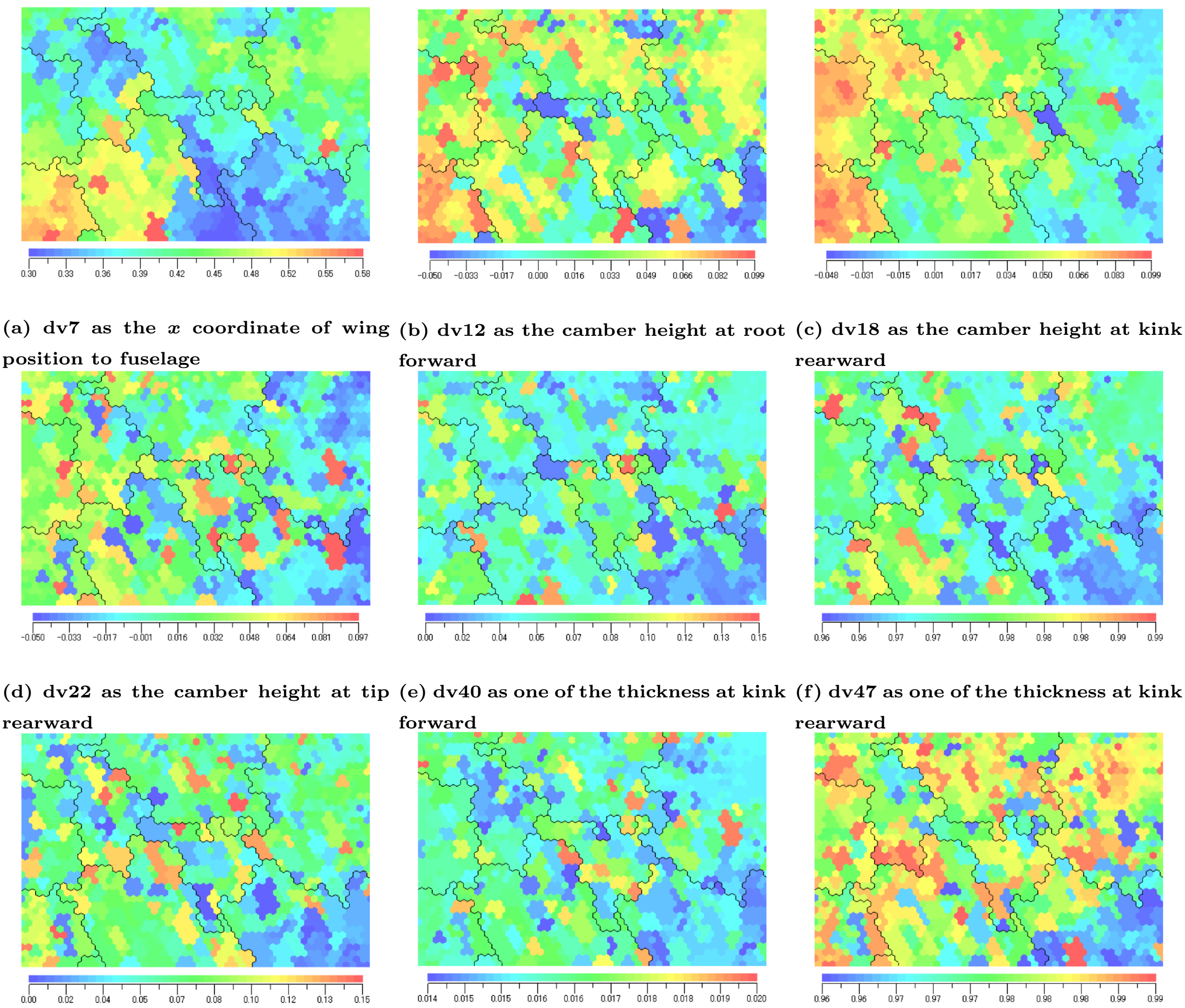


(c) transonic C_D



(d) subsonic C_L

Figure 11. SOMs of the evaluated all solutions colored by the objective functions. The symbol $\times$ denotes the respective extreme non-dominated solutions.



(a) dv7 as the x coordinate of wing position to fuselage

(b) dv12 as the camber height at root forward

(c) dv18 as the camber height at kink rearward

(d) dv22 as the camber height at tip rearward

(e) dv40 as one of the thickness at kink forward

(f) dv47 as one of the thickness at kink rearward

(g) dv54 as one of the thickness at tip forward

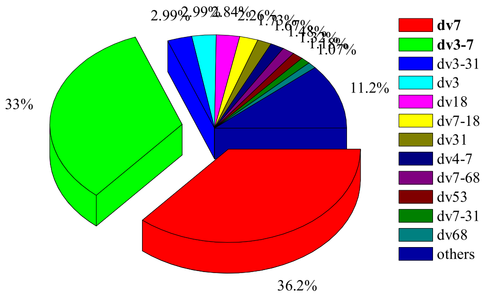
(h) dv55 as one of the thickness at tip forward

(i) dv61 as one of the thickness at tip rearward

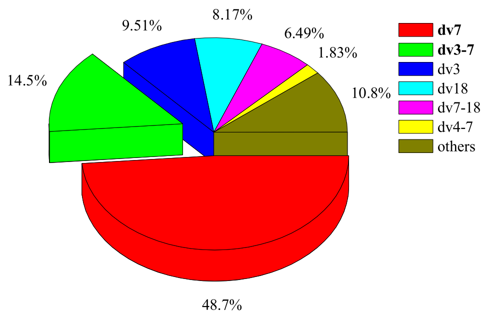
Figure 12. SOMs colored by characteristic design variables.

Table 7. Rules for C1 generated by rough set theory using all-solution data.

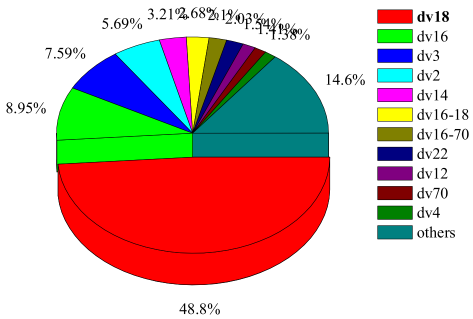
rule	*number of data*
dv7([0.379569,0.463235)) AND dv18([*,0.06781)) AND dv35([0.995614,*)) → C1	20
dv18([*,0.006781)) AND dv22([*,-0.011248)) AND dv42([0.402889,*)) → C1	20
dv18([*,0.006781)) AND dv42([0.402889,*)) AND dv57([0.025121,0.027792)) → C1	19
dv18([*,0.006781)) AND dv25([0.011960,0.012410)) AND dv42([0.402889,*)) → C1	18
dv3([*,55.205799)) AND dv18([*,0.006781)) AND dv24([0.001757,*)) → C1	17
dv18([*,0.006781)) AND dv37([0.001476,0.002020)) AND dv42([0.402889,*)) → C1	17
dv22([*,-0.011248)) AND dv42([0.402889,*)) AND dv60([0.848821,0.870887)) → C1	16
dv3([*,55.205799)) AND dv40([0.040085,0.054506)) AND dv68([0.734865,0.798422)) → C1	16
dv3([*,55.205799)) AND dv40([0.040085,0.054506)) AND dv71([-6.333930,*)) → C1	15
dv18([*,0.006781)) AND dv19([0.156790,0.235667)) AND dv42([0.402889,*)) → C1	15



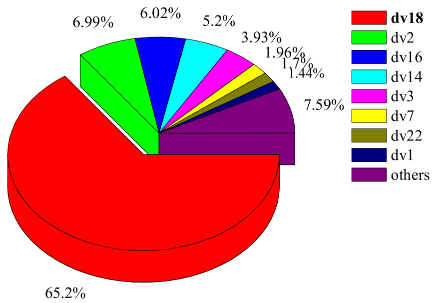
(a) F_1



(b) F_2



(c) F_3



(d) F_4

Figure 13. Proportion of design-variable influence for the objective functions in the all-solution space using ANOVA.

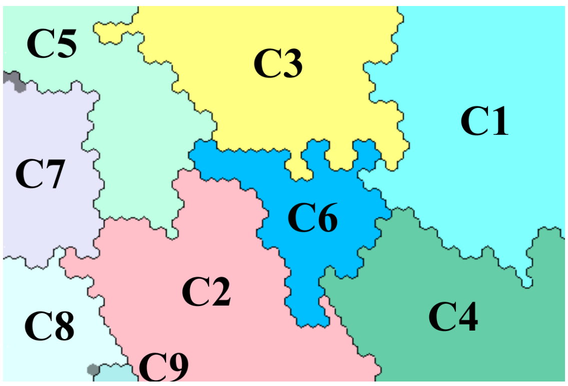


Figure 14. Cluster name on SOM from all solutions in the four dimensional objective function space.

Table 8. **Rules for C3 generated by rough set theory using all-solution data.**

rule	*number of data*
dv28([*,0.336057)) AND dv62([0.009258,*)) AND dv69([-4.622460,*)) → C3	7
dv7([0.379569,0.463235)) AND dv14([*,0.021668)) AND dv32([0.856722,0.881603)) → C3	7
dv7([*,0.379569)) AND dv32([*,0.856722)) AND dv37([0.002020,*)) → C3	7
dv3([62.753700,*)) AND dv34([*,0.007820)) AND dv52([0.001690,0.001762)) → C3	6
dv23([*,0.001821)) AND dv51([0.002018,*)) AND dv52([0.001690,0.001762)) → C3	6
dv50([0.008145,0.008867)) AND dv51([0.002018,*)) AND dv52([0.001690,0.001762)) → C3	5
dv6([302.364990,*)) AND dv7([*,0.379569)) AND dv11([0.264078,*)) → C3	5
dv1([0.183163,0.196521)) AND dv32([*,0.856722)) AND dv48([0.007955,*)) → C3	5
dv18([0.006781,0.041033)) AND dv22([0.032237,*)) AND dv30([*,0.501708)) → C3	5
dv16([0.001324,0.038344)) AND dv33([0.970830,0.982542)) AND dv61([0.983121,*)) → C3	5

Table 9. Rules for C4 generated by rough set theory using all-solution data.

rule	*number of data*
dv18([*,0.006781)) AND dv54([*,0.43352)) AND dv58([*,0.515947)) → C4	14
dv3([*,55.205799)) AND dv51([*,0.001529)) AND dv67([*,-1.647800)) → C4	13
dv40([*,0.040085)) AND dv51([*,0.001529)) AND dv67([*,-1.647800)) → C4	12
dv18([*,0.006781)) AND dv20([0.041139,0.067107)) AND dv58([*,0.515947)) → C4	11
dv20([0.041139,0.067107)) AND dv51([*,0.001529)) AND dv67([*,-1.647800)) → C4	9
dv6([294.427002,302.364990)) AND dv9([*,-1.487150)) AND dv10([*,2.088340)) → C4	9
dv6([294.427002,302.364990)) AND dv10([*,2.088340)) AND dv21([*,0.614866)) → C4	9
dv22([*,-0.011248)) AND dv61([*,0.974320)) AND dv70([-5.336230,-5.236000)) → C4	9
dv31([*,0.598579)) AND dv40([*,0.040085)) AND dv58([*,0.515947)) → C4	8
dv31([*,0.598579)) AND dv57([*,0.025121)) AND dv58([*,0.515947)) → C4	8

Table 10. Rules for C8 generated by rough set theory using all-solution data.

rule	*number of data*
dv29([0.028516,*)) AND dv48([0.007955,*)) AND dv51([0.001529,0.002018)) → C8	7
dv1([0.196521,*)) AND dv29([0.028516,*)) AND dv32([0.881603,*)) → C8	7
dv29([0.028516,*)) AND dv32([0.881603,*)) AND dv65([1.246800,1.424610)) → C8	7
dv29([0.028516,*)) AND dv30([0.501708,0.523840)) AND dv48([0.007955,*)) → C8	6
dv18([0.041033,*)) AND dv30([0.501708,0.523840)) AND dv35([0.995614,*)) → C8	6
dv37([0.002020,*)) AND dv51([0.001529,0.002018)) AND dv61([0.983121,*)) → C8	6
dv7([0.463235,*)) AND dv24([0.001757,*)) AND dv25([*,0.011960)) → C8	4
dv18([0.041033,*)) AND dv20([0.067107,*)) AND dv58([0.515947,0.525239)) → C8	4
dv10([2.088340,3.344390)) AND dv20([0.041139,0.067107)) AND dv48([0.007955,*)) → C8	4
dv20([0.067107,*)) AND dv29([0.028516,*)) AND dv48([0.007955,*)) → C8	4

V. Conclusion

The three data mining techniques have been carried out for the aerodynamic design optimization result of flyback booster wing. These revealed the knowledge in the design space. In addition, the features of three data mining techniques were shown. SOM revealed that 'which' and 'how' design variable influences the objective function. ANOVA showed that 'which' design variable influences. Whereas, rough set theory had different disposition. Rough set theory was useful to narrow down to a detailed design space, when there were many individuals to satisfy a decision attribute, and also there were considerable design variables. However, when there was no considerable design variable and there were many design variables, it was difficult to interpret a rule. Because it was not easy to interpret the physical meaning which the combination of design variables has. Although each data mining could compensate with the respective disadvantages, SOM was an essential data mining technique.

Moreover, data mining was performed to non-dominated and all evaluated solutions, respectively. Consequently, the each design knowledge was similar regarding the tradeoffs, the correlation among the objective functions and design variables, and the influence of design variables. In the present optimization results, the design knowledge from non-dominated-solution data could apply to the design space generated by all solutions, because all solutions included non-dominated solutions. Data mining is essential to understand design space and solve optimization.

References

[1] Obayashi, S., Jeong, S., and Chiba, K., "Multi-Objective Design Exploration for Aerodynamic Configurations," AIAA Paper 2005-4666, 2005.

[2] Jeong, S., Chiba, K., and Obayashi, S., "Data Mining for Aerodynamic Design Space," *Journal of Aerospace Computing, Information, and Communication*, Vol. 2, No. 11, 2005, pp. 452–469.

[3] Holden, C. M. E. and Keane, A. J., "Visualization Methodologies in Aircraft Design," AIAA Paper 2004-4449, 2004.

[4] Chiba, K., Obayashi, S., and Nakahashi, K., "Design Exploration of Aerodynamic Wing Shape for Reusable Launch Vehicle Flyback Booster," *Journal of Aircraft*, Vol. 43, No. 3, 2006, pp. 832–836.

[5] Kohonen, T., *Self-Organizing Maps*, Springer, Berlin, Heidelberg, 1995.

[6] "Eudaptics website," URL: http://www.eudaptics.com [cited 16 June 2004].

[7] Deboeck, G. and Kohonen, T., *Visual Explorations in Finance with Self-Organizing Maps*, London, Springer Finance, 1998.

[8] Pampalk, E., Rauber, A., and Merkl, D., "Content-Based Organization and Visualization of Music Archives," *Proceedings of the 10th International conference on Multimedie*, ACM Press, NY, 2002, pp. 570–579.

[9] Vesanto, J. and Alhoniemi, E., "Clustering of the Self-Organizing Map," *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, 2000, pp. 586–600.

[10] Obayashi, S. and Sasaki, D., "Visualization and Data Mining of Pareto Solutions Using Self-Organizing Map," *EMO 2003, LNCS 2632*, Springer-Verlag Heidelberg, Faro, Portugal, 2003, pp. 796–809.

[11] Chiba, K., Obayashi, S., Nakahashi, K., and Morino, H., "High-Fidelity Multidisciplinary Design Optimization of Aerostructural Wing Shape for Regional Jet," AIAA Paper 2005-5080, 2005.

[12] Jones, D. R., Schonlau, M., and Welch, W. J., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455–492.

[13] Jeong, S., Murayama, M., and Yamamoto, K., "Efficient Optimization Design Method Using Kriging Model," *Journal of Aircraft*, Vol. 42, No. 2, 2005, pp. 413–420.

[14] Pawlak, Z., "Rough Sets," *International Journal of Computer and Information Science*, Vol. 11, No. 5, 1982, pp. 341–356.

[15] Kusiak, A., "Rough Set Theory: A Data Mining Tool for Semiconductor Manufacturing," *IEEE Transactions on Electronics Packaging Manufacturing*, Vol. 24, No. 1, 2001, pp. 44–50.

[16] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*, Boston, MA: Kluwer, 1991.

[17] Shan, N., Ziarko, W., Hamilton, H. J., and Cercone, N., "Using Rough Sets As Tools for Knowledge Descovery," *Proceedings of the 1st International Conference of Knowledge Discovery Data Mining*, U. M. Fayyad and R. Uthurusamy, Eds. Menlo Park, CA, 1995, pp. 263–268.

[18] Pawlak, Z., Slowinski, K., and Slowinski, R., "Rough Classification of Patients after Highly Selective Vagotomy for Duodenal Ulcer," *International Journal of Man-Machine Studies*, Vol. 24, No. 5, 1998, pp. 413–433.

[19] Kusiak, A., Kern, J. A., Kernstine, K. H., and Tseng, T. L., "Autonomous Decision-Making: A Data Mining Approach," *IEEE Transactions on Informatics, Technology in Biomedicine*, Vol. 4, No. 4, 2000, pp. 274–284.

[20] Rosner, B., *Fundamentals of Biostatistics*, Boston, MA: PWS, 1982.

[21] Øhrn, A., *Discernibility and Rough Sets in Medicine: Tools and Applications*, Ph.D. dissertation, Department of Computer and Information Science, Norwegian University of Science and Technology, 1999.

[22] "ROSETTA Technical Reference Manual on Dr. A. Øhrn's website," URL: http://www.idi.ntnu.no/~aleks/thesis/ [cited 2 March 2006].